Ingenieurfakultät Bau Geo Umwelt
Signalverarbeitung in der Erdbeobachtung

Regression-Induced Representation Learning and Its Optimizer: A Novel Paradigm to Revisit Hyperspectral Imagery Analysis

Danfeng Hong

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation

Vorsitzender:		Prof. Dr.-Ing. habil. Richard H. G. Bamler
Prüfer der Dissertation:	1.	Prof. Dr.-Ing. habil. Xiaoxiang Zhu
	2.	Prof. Dr. Jocelyn Chanussot
	3.	Prof. Dr. Gui-Song Xia

Die Dissertation wurde am 23.05.2019 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 07.10.2019 angenommen.

Abstract

Airborne and spaceborne hyperspectral imagery (HSI) is a remotely sensed 3D imaging product of stacking hundreds or thousands of 2D images finely sampled from the continuous wavelength covering the whole electromagnetic spectrum, i.e. 300nm-2500nm. A narrower swath width in spectral domain enables the HSI to discriminate the materials, particularly for those that are extremely similar in the range of visual light, at a more accurate level unachievable by easily-available multispectral or RBG imagery. Over the past decades, unprecedented progress in many challenging tasks of earth observation, such as mineral exploration, precision agriculture, and disaster responses, has been made by the means of HSI acquired by currently operational hyperspectral satellites (e.g., ASTER, Hyperion, CHRIS) and advanced aerial imaging sensors (e.g., DAIS, ROSIS, HyMap, HySpex).

Nevertheless, three crucial issues in HSI – the need for large storage capacity, the spectral variability caused by extrinsic factors (e.g., environmental conditions and instrumental configurations) and the intrinsic deformation of the materials, and small-scale availability due to the limitations of satellite devices itself – hardly make it applicable to a large-scale real scene. Our primary concern is supposed to, therefore, point-to-point addressing the aforementioned problems in order to better brace for upcoming or newly-launched spectroscopy imaging missions, such as German EnMap, NASA's HyspIRI, DLR's DESIS, and Chinese Tiangong-1, Gaofen-5, Zhuhai-1, whose products will be possessed of higher spatial and spectral resolution, wider coverage area, shorter temporal sampling intervals, stronger mapping ability, but also larger storage need and tighter coupling between bands. For this purpose, the thesis will be unfolded from the three main aspects of hyperspectral remote sensing, including **hyperspectral dimensionality reduction**, **spectral unmixing**, and **cross-modality feature fusion and learning**, with five algorithmic contributions to overcoming the *trade-off* between robustness and representation capability of HSI in a regression-based learning paradigm.

In **hyperspectral dimensionality reduction**, the *trade-off* between the explosively growing spectral dimension and the spectral discrimination ability has been an emphatically-focused problem, in that very high dimensionality raises the information redundancy and also introduces more complex noise distribution. Inspired by the statistical robustness of regression technique, a multi-layered regression representation model is developed to improve the discriminative ability of which common regression models are lack 1) by jointly performing dimensionality reduction and classification; 2) by progressively searching several intermediate states of subspaces to approach an optimal mapping; 3) by spectrally embedding manifold structure in each learnt latent subspace in order to preserve the same or similar topological property between the compressed data and the original data.

There is another to-be-considered *trade-off* existed in **spectral unmixing**, that is, spectral variations and accurate unmixing. In this thesis, two feasible solutions to address the spectral variability are introduced by providing new insights into the inverse problem of hyperspectral unmixing. The former assumes to be a low-coherence between real spectral signatures and spectral variabilities and then integrates this attribute into a sparse and dense joint regression model, called the augmented linear mixing model (ALMM). While for the latter, it seeks to unmix the HSI in a to-be-estimated subspace instead of in the original high-dimensional space, and the subspace and abundance maps in unmixing can be jointly optimized with a low-rank attribute embedding.

A rethinking-worthy open problem for the exiting and upcoming hyperspectral imaging missions is *how to use the HSI to contribute to a large area and even global mapping and monitoring*, since there are higher spectral resolution yet lower spatial resolution and smaller coverage from space in these HSIs than those of MSIs. This *trade-off* between HSI and MSI

naturally leads to a challenging issue related to **cross-modality feature fusion and learning**. With this intent, a regression-based cross-modality learning framework is designed, called common subspace learning (CoSpace), to linearly learn a shared latent subspace from hyperspectral-multispectral (HS-MS) correspondences by locally aligning the manifold structure of the two modalities. Through the learned subspace, the HSI's properties can be effectively transferred into the MSI available on a larger scale. Beyond the CoSapce, a semi-supervised learning framework is proposed by learning to simultaneously align the data structures of labeled and unlabeled samples as well as multi-modalities in the form of graph representation.

Moreover, a unified optimizer followed by the alternating direction method of multipliers (ADMM) strategy is developed and generalized to solve the above-mentioned five algorithms.

Besides, these proposed strategies in different hyperspectral tasks have been proven to be superior and effective, from both visually and quantitatively, in comparison with other state-of-the-art methods for a variety of simulated and real data scenarios.

Zusammenfassung

Luft- und raumfahrtgestützte Hyperspektralbilder (HSI) sind ein ferngesteuertes 3D-Bildgebungsprodukt, bei dem hunderte oder tausende 2D-Bilder stapelweise aus der kontinuierlichen Wellenlänge des gesamten elektromagnetischen Spektrums, d.H. von 300 nm bis 2500 nm, abgetastet werden. Eine geringere Streifenbreite im Spektralbereich ermöglicht eine genauere Unterscheidung von Materialien, welche mittels leicht zugänglichen Multispektral- oder RGB-Bildern nicht erreichbar wäre. Dies gilt insbesondere für Materialien, die sich im visuellen Spektrum stark ähneln. In den letzten Jahrzehnten wurden mit Hilfe von "HSI", von derzeit in Betrieb befindlichen hyperspektralen Satelliten (z. B. ASTER, Hyperion, CHRIS) und fortschrittlichen Luftbildsensoren (z.B. DAIS, ROSIS, HyMAP, HYSpex), beispiellose Fortschritte bei vielen anspruchsvollen Aufgaben der Erdbeobachtung erzielt, beispielsweise bei der Mineralexploration, der Präzisionslandwirtschaft und bei der Katastrophenbewältigung.

Dennoch erschweren drei entscheidende Punkte die Anwendbarkeit von HSI für große reale Szenen. Diese umfassen 1) den Bedarf an großer Speicherkapazität, 2) die spektrale Variabilität durch äußere Faktoren (z.B. Umgebungsbedingungen und Gerätekonfigurationen) und 3) die Eigenverformung der Materialien, sowie die Verfügbarkeit im kleinen Maßstab durch die Einschränkungen der Satellitengeräte selbst. Unser Hauptanliegen ist es daher, die oben genannten Probleme punktuell anzugehen, um besser auf bevorstehende oder kürzliche gestartete spektroskopische Bildgebungsmissionen-Missionen, wie German EnMap, HyspIRI der NASA, DESIS des DLR und chinesisches Tiangong-1, Gaofen-5 uund Zuhai-1 vorbereitet zu sein, dessen Produkte eine höhere räumliche und spektrale Auflösung, einen größeren Erfassungsbereich, kürzere zeitliche Abtastintervalle, eine stärkere Kartierungsfähigkeit, aber auch einen größeren Speicherbedarf und eine engere Verflechtung zwischen den Bändern aufweisen werden. Zu diesem Zweck wird diese Dissertation aus den drei Hauptaspekten der hyperspektralen Fernerkundung entwickelt, einschließlich **der Verringerung der hyperspektralen Dimensionalität**, **der spektralen Entmischung** und **der Kombination und des Lernens von Kreuzmodalitäten** mit fünf algorithmischen Beiträgen zur Überwindung des *Kompromisses* zwischen Robustheit und Repräsentationsfähigkeit von HSI in einem regressionsbasierten Lernparadigma.

Bei **der Verringerung der hyperspektralen Dimensionalität** war der *Kompromiss* zwischen der explosionsartig wachsenden spektralen Dimension und der spektralen Diskriminierungsfähigkeit ein nachdrücklich fokussiertes Problem, da eine sehr hohe Dimensionalität die Informationsredundanz erhöht und auch eine komplexere Rauschverteilung einführt. Inspiriert von der statistischen Robustheit der Regressionstechnik wird ein mehrschichtiges Regressionsrepräsentationsmodell entwickelt, um die Diskriminierungsfähigkeit zu verbessern, an der gängige Regressionsmodelle fehlen 1) durch gemeinsames Durchführen der Dimensionsreduktion und -klassifizierung; 2) durch schrittweises Durchsuchen mehrerer Zwischenzustände von Unterräumen, um sich einer optimalen Abbildung anzunähern; 3) durch spektrales Einbetten einer Mannigfaltigkeitsstruktur in jeden erlernten latenten Unterraum, um die gleiche oder eine ähnliche topologische Eigenschaft zwischen den komprimierten Daten und den ursprünglichen Daten zu erhalten.

Bei **der spektralen Entmischung** gibt es einen weiteren *Kompromiss* zwischen spektraler Variationen und genauer Entmischung. In dieser Arbeit werden zwei mögliche Lösungen zur Behebung der spektralen Variabilität vorgestellt, indem neue Erkenntnisse über das inverse Problem der hyperspektralen Entmischung bereitgestellt werden. Ersteres geht von einer niedrigen Kohärenz zwischen echten Spektralsignaturen und Spektralvariabilitäten aus und integriert dieses Attribut in ein spärliches und dichtes Regressionsmodell, das als Augmented Linear Mixing Model (ALMM) bezeichnet wird. Letzteres versucht in einem zu

schätzenden Unterraum statt im ursprünglichen hochdimensionalen Raum zu entmischen, und die Unterraum- und Abundanzkarten beim Entmischen gemeinsam mit einer niederrangigen Attributeinbettung optimiert werden können.

Ein umdenkenswertes offenes Problem für die bestehenden und zukünftigen hyperspektralen Bildgebungsmissionen ist, *wie man mit HSI zu einer großräumigen und sogar globalen Kartierung und Überwachung beitragen kann*, da für HSI zwar eine höhere spektrale Auflösung als für MSI erreicht werden kann, gleichzeitig jedoch eine geringerer räumlicher Auflösung und Abdeckung. Dieser *Kompromiss* zwischen HSI und MSI führt naturgemäß zu einem herausfordernden Problem im Zusammenhang mit **Cross-Modality-Feature-Fusion und Lernen**. Mit dieser Absicht wird ein regressionsbasiertes Cross-Modality-Learning-Framework entwickelt, das als Common Subspace Learning (CoSpace) bezeichnet wird, um einen gemeinsam genutzten latenten Unterraum aus hyperspektralen, multispektralen (HS-MS) -Korrespondenzen durch lokales Ausrichten der vielfältigen Strukturen der beiden linear zu lernen Modalitäten. Durch den erlernten Teilraum können die Eigenschaften des HSI effektiv in das MSI übertragen werden, welches in einem größeren Maßstab verfügbar ist. Über das CoSapce hinaus wird ein semi-überwachtes Lernframework vorgeschlagen, in dem die Datenstrukturen von markierten und nicht markierten Proben sowie Multimodalitäten in Form einer Diagrammdarstellung gleichzeitig abgeglichen werden.

Darüber hinaus wird ein einheitlicher Optimierer, gefolgt von der Strategie der alternierenden Richtungsmethode der Multiplikatoren (ADMM), entwickelt und verallgemeinert, um die oben genannten fünf Algorithmen zu lösen.

Zudem haben sich diese vorgeschlagenen Strategien für unterschiedliche hyperspektrale Aufgaben sowohl visuell als auch quantitativ im Vergleich mit anderen modernen Methoden für eine Vielzahl simulierter und realer Datenszenarien als überlegen und wirksam erwiesen.

List of Abbreviations

Abbreviation	Description
1-D	one-dimensional
2-D	two-dimensional
3-D	three-dimensional
AA	average accuracy
ACMSL	alignment-based cross-modality share learning
ADMM	alternating direction method of multipliers
ALMM	augmented linear mixing model
AVIRIS	Airborne Visible / Infrared Imaging Spectrometer
AutoRULe	auto-reconstructing unsupervised learning
CASI	Compact Airborne Spectrographic Imager
CCF	canonical correlation forests
CDF	cumulative distribution function
CD	coordinate descent
CGDA	collaborative graph-based discriminant analysis
CML	cross-modality learning
CMMFL	concentration-based multi-modality fusion learning
CMs	classification maps
CoSpace	common subspace learning
DADR	discriminant analysis dimensionality reduction
DAIS	Digital Airborne Imaging Spectrometer
DANSER	dictionary-adjusted non-convex sparsity-encouraging regression
DLR	Deutschen Zentrums für Luft- und Raumfahrt
ELMM	extended linear mixing model
FA	factor analysis
FCLUS	fully constrained least squares unmixing
FSDA	feature space discriminant analysis
GBM	generalized bilinear model
GDA	graph-based discriminant analysis
GDN	global data normalization
GED	generalized eigenvalues decomposition
GGE	general graph embedding

Abbreviation	Description
GLP	graph-based label propagation
GSD	ground sampling distance
HDR	hyperspectral dimensionality reduction
HNS	hierarchical neighbor selection
HS	hyperspectral
HSI	hyperspectral imagery
HyMap	Hyperspectral Mapper
ICA	independent component analysis
IFOV	instantaneous field of view
IT	iterative thresholding
ISOMAP	isometric feature mapping
JL	joint learning
JN	joint normalization
J-Play	joint and progressive learning strategy
KCGDA	kernel collaborative graph-based discriminant analysis
KDA	kernelized discriminant analysis
KLD	Kullback-Leibler divergence
KLDA	kernel linear discriminant analysis
KLFDA	kernel local fisher discriminant analysis
KPCA	kernel principle component analysis
KSGDA	kernel sparse graph-based discriminant analysis
LARS	least angle regression
LDA	linear discriminant analysis
LDN	local data normalization
LE	Laplacian eigenmaps
LeMA	learnable manifold alignment
LFDA	local fisher discriminant analysis
LLE	locally linear embedding
LML	local manifold learning
LMM	linear mixing model
LPP	locality preserving projections
LSDR	least-squares dimension reduction
LSL	latent subspace learning
L-SMA	LPP-based supervised manifold alignment
LSQMI	least-squares quadratic mutual information

Abbreviation	Description
LSQMID	least-squares QMI derivative
LTSA	local tangent space alignment
L-USMA	LPP-based unsupervised manifold alignment
MA	manifold alignment
MAP	maximum a posteriori
MIVIS	Multispectral Infrared and Visible Imaging
MMDA	multi-modality data analysis
MS	multispectral
MSI	multispectral imagery
NASA	National Aeronautics and Space Administration
NN	nearest neighbor
NPE	neighborhood preserving embedding
NS	neighbor selection
OA	overall accuracy
OSF	original spectral features
PCA	principle component analysis
PCLSU	partial con-strained least squares unmixing
PGD	proximal gradient descent
P-JDR	PCA-based on joint dimensionality reduction
PPCA	probabilistic principal component analysis
PLMM	perturbed linear mixing model
QP	quadratic programming
RBF	radial basis function
RIRL	regression-induced representation learning
RLMR	robust local manifold representation
RNS	refined neighbor selection
RTT	radiative transfer theory
SAM	spectral angle mapper
SAR	synthetic aperture radar
S-CoSpace	semi-supervised CoSpace
SELD	semi-supervised local discriminant analysis
SELF	semi-supervised local Fisher discriminant analysis
SGDA	sparse graph-based discriminant analysis
SIR	sliced inverse regression
SL	subspace learning

Abbreviation	Description
SLDA	subspace linear discriminant analysis
SMI	squared-loss mutual information
SNR	signal to noise ratio
SPCLUS	scaled partial constrained least squares unmixing
SSDA	semi-supervised discriminant analysis
S-SMA	Semi-supervised supervised manifold alignment
SU	spectral unmixing
SULoRA	subspace unmixing with low-rank attribute embedding
SUnSAL	sparse unmixing by variable splitting and augmented Lagrangian
TDA	topological data analysis
TV	total variation
SVD	singular value decomposition
SVM	support vector machine
SVT	singular value thresholding
UAV	unmanned aerial vehicle
VCA	vertex component analysis

List of Symbols

Abbreviation	Description
α	regularization parameter
β	regularization parameter
γ	regularization parameter
λ	regularization parameter
$\mathbf{\Lambda}$	Lagrangian multiplier
C	class label
Cov	covariance matrix
$\mathbf{D}$	degree matrix
E	expectation
$g^{'}(\bullet)$	derivative of function $g(\bullet)$
h_v	function with respect to the variable v
k	the number of classes
$\mathbf{L}$	Laplacian matrix
ρ	increasing rate of penalty parameter μ
$\mathbf{\Theta}$	latent subspace projection
$\odot$	Schur-Hadamard (term-wise) product
$./$	element-wise (term-wise) division
$p(\bullet)$	probability distribution function
r_i	residual vector
R	residual matrix
S^2	variance
$\mathbf{S}_w$	within-class scatter matrix
$\mathbf{S}_b$	between-class scatter matrix
$\phi_k(\bullet)$	the k nearest neighbor of the variable $\bullet$
t	iteration
$\mathrm{tr}(\bullet)$	trace of matrix $\bullet$
$(\bullet)^{\mathrm{T}}$	transpose of matrix
μ	penalty parameter
μ_{max}	upper bound of penalty parameter μ
η	tolerated errors
v_i	i-th value in projection matrix $\mathbf{V}$

Abbreviation	Description
$\mathbf{V}$	projection matrix
$\mathbf{W}$	adjacency matrix
$\mathbf{x}_i$	spectral signature of i-th pixel (sample)
$\mathbf{X}$	unfolded hyperspectral image
$\overline{\mathbf{X}}$	mean value of the variable $\mathbf{X}$
$\mathbf{y}_i$	class label of i-th pixel (sample)
$\mathbf{Y}$	label matrix that consists of $\mathbf{y}_i$
$\mathbf{Y}_l$	one-hot encoded label matrix
$\mathbf{z}_i$	low-dimensional embedding (vector)
$\mathbf{Z}$	low-dimensional embedding (matrix)
$\partial(\bullet)$	partial derivative of variable $\bullet$
$(\bullet)^{-1}$	inverse of matrix
$\|\bullet\|_{1,1}$	$\mathcal{L}_1$ norm of matrix
$\|\bullet\|_2$	$\mathcal{L}_2$ norm of vector
$\|\bullet\|_{\mathrm{F}}$	Frobenius norm
$\|\bullet\|_*$	nuclear norm

Contents

1 Introduction

1.1 Motivation and Challenges

Hyperspectral imaging, also known as imaging spectroscopy, is a seminal technique of truly achieving the integration of the 1-D spectrum and the 2-D image, which was first-ever to be conceptualized by Goetz1 *et al.* in 1980's [Goetz et al., 1985]. From then on, hyperspectral remote sensing, which is evolved based on imaging spectroscopy, has garnered growing attention from researchers. Unlike those previous optical imaging techniques that sample the spectral space in a discrete (or very sparse) form (e.g., panchromatic, color photography, and multispectral imagery), the hyperspectral imaging systems exploit the senors to collect hundreds or thousands of spectral channels with an approximately continuous spectral sampling at a subtle interval (e.g., 10nm). This makes the hyperspectral remote sensing widely applied in earth observation and environmental surveys. More specifically, the main differences between hyperspectral remote sensing and those remote sensing techniques of low spectral resolution lie in the following three aspects:

- The hyperspectral products are capable of finely discriminating the different classes that belong to the same category, such as Citigroup Pine and American Giant Sequoia, Alunite and Kaolin. While for those traditional optical imaging products (e.g., multispectral imagery), they can only identify some materials with the significant differences in the spectral signatures, such as water, vegetation, soil, etc.

- The higher spectral resolution makes it feasible to some formerly impossible applications, e.g., parameter extraction of biophysics and biochemistry, automatic detection of food safety, which provides new insight into the field of the remote sensing technique.

- Due to the limitations in spectral and spatial resolution of imaging sensors, atmospheric effects, and the interference of soil background, the traditional remote sensing technique was dominated by qualitative analysis. With the emergence of hyperspectral remote sensing, quantitative or semi-quantitative analysis is becoming increasingly possible.

Despite the HSI's merits mentioned above, yet there are still several crucial issues that need to be sufficiently considered in the high-level data analysis with the use of HSI: information redundancy, spectral variability caused by illumination, topography change, atmospheric effects, and complex sensor noises, the need for large storage capacity and high performance computing, and data acquisition over a large area, among others. These drawbacks can be generalized to some specific challenges by raising three important questions about " how" as follows:

- **Overcoming the curse of dimensionality.** As the HSI's dimension gradually increases along the spectral direction, the spectral discrimination ability in identifying the materials would meet the bottleneck and even suffer from the degradation. This might be well explained by many possible factors, such as the coupling between the neighboring spectral bands, more complex noise patterns, the same object with different spectra and different objects with the same spectrum. One challenge is posed to the first "how" question – *how to effectively preserve the task-related information and get rid of the useless information in parallel?*

- **Addressing spectral variability.** Due to the meter-level ground sampling distance (GSD) of hyperspectral imaging, the spectral signatures for most pixels of HSI are acquired in the form of a complex mixture that consists of at least two types of materi-

als, inevitably degrading the performance of spectral identification. Spectral unmixing must be made before the high-level data analysis. However, the spectral variations, such as scaling factors, offsets, low-coherent or incoherent constituents, or complex noises, make it very difficult to accurately unmix these mixed pixels. Thus, the second "how" question corresponding to the challenge in spectral unmixing is *how to accurately estimate the abundance maps of the endmembers in the presence of spectral variability?*

- **Exploring and positioning the HSI's role in future earth observation.** There is an obvious trend in the coming spaceborne earth observation, that is, extraordinary demand on global area data processing and analysis. It is well known, however, that a wealth of spectral bands enable the HSI to distinguish and detect the objects of interest with ease, especially for those spectrally similar classes, but its swath width from the space is completely incomparable to the one of optical broadband (e.g., multispectral) imaging due to the differences of imaging principles and techniques. For that reason, an application-innovative challenge in hyperspectral remote sensing is converted to describing the third "how" question: *how can HSI covering only a limited part of the MSI be explored to help improve the classification (or mapping) of the entire area covered by the MSI?*

1.2 Objectives and Research Focus

With the coming of the "Big Data" era, large-scale remote sensing data management, monitoring and utilizing have developed into the mainstream in the next-generation earth observation. As a central member of the remote sensing community, the HSI is duty-bound to participate in the tasks of global mapping, monitoring, and responding. Hence, the resulting general goal in this thesis is

"developing advanced algorithms to analyzing the hyperspectral data more robustly and efficiently with the potential contributions to improving the classification or mapping tasks in the regional and even global coverage".

Towards this goal, three main research objectives aiming at item-to-item handling the aforementioned challenges have been specified in the following:

– **Objective 1:** *developing novel strategies to reduce the spectral dimension without sacrificing the highly-discriminant information*
Due to the highly-correlated characteristic between spectral bands, the HSI is subjected to the information redundancy, which could hurt the ability to discriminate the materials under certain extremely-conditioned cases. Dimensionality reduction must be conducted before the high-level data analysis starts up. As a result, the first research branch of this thesis is to balance the spectral discrimination and robustness of the results before and after performing dimensionality reduction.

– **Objective 2:** *discovering new prior knowledge against spectral variability for robust hyperspectral unmixing*
In most previously-proposed unmixing models, e.g., the classic linear mixing model (LMM), they generally fail to consider the spectral variability in the process of estimating abundance maps. This leads to the poor unmixing performance by using those LMM-based approaches, since the spectral variability is not ignored or discarded but absorbed by the estimated abundances. To this end, the second investigated branch in the work is supposed to develop the new models linking with the spectral variability from the point of the physically-meaningful view.

– **Objective 3:** *pulling the hyperspectral data into additional data sources or modalities (e.g.,*

multispectral data) to enhance the feature representation ability for preparation of large-scale land cover and land use classification.
Recently, multispectral spaceborne images are freely available on a nearly global scale, thanks to those optical broadband satellites that have been launched, such as Sentinel-2 [Drusch et al., 2012], Landsat-8 [Roy et al., 2014], etc. It should be noted, however, that limited by the spectral bandwidth, it is next to impossible for the multispectral data to distinguish the materials that are spectrally similar only with minute spectral discrepancies. In this connection, the final research branch of this dissertation is transferring the highly-discriminant spectral information from HSI into wider-covered MSI of low spectral resolution only in the training phase, and improving the classification performance of the remaining large-scale MSI under the conditions without the corresponding HSI.

The traditional methodology of hyperspectral data analysis may not be qualified to cope with the above objectives. Consequently, the hyperspectral data analysis has to be revisited to find the technically feasible and theoretically-guaranteed solutions following an effective regression-based representation learning paradigm. Accordingly, the methodological focus of this thesis can be detailed as

- **Solution 1:** Docking to the **Objective 1** – hyperspectral dimensionality reduction task, a **j**oint and **p**rogressive **l**earning str**a**teg**y** (J-Play) is proposed to linearly find an optimal dimension-reduced subspace. The J-Play is made up of two strategies: the joint learning that simultaneously performs subspace learning and regression aims at finding a discriminative subspace by bridging the learned subspace with the label information, while the progressive learning gradually converts the original data space to a potentially optimal subspace through multi-coupled intermediate transformations, tending to find a better solution. Additionally, with the local manifold preservation on each intermediate subspace, the proposed method has demonstrated its robustness and discriminant capability as well as the ability to generalize the out-of-the-sample.

- **Solution 2:** Linking with the **Objective 2**, an **a**ugmented **l**inear **m**ixing **m**odel (ALMM) is proposed to view the spectral unmixing as a special case of bilinear mixing model by incorporating two different regression techniques: sparse regression attempting to accurately estimating the scaled abundance maps in the absence of other spectral variabilities, and dense regression allowing for reconstructing the rest of spectral variabilities except scaling factors with a to-be-updated spectral variability dictionary. The two parts can be organically coupled with a low-coherent assumption to be a joint model.

- **Solution 3:** Different from the **Solution 2** to solve the **Objective 2**, which unmixs the HSI in the original spectral space, a novel **s**ubspace-based **u**nmixing model is developed with **lo**w-**r**ank **a**ttribute embedding, called SULoRA, by jointly estimating subspace projections and regressing the sparse abundance maps to robustify the inverse problems of hyperspectral unmixing against spectral variability.

- **Solution 4:** Connecting to the **Objective 3**, this thesis presents a general but effective **co**mmon sub**space** learning method, CoSpace for short. Similarly to the J-Play in the **Solution 1**, CoSpace also follows the joint learning framework. The main difference lies in that the latent subspace is learned by aligning the class-specific manifold structure of two modalities (MS-HS). Furthermore, through the subspace, the HSI-related properties, e.g., high spectral discrimination, can be effectively transferred to those multispectral out-of-samples, thereby achieving the performance improvement of classification in a larger study scene.

- **Solution 5:** Beyond the **Solution 4**, CoSpace is extended to a semi-supervised version

of cross-modality learning, named as **le**arnable **m**anifold **a**lignment (LeMA). As the name suggests, LeMA aligns the two different modalities not limiting to labeled data but also unlabeled data, by the means of data-driven learning strategy instead of the hand-crafted graph structure. Headed by the learned graph, the decision boundary may be better determined, i.e. using graph-based label propagation.

1.3 Skeleton of the Thesis

This is a *cumulative* dissertation to be unfolded around the general goal of hyperspectral imagery analysis, mainly including seven peer-reviewed papers – one top conference and six journal articles (please see the list of Appendix). The remainder of this thesis is guided as follows:

Chapter 2 starts with the introduction of hyperspectral imaging systems comprising imaging principle, the concept of spectral signals, and the explanation for material mixture and spectral variability as well as the potential applications in the next-generation earth observation. Afterward, this chapter also makes a detailed review of several types of regression techniques and their solvers.

Chapter 3 systematically provides the analysis and discussion of the state-of-the-art methods in hyperspectral data analysis from three main aspects: dimensionality reduction, spectral unmixing, and cross-modality feature fusion and learning. It then ends with clarifying our main contributions of this thesis.

Corresponding to these main contributions, the overview and summary of the seven relevant publications are given in Chapter 4, and the details in each paper can be found in the attached Appendix.

The last Chapter 5 draws the conclusions and looks forward to the promising future work.

2 Basics

This chapter briefly makes a picture of hyperspectral imaging to help the readers who are interested or already in the field of hyperspectral remote sensing quickly accessible into the relevant topics. To begin with, the imaging principle is introduced and then its product is presented in the form of spectral signatures. Next, the material miscibility in HSI is explained by various factors and also spectral variability is pointed out to be ubiquitous. Finally, the role of hyperspectral remote sensing in earth observation and potential applications are clarified.

2.1 Get to Know the Hyperspectral Imaging

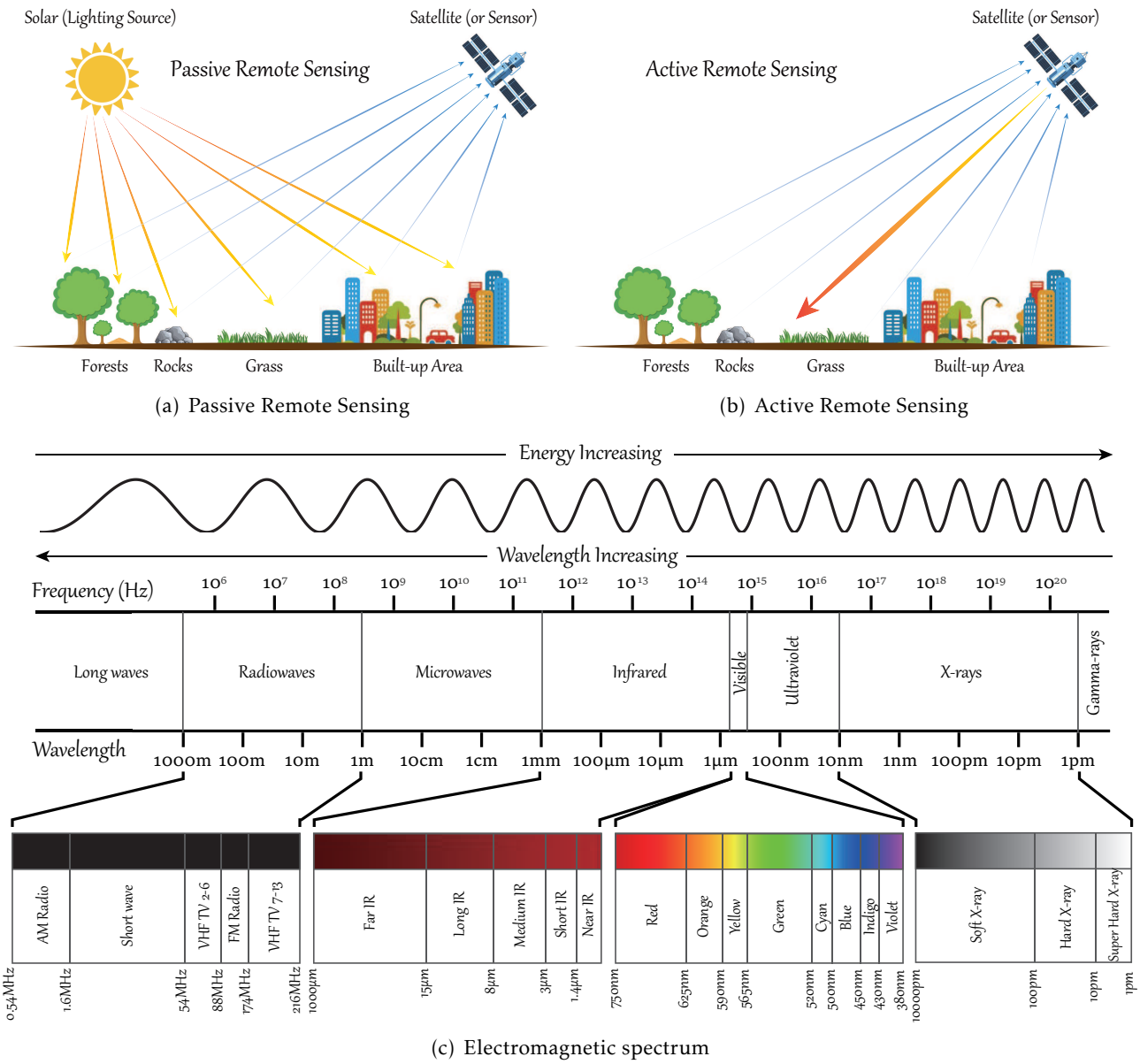
2.1.1 Imaging Principle

Remote sensing [Tsang et al., 1985] is an important means of information acquisition in a contactless fashion. Technically speaking, it falls into "active" remote sensing, emitting the energy or signal by spacecraft or aircraft and receiving the response reflected from the object by the sensor similarly installed in spacecraft or aircraft, and "passive" remote sensing, which directly detects the radiation from the sunlight's reflection on the surface of the Earth [Ulaby et al., 1986]. Figure 2.1 (a) and (b) illustrate the procedures of the two collection patterns of remote sensing data. As a promising category of "passive" remote sensing, hyperspectral imaging [Goetz et al., 1985] judiciously assembles the two techniques of spectroscopy and digital photography in a single system. The resulting product is a 3-D cube by simultaneously scanning the 2-D image plane in spectrally contiguous bands. The HSI holds a complete spectrum, which means that hundreds of (narrow) wavelength bands are collected at each pixel across the electromagnetic spectrum [Turner et al., 2003], i.e. from Gamma-rays, X-rays, the ultraviolet, through the visible and the infrared, to micro-waves, radio-waves, and even long-waves. Figure 2.1 (c) shows a fine partition for the electromagnetic spectrum.

2.1.2 Hyperspectral Sensors

As opposed to broad wavelength imaging techniques, such as RBG or multispectral imaging [Hong et al., 2015, Wu et al., 2019a, 2018], that provide the sparse spectral channels up to ten, imaging spectrometer uses the hyperspectral sensors, which is nothing structurally special with the charge-coupled device (CCD)-like and multispectral scanners but only difference in more spectral channels with the compacted sampling intervals, to record the detailed spectral information approximately being able to go throughout entire electromagnetic spectrum. Up to the present, there has been an incrementally updating in imaging spectrometers from either aircraft or spacecraft, enabling the image quality progressively increasing. According to the different carriers, these sensors can be roughly categorized into two groups – airborne and spaceborne.

The former captures the imagery relatively flexibly due to the self-adapting to the schedule of image acquisition, which is effective to minimize the interference of changeable weather conditions caused by sum illumination, cloud blocking, and other atmospheric effects. The aircraft, also known as an unmanned aerial vehicle (UAV), helicopter, drone, airship, is of great benefit to developing a practical platform due to its flexibility in maintaining, repairing, and re-configuring the devices. A few of popular advanced airborne hyperspectral imagers will be briefly introduced, i.e.



(a) Passive Remote Sensing

(b) Active Remote Sensing

(c) Electromagnetic spectrum

Fig. 2.1. An illustration to clarify the similarities and differences between "active" remote sensing and "passive" remote sensing [Tsang et al., 1985], as shown in (a) and (b). (c) gives a showcase of the electromagnetic spectrum [Turner et al., 2003]: the order from low to high according to frequency is Long-waves, Radio-waves, Micro-waves, Infrared, Visible, Ultraviolet, X-rays, and Gamma-rays, where several highlighted intervals, e.g., Radio-waves, Infrared, Visible, and X-rays are finely partitioned.

- ◇ Airborne Visible / Infrared Imaging Spectrometer (AVIRIS) is a premier airborne equipment used to measure the radiance in the spectral wavelength ranging from 400nm to 2500nm, which has been successively carried on four remote sensing platforms, e.g., NASA ER-2, de Havilland Canada DHC-6 Twin otters, Scaled Composites Model 281 Proteus, and NASA's WB-57.
- ◇ Compact Airborne Spectrographic Imager (CASI) is an instrument of recording the radiance with at most 288 bands in the visible near Infrared (380nm to 1050nm) and offering 25cm spatial resolution. The hyperspectral cameras have contributed to a large number of applications of remote sensing, owing to its finer focus and the high sensitivity to lighting source.
- ◇ Digital Airborne Imaging Spectrometer (DAIS-7915) collects the reflected radiance across a wide range of spectral wavelength: 400nm to 12600nm, 79 channels in total. These channels are captured individually using four different Spectrometers that contain 32 bands (400nm to 1000nm), 8 bands (1500nm to 1800nm), 32 bands (2000nm to

Table 1. An overview of parameter configuration of several representative airborne hyperspectral sensors as well as operational and upcoming spaceborne hyperspectral imaging missions where IFOV means instantaneous field of view. Some details stem from [Ortenberg et al., 2011].

Airborne Hyperspectral Sensors

Sensor	Operator	Spectral Range	Band Number	Spectral Resolution	Spatial Resolution	IFOV	Swath
AVIRIS	NASA	400-2500nm	224	10nm	17m	0.1mrad	11km
CASI	ITRES	380-1050nm	288	2.5nm	0.25-1.5m	0.5mrad	1.5-7km
DAIS-7915	DLR	498-1010nm	32	16nm	3-20m	3.3mrad	2.5km
		1500-1800nm	8	100nm			
		1970-2450nm	32	15nm			
		3000-5000nm	1	2000nm			
		8700-12300nm	6	600nm			
MIVIS	SensTech	1100-1500nm	8	50nm	3-8m	2mrad	2.85km
		1900-2500nm	64	9nm			
		8200-12700nm	10	35-45nm			
HyMap	HyVista	400-2500nm	128	15-20nm	3-10m	2.5mrad	1.7-6km
		400-800nm	20	20nm			

Spaceborne Hyperspectral Sensors

Sensor (Satellite)	Altitude	Spectral Range	Band Number	Spectral Resolution	Spatial Resolution	IFOV	Swath
HIS (SIMSA)	523km	430-2400nm	220	20nm	25m	47.8urad	7.7km
Hyperion (EO-1)	705km	400-2500nm	220	10nm	30m	42.5urad	7.5km
CHRIS (PROBA)	580km	400-1050nm	19	1.25-11nm	25m	43.1urad	17.5km
MODIS (TERRA)	705km	400-1440nm	36	10-50nm	250-1000m	2000urad	2330km
HypSEO (MITA)	620km	400-2500nm	210	10nm	20m	40urad	20 km
Global Imager (ADEOS-2)	802km	380-1195nm	36	10-1000nm	250-1000m	310-1250urad	1600km
EnMAP	675km	420-1030nm	92	5-10nm	30nm	30urad	30km
		950-2450nm	108	10-20nm			
HyspIRI	700km	380-2500nm	200	10nm	60m	80urad	145km

2500nm) and 1 band (3000nm to 5000nm), and 6 bands (8000nm to 12600nm), respectively.

- ◇ Multispectral Infrared and Visible Imaging (MIVIS): Similar to DAIS-7915, MIVIS, which is a concurrent hyperspectral imaging system that operates from the visible to Thermal infrared ranges between 1100nm to 127000nm, covers three different wavelength ranges with 102 spectral channels.
- ◇ Hyperspectral Mapper (HyMap) is manufactured in Australia, yielding four spectrometers covering the spectral ranges of 400nm to 2500nm at a GSD of 5m. It is a well-known hyperspectral sensors that have been widely recognized in commercial circles.

More specifically, table 1 lists the parameter configuration of the above-mentioned hyperspectral sensors in terms of the operator, spectral coverage, the number of spectral bands, spectral resolution, spatial resolution, and the instantaneous field of view. Furthermore, some operational and upcoming spaceborne hyperspectral missions (satellites) are also summarized in Table 1 with more detailed characteristics.

2.1.3 Scanning Techniques of Hyperspectral Acquisition

From the sampling point of the perspective, there are five types of basic ways to acquire the hyperspectral cube in hyperspectral imaging, they are point scanning, spatial scanning, spectral scanning, non-scanning (snapshot hyperspectral imaging), and spatio-spectral joint scanning, respectively. Figure 2.2 visualizes the five different scanning techniques in a 3-D

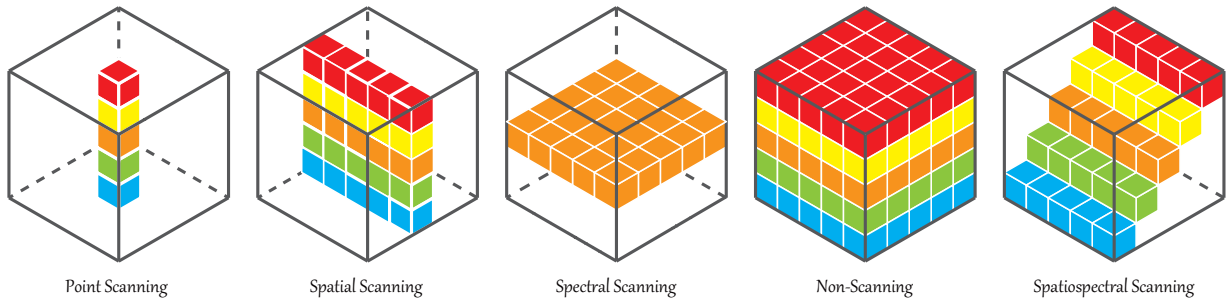


Fig. 2.2. An evolutionary process of scanning techniques in hyperspectral imaging: five toy examples, from left to right, corresponding to point scanning, spatial scanning, spectral scanning, non-spanning, and spatio-spectral scanning, respectively.

toy sample of the hyperspectral cube.

- **Point Scanning**, also known as whisk broom scanning, is a well-known technique of passive remote sensing from aircraft or spacecraft, which has been extensively applied to obtain the aerial and satellite imagery. This possible reason to interpret this phenomenon is that the single detector in the whisk broom scanner only allows one pixel access to the lighting source each time. Although the resulting satellite products hold a high spatial resolution, yet such costly moving strategy burdens the sensor. Figure 2.2 illustrates the imaging process.
- **Spatial Scanning** is a system of line-based scanning that uses the 2-D aperture sensor to obtain slit-like spectra. The line scanning system is operated with a push broom scanner, which can be viewed as a variant of a whisk broom scanner. Thanks to the wider receptive field and longer scanning time, the push broom strategy tends to capture more diversified light.
- **Spectral Scanning** is also part of line scanning, yet the main difference with spatial scanning is the spectrally scanning direction, which can be interpreted as a kind of spectral band-pass filters.
- **Non-Scanning** outputs the entire hyperspectral cube in one shot because of without any scanning operation. This greatly shortens the acquisition time of the image and meanwhile effectively avoids the motion artifacts caused by scanning. However, this is also a two-edged sword, since the snapshot benefits need to be supported by expensively computational cost.
- **Spatiospectral Scanning** overcomes the drawbacks of the above line-based scanning that only considers either spectrally or spatially moving direction at one time. By taking advantage of the dispersion technique, the scanning system is of benefit to generate the hyperspectral product of high spatial-spectral resolution.

It is worth mentioning that the specific application requirements guide the selection of scanning techniques.

2.1.4 Spectral Signature

Loosely speaking, spectral signature refers to the electromagnetic energy that is scattered, absorbed, transited and emitted from the surface of objects on the Earth, theoretically across any range of wavelengths. In hyperspectral imaging, the HSI is gathered with pixels each of which corresponds to a spectral signature that is quantified by vectors. The vector is a combinatorial radiance or reflectance, whose size is identical to the number of sampled spectral bands.

In view of the different reflectivity and absorptivity to various surface features on the ground, such as water, soil, forests and complex classes of land cover or land use, the detailed spectral signatures collected by hyperspectral sensors are capable of discriminating

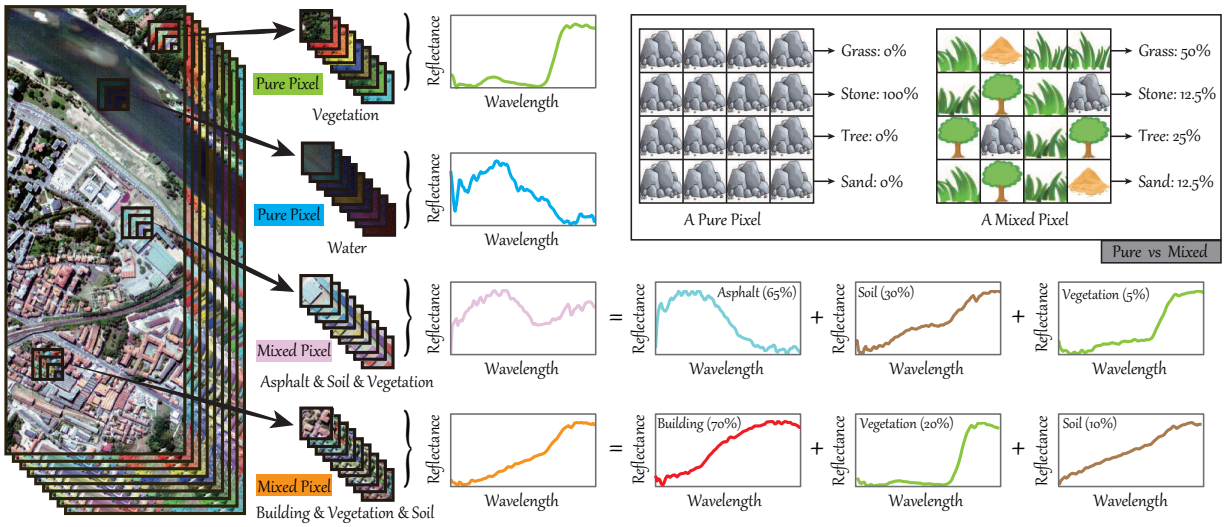


Fig. 2.3. A showcase in a real hyperspectral scene (Pavia City Centre) to quickly look at the concept of the hyperspectral image, spectral signature, and material mixture as well as pure pixel (endmember) and mixed pixel. The spectral signatures in the hyperspectral data are, as often as not, exhibited in the form of the reflectance, aiming to make the pixel spectral profiles comparable to some known materials. In the studied scene, the pure pixels correspond to two spectral reflectance curves of vegetation and water, respectively, while the mixed ones illustrate the case of spectral mixing, i.e. these mixed pixels consist of three components with different proportion. Furthermore, the right upper of the figure also gives two toy examples to explain the material miscibility.

and identifying the spectrally similar classes by capturing more subtle differences from the geometrically similar spectral shape. HSI can be usually seen as a stack of 2-D images continuously acquired in the spectral direction. Figure 2.3 shows several examples of visualizing spectral profiles in a real hyperspectral scene.

Besides, to correct the illumination, sensor devices, atmospheric effects, and solar and topographic compensation in the collection of remotely sensed digital images, radiometric calibration is an essential step in the data processing flow, yielding the calibrating spectral signatures.

2.1.5 Material Miscibility and Spectral Variability

Material mixing frequently occurs in hyperspectral imaging due to the inadequate spatial resolution in the image domain, or worse yet, intimate nonlinear interaction, which makes the recorded spectral signature commonly mixed at each pixel. This mixing behavior can be grouped into macroscopic mixing and microscopic mixing. Just as its name implies, the macroscopic mixture is an outcome by microscopically mixing multiple material components that come from the outside of the materials (or pixels), while for the microscopic mixture, the mixing process happens inside the materials (or pixels) in a nonlinear fashion. A showcase of material miscibility in a real city scenario with an illustration of toy examples (top-right corner) is given in Figure 2.3, where the mixing behavior happens in a pixel level and thus there are, more often than not, pure pixels and mixed pixels in a real-world hyperspectral scene. For the former, only one material exists in the real ground area, which means that its percentage or abundance is 100%. Whereas the latter pixels are usually made up of two and more materials at a given GSD, hence their proportion can be computed according to the actual ground meters, e.g., there are four materials involved in a HSI's pixel: *Grass*, *Stone*, *Tree*, *and Sand* with the percentages of 50%, 12.5%, 25%, and 12.5%, respectively, as shown in Figure 2.3.

Due to the existence of material mixing behavior in HSI, the variation in the spectral signature of material is inevitable. As shown in Figure 2.4, there is a visual example to clarify the spectral variability in a real hyperspectral scene. The factors may be multifarious, possibly

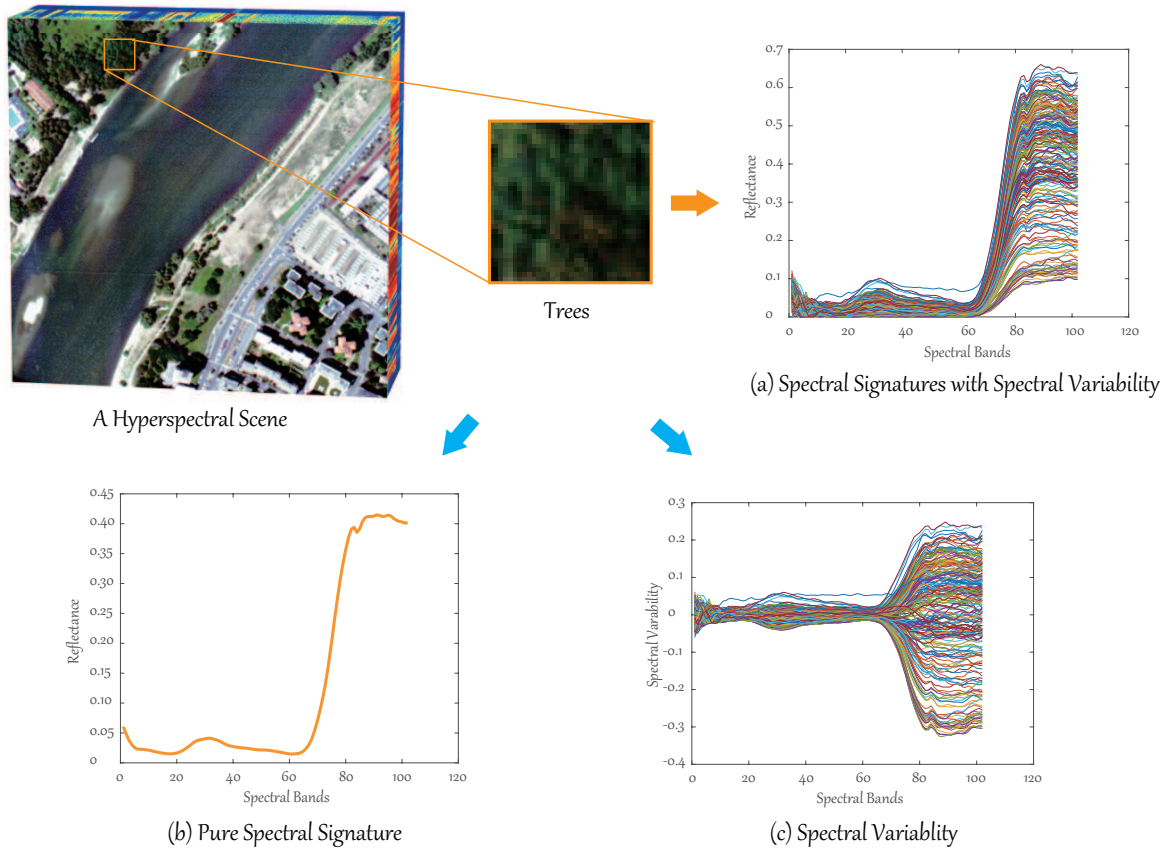


Fig. 2.4. A visual example of spectral variability in a real hyperspectral scene. A sub-area covering the trees is cropped to show the spectral variations in (a). (b) gives a smooth pure spectral signature for trees captured from the lab. It is clear to see from (c) the spectral variability between (a) and (b).

resulting from environmental, atmospheric, instrumental, physical or chemical effects. The spectral scaling, as a principal spectral variability, is frequently occurring, as the illumination conditions, which are sensitive to the elevation and azimuth of the lighting source, result in the deformations of the topography and the changes of roughness in materials. Another important factor to bring the spectral variability is an atmospheric impact on the reflection, scattering or absorption of the electromagnetic energy when encountering various gases (e.g., carbon dioxide, oxygen), aerosol particles, water vapour, dust, to name a few. As explained above, the intimately mixing is also a leading source of spectral variability, which is given rise to the microscopically multiple scatting between-in the material. As a result, the robust estimation techniques [Hong et al., 2016a, 2014b] are needed to address the challenges.

2.1.6 Applications and Role of Hyperspectral Remote Sensing in Earth Observation

Compared to on-site exploration, hyperspectral sensors record the information without the need to contact with the objects of interest. Together with the rich spectral information of HSI, hyperspectral remote sensing has gained growing attention in a wide range of applications, not limiting to remote sensing, but including

- ◇ *Atmospheric and Hydrological Monitoring:* Hyperspectral images have been viewed as a powerful tool to detect the changes, e.g., in estimating aerosol density, tracking pollution sources, mapping hydrological structure, analyzing gas constituents, and evaluating water quality.
- ◇ *Food Detecting:* Recently, food security has been a successful application of hyperspec-

tral imaging [Feng and Sun, 2012, Gowen et al., 2007]. For example, the HSI can be used to inspect the freshness of the fruit and the concentrations of pesticides, owing to its spectral information that goes beyond the visible spectrum.

- ◇ *Forensic Medicine:* Hyperspectral technique is apt to discover some tiny marks that are easy to be ignored, such as remaining bloodstain, fiber differences, yielding the great support in the criminal cases.
- ◇ *Medical Diagnose:* Highly spectral resolution makes it possible to timely detect the disease and obtain early treatment.
- ◇ *Energy Exploitation:* Hyperspectral data are widely applied in the detection of oil and toxic gas seeps, and it, on the other hand, also has the potential of exploiting the onshore and offshore petroleum, natural gas, minerals, and other energy sources.
- ◇ *Ecological Research:* Hyperspectral images have been proven to be effective for the biodiversity investigation in the forest-covering area [Ghiyamat and Shafri, 2010] and the biomass and carbon estimation [Dube and Mutanga, 2016, Karila et al., 2019], i.e. by the means of the HSI-based classification.
- ◇ *Urban Planning and Management:* Currently, a large number of researches have shown the HSI's superiority and effectiveness in a precise urban mapping (or classification) and change detection. This provides the researchers with a good foundation for the follow-up urban planning and management.

Still returning to the hyperspectral remote sensing of earth observation, the low spatial resolution and small-scale data collection have been two main factors to limit the HSI to be a dominant role, in spite of great benefits to various applications. Fortunately, the Sentinel-1 SAR and Sentinel-2 multispectral satellites in operation allow largely and even globally SAR and multispectral data of high spatial resolution to be available. This naturally might determine the role of the HSI that can become a significant complementary source to contribute to the large-scale earth observation tasks. That is to say, however, that the HSI is dispensable; on the contrary, its highly discriminative spectral information is a key to unlock the bottleneck that SAR, multispectral, or other data sources fail to classify or recognize the materials with fine-grained differences. As a result, this thesis not only presents the improvements targeting at some traditional challenges in hyperspectral data analysis, but also casts an interesting question related to cross-modality data analysis and proposes two advanced solutions.

2.2 Regression Techniques and Their Optimizers

Popularly speaking, the regression technique refers to utilizing the mathematically statistical method to measure or model the relations between dependent and independent variables. According to the causality of describing the two variables, the regression technique can be further divided into linear regression analysis and nonlinear regression analysis. Among them, the linear regression is a frequently-used approach in practice, due to its easy-to-use style and ability to generalize well, while the nonlinear one is used to deal with more complex nonlinear relationships and its solution is usually obtained by solving an approximate linear regression problem. Combining with the main focus of this thesis in hyperspectral data analysis, the following subsections will emphatically give priority to the linear regression-based techniques.

2.2.1 Linear Regression

From the machine learning perspective, the linear regression, a typical supervised learning technique, aims at learning a function or model that could be any of a line, a

plane or a higher dimensional hyperplane by a linearized combination of different attributes. The learned model is expected to minimize the errors between the predicted and real values, thereby better generalizing the out-of-sample. Given a pair-wise training set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_i, \mathbf{y}_i), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$ that contains the training samples $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m \in \mathbb{R}^{b\times m}$ with b dimensions (or bands) by m pixels (or samples) and corresponding class labels $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^m \in \mathbb{R}^{1\times m}$, $\mathbf{y}_i \in \{C_1, C_2, \ldots, C_k\}$, where k denotes the number of classes, the regression function or model $h_v(\bullet)$ can be written as

$$h_v(\mathbf{x}_i) = v_0 + v_1 x_{i1} + v_2 x_{i2} + \cdots + v_b x_{ib}, \tag{2.1}$$

making the to-be-estimated $h_v(\mathbf{x}_i)$ approach to $\mathbf{y}_i$. Eq. (2.1) can be also represented with vector as $h_v(\mathbf{x}_i) = \mathbf{V}\mathbf{x}_i$, or with matrix as $h_{\mathbf{V}}(\mathbf{X}) = \mathbf{V}\mathbf{X}$, where

$$\mathbf{V} = [v_0, v_1, v_2, \ldots, v_m] \in \mathbb{R}^{1\times b}, \qquad \mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}^{b\times m}, \tag{2.2}$$

and $\mathbf{x}_0 = [1, 1, 1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^{b\times 1}$. In order to assess the quality of the variable $\mathbf{V}$, we need to define a following loss function $\mathcal{J}$

$$\mathcal{J}(\mathbf{V}) = \frac{1}{2}\sum_{i}^{m}(h_v(\mathbf{x}_i) - \mathbf{y}_i)^2 = \frac{1}{2}\|\mathbf{V}\mathbf{X} - \mathbf{Y}\|_{\mathrm{F}}^2 = \frac{1}{2}(\mathbf{X}\mathbf{V} - \mathbf{Y})^{\mathrm{T}}(\mathbf{X}\mathbf{V} - \mathbf{Y}), \tag{2.3}$$

where $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm. There are many strategies in minimizing $\mathcal{J}(\mathbf{V})$ with the respect to the variable $\mathbf{V}$, written as $\min_{\mathbf{V}} \mathcal{J}(\mathbf{V})$, such as least squares or gradient descend, which are two commonly-used and effective algorithms.

Solution 1 – least squares: To facilitate the derivation, Eq. (2.3) can be unfolded as

$$\begin{aligned}\mathcal{J}(\mathbf{V}) &= \frac{1}{2}(\mathbf{X}\mathbf{V} - \mathbf{Y})^{\mathrm{T}}(\mathbf{X}\mathbf{V} - \mathbf{Y}) \\ &= \frac{1}{2}\left[\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V} - \mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{W} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right] \\ &= \frac{1}{2}\left[\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V} - 2\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right].\end{aligned} \tag{2.4}$$

If and only if the input matrix $\mathbf{X}$ is full rank ($m \gg b$), we then have the derivation of Eq. (2.3):

$$\frac{\partial\mathcal{J}(\mathbf{V})}{\mathbf{V}} = \mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{2.5}$$

Let Eq. (2.5) be equal to zeros, thus the variable $\mathbf{V}$ has an analytical solution of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ when $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible .

Solution 2 – gradient descent: In more general cases, the gradient descend is used to solve the problem (2.3) by searching the minimum. Note that due to the sensitivity to the initial point, the gradient descent algorithm could fall into a local minimum. More specifically, the gradient descend follows the following procedures:

- ◇ 1) initializing the variable $\mathbf{V}$ with randomization or zero vector;
- ◇ 2) updating the variable $\mathbf{V}$, making the value of $\mathcal{J}(\mathbf{V})$ reduced towards the direction of gradient descent, according to the rule:

$$\mathbf{V} \leftarrow \mathbf{V} - \alpha\frac{\partial\mathcal{J}(\mathbf{V})}{\mathbf{V}}, \tag{2.6}$$

where α is the predetermined step-size.

In this thesis, the full rank assumption of the matrix $\mathbf{X}$ is satisfied, thus the **Solution 1** is preferable.

2.2.2 Ridge (or Dense) Regression

In reality, multicollinearity exits extensively between the data due to certain highly correlated vectors (or columns) in the matrix, particularly when the matrix is approaching singularity. To avoid the trivial solution or overfitting issue, the aforementioned ill-posed problem can be steadily and reliably solved by adding a regularization term parameterized by λ. This leads to a least-squares optimization problem with Tikhonov regularization, which can be formulated as

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{V}\|_{\mathrm{F}}^2, \tag{2.7}$$

whose closed-form solution is

$$\mathbf{V} \leftarrow (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{2.8}$$

Similarly, the variable $\mathbf{V}$ in Eq. (2.7) can be optimized by gradient descent method with a modified update rule:

$$\mathbf{V} \leftarrow \mathbf{V}(1 - \alpha\lambda) - \alpha\frac{\partial \mathcal{J}(\mathbf{V})}{\mathbf{V}}, \tag{2.9}$$

where the penalty parameter λ controls the model's complexity. As λ increases gradually, the absolute value of each element in the variable $\mathbf{V}$ tends to uninterruptedly decrease, further yielding a growing deviation relative to the actual $\mathbf{V}$. This process could lead to a well-known underfitting, conversely call overfitting. Therefore, a proper λ may assist the model to reach a balance between robustness and fitting ability. In addition, due to the characteristic of the $\mathcal{L}_2$-norm, that is, each element in the estimated $\mathbf{V}$ is a contributor to the data fitting, the ridge regression is also called as dense regression or representation [Jiang and Lai, 2015].

2.2.3 Lasso (or Sparse) Regression

In ridge regression, the $\mathcal{L}_2$-norm constraint shrinks the to-be-estimated coefficients of the variable $\mathbf{V}$ to a value close to zero but not exactly zero, which brings the difficulties in model understanding to a great extent, or in other words, is lack of physical meaning to explicitly guide the variable (or feature) selection. What's more, since the ridge regression has to estimate all elements of the variable $\mathbf{V}$, even though a very small number, it still yields a computationally expensive cost. A straightforward and effective way to address the two problems is the least absolute shrinkage and selection operator (Lasso). By using $\mathcal{L}_1$-norm penalty constraint in place of $\mathcal{L}_2$-norm term in ridge regression, also known as sparse regression [Bioucas-Dias et al., 2012], the resulting Lasso regression can be represented in the form of a matrix as

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{V}\|_{1,1}, \tag{2.10}$$

where $\|\mathbf{X}\|_{1,1}$ is defined as the sum of absolute values of each element of the matrix $\mathbf{V}$, formalized by $\sum_{i=1}^{m}\|\mathbf{x}_i\|_1$. Such $\mathcal{L}_1$-norm may make many coefficients accurately converged to zeros and reduce the effects of multicollinearity. Especially in the case – small-size samples but high feature (variable) dimension, Lasso has been proven to be effective for high-dimensional statistic analysis by reducing the variations of the model and meanwhile increasing its regression precision.

Considering the non-differentiable points in $\mathcal{L}_1$-norm, the solvers, whether it be least-squares or gradient descend, lose their functions. There have been some tailored optimization algorithms proposed for solving the Lasso problem, such as coordinate descent (CD),

Algorithm 1 ADMM-based solver to Lasso (sparse) regression

Input: $\mathbf{X}$, $\mathbf{Y}$, and regularization parameter λ, *maxIter*
Output: the transformation vector $\mathbf{V}$
Initialize: $\mathbf{Z}^1 = \mathbf{V}^1 = \mathbf{0}$, $\mu^1 = 10^{-3}$, $\mu_{\max} = 10^6$, $\rho = 1.5$, $t = 1$, $\zeta = 10^{-4}$

while not converged or $t > maxIter$ **do**
  Fix other variables to update $\mathbf{V}^{t+1}$ by solving a least-squares problem with Tikhonov regularization
  Fix other variables to update $\mathbf{Z}^{t+1}$ by Eqs. (2.11) and (2.12)
  Update Lagrange multipliers by $\mathbf{\Lambda}^{t+1} \leftarrow \mathbf{\Lambda}^t + \mu^t(\mathbf{Z}^{t+1} - \mathbf{V}^{t+1})$
  Update penalty parameter by $\mu^{t+1} = \min(\rho\mu^t, \mu_{\max})$
  Check the convergence condition:
  if $\|\mathbf{V}^{t+1} - \mathbf{Z}^{t+1}\|_{\mathrm{F}} < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
    Break;
  end if
end while

least angle regression (LARS), proximal gradient descent (PGD), quadratic programming (QP), etc. These methods basically follow the iterative strategy and differ primarily in the heuristic mode, i.e. CD walks along the coordinate direction and LARS seeks to find the next-step points by maximizing the correlations with Cosine distance (corresponding to the least angle). Inspired by the iterative thresholding (IT) [Wright et al., 2009] that splits the problem (2.10) into two subproblems and then the sparse part ($\mathcal{L}_1$-norm: $\|\mathbf{V}\|_{1,1}$) can be updated with a well-known *soft-thresholding (shrinkage)* operator [Chen et al., 2001]:

$$\mathbf{S} \leftarrow \max\{\mathbf{0}, \|\mathbf{V} - \mathbf{\Lambda}/\mu\|_1 - \lambda/\mu\} sign(\mathbf{V} - \mathbf{\Lambda}/\mu), \tag{2.11}$$

where $sign(\bullet)$ is defined by

$$sign(\bullet) = \begin{cases} 1\ , & \bullet \geq 0 \\ -1, & \bullet < 0, \end{cases} \tag{2.12}$$

the sparse regression problem in Eq. (2.10) can be effectively solved by embedding the *soft-thresholding (shrinkage)* operator into a general optimization framework based on the alternating direction method of multipliers (ADMM) [Bioucas-Dias and Figueiredo, 2010]. The general form of ADMM-based optimization problem is

$$\min_{\mathbf{V}} f(\mathbf{V}) + g(\mathbf{V}), \tag{2.13}$$

where $\mathbf{V}$ is the model parameter; $f(\mathbf{V})$ and $g(\mathbf{V})$ denote the loss function and the regularization term (i.e. $g(\mathbf{V}) = \|\mathbf{V}\|_{1,1}$), respectively. Separating the $g(\mathbf{V})$ from the overall objective function, the Eq. (2.13) can be then rewritten by replacing the variable $\mathbf{V}$ of $g(\mathbf{V})$ with a new variable $\mathbf{Z}$:

$$\min_{\mathbf{V}} f(\mathbf{V}) + g(\mathbf{Z}), \ \text{s.t.}\ \mathbf{V} - \mathbf{Z} = \mathbf{0}. \tag{2.14}$$

To relax the equality constraints of Eq. (2.14), the corresponding augmented Lagrangian function can be written as

$$\mathscr{L}_D(\mathbf{V}, \mathbf{Z}, \mathbf{\Lambda}) = f(\mathbf{V}) + g(\mathbf{Z}) + \mathbf{\Lambda}^{\mathrm{T}}(\mathbf{V} - \mathbf{Z}) + \frac{\mu}{2}\|\mathbf{V} - \mathbf{Z}\|_{\mathrm{F}}^2, \tag{2.15}$$

where $\mathbf{\Lambda}$ is the Lagrangian multiplier and $\mu > 0$ is the penalty factor. Thereby, the variables

of Eq. (2.15) can be updated in $(t+1)$-th iteration by solving the following subproblems:

$$\begin{aligned}
\mathbf{V}^{t+1} &:= \arg\min_{\mathbf{V}} f(\mathbf{V}) + \mathbf{\Lambda}^{\mathrm{T}}(\mathbf{V} - \mathbf{Z}^t) + \frac{\mu}{2}\|\mathbf{V} - \mathbf{Z}^t\|_{\mathrm{F}}^2 \\
\mathbf{Z}^{t+1} &:= \arg\min_{\mathbf{Z}} g(\mathbf{Z}) + \mathbf{\Lambda}^{\mathrm{T}}(\mathbf{V}^{t+1} - \mathbf{Z}) + \frac{\mu}{2}\|\mathbf{V}^{t+1} - \mathbf{Z}\|_{\mathrm{F}}^2 \\
\mathbf{\Lambda}^{t+1} &:= \mathbf{\Lambda}^t + \mu(\mathbf{V}^{t+1} - \mathbf{Z}^{t+1}).
\end{aligned} \tag{2.16}$$

Finally, these iterative procedures will be repeated until a stopping criterion is satisfied. **Algorithm 1** details the iterative procedures for solving the Lasso (or sparse) regression problem.

2.2.4 Low-rank Regression

Although the sparsity plays a role in selecting the dominant features (or attributes), yet it is more like a *hard* process by rudely removing other features that are correlated with the target feature rather than structurally considering the attribute dependencies. To this end, the low-rank regression [Su et al., 2015] was developed to find a low-rank transformation vector. Through such a transformation, those correlated samples expect to be collaboratively represented in a grouping fashion. The resulting regression is capable of effectively capturing the correlations of intrinsic structure between samples. Mathematically, the process can be modeled as follows:

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{V}\|_*, \tag{2.17}$$

where $\|\bullet\|_*$ denotes the nuclear norm, which represents the sum of singular values of a given matrix. This term can be estimated via a so-called singular value thresholding (SVT) operator [Liu et al., 2013]:

- ◇ Step 1. Given a matrix $\mathbf{M}$ with rank r, the singular value decomposition (SVD) is first performed by
$$\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}, \quad \mathbf{S} = \mathrm{diag}(\{s_k\}_{1\le k\le r}). \tag{2.18}$$
- ◇ Step 2. For each $\tau \ge 0$, the soft-thresholding operator $\mathcal{D}_\tau$ is defined as
$$\mathcal{D}(\mathbf{M}) := \mathbf{U}\mathcal{D}_\tau(\mathbf{S})\mathbf{V}, \quad \mathcal{D}_\tau(\mathbf{S}) = \mathrm{diag}(\{s_k - \tau\}^+). \tag{2.19}$$
- ◇ Step 3. Using Eq. (2.19), $\|\mathbf{M}\|_*$ can be computed by $\|\mathcal{D}_\tau(\mathbf{S})\|_{1,1}$.

Likewise, the SVT operator can be also integrated into the ADMM optimization framework by effectively splitting the non-smooth or non-convex original problem into several smooth and convex subproblems. By replacing the variable $\mathbf{V}$ of the term $\|\mathbf{V}\|_*$ with an auxiliary variable $\mathbf{Z}$, an equivalent from of Eq. (2.17) can be written as

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{Z}\|_*, \ \text{s.t.} \ \mathbf{V} - \mathbf{Z} = \mathbf{0}, \tag{2.20}$$

we therefore have the following augmented Lagrangian function by introducing the Lagrangian multiplier $\mathbf{\Lambda}$ and an updatable penatly parameter μ:

$$\min_{\mathbf{V}} \frac{1}{2}\|\mathbf{Y} - \mathbf{V}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{Z}\|_* + \mathbf{\Lambda}^{\mathrm{T}}(\mathbf{V} - \mathbf{Z}) + \frac{\mu}{2}\|\mathbf{V} - \mathbf{Z}\|_{\mathrm{F}}^2, \tag{2.21}$$

whose detailed optimization flow is summarized in **Algorithm 2**.

Algorithm 2 ADMM-based solver to low-rank regression

Input: $\mathbf{X}$, $\mathbf{Y}$, and parameters: λ, $\mu_{\max} = 10^6$, $\zeta = 10^{-4}$, $\rho = 1.5$, $t = 1$, *maxIter*
Output: the transformation vector $\mathbf{V}$
Initialize: $\mathbf{Z}^1 = \mathbf{V}^1 = \mathbf{0}$, $\mathbf{\Lambda}^1 = \mathbf{0}$, $\mu^1 = 10^{-3}$

while not converged or $t >$ *maxIter* **do**
  Fix other variables to update $\mathbf{V}^{t+1}$ by solving a least-squares problem with Tikhonov regularization
  Fix other variables to update $\mathbf{Z}^{t+1}$ using SVT operator as shown in Eqs. (2.18) and (2.19)
  Update Lagrange multipliers by $\mathbf{\Lambda}^{t+1} \leftarrow \mathbf{\Lambda}^t + \mu^t(\mathbf{Z}^{t+1} - \mathbf{V}^{t+1})$
  Update penalty parameter by $\mu^{t+1} = \min(\rho\mu^t, \mu_{\max})$
  Check the convergence condition:
  if $\|\mathbf{V}^{t+1} - \mathbf{Z}^{t+1}\|_{\mathrm{F}} < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
    Break;
  end if
end while

2.2.5 Joint Regression

Latent subspace learning (LSL) provides us a new insight to investigate the regression techniques. The main idea is to convert the regression problem in the high dimensional space into one in a latent low-dimensional subspace. The benefits of the scheme are two-fold. On the one hand, it prevents, to a larger extent, overfitting of the input data compared to the previous techniques. Intuitively speaking, the subspace is more robust to noises and outliers than original high dimensional space. On the other hand, it excavates the intrinsic attributes of the data more effectively and efficiently, since it is obvious that the data with the lower dimension make it easier for the statistical regularity to be discovered. One representative joint regression that performs subspace learning and linear regression simultaneously [Ji and Ye, 2009] can be modeled with an expected output of $\mathbf{\Theta X}$:

$$\min_{\mathbf{V},\mathbf{\Theta}} \frac{1}{2}\|\mathbf{Y}_l - \mathbf{V\Theta X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{V}\|_{\mathrm{F}}^2, \ \text{s.t.} \ \mathbf{\Theta\Theta}^{\mathrm{T}} = \mathbf{I}, \tag{2.22}$$

where the $\mathbf{Y}_l \in \mathbb{R}^{k\times m}$ is defined as the one-hot encoded matrix of $\mathbf{Y}$, i.e.

$$\mathbf{Y}_l = \begin{bmatrix} 1\ 0\ \ldots\ 0\ \ldots\ 0 \\ 0\ 1\ \ldots\ 0\ \ldots\ 0 \\ \quad \ldots\ \ldots\ \ldots \\ 0\ 0\ \ldots\ 1\ \ldots\ 0 \\ \quad \ldots\ \ldots\ \ldots \\ 0\ 0\ \ldots\ 0\ \ldots\ 1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ \ldots \\ j \\ \ldots \\ k \end{matrix} \qquad when \quad \mathbf{Y} = [1, 2, \ldots, j, \ldots, k], \tag{2.23}$$

and $\mathbf{\Theta} \in \mathbb{R}^{d\times b}$ (d is the dimension of the latent subspace) denotes the latent subspace projections, respectively. Please note that the herein variable $\mathbf{V} \in \mathbb{R}^{k\times d}$ is a transformation matrix linking the latent subspace with the encoded label information.

Considering the nonconvexity of the Eq. (2.22), an iterative optimization strategy is adopted to alternatively solve the convex subproblems with respect to each variable $\mathbf{V}$ and $\mathbf{\Theta}$. The

Algorithm 3 ADMM-based solver with respect to the variable $\boldsymbol{\Theta}$

Input: $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{V}$, and parameters: $\mu_{\max} = 10^6$, $\rho = 1.5$, $t = 1$, $\zeta = 10^{-4}$, *maxIter*
Output: the transformation vector $\boldsymbol{\Theta}$
Initialize: $\boldsymbol{\Theta}^1 = \mathbf{G}^1 = \mathbf{0}$, $\mathbf{J}^1 = \boldsymbol{\Theta}\mathbf{X}$, $\boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \mathbf{0}$, $\mu^1 = 10^{-3}$

while not converged or $t >$ *maxIter* **do**
  Fix other variables to update $\mathbf{J}^{t+1}$ by $(\mathbf{V}^{\mathrm{T}}\mathbf{V} + \mu^t\mathbf{I})^{-1}(\mathbf{V}^{\mathrm{T}}\mathbf{Y} + \mu^t\boldsymbol{\Theta}^t\mathbf{X} - \boldsymbol{\Lambda}_1^t)$
  Fix other variables to update $\boldsymbol{\Theta}^{t+1}$ by $(\mu^t\mathbf{J}^{t+1}\mathbf{X}^{\mathrm{T}} + \boldsymbol{\Lambda}_1^t\mathbf{X}^{\mathrm{T}} + \mu^t\mathbf{G}^t + \boldsymbol{\Lambda}_2^t) \times (\mu^t\mathbf{X}\mathbf{X}^{\mathrm{T}} + \mu^t\mathbf{I})^{-1}$
  Fix other variables to update $\mathbf{G}^{t+1}$ by $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{svd}(\boldsymbol{\Theta}^{\mathbf{t+1}} - \boldsymbol{\Lambda}_2^{\mathbf{t}}/\mu^{\mathbf{t}})$, $\mathbf{G} = \mathbf{U}\mathbf{I}_{n\times m}\mathbf{V}$.
  Update Lagrange multipliers by
$$\boldsymbol{\Lambda}_1^{t+1} \leftarrow \boldsymbol{\Lambda}_1^t + \mu^t(\mathbf{J}^{t+1} - \boldsymbol{\Theta}^{t+1}\mathbf{X}), \ \ \boldsymbol{\Lambda}_2^{t+1} \leftarrow \boldsymbol{\Lambda}_2^t + \mu^t(\boldsymbol{\Theta}^{t+1} - \mathbf{G}^{t+1})$$
  Update penalty parameter by $\mu^{t+1} = \min(\rho\mu^t, \mu_{\max})$
  Check the convergence condition:
  if $\|\mathbf{J}^{t+1} - \boldsymbol{\Theta}^{t+1}\mathbf{X}\|_{\mathrm{F}} < \zeta$ and $\|\mathbf{G}^{t+1} - \boldsymbol{\Theta}^{t+1}\|_{\mathrm{F}} < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
    Break;
  end if
end while

subproblem for the variable $\mathbf{V}$ is straightforward to derive that

$$\mathbf{V} = (\mathbf{Y}_l\mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}})(\boldsymbol{\Theta}\mathbf{X}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Theta}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}. \tag{2.24}$$

When the variable $\mathbf{V}$ is fixed, the optimization problem of the variable $\boldsymbol{\Theta}$ is

$$\min_{\boldsymbol{\Theta}} \frac{1}{2}\|\mathbf{Y}_l - \mathbf{V}\boldsymbol{\Theta}\mathbf{X}\|_{\mathrm{F}}^2, \ \text{s.t.} \ \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}} = \mathbf{I}. \tag{2.25}$$

To facilitate the effective use of ADMM optimizer, the Eq. (2.26) is converted to the corresponding augmented Lagrangian version by introducing two additional auxiliary variables $\mathbf{J}$ and $\mathbf{G}$ in place of $\boldsymbol{\Theta}\mathbf{X}$ and $\boldsymbol{\Theta}$, respectively, as follows:

$$\min_{\boldsymbol{\Theta},\mathbf{J},\mathbf{G}} \frac{1}{2}\|\mathbf{Y}_l - \mathbf{V}\mathbf{J}\|_{\mathrm{F}}^2 + \boldsymbol{\Lambda}_1^{\mathrm{T}}(\mathbf{J} - \boldsymbol{\Theta}\mathbf{X}) + \frac{\mu}{2}\|\mathbf{J} - \boldsymbol{\Theta}\mathbf{X}\|_{\mathrm{F}}^2 + \boldsymbol{\Lambda}_2^{\mathrm{T}}(\mathbf{G} - \boldsymbol{\Theta}) + \frac{\mu}{2}\|\mathbf{G} - \boldsymbol{\Theta}\|_{\mathrm{F}}^2, \ \text{s.t.} \ \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I}, \tag{2.26}$$

which can be effectively solved by implementing the **Algorithm 3**. Please refer to [Hong et al., 2019d] for more details regarding the solution of joint regression.

3 State-of-the-art in Hyperspectral Data Analysis

This chapter revisits the hyperspectral data analysis from the aspects of elaborating the state-of-the-art methodology, mainly focusing on three primary topics – hyperspectral dimensionality reduction (HDR), spectral unmixing (SU), and multi-modality data analysis (MMDA). More specifically,

- ◇ according to the different usage strategies of label information to lessen the highly spectral correlations with the expectancy of preserving the desirable information as discriminatory as possible, the **first topic** can be sectionalized into three parts: unsupervised HDR, supervised HDR, and semi-supervised HDR;
- ◇ although the spectral profiles are assumed to be linearized mixture in most cases, yet there still exist some more complex scenarios that need to be nonlinearly unmixed and analyzed in a real-world application. Correspondingly, we review the **second topic** from perspectives of linear and nonlinear mixing models;
- ◇ theoretically, "perfect" data source that can independently cope with various challenges or tasks is unattainable. Therefore, the union of two or more complementary modalities could equivalently achieve the same goals to some extent. In the **third topic**, many advanced types of research on the use of multi-modalities are investigated with the applications to HSI-related tasks, as well as some tentative works related to cross-modality fusion and learning are introduced by using hyperspectral data.

3.1 Hyperspectral Dimensionality Reduction

HSI can be considered as a set of 2-D images with hundreds of spectral channels, enabling them to detect the objects or identify the materials of interest easier. However, information overload and redundancy bring great challenges to data storage and representation capabilities. For example, *the curse of dimensionality* [Indyk and Motwani, 1998] is often accompanied by the hyperspectral data, inevitably leading to the performance degradation with the explosive increase of the spectral dimensions. This might be explained by the highly coupled and correlated spectral bands. To alleviate the effects, dimensionality reduction is usually performed as an important preprocessing step prior to high-level data analysis, which has received increasing attention in the hyperspectral field. Roughly speaking, the dimensionality reduction technique can be categorized into feature selection and feature transformation.

The feature selection of hyperspectral data, also called as band selection, screens out a subset of hyperspectral bands from original spectral bands by maximizing or minimizing certain criteria, e.g., information entropy [Koller and Sahami, 1996], correlations [Hall, 1999], maximum likelihood [Riedmann and Milton, 2003], and rough set-based feature selection [Patra et al., 2015], in order to reduce the redundancy between those adjacent bands as much as possible. Its advantage is obvious that the selected bands maintain the physical meaning of the original bands with the highest possibility, while its disadvantage is also not negligible, that is, in hyperspectral imaging, the spectral information in each band is recorded by coupling surrounding band information rather than being independent each other. This leads to an incomplete spectral information separation.

An alternative strategy is the transformation-based feature extraction that compresses the HSI to a low-dimensional subspace through a learnable projection or transformation matrix. In this process, the coupling characteristics of between-in bands can be removed. Therefore, a comprehensive survey is made to give the readers a big picture of the transformation-based HDR techniques. Depending on different learning strategies, in the first topic, HDR

techniques are progressively unfolded in sequential order of unsupervised, supervised, and semi-supervised learning.

3.1.1 Priority-driven Unsupervised Dimensionality Reduction

Although the surprisingly improving capability in data collection and storage has shown the possibility in large-scale and high-performance computing, yet the dimensionality reduction techniques have been a vibrant field in either the general data science or the particular hyperspectral data analysis. Unsupervised dimensionality reduction, as one of the main focuses on HDR, has been developed through decades of research to make the hyperspectral data slightly redundant along the spectral direction. There are two main streams in the unsupervised dimensionality reduction of HSI. One follows some statistics assumptions, i.e. principal component analysis (PCA) [Jolliffe, 2011], factor analysis (FA) [Thompson, 2007], independent component analysis (ICA) [Comon, 1994].

Statistical Analysis

The classic PCA is a widely-used and user-friendly data compression method that extracts the principal linearized components by rotating the original coordinate to a new system. In the new coordinate system whose axis directions are determined by maximizing the variances, the transformed data are expected to be linearly independent between the variables, which can be formulated as

$$\mathbf{Z} = \mathbf{V}\mathbf{X}, \tag{3.1}$$

where $\mathbf{Z}$ denotes the dimension-reduced subspace, $\mathbf{X}$ is the corresponding high-dimensional matrix representation, e.g., unfolded hyperspectral data, and the variable $\mathbf{V}$ represents the to-be-estimated mapping or projection. In PCA, the principal component orientations of maximum difference can be measured by computing the covariance matrix of the input matrix. Given two random variables $\mathbf{A}$ and $\mathbf{B}$ with the dimensions (b) being equal to 2 and the number of samples m, their covariance matrix is defined as

$$\begin{aligned} Cov(\mathbf{A},\mathbf{B}) &= \frac{1}{b-1}\sum_{i=1}^{b}(\mathbf{A}_i-\mu_A)(\mathbf{B}_i-\mu_B) \\ &= E\left[(\mathbf{A}-E(\mathbf{A}))(\mathbf{B}-E(\mathbf{B}))\right] \\ \mu_A &= \frac{1}{b}\sum_{i=1}^{b}\mathbf{A}_i \\ \mu_B &= \frac{1}{b}\sum_{i=1}^{b}\mathbf{B}_i, \end{aligned} \tag{3.2}$$

where μ_A and μ_B are the mean values with respect to the variables $\mathbf{A}$ and $\mathbf{B}$, respectively, and E stands for the expectation. Note that a greater absolute value of $Cov(\mathbf{A},\mathbf{B}) \in (-1,1)$ indicates a higher correlation. However, when the variable dimension turns into 1-D, the covariance matrix then degrades to the variance ($\mathbf{S}(\mathbf{A})$), e.g.,

$$\begin{aligned} S(\mathbf{A})^2 = Cov(\mathbf{A},\mathbf{A}) &= \frac{1}{b-1}\sum_{i=1}^{b}(\mathbf{A}_i-\mu_A)^2 \\ &= E\left[(\mathbf{A}-E(\mathbf{A}))\right]. \end{aligned} \tag{3.3}$$

And so on, for higher dimensions (≥ 3), e.g., 3-D data ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$), the covariance matrix can

be represented as follows:

$$Cov(\mathbf{A},\mathbf{B},\mathbf{C}) = \begin{bmatrix} Cov(\mathbf{A},\mathbf{A}) & Cov(\mathbf{A},\mathbf{B}) & Cov(\mathbf{A},\mathbf{C}) \\ Cov(\mathbf{B},\mathbf{A}) & Cov(\mathbf{B},\mathbf{B}) & Cov(\mathbf{B},\mathbf{C}) \\ Cov(\mathbf{C},\mathbf{A}) & Cov(\mathbf{C},\mathbf{B}) & Cov(\mathbf{C},\mathbf{C}) \end{bmatrix}. \tag{3.4}$$

More specifically, PCA can be performed by the following steps:

1) the input data $\mathbf{X}$ are centralized by subtracting their mean value, that is,

$$\overline{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i. \tag{3.5}$$

2) Using Eq. (3.2), the covariance matrix $\mathbf{C}$ of input data $\mathbf{X}$ is

$$\begin{aligned} \mathbf{C} &= \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{\mathbf{x}}_i)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \ldots \mathbf{x}_i^{\mathrm{T}} \end{aligned} \tag{3.6}$$

3) Suppose the projections or transformations be $\mathbf{V}$, the variance of projected hyperplane can be given by

$$\begin{aligned} S^2 &= \frac{1}{m}\sum_{i=1}^{m}(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}} \\ &= \mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V}. \end{aligned} \tag{3.7}$$

4) The projection matrix can be estimated by maximizing the scatter matrix, namely Eq. (3.7), hence the optimization problem is written as

$$\arg\max_{\mathbf{V}} \mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V}, \;\; s.t. \;\; \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}. \tag{3.8}$$

5) The Eq. (3.9) can be converted to an unconstrained optimization problem using Lagrange multiplier method:

$$\arg\max_{\mathbf{V}} \mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V} - \alpha(\mathbf{V}\mathbf{V}^{\mathrm{T}} - \mathbf{I}), \tag{3.9}$$

which is taken the derivative for the variable $\mathbf{V}$. Let the derivative be zeros, then we have $\mathbf{C}\mathbf{V} = \alpha\mathbf{V}$, which can be deduced by a generalized eigenvalues decomposition (GED).

6) Finally, rank eigenvalues from large to small and select k eigenvectors corresponding to the k maximum eigenvalues as dimension-reduced data.

Over the past decades, a leaf-style development on PCA-based approaches has been made in HDR. [Rodarmel and Shan, 2002] introduced the PCA into the hyperspectral community, achieving an effective image classification. Considering the nonlinearity of HSI, [Fauvel et al., 2009] performed the hyperspectral data classification by projecting the original data into a higher-dimensional kernel-induced space. In [Licciardi et al., 2012], authors developed a nonlinear PCA for hyperspectral feature extraction with a quantitative comparison with the traditional PCA approach. Some advanced extensions of PCA-based schemes have been successively proposed for hyperspectral feature extraction and dimensionality reduction, such as folded-PCA [Zabalza et al., 2014], probabilistic PCA [Xia et al., 2014], robust PCA [Sun and Du, 2018].

Another representative statistics-based dimensionality reduction approach is ICA. Superior to PCA, ICA can model the data of non-Gaussian distribution. Given an observed signal $\mathbf{X}$, ICA assumes that the observed signal is a linear combination of a set of independent signals ($\mathbf{S}$). Let $\mathbf{A}$ be the combinational coefficients, the resulting expression of ICA is

$$\mathbf{X} = \mathbf{AS}, \tag{3.10}$$

this is a classic blind source separation problem, which means that the variables $\mathbf{A}$ and $\mathbf{S}$ need to estimated simultaneously. According to Eq. (3.10), the ICA's dimensionality reduction model can be obtained as

$$\mathbf{S} = \mathbf{WX}, \ \ s.t. \ \ \mathbf{W} = \mathbf{A}^{-1}, \tag{3.11}$$

where $\mathbf{W}$ represents the inverse matrix or generalized inverse matrix of the matrix $\mathbf{A}$. In this problem, the probability density function of $i\text{-}th$ source signal is assumed to be $p(\mathbf{S}_i)$, thus the joint probability distribution is expressed by

$$p(\mathbf{S}) = \prod_{i=1}^{b} p(\mathbf{S}_i). \tag{3.12}$$

By bringing the Eq. (3.11) into the Eq. (3.12), we then have

$$p(\mathbf{X}) = p(\mathbf{WX})|\mathbf{W}| = |\mathbf{W}| \prod_{i=1}^{b} p(\mathbf{W}_i\mathbf{X}), \tag{3.13}$$

where sigmoid function has been proven to be a good choice of cumulative distribution function (CDF), thereby $p(\mathbf{S})$ is the derivation of the sigmoid function:

$$p(\mathbf{S}) = g'(\mathbf{S}) = \frac{e^{\mathbf{S}}}{(1+e^{\mathbf{S}})^2}. \tag{3.14}$$

To meet the independence between each $\mathbf{S}_i$, the formulation of the likelihood function with respect to $\mathbf{W}$ is

$$L(\mathbf{W}) = \sum_{i=1}^{m} \left(\sum_{j=1}^{b} \log g'(\mathbf{W}_j\mathbf{x}_i) + \log|\mathbf{W}| \right). \tag{3.15}$$

The Eq. (3.15) can be effectively solved via stochastic gradient descent method, hence the

iterative formula of the variable $\mathbf{W}$ is

$$\mathbf{W} := \mathbf{W} + \alpha \left(\begin{bmatrix} 1 - 2g(\mathbf{W}_1 \mathbf{x}_i) \\ 1 - 2g(\mathbf{W}_2 \mathbf{x}_i) \\ \cdots \\ 1 - 2g(\mathbf{W}_b \mathbf{x}_i) \end{bmatrix} \mathbf{x}_i^{\mathrm{T}} + (\mathbf{W}^{\mathrm{T}})^{-1}) \right), \tag{3.16}$$

where α denotes the step length of gradient descent.

In light of the superiority in modeling the hyperspectral data, ICA has played a significant role in hyperspectral data processing and analysis, especially in HDR. [Wang and Chang, 2006] proposed to treat the ICA as a tool of dimensionality reduction with the application to hyperspectral image analysis. The work presented in [Dalla Mura et al., 2011] jointly utilized extended morphological attribute profiles and ICA for hyperspectral image classification. Moreover, some ICA-based extensions and variants have been proposed for a wide range of applications, e.g., kernel ICA [Khan et al., 2009], discriminant ICA [Villa et al., 2011], ICA with edge-preserving filtering [Xia et al., 2016], randomized ICA [Jayaprakash et al., 2018], etc.

In addition to the groups of PCA and ICA, there are the other types of tensor-based modeling methods in HDR. The researchers in [Renard and Bourennane, 2009] conducted the dimensionality reduction by directly considering the hyperspectral data as a 3-D tensor structure rather than an unfolded 2-D matrix. A similar research [Karami et al., 2012] explored the wavelet transform and tucker decomposition for hyperspectral image compression. Further, a patch-wise tensor decomposition with the low-rank constraint [Du et al., 2017] was proposed to reconstruct the hyperspectral images.

Graph Embedding

Graph embedding uncovers the intrinsic structure of the data and assumes to be the existence of a low-dimensional manifold that shares the same or similar structure with the high-dimensional data. In practice, the assumption can be achieved by preserving the locally neighboring relationships of each point within the data. From this respect, the general embedding is a more suitable technique to conduct the HDR, since it is capable of capturing the underlying topology of the data that lies in the more complex real world. Recently, there exist massive related approaches in the task of dimensionality reduction, such as ISOMAP [Balasubramanian and Schwartz, 2002], locally linear embedding (LLE) [Huang et al., 2019, Roweis and Saul, 2000], Laplacian eigenmaps (LE) [Belkin and Niyogi, 2003], and their linearized versions: locality preserving projections (LPP) [He and Niyogi, 2004] and neighborhood preserving embedding (NPE) [He et al., 2005]. Before revisiting these methods in detail, a general graph embedding (GGE) framework presented in [Yan et al., 2007] is first introduced and formulated as follows:

$$\min_{\mathbf{Z}} \sum_{i \neq j} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \mathbf{W}^{i,j} = \min_{\mathbf{Z}} \mathrm{tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}), \ \text{s.t.} \ \mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}, \tag{3.17}$$

where the $\mathbf{L}$ is the Laplacian matrix computed by $\mathbf{D} - \mathbf{W}$; each element of the diagonal matrix $\mathbf{D}$ is defined as $D_{ii} = \sum_{i \neq j} \mathbf{W}^{i,j}$, while $\mathbf{W}$ represents the affinity (or adjacency) matrix. The goal of this graph embedding model is to find or discover a low-dimensional representation $\mathbf{Z}$ that is able to capture the high-dimensional manifold and preserve it in a low-dimensional space. Generally, GGE performs the dimensionality reduction with three main steps:

1) Pair-wise similarity computation for neighbor selection,
2) Affinity matrix or weights (graph structure) generation, and

3) Calculation of low-dimensional embedding.

In step 1), it is nothing special that the similarities are usually measured, e.g., by Euclidean distance, spectral angle mapper (SAM). The main difference should lie in step 2), i.e. the affinity matrix ($\mathbf{W}_{(\mathrm{LPP})}$) of LPP or LE can be constructed by using a radial basis function (RBF):

$$\mathbf{W}_{(\mathrm{LPP})}^{i,j} = \begin{cases} \exp\dfrac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}, & \text{if } \mathbf{x}_j \in \phi_k(\mathbf{x}_i), \\ 0, & \text{otherwise}, \end{cases} \tag{3.18}$$

where $\phi_k(\mathbf{x}_i)$ denotes the k nearest neighbor (knn) of $\mathbf{x}_i$ and σ is the standard derivation; while the LLE or NPE-based weight matrix can be defined by exploiting the regression technique connecting each given point with its k surrounding neighbors. By solving the following optimization problem to obtain the reconstruction coefficients ($\mathbf{A}$)

$$\min_{\mathbf{A}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j \in \phi_k(\mathbf{x}_i)} \mathbf{A}_{ij} \mathbf{x}_j \right\|_2^2, \ \text{s.t.} \sum_{j \in \phi_k(\mathbf{x}_i)} \mathbf{A}_{i,j} = \mathbf{1}, \tag{3.19}$$

the affinity matrix ($\mathbf{W}_{(\mathrm{LLE})}$) can be then derived to be

$$\mathbf{W}_{(\mathrm{LLE})}^{i,j} = \begin{cases} \mathbf{A}_{i,j} + \mathbf{A}_{j,i} - \mathbf{A}_{i,j}\mathbf{A}_{j,i}, & \text{if } \mathbf{x}_j \in \phi_k(\mathbf{x}_i); \\ 0, & \text{otherwise}. \end{cases} \tag{3.20}$$

Once the affinity matrix is given, the closed-form solution of the low-dimensional embedding in step 3) is obtained by solving a GED. More specifically, the calculation of embedding coordinates of LEE or LE is equivalent to solving the problem (3.17), while one of LLE or NPE needs to minimize the following embedding function:

$$\min_{\mathbf{Z}} \sum_{i=1}^{n} \left\| \mathbf{z}_i - \sum_{j \in \phi_k(\mathbf{z}_i)} \mathbf{A}_{ij} \mathbf{z}_j \right\|_2^2, \ \text{s.t.} \ \sum_{i=1}^{n} \mathbf{z}_i = \mathbf{0}, \ \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} = \mathbf{I}. \tag{3.21}$$

From the viewpoint of the GGE framework, the Eq. (3.21) can be rewritten to be a graph embedding problem:

$$\begin{aligned} &\min_{\mathbf{Z}} \sum_{i=1}^{n} \left\| \mathbf{z}_i - \sum_{j \in \phi_k(\mathbf{Z}_i)} \mathbf{A}_{ij} \mathbf{z}_j \right\|_2^2 \\ =&\min_{\mathbf{Z}} \sum_{i=1}^{n} \sum_{j \in \phi_k(\mathbf{Z}_i)} \left\| \mathbf{z}_i - \mathbf{z}_j \right\|_2^2 \mathbf{W}_{(\mathrm{LLE})}^{i,j} \\ =&\min_{\mathbf{Z}} \mathrm{tr}(\mathbf{Z}\mathbf{L}_{(\mathrm{LLE})}\mathbf{Z}^{\mathrm{T}}) \ \text{s.t.} \ \mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}, \end{aligned} \tag{3.22}$$

where $\mathbf{L}_{(\mathrm{LLE})}$ is the corresponding Laplacian matrix of Eq. (3.20), computed by $\mathbf{D}_{(\mathrm{LLE})} - \mathbf{W}_{(\mathrm{LLE})} = (\mathbf{I} - \mathbf{A})^{\mathrm{T}}(\mathbf{I} - \mathbf{A})$.

In hyperspectral remote sensing, a variety of modified or improved graph embedding algorithms have been proposed, making it applicable to the different challenges and tasks, particularly for HDR [Lunga et al., 2014]. In [Ma et al., 2010], a novel hyperspectral image classification framework was designed by integrating the local manifold learning (LML) techniques with knn classifier. To alleviate the effects of multicollinearity when calculating the affinity matrix, [Hong et al., 2016c] proposed a robust scheme for neighbor selection for LML. The same investigators extended their work to a spatial-spectral joint embedding for

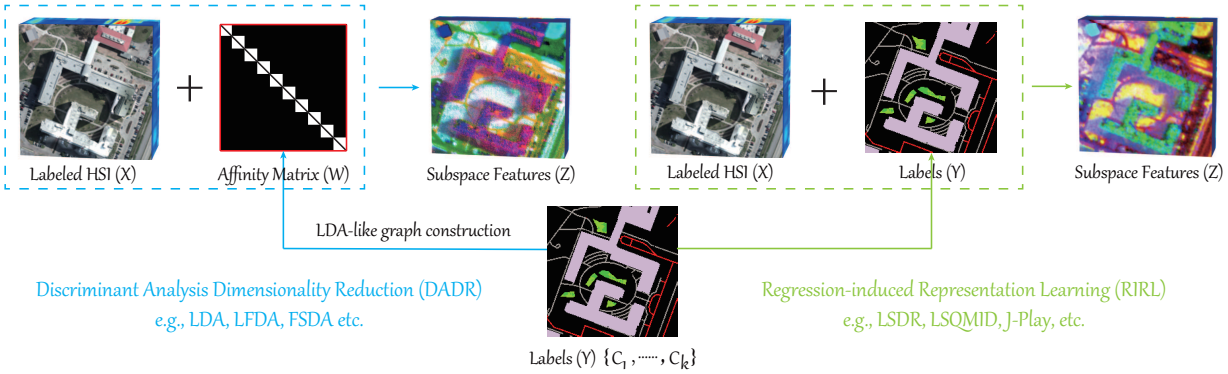


Fig. 3.1. An illustration for supervised HDR with two different strategies. A main difference lies in the use form of label information, i.e. DADR: affinity matrix, RIRL: labels or its one-hot encoding (e.g., J-Play).

HDR [Hong et al., 2017b]. To our knowledge, the LML approaches are easy to be a stick in computing the large-scale eigen decomposition (or spectral decomposition). To this end, the authors of [Hong et al., 2016b] developed a hierarchical LML and implemented a large-scale HDR by utilizing the cluster centers as the input of LLE. Furthermore, there are some latest researches in the manifold or graph embedding of hyperspectral data [Hu et al., 2019, Liao et al., 2018, Ma et al., 2016b, Pan et al., 2017, Yang and Crawford, 2016, Zhang et al., 2019, 2013a].

3.1.2 Category-guided Supervised Dimensionality Reduction

Unlike unsupervised HDR techniques that rely on modeling diverse prior knowledge of HSI, supervised methods are capable of extracting class-separable features more effectively, owing to the use of label information. It is well-known that dimensionality reduction is also referred to *subspace learning (SL)*, in which two main streams – discriminant analysis dimensionality reduction (DADR), e.g., subspace LDA (SLDA) [Yang and Yang, 2003], local fisher discriminant analysis (LFDA) [Sugiyama, 2007], feature space discriminative analysis (FSDA) [Imani and Ghassemian, 2015], and regression-induced representation learning (RIRL), e.g., least-squares dimension reduction (LSDR), [Sainui and Sugiyama, 2013] least-squares quadratic mutual information (LSQMI) [Suzuki and Sugiyama, 2013], joint & progressive learning strategy (J-Play) [Hong et al., 2018], – are emphatically investigated and compared by clarifying their similarities and differences as well as pros and cons, as briefly illustrated in Figure 3.1.

Discriminant Analysis Dimensionality Reduction (DADR)

Generally speaking, DADR seeks to find an optimal projection or transformation matrix $\mathbf{P} \in \mathbb{R}^{p\times d}$ (d is the dimension of the to-be-estimated subspace) by optimizing certain class-relevant separation criterion associated with the label information. In this process, the estimated subspace $\mathbf{Z} \in \mathbb{R}^{d\times n}$ that consists of a series of vector $\mathbf{z}_i$ can be obtained by projecting the samples $\mathbf{X}_n = \{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{p\times n}$ onto a decision boundary, which can be generally expressed as $\mathbf{Z} = \mathbf{P}^{\mathrm{T}}\mathbf{X}$. Each scalar $\mathbf{z}_i$ in $\mathbf{Z}$ can be collected by $\mathbf{P}^{\mathrm{T}}\mathbf{x}_i$. Depending on the different type of label embedding, DADR can be subdivided into LDA and its variants, graph-based discriminant analysis (GDA) and its extensions, and kernelized discriminant analysis (KDA).

▹ *LDA and Its Variants:* The traditional LDA linearly transforms the original data into a discriminative subspace by maximizing the Fisher's ratio in the form of generalized Rayleigh quotient, that is, simultaneously minimizing the intra-class scatter and maximizing inter-class scatter. Given a pair-wise training set $\{(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_i,\mathbf{y}_i),\ldots,(\mathbf{x}_m,\mathbf{y}_m)\}$ that contains the training samples $\mathbf{X}_m = \{\mathbf{x}_i\}_{i=1}^m \in \mathbb{R}^{p\times m}$ with p bands by m pixels (or samples) and corresponding class labels $\mathbf{Y}_m = \{\mathbf{y}_i\}_{i=1}^m \in \mathbb{R}^{1\times m}$, $\mathbf{y}_i \in \{C_1, C_2, \ldots, C_k\}$, where k denotes the number of

classes, the objective function of multi-class LDA to estimate the linear projection matrix $\mathbf{P}$ can be written as follows.

$$\max_{\mathbf{P}} \frac{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_b\mathbf{P})}{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_w\mathbf{P})}, \tag{3.23}$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are defined as the within-class scatter matrix and the between-class scatter matrix, respectively. With the constraint of $\mathbf{P}^{\mathrm{T}}\mathbf{S}_w\mathbf{P} = \mathbf{I}$, the optimization problem in Eq. (3.23) can be equivalently converted to one of $\mathbf{S}_b\mathbf{P} = \lambda\mathbf{S}_w\mathbf{P}$ by introducing the Lagrange multiplier λ. The closed-form solution to the simplified optimization problem can be deduced by a GED.

Due to the sensitivity to complex high-dimensional noises caused by the environmental and instrumental factors and the availability of labeled samples, the original LDA inevitably suffers from an ill-posed statistical degradation, especially in the case of small-scale samples. The degraded reasons mainly lie in the singularity of two scatter metrics ($\mathbf{S}_w$ and $\mathbf{S}_b$), thereby easily leading to the overfitting problem. To improve the stability and generalization, the regularized LDA was proposed by additionally adding a l_2-norm constraint on $\mathbf{S}_w$ parameterized by γ as $\mathbf{S}_w^{\mathrm{reg}} = \mathbf{S}_w + \gamma\mathbf{I}$. By replacing the $\mathbf{S}_w$ in Eq. (3.23) with the regularized $\mathbf{S}_w^{\mathrm{reg}}$, the solution in the regularized LDA can be still obtained by the GED solver.

Considering the local neighborhood relations between samples in the process of model learning, LFDA breaks through the bottleneck of those LDA-based methods by assuming that the data are distributed in the nonlinear manifolds rather than a homogeneous Gaussian space. For this purpose, LFDA is capable of effectively excavating the locally underlying structure of the data that lies in the real world. Essentially, LFDA can be regarded as a weighted LDA by locally weighing the $\mathbf{S}_w$ and $\mathbf{S}_b$ matrices. Therefore, the two modified scatter matrices, denoted as $\widetilde{\mathbf{S}}_w$ and $\widetilde{\mathbf{S}}_b$, can be formulated as

$$\begin{aligned}
\widetilde{\mathbf{S}}_w &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{W}_w^{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}, \\
\widetilde{\mathbf{S}}_b &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{W}_b^{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}},
\end{aligned} \tag{3.24}$$

where the two weights ($\mathbf{W}_w$ and $\mathbf{W}_b$) denote the sample-wise similarities. There are several widely-used strategies in calculating such similarity matrix symbolized by $\mathbf{W}$. A simple but effective one is given by $\mathbf{W}^{i,j} = 1$, if $\mathbf{x}_j \in \phi_k(\mathbf{x}_i)$, where $\phi_k(\mathbf{x}_i)$ represents the k-nearest-neighbor of $\mathbf{x}_i$; otherwise, $\mathbf{W}^{i,j} = 0$.

Similar to SLDA that first projects the original data into a subspace and then LDA is performed in the transformed subspace, FSDA starts with maximizing the between-spectral scatter matrix ($\mathbf{S}_f$) to enhance the differences along the spectral dimension, and similarly the LDA is further used for extracting the representations of class separability from the feature domain. In the first step, let $\mu_{i,j}$ be the average value of the j-th class and the i-th spectral band, then we have the definition of $\mathbf{S}_f$ as follows:

$$\mathbf{S}_f = \frac{1}{2}\sum_{i=1}^{p}(\mathbf{h}_i - \overline{\mathbf{h}})(\mathbf{h}_i - \overline{\mathbf{h}})^{\mathrm{T}}, \tag{3.25}$$

where $\mathbf{h}_i = [\mu_{i,1}, \mu_{i,2}, ..., \mu_{i,k}]$ is the spectral representation in the feature space and $\overline{\mathbf{h}} = \frac{1}{p}\sum_{i=1}^{p}\mathbf{h}_i$. The primary transformation ($\mathbf{P}_f$) that aims at improving the spectral discriminant

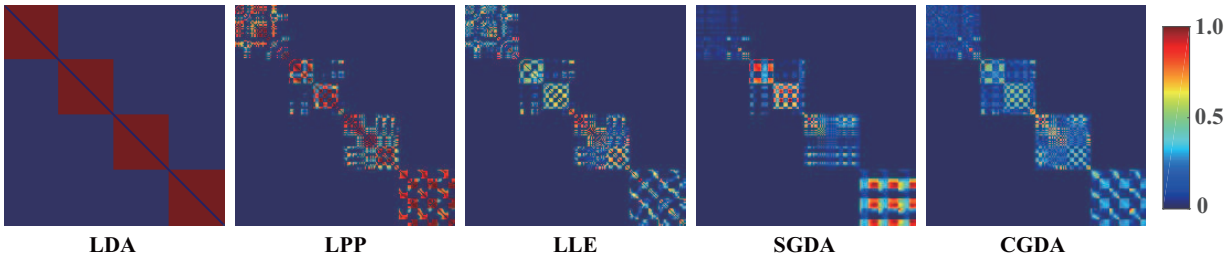


Fig. 3.2. Four types of affinity matrices ($\mathbf{W}$) used in five different approaches: LDA, LPP, LLE, SGDA, and CGDA, respectively, where the connectivity (or edge) of $\mathbf{W}$ is computed within each class.

can be estimated with maximizing the trace term of $\mathbf{S}_f$ as

$$\max_{\mathbf{P}_f} \mathrm{tr}(\mathbf{P}_f^{\mathrm{T}} \mathbf{S}_f \mathbf{P}_f). \tag{3.26}$$

Using the obtained $\mathbf{P}_f$, the latent representation in the feature space $\mathbf{g}_i = \mathbf{P}_f^{\mathrm{T}} \mathbf{h}_i,\ i = 1, 2, ..., p$ can be further fed into the next-step LDA.

▹ *GDA and Its Extensions:* As introduced in the previous chapter, the GDA methods similarly follow the GGE framework. Obviously, the extracted features $\mathbf{Z}$ in the GGE framework are determined by the construction of $\mathbf{W}$ to a great extent. Thus, we will highlight several kinds of representative affinity matrices corresponding to the different graph embedding approaches, i.e. LDA, LE and its linearized LPP, LLE, sparse GDA (SGDA) [Ly et al., 2014b], and collaborative GDA (CGDA) [Ly et al., 2014a]. Figure 3.2 visualizes the affinity matrices given by five different strategies in a four-class case.

- ◇ **LDA-like affinity matrix:** In essence, LDA is vested in a special case of GGE framework with $\mathbf{D}_{(\mathrm{LDA})} = \mathbf{I}$, whose affinity matrix can be represented as

$$\mathbf{W}_{(\mathrm{LDA})}^{i,j} = \begin{cases} 1/N_k, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \in C_k; \\ 0, & \text{otherwise,} \end{cases} \tag{3.27}$$

  where N_k is the number of samples belonging to k-th class.
- ◇ **LPP or LE-based affinity matrix:** One is to be constructed in a kernel space with a higher dimension via similarity measurement, i.e. extensively using Eq. (3.18).
- ◇ **LLE-based affinity matrix:** Different from the hand-crafted graph, LLE reconstructs each given sample with its k-nearest neighbors by exploiting the linear regression techniques, as shown in Eq. (3.20).
- ◇ **SGDA and CGDA-guided affinity matrix:** Similarly to LLE, the affinity matrix can be estimated using the data-driven representation learning, i.e. sparse and collaborative representations. Accordingly, the two learning strategies can be equivalent to respectively solving the constrained l_1-norm optimization problem:

$$\min_{\mathbf{W}} \|\mathbf{W}\|_{1,1} \quad \text{s.t.} \|\mathbf{X}_m \mathbf{W} - \mathbf{X}_m\|_{\mathrm{F}}^2 \leq \epsilon, \tag{3.28}$$

  and the l_2-norm optimization problem:

$$\min_{\mathbf{W}} \|\mathbf{W}\|_{\mathrm{F}}^2 \quad \text{s.t.} \|\mathbf{X}_m \mathbf{W} - \mathbf{X}_m\|_{\mathrm{F}}^2 \leq \epsilon. \tag{3.29}$$

  The aforementioned affinity matrices can be unified to the GGE framework of Eq. (3.17).

  In addition to SGDA and CGDA (the two baselines), [Huang et al., 2015] learned a set of sparse coefficients on manifolds and then preserved the sparse manifold structure in

the embedded space. The work in [Xue et al., 2015] extended the existing SGDA to the spatial-spectral graph embedding to address the issues of the spatial variability and spectral multimodality. This requires the embedding of the intrinsic geometric structure of the data motivate to develop a Laplacian regularizer CGDA [Li and Du, 2016] to further improve the graph's confidence. A well-done work proposed in [Li et al., 2016] simultaneously integrated the sparsity and low-rankness into the graph for capturing a more robust structure of the data locally and globally. Furthermore, [Pan et al., 2017] further improved the above work by unfolding the HS data with the form of a tensor.

▹ *KDA:* In reality, the HSI usually exhibits a highly nonlinear data distribution, which may result in difficulties in effectively identifying the materials. The solution to this issue is making use of a so-called kernel trick [Müller et al., 2001] that can map the data of the input space into a new Hilbert space with a higher feature dimension. In the kernel-induced space, the complex nonlinearity of the HS data can be well analyzed in a linearized system. Comparatively, the input to KDA is an inner product of original data pairs, defined as $k(\mathbf{x}_i, \mathbf{x}_j)$ which can be given by Eq. (3.18). By introducing the kernel Gram matrix $\mathbf{K}$ with $\mathbf{K}_{i,j} = \mathbf{\Phi}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{\Phi}(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$, most of previous LDA-based methods can be simply extended to the corresponding kernelized versions, i.e. KLDA and KLFDA can calculate their projections $\mathbf{P}$ by solving a GED problem of

$$\mathbf{KLKP} = \lambda(\mathbf{KBK} + \gamma\mathbf{I})\mathbf{P}. \tag{3.30}$$

Note that $\mathbf{B} = \mathbf{I}$ in KLDA, while $\mathbf{L} = \mathbf{L}_w$ and $\mathbf{B} = \mathbf{L}_b$ are computed by $\mathbf{D}_w - \mathbf{W}_w$ and $\mathbf{D}_b - \mathbf{W}_b$ in the kernel space, respectively, for KLFDA. Furthermore, for KSGDA and KCGDA, the main difference lies in the computation of the adjacency matrix, which can be performed in the kernel space by solving the general kernel coding problem as follows:

$$\min_{\mathbf{W}} \mathbf{\Omega}(\mathbf{W}) \ \text{ s.t.} \|\mathbf{\Phi}(\mathbf{X}_m)\mathbf{W} - \mathbf{\Phi}(\mathbf{X}_m)\|_{\mathrm{F}}^2 \leq \epsilon, \tag{3.31}$$

where $\mathbf{\Omega}(\mathbf{W})$ can be selected to be either sparsity-prompting term $\|\mathbf{W}\|_{1,1}$ of KSGDA or dense (or collaborative) term $\|\mathbf{W}\|_F^2$ of KCGDA.

Regression-induced Representation Learning (RIRL)

RIRL provides a new insight from the point of regression view to model the dimensionality reduction behavior by bridging the training samples with the corresponding labels rather than indirectly using the label information in the form of graph or affinity matrix in DADR-based methods.

▹ *Least-Squares Dimension Reduction (LSDR):* We begin with sliced inverse regression (SIR) [Li, 1991], which is a landmark in supervised dimensionality reduction techniques. It assumes that the pair-wise data pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$ are conditionally independent on the to-be-estimated subspace features $\{\mathbf{z}_i\}_{i=1}^m$, formulated as $(\mathbf{X} \perp \mathbf{Y}) \mid \mathbf{Z}$. Following this rule, the LSDR proposed by Suzuki and Sugiyama [Suzuki and Sugiyama, 2013] attempts to find a maximizer of the squared-loss mutual information (SMI) to satisfy the above independent assumption. The projections $\mathbf{P}$ for LSDR can be searched by optimizing the following maximization problem:

$$\max_{\mathbf{P}} \mathrm{SMI}(\mathbf{Z}, \mathbf{Y}) \ \text{ s.t. } \ \mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{I}, \tag{3.32}$$

and the SMI to measure a statistical dependence between two discrete variables is defined as

$$\mathrm{SMI}(\mathbf{Z}, \mathbf{Y}) = \sum_{\mathbf{z}\in\mathbf{Z}} \sum_{\mathbf{y}\in\mathbf{Y}} p(\mathbf{z})p(\mathbf{y}) \left(\frac{p(\mathbf{z}, \mathbf{y})}{p(\mathbf{z})p(\mathbf{y})} - 1 \right)^2, \tag{3.33}$$

where $p(\bullet)$ is the probability distribution function.

▹ *Least-Squares Quadratic Mutual Information (LSQMI):* Limited by the sensitivity of MI to outliers, authors of [Sainui and Sugiyama, 2013] designed a more robust LSQMI with the basis of QMI criterion, hence let us define the QMI as

$$\mathrm{QMI}(\mathbf{Z},\mathbf{Y}) = \sum_{\mathbf{z}\in\mathbf{Z}}\sum_{\mathbf{y}\in\mathbf{Y}} \left(p(\mathbf{z},\mathbf{y}) - p(\mathbf{z})p(\mathbf{y})\right)^2. \tag{3.34}$$

Similarly, we solve the Eq. (3.32)-like optimization problem by replacing SMI with QMI.

▹ *Least-Squares QMI Derivative (LSQMID):* Due to the difficulty in accurately computing the derivative of QMI estimator, LSQMI was further extended to a computationally effective LSQMID by estimating the derivative of QMI instead of QMI itself [Tangkaratt et al., 2017]. In this work, authors have demonstrated more accurate and efficient derivative computation of QMI.

▹ *Latent Subspace Learning (LSL):* Another MI-free estimation group is LSL. One representative LSL performs dimensionality reduction and classification simultaneously in joint learning (JL) fashion [Ji and Ye, 2009]. With an expected output $\mathbf{\Theta X}_m$, the process can be modeled as

$$\min_{\mathbf{P},\mathbf{\Theta}} \|\mathbf{Y}_l - \mathbf{P\Theta X}_m\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \;\; \text{s.t.} \;\; \mathbf{\Theta\Theta}^{\mathrm{T}} = \mathbf{I}, \tag{3.35}$$

where $\mathbf{Y}_l \in \mathbb{R}^{k\times n}$ and $\mathbf{\Theta} \in \mathbb{R}^{d\times p}$ are defined as the one-hot encoded label matrix and the latent subspace projections, respectively. $\mathbf{P} \in \mathbb{R}^{k\times d}$ denotes the regression matrix that connects the learned subspace and the label information. In [Ji and Ye, 2009], the model's solution has been proven to be a closed-form. Moreover, the work in [Hong et al., 2019d] explored a LDA-like graph as a regularizer to learn a spectrally discriminative feature representation, thus Eq. (3.35) becomes

$$\min_{\mathbf{P},\mathbf{\Theta}} \|\mathbf{Y}_l - \mathbf{P\Theta X}_m\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\mathbf{\Theta X}_m\mathbf{L}\mathbf{X}_m^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}), \tag{3.36}$$

with the orthogonal constraint $\mathbf{\Theta\Theta}^{\mathrm{T}} = \mathbf{I}$.

3.1.3 Semi-supervised Strategy of Dimensionality Reduction

Loosely speaking, the semi-supervised learning refers to simultaneously using the labeled and unlabeled data to learn a more powerful model, thereby exceeding the performance of only using either labeled or unlabeled samples. In machine learning and computer vision communities, the semi-supervised learning has been studied by the researchers for quite a while. Yet it is relatively less investigated in hyperspectral data processing and analysis, particularly in HDR. Up to the present, there have been some tentative researches in semi-supervised HDR. These methods are mostly developed under the framework of semi-supervised discriminant analysis (SSDA) [Cai et al., 2007] that utilizes massive unlabeled samples to improve the class separability obtained using few labeled samples.

SSDA assumes the class consistency, which means nearby points in the feature space should share the same label in classification tasks or similar low-dimensional embeddings in dimensionality reduction tasks. The prior assumption can be formulated by imposing a regularizer into LDA, resulting in the following optimization problem:

$$\max_{\mathbf{P}} \frac{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_b\mathbf{P})}{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_w\mathbf{P}) + \alpha\mathcal{J}(\mathbf{P})}, \tag{3.37}$$

where the regularizer term $\mathcal{J}(\mathbf{P})$ is defined as

$$\mathcal{J}(\mathbf{P}) = \sum_{i,j} \|\mathbf{P}^{\mathrm{T}}\mathbf{x}_i - \mathbf{P}^{\mathrm{T}}\mathbf{x}_j\|_2^2 \mathbf{W}^{i,j}_{(\mathrm{SDA})}. \tag{3.38}$$

The Eq. (3.39) can be further derived to be

$$\begin{aligned}
\mathcal{J}(\mathbf{P}) &= \sum_{i,j} \|\mathbf{P}^{\mathrm{T}}\mathbf{x}_i - \mathbf{P}^{\mathrm{T}}\mathbf{x}_j\|_2^2 \mathbf{W}^{i,j}_{(\mathrm{SDA})} \\
&= 2\sum_{i} \mathbf{P}^{\mathrm{T}}\mathbf{x}_i \mathbf{D}^{i,i}_{(\mathrm{SDA})} \mathbf{x}_i^{\mathrm{T}} - 2\sum_{i,j} \mathbf{P}^{\mathrm{T}}\mathbf{x}_i \mathbf{W}^{i,j}_{(\mathrm{SDA})} \mathbf{x}_j^{\mathrm{T}}\mathbf{P} \\
&= 2\mathbf{P}^{\mathrm{T}}\mathbf{X}(\mathbf{D}_{(\mathrm{SDA})} - \mathbf{W}_{(\mathrm{SDA})})\mathbf{X}^{\mathrm{T}}\mathbf{P} \\
&= 2\mathbf{P}^{\mathrm{T}}\mathbf{X}\mathbf{L}_{(\mathrm{SDA})}\mathbf{X}^{\mathrm{T}}\mathbf{P},
\end{aligned} \tag{3.39}$$

where the weight matrix $\mathbf{W}_{(\mathrm{SDA})}$ can be given by

$$\mathbf{W}^{i,j}_{(\mathrm{SDA})} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in \phi_k(\mathbf{x}_i) \textit{ or } \text{if } \mathbf{x}_i \in \phi_k(\mathbf{x}_j), \\ 0, & \text{otherwise.} \end{cases} \tag{3.40}$$

With the Eq. (3.39), the objective function of SDA in Eq. (3.37) becomes

$$\max_{\mathbf{P}} \frac{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}\mathbf{S}_b\mathbf{P})}{\mathrm{tr}(\mathbf{P}^{\mathrm{T}}(\mathbf{S}_w + 2\alpha\mathbf{X}\mathbf{L}_{(\mathrm{SDA})}\mathbf{X}^{\mathrm{T}})\mathbf{P}}. \tag{3.41}$$

Inspired by SSDA, a general but effective work integrating LDA with LPP, called semi-supervised local discriminant analysis (SELD), was proposed in [Liao et al., 2013] for semi-supervised hyperspectral feature extraction. In [Shao and Zhang, 2014], the semi-supervised local Fisher discriminant analysis (SELF) was improved with the sparse preservation embedding to effectively reduce the dimensions of hyperspectral data. [Ma et al., 2015] followed a graph-based semi-supervised learning paradigm by the attempt to preserve the potentially global data structure that lies in the whole high-dimensional space, yielding the HDR and classification, where the graphs are constructed by different local manifold learning approaches. Ma *et al.* further extended their work to a spatial-spectral version of graph-based semi-supervised manifold embedding [Ma et al., 2016a], thus leading to a smoother classification result of hyperspectral image. [Wu and Prasad, 2018] proposed a similar approach to achieving the discriminative dimensionality reduction of HSI in a semi-supervised fashion. The main difference lies in the use of pseudo-labels instead of the similarity measurement in LPP. However, these approaches mainly benefit from the fixed graph structure generated manually or given from other algorithms (e.g., local manifold learning [Ma et al., 2015], the Dirichlet process mixing model [Wu and Prasad, 2018]). This type of graph construction strategy tends to exhibit weak generalization of the features, further causing a performance bottleneck.

Apart from the SSDA-related methods, there is certainly the other type of SSL algorithms in HDR. For example, [Zhang et al., 2013b] proposed to apply the local scaling cut criterion for semi-supervised dimensionality reduction of hyperspectral image. The researchers of [Liao et al., 2012] performed a semi-supervised hyperspectral classification over urban areas by integrating the directional morphological profiles and SDA, attempting to further improve the classification performance. Interestingly, [Su et al., 2012a] took advantage of the clustering technique based on divergence in order to learn an orthogonal projection and achieved a semi-supervised HDR. Motivated by probabilistic principal component analysis (PPCA),

[Xia et al., 2014] developed a semi-supervised PPCA approach for hyperspectral image classification. With a different strategy, [Wang et al., 2014] propagated the label information in a spatial-spectral fashion, yielding a semi-supervised classification of HSI. In [Wang et al., 2016], the issue of dimensionality reduction is viewed as a non-negative matrix factorization, and they aimed at learning a sparsity-promoting projection matrix in the process of HDR. Very recently, [Hong et al., 2019b] proposed to propagate the labels on a dynamic graph for semi-supervised HDR by the means of data-driven graph learning strategy.

Furthermore, many hyperspectral band selection methods have been also investigated in semi-supervised HDR. A work related to graph-based SSL was proposed in [Chen et al., 2010] by weighing the different bands to adaptively select the spectral bands. A similar work was presented in [Su et al., 2012b] by the means of adaptive affinity propagation with the measurement of SAM for semi-supervised band selection.

3.2 Spectral Unmixing

An abundant of spectral information provided in the hyperspectral image allows identifying the materials through subtle spectral discrepancies. Nevertheless, the material easy-mixing, as mentioned earlier, makes the spectral profiles of different materials difficult to be distinguished. Same with dimensionality reduction, spectral unmixing is, as often as not, an essential step before the high-level data analysis is made. Spectral unmixing refers to a procedure that decomposes the observed pixel spectrum of the hyperspectral image into a series of constituent spectral signals (or *endmembers*) of pure materials and a set of corresponding abundance fractions (or *abundance maps*). Depending on the different types of material mixing, unmixing methods can be appropriately characterized based on the linear mixing model and nonlinear mixing model. Figure 3.3 illustrates the different mixing scenarios: (a) linear mixing; (b) and (c) nonlinear mixing.

3.2.1 Linear Mixing Model and Its Variants

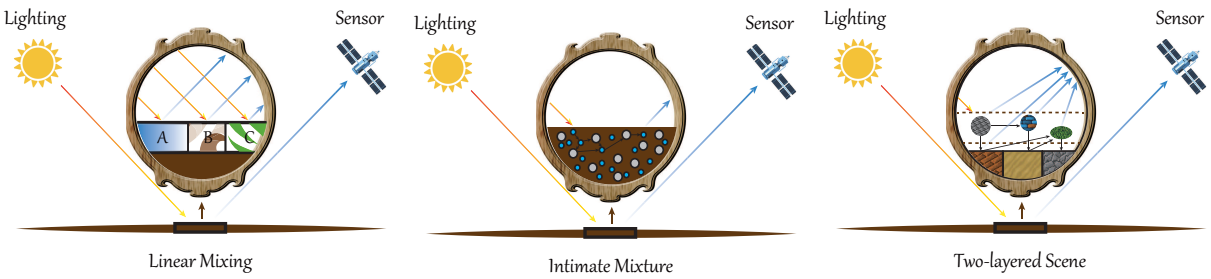
When the materials are mixed at a macroscopic scale, which means the incident light only interacts with one certain material, the linear mixing usually occurs in the sensor receiver side, as depicted in Figure 3.3(a). Assuming the absence of any spectral, spatial, and temporal deformations as well as microscopic interactions between the materials, such as multiple scatting and intimate mixing, are negligible, the mixed spectral profile of each pixel in a hyperspectral scene is well measured by a *linear mixing model* (LMM) [Bioucas-Dias et al., 2012].

A. LMM

Let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_i, ..., \mathbf{y}_N] \in \mathbb{R}^{D\times N}$ be a stretched hyperspectral image with D bands by N pixels and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P] \in \mathbb{R}^{D\times P}$ be the endmembers with the dimension of $D \times P$. $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_N] \in \mathbb{R}^{P\times N}$ represents the abundance maps, whose each column vector denotes the fractional abundance at each pixelof HSI. $\mathbf{R} = [\mathbf{r}_1, ..., \mathbf{r}_i, ..., \mathbf{r}_N] \in \mathbb{R}^{D\times N}$ is the residual with matrix form, including various noises, reconstruction errors of the model, and among others. Ideally, the measured spectrum in a given pixel is denoted as $\mathbf{y}_i \in \mathbb{R}^{D\times 1}$, which can be linearly approximated by a set of endmember spectra associated with the corresponding fractions. Therefore, the resulting LMM can be written as

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{r}_i, \tag{3.42}$$

where each $\mathbf{a}_i$ and $\mathbf{x}_i$ should be non-negative so that the physical conditions are satisfied in reality. What's more, the abundance $\mathbf{x}_i$, as the name suggests, represents the proportions



(a) linear mixing (b) nonlinear: intimate mixing (c) nonlinear: multilayered case

Fig. 3.3. Linear and nonlinear mixing scenarios: (a) linear mixing. (b) nonlinear mixing of intimate mixture. (c) nonlinear mixing of multilayered scattering: a two-layered case.

occupied by the different endmembers. This makes $\mathbf{x}_i$ be subject to a sum-to-one constraint as well. Thus, Eq. (3.42) with the necessary constraints can be rewritten as

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{r}_i, \ \text{s.t.} \ \mathbf{A} \succeq \mathbf{0}, \ \mathbf{x}_i \succeq 0, \ \sum_{i=1}^{N} \mathbf{x}_i = 1. \tag{3.43}$$

By collecting all pixels, we have a compact matrix form of Eq. (3.43):

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{R}, \ \text{s.t.} \ \mathbf{A} \succeq \mathbf{0}, \ \mathbf{X} \succeq \mathbf{0}, \ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \tag{3.44}$$

We will emphatically introduce several state-of-the-art unmixing methods based on LMM, they are fully constrained least squares unmixing (FCLSU) [Heinz et al., 2001], partial constrained least squares unmixing (PCLSU) [Heylen et al., 2011], sparse unmixing by variable splitting and augmented Lagrangian (SUnSAL) [Bioucas-Dias and Figueiredo, 2010], scaled partial constrained least squares unmixing (SPCLSU) [Veganzones et al., 2014], dictionary-adjusted non-convex sparsity-encouraging regression (DANSER) [Fu et al., 2016], extended linear mixing model (ELMM) [Drumetz et al., 2016], and perturbed linear mixing model (PLMM) [Thouvenin et al., 2016].

▹ *1) FCLSU:* In practice, the endmembers ($\mathbf{A}$) can be pre-extracted from the given hyperspectral scene by the means of some endmember extraction methods, e.g., vertex component analysis (VCA) [Nascimento and Dias, 2005]. When the endmember matrix is given, the problem of estimating the abundance maps ($\mathbf{X}$) is degraded to solve a least-square regression problem, we then have the FCLSU:

$$\min_{\mathbf{X}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 \quad \text{s.t.} \ \mathbf{X} \succeq \mathbf{0}, \ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \right\}. \tag{3.45}$$

Due to the presence of spectral variability, FCLSU usually yields poor unmixing performance. This might result from strong sum-to-constraint. A typical solution to this issue is to relax the sum-to-one constraint to less or larger than one or to extremely ignore this constraint.

▹ *2) PCLSU:* According to the aforementioned solution, the resulting PCLSU becomes the following form:

$$\min_{\mathbf{X}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 \ \text{s.t.} \ \mathbf{X} \succeq \mathbf{0}. \right\}. \tag{3.46}$$

The estimated variable $\mathbf{X}$ in Eq. (3.46) might be any scales, owing to a badly-conditioned observed matrix $\mathbf{Y}$. To alleviate the effects of the ill-posed problem, meaningfully physical assumptions have to be added in the form of regularization.

▹ *3) SUnSAL:* As observed, the abundances on each endmember are theoretically supposed to be sparse. Bioucas-Dias *et al.* embedded this property into LMM and achieved a powerful SUnSAL algorithm. The resulting optimization problem can be written as follows

$$\min_{\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{X}\|_{1,1} \text{ s.t. } \mathbf{X} \geq \mathbf{0},\ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \right\}, \tag{3.47}$$

where $\|\mathbf{X}\|_{1,1} \equiv \sum_{k=1}^{N} \|\mathbf{x}_k\|_1$ is denoted as an approximation of sparsity-promoting term.

In view of effectiveness of SUnSAL, SUnSAL's variations have been subsequently proposed in recent years, such as SUnSAL with total variation spatial regularization (SUnSAL-TV) [Iordache et al., 2012], collaborative sparse regression (CLSUnSAL) [Iordache et al., 2014], etc. We have to admit, however, that these advanced methods are still subject to the framework of LMM that is sensitive to spectral variabilities.

B. ELMM

ELMM aims to modeling the principle spectral variability (scaling factors) to allow a pixel-wise variation at each endmember:

$$\mathbf{y}_i = \mathbf{A}\mathbf{S}_i\mathbf{x}_i + \mathbf{r}_i, \tag{3.48}$$

where $\mathbf{S}_i \in \mathbb{R}^{P\times P}$ is a diagonal matrix with the non-negative constraint ($\mathbf{S}_i \geq 0$). A matrix form of Eq. (3.48) can be repented as

$$\mathbf{Y} = \mathbf{A}(\mathbf{S} \odot \mathbf{X}) + \mathbf{R}, \tag{3.49}$$

here $\mathbf{S} \in \mathbb{R}^{P\times N}$ is a full matrix collecting the scaling factors from all pixels whose i^{th} column is $\mathbf{S}_i$. The operator $\odot$ is denoted as the Schur-Hadamard (termwise) product.

▹ *1) Unmixing under the ELMM:* Intuitively, the optimization problems in (3.48) and (3.49) are hardly to be analytically solved. In **?**, a trick is employed by splitting the coupled variables ($\mathbf{S}$ and $\mathbf{X}$), then we have

$$\min_{\mathbf{X},\mathbf{S}\geq 0,\underline{\mathbf{A}}} \left\{ \sum_{k=1}^{N} (\|\mathbf{y}_k - \mathbf{A}_k\mathbf{x}_k\|_2^2 + \lambda_S\|\mathbf{A}_k - \mathbf{A}_0\mathbf{S}_k\|_{\mathrm{F}}^2) \right\}, \tag{3.50}$$

where $\mathbf{A}_0$ is the reference endmember spectrum, $\underline{\mathbf{A}} = \{\mathbf{A}_i\}$ is a collection of pixel-dependent endmember matrices, and λ_S plays a balance role between the two separated terms. Eq.(3.50) can be alternatively optimized with respect to each variable by alternating minimization strategy [Kim and Park, 2008].

▹ *2) SPCLSU:* Prior to ELMM, scaling factors have been investigated in a simple way, that is SPCLSU in which endmembers are reasonably assumed by sharing the same scale as the scaling factors are strongly associated with topography. SPCLSU actually conducts a PCLSU in the beginning, and then normalizes the abundance maps to meet sum-to-one. This is a simple but effective strategy, which is also involved in our proposed method.

C. PLMM

As the name suggested, PLMM attempts to describe the spectral variability as additive perturbation information. Both the pixel-wise and the corresponding matrix form of PLMM can be expressed, respectively

$$\mathbf{y}_i = (\mathbf{A} + \mathbf{\Delta}_i)\mathbf{x}_i + \mathbf{r}_i, \tag{3.51}$$

and

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \underbrace{[\mathbf{\Delta}_1\mathbf{x}_1|...|\mathbf{\Delta}_i\mathbf{x}_i|...|\mathbf{\Delta}_N\mathbf{x}_N]}_{\mathbf{\Delta}} + \mathbf{R}, \tag{3.52}$$

where $\mathbf{\Delta}$ is $[\mathbf{\Delta}_1\mathbf{x}_1|\ldots|\mathbf{\Delta}_i\mathbf{x}_i|...|\mathbf{\Delta}_N\mathbf{x}_N]$ denotes the perturbation information of the endmembers.

▹ *1) Unmixing under the PLMM:* The optimization problem corresponding to PLMM-based unmixing can be given as

$$\min_{\mathbf{A},\mathbf{\Delta},\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{X}-\mathbf{\Delta}\|_{\mathrm{F}}^2 + \alpha\mathbf{\Phi}(\mathbf{X}) + \beta\mathbf{\Psi}(\mathbf{A}) + \gamma\mathbf{\Upsilon}(\mathbf{\Delta}) \right\}, \tag{3.53}$$

where $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Upsilon}$ parameterized by α, β, and γ, are penalties with respect to variables $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{\Delta}$, receptively. Notably, $\mathbf{\Upsilon}$ term is modeled by a Frobenius norm.

▹ *2) DANSER:* Likewise being generalized to PLMM framework, DANSER adopts a sparsity-encouraging regression technique for a dictionary-based spectral unmixing, where a perturbation-like information is explored to measure the mismatch between spectral dictionary and observed endmembers. This model, namely DANSER, is formulated by

$$\min_{\mathbf{A}',\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{A}'\mathbf{X}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2 + \beta\|\mathbf{X}\|_{2,p}^p \ \text{ s.t. } \mathbf{X}\geq\mathbf{0} \right\}, \tag{3.54}$$

where $\mathbf{A}'$ represents the corrupted endmember matrix obtained by perturbing the variable $\mathbf{A}$.

Besides, there are some very-recently-proposed models providing the spectral variability with different prior assumptions in hyperspectral image unmixing, such as low-coherent modeling between the ideal spectral signatures and spectral variabilities [Hong et al., 2017a, 2019c, Hong and Zhu, 2018], tensor factorization using total variation regularization [Xiong et al., 2018], adaptive bundles and double sparsity [Uezato et al., 2019], non-convex sparsity and non-local smoothness [Yao et al., 2019], endmember bundles and group sparsity [Drumetz et al., 2019], etc.

3.2.2 Nonlinear Mixing Models

Compared to LMM, nonlinear mixing models are inclined to investigate and analyze the physical interactions between multiple materials due to the reflection and scattering of light. These interactions might happen at a microscopic scale (intimate mixing) or multilayered level, as illustrated in Figures 3.3(b) and (c). Accordingly, two groups of nonlinear mixing models: intimate mixtures and bilinear models, are detailed as follows.

Intimate Mixtures

Intimately mixing is caused mainly due to the complex interactions between the inside of materials. Currently, there has been a well-established mathematical theory, called radiative transfer theory (RTT) [Chandrasekhar, 2013], in modeling the energy transferring between the photons interacts of the materials. Nevertheless, simultaneously estimating the spectral signatures and the corresponding material densities with the RTT in a nonlinear scene needs to solve an extremely ill-posed problem, which is hardly possible to be achieved under the conditions of limitedly available scene parameters. For this reason, researchers have found three approximated models – Hapke model [Hapke, 1981], Kulbelka-Munk model [Kubelka and Munk, 1931], and Shkuratov formulation [Shkuratov et al., 1999], to approach the analytical solution of the RTT. These models have been successfully applied to address many practical challenges and tasks, e.g., in chemistry, topology and illumination analysis, mineral exploration, and so on.

Although the above three kinds of models weaken the complexity of RTT, yet such complete physical models are still too complex to be widely used in the real case. Kernel strategy is an effective tool being able to fully take the intimate mixtures into consideration [Broadwater

et al., 2009]. For that, several kernel-inspired unmixing methods [Broadwater and Banerjee, 2010, 2011] have been proposed, making it feasible to model the nonlinear degrees by using RBF, polynomial formulation or some physics-induced kernel functions [Broadwater and Banerjee, 2009].

In addition, to alleviate the effects of the microscopically mixing materials, it makes sense to regard the homogeneous composition in the spatial domain as a pure endmember as long as the sensor resolution can be reasonably assumed to be the same with the objects of interest. Following this way, the intimate mixtures in a hyperspectral scene can be seen as a microscopically mixed product [Dobigeon et al., 2014].

Bilinear Models

[Borel and Gerstl, 1994] gave a nice and easy-understanding illustration to clarify the mixing procedures of the multilayered model, where the combination of material mixing at a macroscopic level is enumerable, as shown in Figure 3.3(c). One common thing may meet the nonlinear mixing behavior in reality, that is, the incident light is scattered from a given material and encounters other materials again before it is received by the sensor. This is a quite common case that often happens in the forest-covered areas. In mathematics, this process can be modeled in the bilinear fashion:

$$\begin{aligned} \mathbf{y}_i &= \sum_{q=1}^{P} \mathbf{a}_q \mathbf{x}_{q,i} + \sum_{u=1}^{P-1} \sum_{v=u+1}^{P} (\mathbf{a}_u \odot \mathbf{a}_v) \mathbf{x}_{u,v,i} + \mathbf{r}_i, \\ &\text{s.t. } \mathbf{a}_q \succeq \mathbf{0},\ \mathbf{a}_u \succeq \mathbf{0},\ \mathbf{a}_v \succeq \mathbf{0},\ \mathbf{x}_{q,i} \succeq \mathbf{0},\ \sum_{q=1}^{P} \mathbf{x}_{q,i} = 1,\ \mathbf{x}_{u,v,i} \in (0,1), \end{aligned} \tag{3.55}$$

where the term $\mathbf{a}_u \odot \mathbf{a}_v$ is defined as

$$\mathbf{a}_u \odot \mathbf{a}_v = \begin{pmatrix} \mathbf{a}_{1,u} \\ \mathbf{a}_{2,u} \\ \cdots \\ \mathbf{a}_{D,u} \end{pmatrix} \odot \begin{pmatrix} \mathbf{a}_{1,v} \\ \mathbf{a}_{2,v} \\ \cdots \\ \mathbf{a}_{D,v} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_{1,u}\mathbf{a}_{1,v} \\ \mathbf{a}_{2,u}\mathbf{a}_{2,v} \\ \cdots \\ \mathbf{a}_{D,v}\mathbf{a}_{D,v} \end{pmatrix}. \tag{3.56}$$

In Eq. (3.55), the nonlinear interaction behavior across the materials is modeled in the second term. Additionally, there are some alternatives in describing the constraints, i.e. [Somers et al., 2009] proposed the abundances and nonlinear mixing coefficients should follow the improved sum-to-one constraint as

$$\sum_{q=1}^{P} \mathbf{x}_{q,i} + \sum_{u=1}^{P-1} \sum_{v=u+1}^{P} \mathbf{x}_{u,v,i} = 1. \tag{3.57}$$

Unlikely the model shown in Eq. (3.55), [Fan et al., 2009] represented the coefficients $\mathbf{x}_{u,v,i}$ with a function of abundances $\mathbf{x}_{q,i}$, denoted as $\mathbf{x}_{u,v,i} = \mathbf{x}_{u,i}\mathbf{x}_{v,i}$. Hence, the improved bilinear

model, called Fan Model (FM), is given in the following:

$$\begin{aligned}
\mathbf{y}_i &= \sum_{q=1}^{P} \mathbf{a}_q \mathbf{x}_{q,i} + \sum_{u=1}^{P-1} \sum_{v=u+1}^{P} (\mathbf{a}_u \odot \mathbf{a}_v) \mathbf{x}_{u,i} \mathbf{x}_{v,i} + \mathbf{r}_i, \\
&\text{s.t. } \mathbf{a}_q \succeq \mathbf{0},\ \mathbf{a}_u \succeq \mathbf{0},\ \mathbf{a}_v \succeq \mathbf{0},\ \mathbf{x}_{q,i} \succeq \mathbf{0},\ \mathbf{x}_{u,i} \succeq \mathbf{0},\ \mathbf{x}_{v,i} \succeq \mathbf{0}, \\
&\sum_{q=1}^{P} \mathbf{x}_{q,i} = 1,\ \sum_{q=1}^{P} \mathbf{x}_{u,i} = 1,\ \sum_{q=1}^{P} \mathbf{x}_{v,i} = 1,
\end{aligned} \tag{3.58}$$

which is clear, however, that the proposed model does not have the ability to generalize the LMM, as the quantity of nonlinear interactions in each pixel is only restricted in two kinds of materials. To effectively address this issue, a generalized bilinear model (GBM) was proposed in [Halimi et al., 2011] by combining the advantages of Eqs. (3.55) and (3.58):

$$\begin{aligned}
\mathbf{y}_i &= \sum_{q=1}^{P} \mathbf{a}_q \mathbf{x}_{q,i} + \sum_{u=1}^{P-1} \sum_{v=u+1}^{P} (\mathbf{a}_u \odot \mathbf{a}_v) \mathbf{x}_{u,i} \mathbf{x}_{v,u,i} \mathbf{x}_{v,i} + \mathbf{r}_i, \\
&\text{s.t. } \mathbf{a}_q \succeq \mathbf{0},\ \mathbf{a}_u \succeq \mathbf{0},\ \mathbf{a}_v \succeq \mathbf{0},\ \mathbf{x}_{q,i} \succeq \mathbf{0},\ \mathbf{x}_{u,i} \succeq \mathbf{0},\ \mathbf{x}_{v,i} \succeq \mathbf{0}, \\
&\sum_{q=1}^{P} \mathbf{x}_{q,i} = 1,\ \sum_{q=1}^{P} \mathbf{x}_{u,i} = 1,\ \sum_{q=1}^{P} \mathbf{x}_{v,i} = 1,\ \mathbf{x}_{u,v,i} \in (0,1).
\end{aligned} \tag{3.59}$$

Most recently, according to the powerful fitting or learning ability of deep learning techniques, many deep models have been proposed to nonlinearly unmix the hyperspectral data one after another. In early stage, [Guo et al., 2015] tried to use a cascade autoencoder network to simultaneously denoise and unmix the hyperspectral image. By adding a partial non-negative constraint, they further improved the autoencoder-based unmixing model [Qu et al., 2017]. Subsequently, the investigators of [Palsson et al., 2018] purposefully designed three special layers in the autoencoder network corresponding to the non-negativity, sum-to-one constraint, and denoise removal (spectral variability), respectively, with the application to the pixel-wise spectral unmixing. In the meanwhile, a stacked non-negative sparse autoencoder was presented for robust hyperspectral unmixing [Su et al., 2018]. [Su et al., 2019] continuously increased the number of layers, yielding a deep unmixing network based on autoencoders. Beyond that, [Hong et al., 2019a] proposed to model the physically meaningful endmembers into the network learning, yielding a weakly supervised unmixing network.

3.3 Multi-Modality Data Analysis

Owing to the innovation and advancement in the imaging system, the availability of data becomes diversity. In particular, with the rapid development of remote sensing techniques, the data are able to be largely captured by different sensors from aircraft and spacecraft. The sharp increase in the diversity and complexity of data collection, however, brings a serious difficulty in processing and analyzing this kind of multimodal data effectively and efficiently. Conversely, it also provides us a new opportunity to use the multimodal data in a complementary way, making it possible to further improve the learning ability of the model. As the proverb says, two heads are better than one. The unimodal data often fail to make a desirable decision, because much of the important information is missing. This motivates us to synthetically leverage the complementary information of multiple data sources in order to jointly contribute to the complex remote sensing tasks. For example, in land cover

and land use classification, there exists a miscellany of buildings in a city scene, which is expected to be classified accurately. In this case, optical remote sensing imagery, e.g., RGB, or multispectral and even hyperspectral data, fails to accomplish this goal to some extent, due to the lack of height information retrieval. What is possible at this time if the Ladar or SAR data are involved in this learning system.

In recent years, some exploratory researches have been made to enhance the representation capability of the learned model by introducing multi-source or multi-temporal remote sensing data. Hyperspectral data have, which is characterized by rich spectral information, received enormous attention in multi-modality data analysis. Targeting at the different application background, the fusion and learning strategies can be roughly categorized into two parts: concentration-based multi-modality fusion learning (CMMFL) and alignment-based cross-modality share learning (ACMSL). A showcase of the differences between the two strategies is clarified in Figure 3.4.

3.3.1 Concentration-based Multi-Modality Fusion Learning

A general but effective way in handling the issue of image or feature-level fusion is concentrating the multi-source data into a stacked vector. Given the inputs of two different modalities $\mathbf{X}_A = [\mathbf{x}_{A,1}, \mathbf{x}_{A,2}, \ldots, \mathbf{x}_{A,m}]$ and $\mathbf{X}_B = [\mathbf{x}_{B,1}, \mathbf{x}_{B,2}, \ldots, \mathbf{x}_{B,m}]$, the output in the form of stack ($\mathbf{X}_C$) can be then expressed as

$$\mathbf{X}_C = [\mathbf{x}_{C,1}, \mathbf{x}_{C,2}, \ldots, \mathbf{x}_{C,m}] = \left[\begin{pmatrix}\mathbf{x}_{A,1}\\ \mathbf{x}_{B,1}\end{pmatrix}, \begin{pmatrix}\mathbf{x}_{A,2}\\ \mathbf{x}_{B,2}\end{pmatrix}, \ \ldots\ , \begin{pmatrix}\mathbf{x}_{A,m}\\ \mathbf{x}_{B,m}\end{pmatrix}\right]. \tag{3.60}$$

Then, the stacked features can be fed into certain classifiers or feature learning models. Figuratively speaking (see the Figure 3.4(a)), the stacking operation might happen in any stages, i.e. it can be performed in the image-level, as shown in Eq. (3.60), and the information is also fused in the feature-level, formulated by

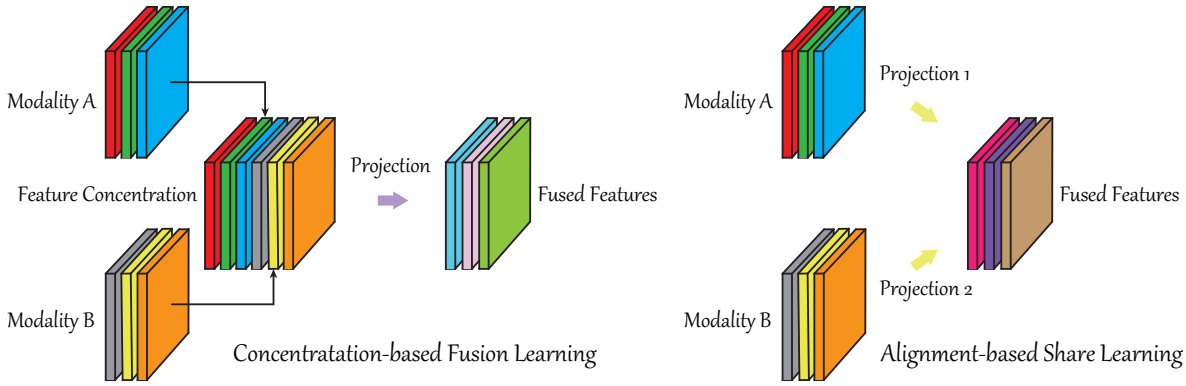
$$\mathbf{F}_C = [\mathbf{f}_{C,1}, \mathbf{f}_{C,2}, \ldots, \mathbf{f}_{C,m}] = \left[\begin{pmatrix}\mathbf{f}_{A,1}\\ \mathbf{f}_{B,1}\end{pmatrix}, \begin{pmatrix}\mathbf{f}_{A,2}\\ \mathbf{f}_{B,2}\end{pmatrix}, \ \ldots\ , \begin{pmatrix}\mathbf{f}_{A,m}\\ \mathbf{f}_{B,m}\end{pmatrix}\right], \tag{3.61}$$

where $\mathbf{F}_C$ denotes the fused feature vectors and its each component can be represented as

$$\mathbf{f}_{C,i} = g_{fe}(\mathbf{X}_{C,i}), \tag{3.62}$$

where g_{fe} is defined as the feature extractor. Of course, the fusion process occurs in the decision-level as well.

In this topic, there have been many representative methods successfully developed and applied for a wide variety of HSI-related applications. Stacked features extracted both spatially and spectrally were adopted in [Camps-Valls et al., 2006] and then were fed into the kernelized support vector machines (SVMs) for hyperspectral image classification. [Chen et al., 2009] performed the hyperspectral data classification by exploiting stacked generalization to combine shape features and spectral information into the SVMs. The authors of [Ghamisi et al., 2014] investigated an automatic spectral-spatial hyperspectral classification with a stacking combination of attribute profiles and its extracted features. Inspired by the graph-based embedding framework, [Liao et al., 2015] concentrated the morphological features extracted from hyperspectral and Lidar data in low-dimensional embedding space. The experimental results demonstrated the effectiveness of this type of stacking strategy. In the previous IEEE GRSS data fusion contest of multimodal land use classification, [Yokoya



(a) CMMFL (b) ACMSL

Fig. 3.4. An illustration to clarify the differences of two different multi-modality data analysis strategies. (a) CMMFL: the features are stacked to jointly learn the fused features via one Projection. (b) ACMSL: the fused features are obtained by learning two different projections (Projections 1 and 2) from two corresponding modalities, respectively.

et al., 2018] won the first place by using multiple stacked features for local climate zones classification. [Kang et al., 2018] augmented the features of the hyperspectral images towards multi-scale and multi-direction generated by Gabor filters [Hong et al., 2014a] as a new input of deep networks. Similarly, [Hang et al., 2019] constructed a cascaded recurrent neural network for hyperspectral image classification. Quite recently, a novel object detector based on cumulative spatial-frequency channel features has been proposed by [Wu et al., 2019b] to yield a robust optical remote sensing imagery detection.

Similarly, there is an undeniable success of using concentration-based fusion strategies in the currently popular deep learning techniques. They give the different names to the fusion of image-level and feature-level, corresponding to early fusion and later fusion, respectively. For example, [Audebert et al., 2016] semantically segmented the earth observation data with multi-modal and multi-scale deep networks. Further, the same authors fused the multimodal data that go beyond the RGB in a FuseNet [Hazirbas et al., 2016] architecture with a residual structure for urban scene parsing [Audebert et al., 2018]. More forcefully, in [Ghamisi et al., 2017], multiple hand-craft features were first extracted by simultaneously collecting the spectrally discriminative features and spatially morphological features of the hyperspectral data as well as Lidar multiextinction profiles with a final deep fusion module, achieving a new multi-modality fusion paradigm. A similar strategy was adopted in the literature [Wu et al., 2018] where multi-scaled and rotation-invariant features learned from a VGG network are input into an ensemble classifier to robustly detect the geospatial objects.

3.3.2 Alignment-based Cross-Modality Share Learning

The completeness of data correspondence is the prerequisite behind the advantages of CMMFL-based methods. This compulsory requirement undoubtedly leads to a poor fit for cross-modality-related tasks [Ngiam et al., 2011]. In contrast to the multi-modality learning (take bi-modality as an example), the cross-modality learning (CML) trains on single modality and tests on bi-modality, or *vice versa* (train on bi-modality and test on single modality).

Such a cross-modality learning problem exists widely in real-world remote sensing tasks. An explicit evidence is the large-scale land cover and land use classification of jointly using hyperspectral and multispectral data. It is obvious that as currently operational optical broadband (multispectral) satellites (e.g. Sentinel-2 or Landsat-8) enable the multispectral data freely available on a nearly global scale. Rather, the geospatial coverage of the hyperspectral data is extremely narrower than the one of multispectral imaging, due to the limitations of satellite imaging techniques and a larger requirement in storage capability.

Nevertheless, we may be able to expect a small amount of such data available. This is a typical CML problem related to transfer learning.

Common or shared subspace learning is an effective solution to address this CML's issue, where manifold alignment (MA) is a relatively mature algorithm group by aligning multiple modalities into a latent subspace, thereby yielding an effective knowledge transfer. The key idea of MA can be generalized as learning a common (or shared) subspace where different data sources can be aligned to learn a joint feature representation. Figure 3.4(b) gives an illustrative vision example. Intuitively, MA [Wang and Mahadevan, 2009] follows similar steps with manifold learning techniques of graph-based embedding in the following:

1) In supervised models, the aligned graph structure is computed based on labels, while for unsupervised models, the weights between samples can be automatically generated by similarity measurement.
2) Once the weighted graph structure is given, the aligned subspace can be naturally estimated by calculating a low-dimensional manifold embedding.

By preserving a joint manifold structure that consists of a label-inspired aligned graph and weighted graph measured by the similarity between unlabeled data, semi-supervised MA allows different data sources to be better transformed into a shared latent subspace The model in [Wang and Mahadevan, 2011] is a classic semi-supervised MA approach, which fuses three different properties into a joint manifold structure, that is, similarity matrix $\mathbf{W}_s$:

$$\mathbf{W}_s = \begin{pmatrix} \mathbf{W}_s^{1,1} & \mathbf{W}_s^{1,2} & \dots & \mathbf{W}_s^{1,K} \\ \mathbf{W}_s^{2,1} & \mathbf{W}_s^{2,2} & \dots & \mathbf{W}_s^{2,K} \\ \dots & \dots & \dots & \dots \\ \mathbf{W}_s^{K,1} & \mathbf{W}_s^{K,2} & \dots & \mathbf{W}_s^{K,K} \end{pmatrix}, \tag{3.63}$$

dissimilarity matrix $\mathbf{W}_d$:

$$\mathbf{W}_s = \begin{pmatrix} \mathbf{W}_d^{1,1} & \mathbf{W}_d^{1,2} & \dots & \mathbf{W}_d^{1,K} \\ \mathbf{W}_d^{2,1} & \mathbf{W}_d^{2,2} & \dots & \mathbf{W}_d^{2,K} \\ \dots & \dots & \dots & \dots \\ \mathbf{W}_d^{K,1} & \mathbf{W}_d^{K,2} & \dots & \mathbf{W}_d^{K,K} \end{pmatrix}, \tag{3.64}$$

and $\mathbf{W}_t$ that describes the topology structure of individual data by the means of *knn*, is defined by

$$\mathbf{W}_t = \begin{pmatrix} \mathbf{W}_t^{1,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_d^{2,2} & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{W}_t^{K,K} \end{pmatrix}, \tag{3.65}$$

where K stands for the number of data source, and $\mathbf{W}_s^{i,j}$ can be computed, respectively, by

$$\mathbf{W}_s^{i,j} = \begin{cases} 1, & \text{if } \mathbf{x}_A^i \text{ and } \mathbf{x}_B^j \text{ are from the same class;} \\ 0, & \text{otherwise,} \end{cases} \tag{3.66}$$

while $\mathbf{W}_d^{i,j}$ is

$$\mathbf{W}_d^{i,j} = \begin{cases} 1, & \text{if } \mathbf{x}_A^i \text{ and } \mathbf{x}_B^j \text{ are from the different class;} \\ 0, & \text{otherwise,} \end{cases} \tag{3.67}$$

and $\mathbf{W}_t^{i,j}$ can be given by

$$\mathbf{W}_t^{i,j} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}, & \text{if } \mathbf{x}_B^j \in \phi_k(\mathbf{x}_A^i); \\ 0, & \text{otherwise.} \end{cases} \tag{3.68}$$

When the joint manifold structure is ready to go, the cost function of semi-supervised MA is composed of three scalars:

$$A = \frac{1}{2} \sum_{A=1}^{K} \sum_{B=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\mathbf{P}_A^{\mathrm{T}} \mathbf{x}_A^i - \mathbf{P}_B^{\mathrm{T}} \mathbf{x}_B^j\|^2 \mathbf{W}_s^{i,j}, \tag{3.69}$$

$$B = \frac{1}{2} \sum_{A=1}^{K} \sum_{B=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\mathbf{P}_A^{\mathrm{T}} \mathbf{x}_A^i - \mathbf{P}_B^{\mathrm{T}} \mathbf{x}_B^j\|^2 \mathbf{W}_d^{i,j}, \tag{3.70}$$

and

$$C = \frac{1}{2} \sum_{t=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{m} \|\mathbf{P}_t^{\mathrm{T}} \mathbf{x}_t^i - \mathbf{P}_t^{\mathrm{T}} \mathbf{x}_t^j\|^2 \mathbf{W}_t^{i,j}. \tag{3.71}$$

As a result, the overall cost function can be written as

$$\mathcal{L} = \frac{(A + C)}{B}. \tag{3.72}$$

By minimizing the Eq. (3.72), we then can obtain the corresponding solution (the to-be-estimated projection $\mathbf{P}$) by solving a GED solver.

Different from the CMMFL that aims to learning the stacked multimodal representation, ACMSL is able to adaptively shuttle back and forth between the different modalities or domains by the means of the learned common subspace. The former is suitable for the case of holding complete paired data between the multimodalities, which may be met only in a small-scale case. Yet the latter ACMSL would have a good fit for the case of large-scale classification or mapping, due to the shared learning strategy. In [Matasci et al., 2011], the hyperspectral data in the source and target domains are linearly projected into a shared subspace in which the spectral gap between the two domains is expected to be reduced for hyperspectral image classification. Inspired by the idea of semi-supervised MA, as mentioned above, [Tuia et al., 2014] attempted to align the multi-view remote sensing images on manifolds by fully allowing for the view variants between the images captured from the different angles. This seems to play a registration-like role to avoid the effects of multi-view variants to some extent. A similar work was proposed in [Matasci et al., 2015], which is developed based on an excising transfer component analysis [Pan et al., 2011], making it applicable for remote sensing image classification. [Tuia and Camps-Valls, 2016] proposed to nonlinearly align the multimodal data in a higher dimensional kernel-induced space instead of in the original space. Integrated with semi-supervised MA, authors of [Hu et al., 2019] presented a novel graph construction strategy with the use of mathematically topological data analysis (TDA) for learning the fusion of hyperspectral and polarimetric SAR images. Moreover, [Liu et al., 2019] designed a two-stream convolutional neural network for the fusion of spatiotemporal images.

It should be noted, however, that due to considerable heterogeneity of multimodal data, either CMMFL-based or those previously-proposed ACMSL methods fails to activate the connections across modalities, yielding a relatively weak transferability of multi-modality. To this end, [Hong et al., 2019d] proposed to learn a common subspace by aligning the

multimodalities on a latent subspace where the features are apt to be better blended. Beyond the pure supervised models, the same investigators [Hong et al., 2019e] further explored the potential of the common subspace learning and proposed a semi-supervised learning framework to align the data structure on a learnable manifold space.

4 Summary of the Work

Linking up with one general goal of this thesis as well as its spin-off three research objectives, as mentioned in Chapter 1, six solutions are contrapuntally proposed to address the corresponding challenges in six peer-reviewed articles by the author (full as the first author), published in one top conference paper and five journal papers, respectively. Accordingly, this chapter draws a brief summary of these articles by highlighting the following sixfold contributions.

- **Contribution 1:** To alleviate the effects of non-uniform distribution of HSI and multicollinearity of affinity matrix computation in the traditional manifold embedding methods, a robust manifold representation learning is proposed for nonlinearly spatial-spectral HDR in section 4.1.
- **Contribution 2:** Considering the trade-off between the robustness and discrimination in HDR, a joint & progressive learning strategy (J-Play) is developed in section 4.2 for supervised hyperspectral image classification.
- **Contribution 3:** By deeply investigating the statistical characteristics between spectral signatures and spectral variabilities, a low-coherent prior is modeled into the LMM to yield the proposed augmented linear mixing model (ALMM). This resulting model addresses the spectral variability by separating the variabilities from the original spectral reconstruction problem (see section 4.3).
- **Contribution 4:** A novel subspace-based insight to see the issue of spectral unmixing, called SULoRA, is provided in section 4.4 by jointly estimating the low-rank subspace projections and abundance maps in the process of unmixing.
- **Contribution 5:** The first attempt in methodology is made to survey the role of hyperspectral data in the large-scale earth observation tasks. Section 4.5 will introduce a specific case of multispectral image classification in a large area with the aid of a partially overlapped HSI. This process will be achieved by a simple but effective common subspace learning (CoSpace) algorithm.
- **Contribution 6:** In section 4.6, a follow-up work with regard to the same problem mentioned in section 4.5 is presented to adaptively learn an aligned cross-modal representation on the learnable manifolds in a semi-supervised fashion, named learnable manifold alignment (LeMA).

4.1 Robust Local Manifold Representation for HDR

Appendix A aims at addressing two great challenges that traditional manifold learning methods are facing in the topic of HDR. It is well-known that local manifold learning (LML) is mainly characterized by affinity matrix construction, which is composed of two steps: neighbor selection and computation of affinity weights. More specifically,

1) The neighbor selection is sensitive to complex spectral variability due to non-uniform data distribution, illumination variations, and sensor noise;
2) The computation of affinity weights is challenging due to highly correlated spectral signatures in the neighborhood.

To this end, a novel manifold learning methodology based on locally linear embedding (LLE) is proposed in this work through learning a robust local manifold representation (RLMR). More specifically, a hierarchical neighbor selection (HNS) is designed to progressively eliminate the effects of complex spectral variability using joint normalization (JN) and to robustly compute affinity (or reconstruction) weights reducing multicollinearity via refined neighbor

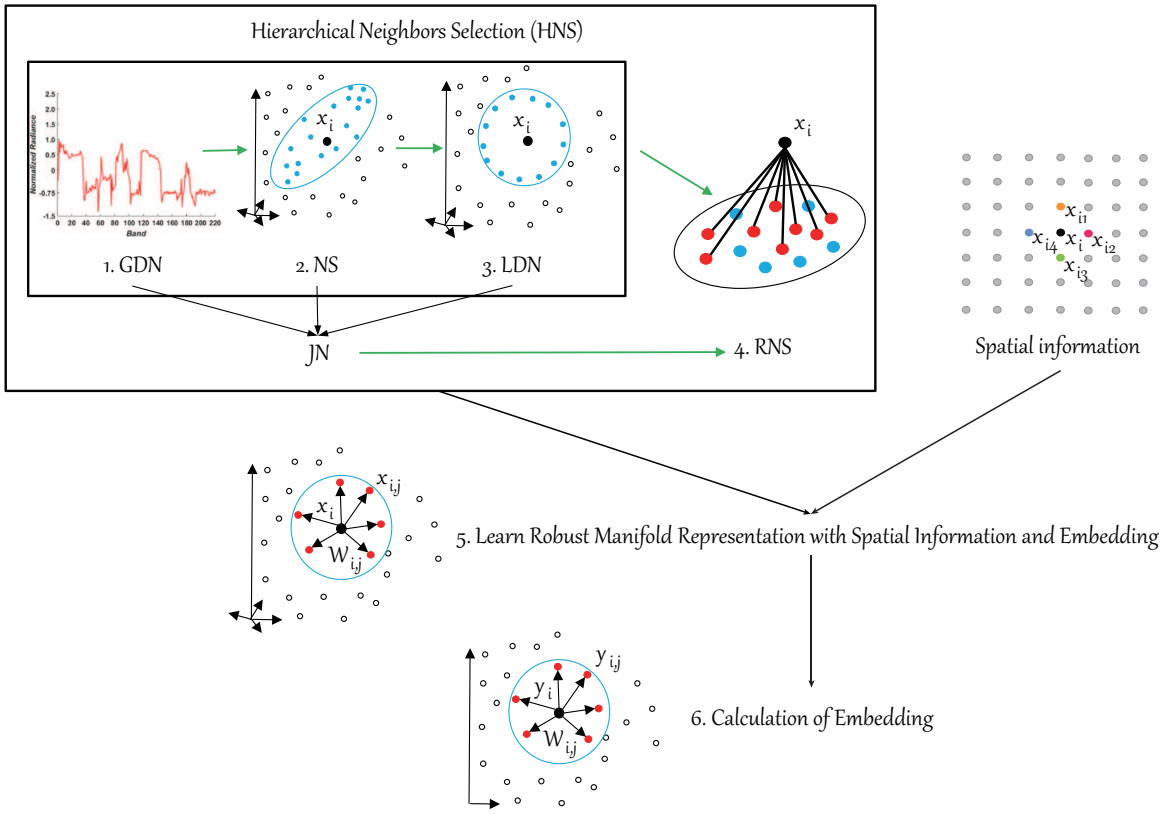


Fig. 4.1. The workflow of the proposed RLMR algorithm.

selection (RNS). Additionally, an idea that jointly embeds the spatial-spectral information is introduced into the proposed manifold learning methodology to further improve the robustness of affinity calculations. An illustrative workflow of the proposed RLMR is shown in Figure 4.1, and the specific procedures are detailed as follows:

- ◇ *Step 1.* Global data normalization (GDN) is performed to deal with spectral variability modeled by scaling and shifting.
- ◇ *Step 2.* Neighbor selection (NS) coarsely selects local neighbors of the target pixel.
- ◇ *Step 3.* Local data normalization (LDN) is applied to make local data distribution more uniform and isotropic and further eliminate locally spectral variability.
- ◇ *Step 4.* RNS aims at mitigating multicollinearity in local manifold space, making it possible to obtain a relatively accurate and intrinsic structure of the underlying manifold.
- ◇ *Step 5.* Computation of reconstruction weights with contextual information jointly embeds spectral and spatial information for a robust calculation of the reconstruction weights.
- ◇ *Step 6.* Calculation of embedding obtains the low-dimensional feature representation by embedding robust local manifold properties into the low-dimensional space.

4.1.1 Hierarchical Neighbors Selection

To solve the first challenge regarding the sensitivity to complex spectral variability, HNS is purposefully proposed by combining the JN and the RNS, as shown in Figure 4.2.

A. JN

Data normalization aims at reducing the effect of numerous variations and improving the performance of subsequent algorithms. Generally, data normalization includes GDN and LDN. The purpose of GDN is to mitigate illumination variations and modify the global data distribution so that it is more uniform and isotropic, enabling them to be measured

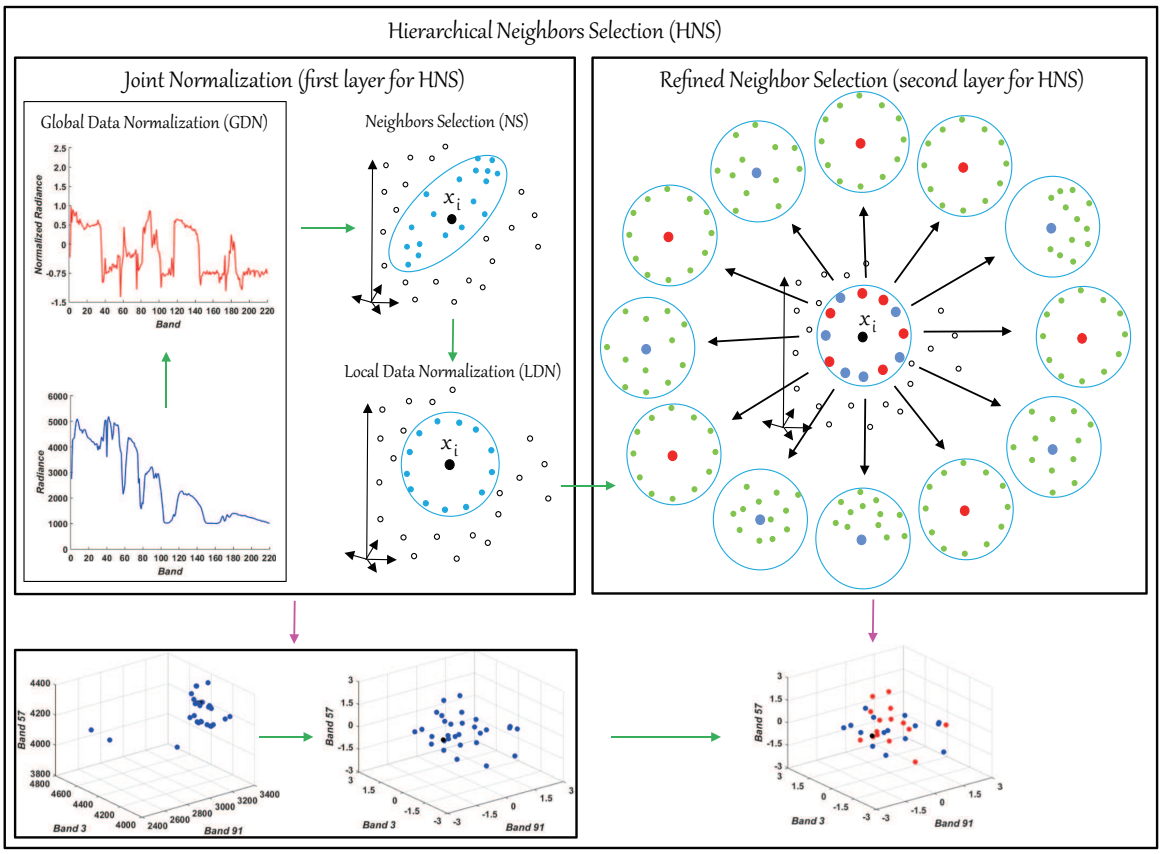


Fig. 4.2. A detailed diagram of hierarchical neighbors selection, including 2-D and 3D visualization.

in the same, or similar, level or unit. Therefore, GND should be performed on the whole hyperspectral image. Unlike GDN, LDN tends to uniformize the mean and variance of the local neighborhood, which is good for eliminating the local mutation in the spectral domain. Owing to the merits of GDN and LDN, JN is an appropriate approach to effectively address the issues of spectral variability and non-uniform data distribution. There is a step-wise implementation in the following.

(1) GDN: it performs the following computations:

$$\begin{aligned} \mathbf{x}_i^{ns} &= \frac{\mathbf{x}_i^o - c_i^o}{s_i^o}, \\ \mathbf{x}_i^g &= (\mathbf{x}_i^{ns} - \mathbf{c}^{ns})./\mathbf{s}^{ns}, \end{aligned} \tag{4.1}$$

where the operator "./" means the element-wise division; $\mathbf{x}_i^o$ denotes the *i-th* original spectral signature and c_i^o and s_i^o are defined as the mean value and variance of $\mathbf{x}_i^o$, respectively, while the $\mathbf{x}_i^{ns}$ is specified as the normalized representation of $\mathbf{x}_i^o$. By collecting all spectral signatures, we then have the matrix form of normalized spectral signatures $\mathbf{X}^{ns}$, whose mean and variance values can be represented as $\mathbf{c}^{ns}$ and $\mathbf{s}^{ns}$.

(2) LDN: After selecting coarse neighbors for each data point using the Euclidean distance, LDN is exploited to make the data distribution more uniform and isotropic in a local manifold space, which can be formulated as

$$\mathbf{x}_{i,j}^l = \begin{cases} (\mathbf{x}_i^g - \mathbf{c}_i^g)./\mathbf{s}_i^g, & j = 0 \\ (\mathbf{x}_{i,j}^g - \mathbf{c}_i^g)./\mathbf{s}_i^g, & j = 1, 2, \ldots, K, \end{cases} \tag{4.2}$$

where the $\mathbf{c}_i^g$ and $\mathbf{s}_i^g$ stand for the mean and variance of the globally normalized spectral

features ($\mathbf{X}_i^g = \left[\mathbf{x}_i^g, \mathbf{x}_{i,1}^g, \ldots, \mathbf{x}_{i,j}^g, \ldots, \mathbf{x}_{i,K}^g\right]$) obtained by Eq. (4.1) in the *i-th* data point and its K neighbors, respectively. $\mathbf{X}_i^l = \left[\mathbf{x}_i^l, \mathbf{x}_{i,1}^l, \ldots, \mathbf{x}_{i,j}^l, \ldots, \mathbf{x}_{i,K}^l\right]$ is the final output of spectral features for a given *i-th* data point and its surroundings by JN.

B. RNS

After JN, the influence of spectral variability has been mitigated to a great extent, but multicollinearity still exists among neighbors. Multicollinearity leads to an inaccurate estimation of the affinity matrix, thereby degrading the quality of the local manifold structure. To address this issue, refined neighbor selection (RNS) is performed as the second layer of HNS. RNS can mitigate the effects of multicollinearity by matching locally neighboring manifold structures to reduce information redundancy. The first step is to construct the local structure features $\mathbf{F}_p^{local}$ for each data point p by the means of its neighbor's information $\mathbf{X}_p^l = \left[\mathbf{x}_{p,1}^l, \ldots, \mathbf{x}_{p,j}^l, \ldots, \mathbf{x}_{p,K}^l\right]$, thereby the $\mathbf{F}_p^{local}$ can be formed by the distance property between the feature of p with those of its neighbors using a RBF function:

$$\begin{aligned} \mathbf{F}_{p,j}^{local} &= e^{(-\|\mathbf{x}_p^l - \mathbf{x}_{p,j}^l\|_2^2)}, \\ \mathbf{F}_p^{local} &= \left[F_{p,j}^{local}, \ldots, F_{p,j}^{local}, \ldots, F_{pK}^{local}\right], \end{aligned} \tag{4.3}$$

and the second step is to to screen out new local neighbors that hold similar data distribution (local manifold structure) using a Kullback-Leibler divergence (KLD). Thus, the differences of the above local features, defined as $\mathbf{d}^f = \left[d_1^f, \ldots, d_q^f, \ldots, d_K^f\right]$ between the point p and its certain neighbor q can be measured as:

$$d_q^f = KLD(\mathbf{F}_p^{local} \| \mathbf{F}_q^{local}) + \alpha KLD(\mathbf{F}_q^{local} \| \mathbf{F}_p^{local})), \tag{4.4}$$

where

$$\begin{aligned} KLD(\mathbf{F}_p^{local} \| \mathbf{F}_q^{local}) &= \sum_{j=1}^{K} F_{p,j}^{local} \times log_2\left(\frac{F_{p,j}^{local}}{F_{q,j}^{local}}\right), \\ KLD(\mathbf{F}_q^{local} \| \mathbf{F}_p^{local}) &= \sum_{j=1}^{K} F_{q,j}^{local} \times log_2\left(\frac{F_{q,j}^{local}}{F_{p,j}^{local}}\right). \end{aligned} \tag{4.5}$$

where α is a penalty parameter balancing the two terms described in Eq. (4.5). Using the Eq. (4.4), the neighbors with the k smallest values are chosen from the coarse neighbors as the new neighbors of the data point p, namely $\mathbf{X}_p^{nl} = \left[\mathbf{x}_{p,1}^{nl}, \ldots, \mathbf{x}_{p,j}^{nl}, \ldots, \mathbf{x}_{p,K}^{nl}\right]$. k is the final number of neighbors for each point, and we make the value of K equal to twofold k.

4.1.2 Spatial-Spectral Contextual Information Embedding

To further improve the robustness of the calculation of reconstruction weights, the spatial information is incorporated into linear reconstructions. We assume that spatially neighboring spectral pixels can be explained by the same or similar reconstruction weights [Chen et al., 2011], if spatially neighboring pixels include similar spectral components. Following this, the calculation of reconstruction weights can be re-formulated with the spatial-spectral contextual constraint that the reconstruction weights of the target pixel are approximately

equal to the average of those of its neighboring pixels, by

$$\mathbf{a}_i^0 = \arg\min_{\mathbf{a}_i^0} \sum_{s=0}^{4} \left\| \mathbf{x}_{i,s}^{nl} - \mathbf{X}_i^{nl} \mathbf{a}_i^s \right\|_2^2 \text{ s.t. } \|\mathbf{X}_i^{nl}(4\mathbf{a}_i^0 - \sum_{s=1}^{4} \mathbf{a}_i^s)\|_2^2 \leq \eta, \ (\mathbf{a}_i^s)^{\mathrm{T}} \mathbf{a}_i^s = 1, \ s = 0, 1, \ldots, 4, \quad (4.6)$$

where $\{\mathbf{x}_{i,s}^{nl}\}_{s=0}^4$ are the target spectral pixel and its four spatial neighbors, respectively, and $\{\mathbf{a}_i^s\}_{s=0}^4$ are the corresponding reconstruction weights. η is a tiny real number (e.g., 10^{-3}) that represents the tolerant errors.

To optimize the problem of Eq. (4.6), we rewrite it as a joint optimization problem:

$$\mathbf{a}_i^0 = \arg\min_{\mathbf{a}_i^0} \sum_{s=0}^{4} \left\| \widehat{\mathbf{X}}_i^{nl} - \mathbf{L}\widehat{\mathbf{A}}_i \right\|_{\mathrm{F}}^2, \quad (4.7)$$

which is subject to $\mathbf{C}\widehat{\mathbf{A}}_i = [1 \ 1 \ 1 \ 1 \ 1]^{\mathrm{T}}$,

$$\mathbf{L} = \begin{bmatrix} 4\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} \\ \mathbf{X}_i^{nl} & & & & \\ & \mathbf{X}_i^{nl} & & & \\ & & \mathbf{X}_i^{nl} & & \\ & & & \mathbf{X}_i^{nl} & \\ & & & & \mathbf{X}_i^{nl} \end{bmatrix}, \quad (4.8)$$

and

$$\widehat{\mathbf{A}}_i = \begin{bmatrix} \mathbf{a}_i^0 \\ \mathbf{a}_i^1 \\ \mathbf{a}_i^2 \\ \mathbf{a}_i^3 \\ \mathbf{a}_i^4 \end{bmatrix}, \quad \widehat{\mathbf{X}}_i^{nl} = \begin{bmatrix} \mathbf{0} \\ \mathbf{x}_{i,0}^{nl} \\ \mathbf{x}_{i,1}^{nl} \\ \mathbf{x}_{i,2}^{nl} \\ \mathbf{x}_{i,3}^{nl} \\ \mathbf{x}_{i,4}^{nl} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{e} & & & & \\ & \mathbf{e} & & & \\ & & \mathbf{e} & & \\ & & & \mathbf{e} & \\ & & & & \mathbf{e} \end{bmatrix}, \quad (4.9)$$

where $\mathbf{e}$ represents the unit vector with the size of $1 \times k$ and the parameter β is to balance the importance between error item and constraints. By relaxing the constraint with the Lagrange multiplier (λ), the analytical solution of Eq. (4.7) can be obtained using matrix derivation operation as

$$\mathbf{a}_i^0 = (\mathbf{L}^{\mathrm{T}}\mathbf{L} + \lambda\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}(\mathbf{L}^{\mathrm{T}}\widehat{\mathbf{X}}_i^{nl} + \lambda\mathbf{C}^{\mathrm{T}}\mathbf{e}). \quad (4.10)$$

4.1.3 Performance Assessment: A Case of Classification

Classification is explored as a potential application for validating the proposed algorithm. Experimental results are validated in comparison with other state-of-the-art HDR methods using two common classifiers: nearest neighbor (NN) based on the Euclidean distance and linear SVMs, on two hyperspectral datasets (Indian Pines [Baumgardner et al., 2015] and Houston2013 [Pacifici et al., 2013]).

A. Classification Results

Table 2. Quantitative performance comparison of nine HDR methods using two classifiers (NN and linear SVMs) under two different sampling strategies (random sampling and region-based sampling) in terms of OA and AA on the two used hyperspectral datasets (Indine Pines and Houston2013). The optimal parameters for all algorithms are determined by 10-fold cross-validation on the training set. The parameter v denotes the variance of Gaussian kernel only for KPCA; OSF and LTSA are the acronym of original spectral features and local tangent space alignment, respectively.

Indine Pines Dataset

Methods	Optimal Parameters	Random Sampling				Region-based Sampling			
		OA (%)		AA (%)		OA (%)		AA (%)	
		NN	SVM	NN	SVM	NN	SVM	NN	SVM
OSF	/ (/)	64.74	72.72	44.78	56.67	73.86	76.04	47.39	61.87
PCA	$d = 50$ ($d = 50$)	64.62	72.66	44.74	56.64	70.60	79.50	47.82	58.38
KPCA	$d = 50$, $v = 10$ ($d = 60$, $k = 10$)	66.95	76.03	48.79	61.25	72.16	80.88	50.36	63.52
LLE	$d = 60$, $k = 40$ ($d = 40$, $k = 50$)	68.49	75.51	47.45	59.55	71.47	72.51	47.23	62.49
LE	$d = 60$, $k = 7$ ($d = 80$, $k = 3$)	59.57	68.19	40.92	52.73	56.93	65.06	36.59	52.85
LTSA	$d = 60$, $k = 70$ ($d = 40$, $k = 70$)	71.22	81.12	51.63	66.09	75.49	84.93	52.79	64.51
JN	$d = 70$, $k = 40$ ($d = 90$, $k = 60$)	72.99	82.92	52.20	66.35	76.52	83.03	52.83	66.95
HNS	$d = 70$, $k = 40$ ($d = 100$, $k = 50$)	77.45	85.61	53.61	67.62	78.75	85.04	54.73	68.03
RLMR	$d = 50$, $k = 80$ ($d = 40$, $k = 90$)	**85.84**	**90.83**	**55.24**	**68.21**	**87.06**	**90.93**	**56.92**	**69.24**

Houston Dataset

Methods	Optimal Parameters	NN		Methods	Optimal Parameters	SVM	
		OA (%)	AA (%)			OA (%)	AA (%)
OSF	/	72.83	76.16	OSF	/	74.68	77.84
PCA	$d = 50$	72.85	76.19	PCA	$d = 30$	74.78	77.79
KPCA	$d = 50$, $v = 10$	73.80	77.79	KPCA	$d = 30$,$v = 10$	75.12	78.14
LLE	$d = 40$, $k = 50$	74.23	77.49	LLE	$d = 60$,$v = 40$	75.33	78.03
LE	$d = 60$, $k = 20$	66.70	70.66	LE	$d = 20$,$v = 30$	70.71	72.98
LTSA	$d = 40$, $k = 50$	75.40	78.75	LTSA	$d = 30$,$v = 50$	76.04	79.18
JN	$d = 60$, $k = 50$	77.45	80.69	JN	$d = 70$,$v = 60$	77.86	80.12
HNS	$d = 80$, $k = 70$	78.52	81.75	HNS	$d = 90$,$v = 60$	78.98	82.01
RLMR	$d = 70$, $k = 50$	**80.87**	**82.77**	RLMR	$d = 90$,$v = 100$	**81.13**	**82.79**

In the first dataset, classification accuracy in the use of different dimensionality reduction methods is evaluated and compared, while two kinds of strategies are applied in selecting the training and test samples: random sampling and region-based sampling. Table 2 lists the performance comparison of different algorithms under the two different sampling conditions in terms of overall accuracy (OA) and average accuracy (AA). Correspondingly, Figure 4.3 also shows the classification maps (CMs) of nine HDR methods using the two classifiers under two different sampling strategies of training samples.

In particular, the CMs obtained by RLMR are smoother than those of other methods in locally spatial regions due to the embedding of spatial information. This demonstrates the effectiveness of three kinds of proposed technical components of the RLMR, i.e. JN, RNS, and spatial-spectral information embedding. Furthermore, one can be also observed from

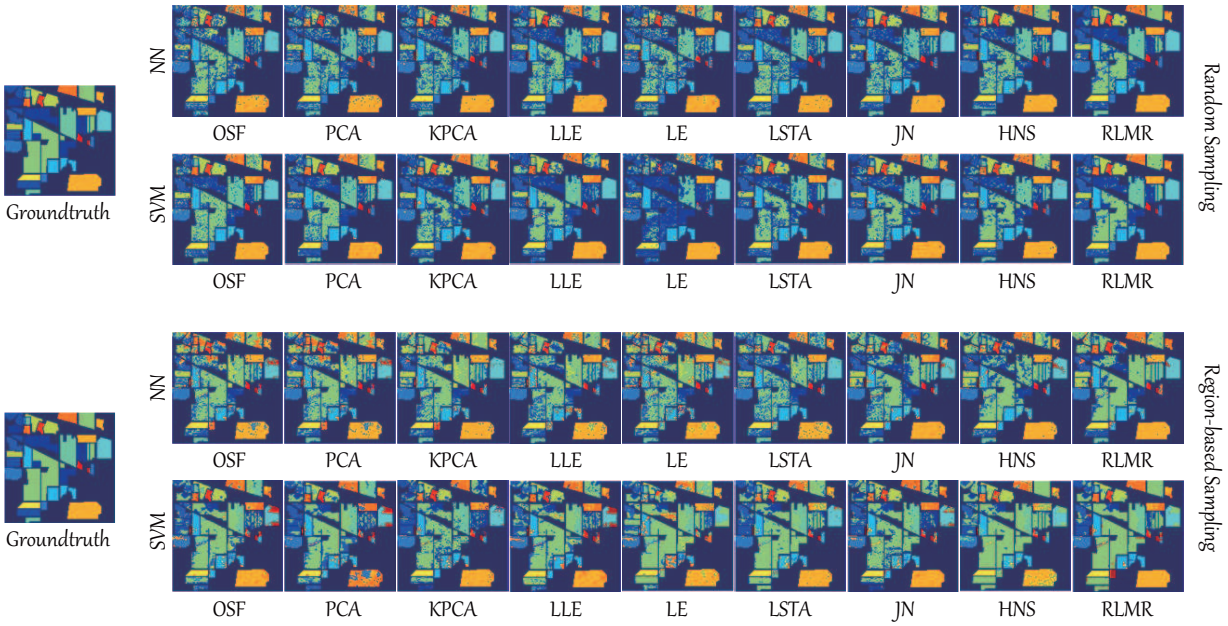
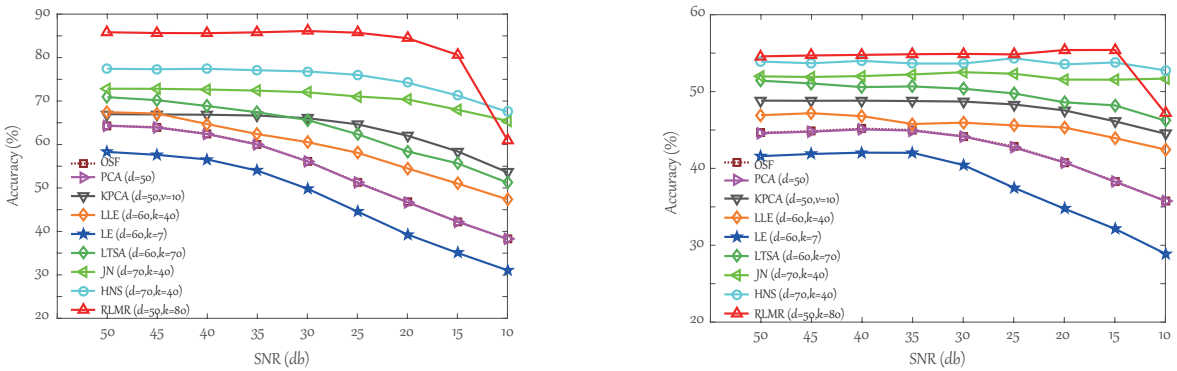


Fig. 4.3. Classification maps of nine HDR methods for the Indian Pines dataset using NN and SVM classifiers under two different sampling strategies of training samples: random sampling and region-based sampling.



(a) random sampling (b) region-based sampling

Fig. 4.4. Robustness analysis of all compared methods with different SNRs on the Indine Pines dataset in terms of classification accuracy. (a) classification results of random sampling. (b) classification results of region-based sampling.

Table that the performance of JN, HNS, and RLMR is progressively increased, owing to the contributions of normalization, RNS, and spatial information, respectively.

For the Houston2013 dataset, the quantitative performance comparison is given in Table 2 with the fixed training and test sets. Similarly, Figure 4.5 visually shows the corresponding CMs corresponding to the results of NN and SVM, respectively. Note that a general framework for the out-of-samples extension [Bengio et al., 2004] is used in this paper to approximate the large-scale LML-based HDR. As shown in the false-color image of Figure 4.5, the east side of the scene is covered with shadows of clouds, resulting in the performance degradation of those previous HDR algorithms, while the results of the proposed RLMR are rather robust against this variability using both NN and SVM classifiers.

B. Analysis of Sensitivity and Robustness against Noise

The sensitivity of parameters is investigated and discussed by varying the various parameters, e.g., the number of neighbors (k), the dimension of subspace (d), and the variance (v) of kernel only for Kernel PCA (KPCA). Please see the **Appendix A** for more specific results with the changes of these parameters. It is worth noting that due to the robustness of our proposed method (RLMR), its results remain stable with the increase in the num-

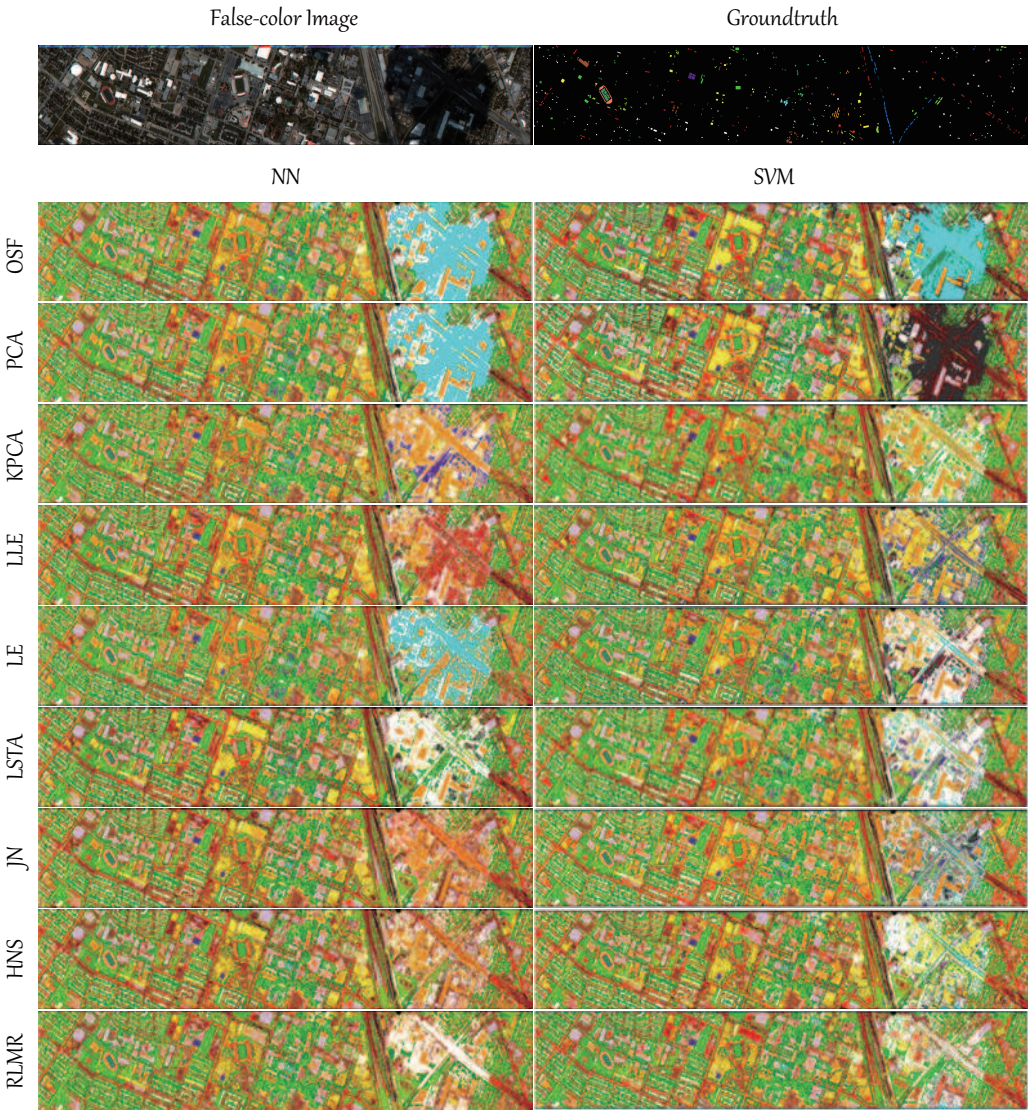


Fig. 4.5. Classification maps of Houston2013 dataset using all HDR methods with two different classifiers (NN and SVM).

ber of neighbors k and reduced dimensionality d. Conversely, the performances of JN and HNS are progressively degrading with the change of parameters; particularly in a situation with a large k, the classification accuracies even degrade to a level similar to classical LML methods.

To validate the robustness of the proposed RLMR, a further experiment is performed, which adds noise with a different signal-to-noise-ratio (SNR) into the first dataset. Figure 4.4 shows the classification accuracies under the two sampling strategies: (a) random sampling, (b) region-based sampling. As the SNR decreases, the performance of JN, HNS, and RLMR are comparatively stable and superior compared to those of classical LML methods, PCA, KPCA, and original spectral features. This demonstrates the robustness of the proposed method against noise and implies its effectiveness for low SRN hyperspectral images.

4.2 Joint & Progressive Learning of Hyperspectral Data

Despite the fact that nonlinear subspace learning techniques (e.g., manifold learning) have successfully applied to low-dimensional data representation, yet there is still room for improvement in explainability (explicit mapping), generalization (out-of-samples), and cost-effectiveness (linearization). Plus, limited by the imaging devices and environment, aerial

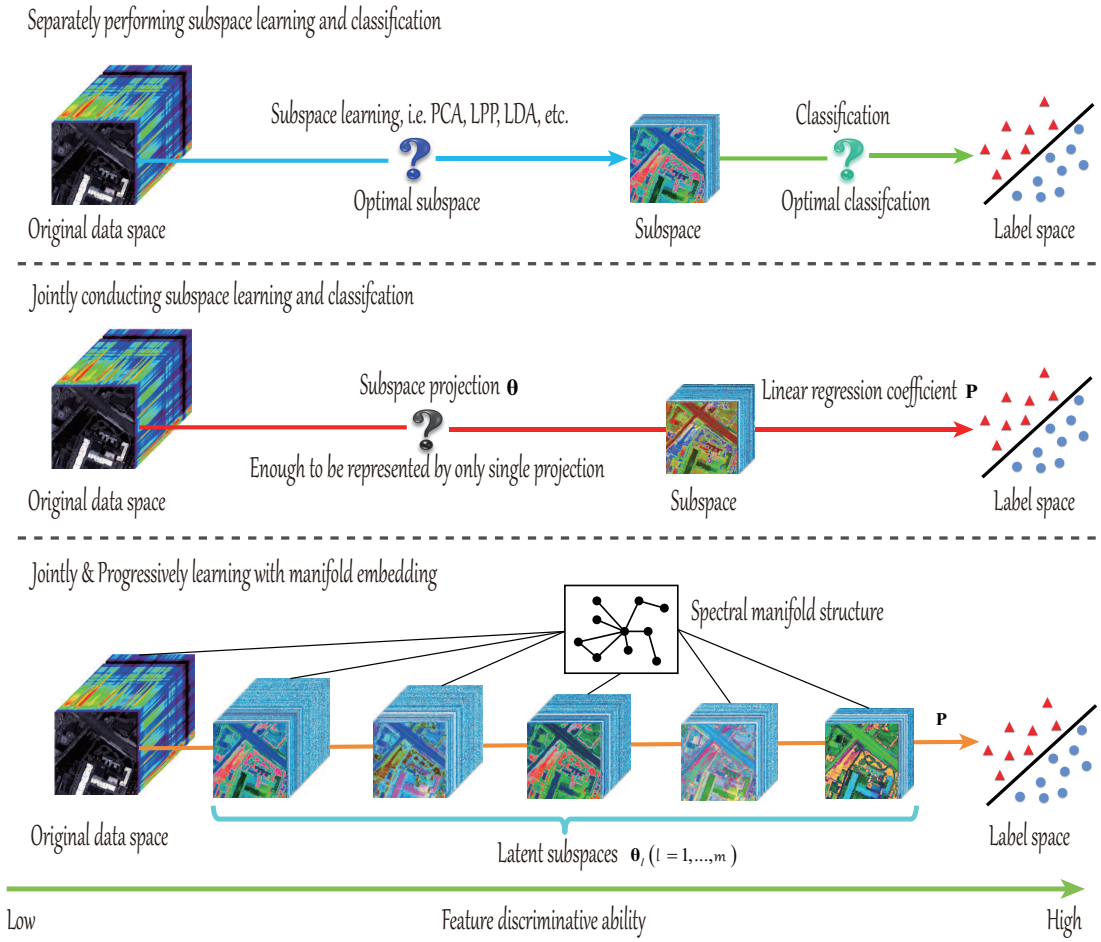


Fig. 4.6. The motivation interpolation from separately performing subspace learning and classification to joint learning to joint & progressive learning again. The subspace learned from the proposed model demonstrates higher feature discrimination as clarified by the green bottom line.

or spaceborne HSI inevitably suffers from the quality degradation in a more complex way (spectral variability), making it difficult to spectrally discriminate the materials (pixel-wise classification). Fortunately, we found that such spectral variability fails to be fitted well by linear-based reconstruction models. In light of the discovery, we progressively search the potential optimal subspace through multiple coupled linear transformations, and meanwhile the intrinsic structure of the data (e.g., manifold prior) can be effectively preserved in the process of subspace search. (See **Appendix B**)

Towards this goal, we develop a novel learning strategy, namely joint & progressive learning, with the application to HDR, by

1) jointly performing multiple subspace learning and classification to find a latent subspace where samples are expected to be better classified;
2) progressively learning multi-coupled projections to linearly approach the optimal mapping bridging the original space with the most discriminative subspace;
3) locally embedding manifold structure in each learnable latent subspace.

The motivation of the proposed method can be interpolated step by step by Figure.

4.2.1 HDR from the View of Subspace Learning

A. General Remark

Subspace learning is to find a low-dimensional space where we expect to maximize certain properties of the original data, e.g. variance (PCA), discriminative ability (LDA), and graph

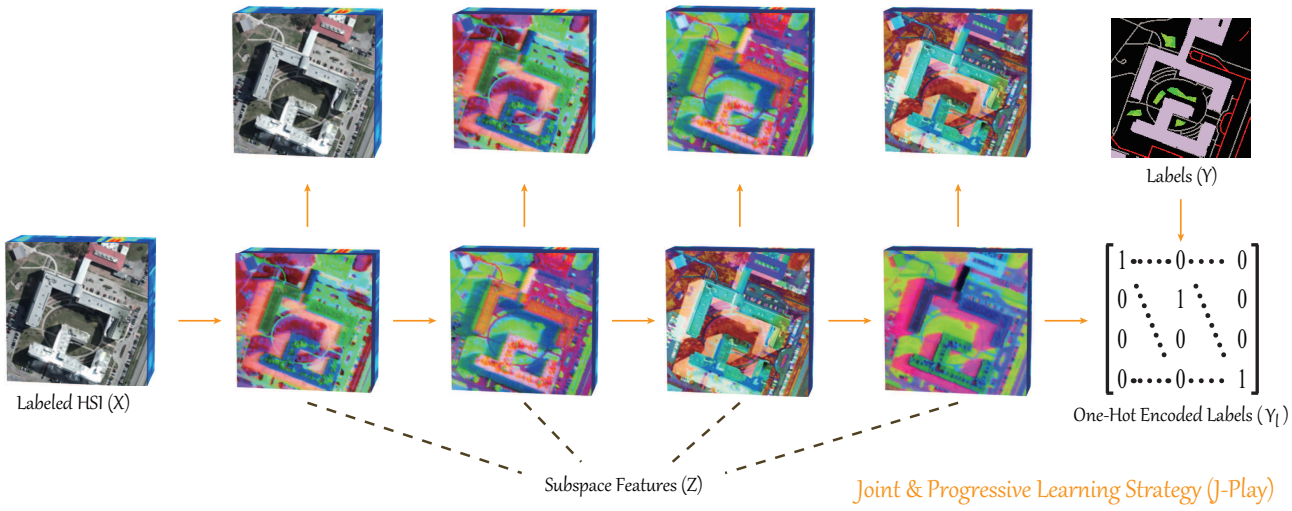


Fig. 4.7. The illustration of the proposed J-Play framework.

structure (graph embedding). [Yan et al., 2007] summarized these subspace learning methods in a GGE framework, which can be formulated as

$$\min_{\mathbf{Z}} \operatorname{tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}) \ \text{ s.t. } \ \mathbf{Z}\mathbf{D}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}, \tag{4.11}$$

whose linearized version can be reformulated on the basis of Eq. (4.11) by substituting $\mathbf{\Theta X}$ for $\mathbf{Z}$:

$$\min_{\mathbf{\Theta}} \operatorname{tr}(\mathbf{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}) \ \text{ s.t. } \ \mathbf{\Theta}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}. \tag{4.12}$$

As opposed to the GGE framework, a regression-based joint learning model [Ji and Ye, 2009] can explicitly bridge the learned latent subspace and labels by

$$\min_{\mathbf{P},\mathbf{\Theta}} \frac{1}{2}\|\mathbf{Y}_l - \mathbf{P}\mathbf{\Theta}\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \ \text{ s.t. } \ \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}. \tag{4.13}$$

More details regarding the above models can be found in Chapter 2.

B. Problem Formulation

On the basis of Eq. (4.13), we further extend the framework in a progressive learning strategy, which can be formulated to be the following constrained optimization problem

$$\begin{aligned} \min_{\mathbf{P},\{\mathbf{\Theta}_l\}_{l=1}^m} & \frac{1}{2}\sum\nolimits_{l=1}^{m}\|\mathbf{X}_{l-1} - \mathbf{\Theta}_l^{\mathrm{T}}\mathbf{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{P}\mathbf{\Theta}_m \ldots \mathbf{\Theta}_l \ldots \mathbf{\Theta}_1\mathbf{X}\|_{\mathrm{F}}^2 \\ & + \frac{\beta}{2}\sum\nolimits_{l=1}^{m}\operatorname{tr}(\mathbf{\Theta}_l\mathbf{X}_{l-1}\mathbf{L}\mathbf{X}_{l-1}^{\mathrm{T}}\mathbf{\Theta}_l^{\mathrm{T}}) + \frac{\gamma}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \\ & \text{s.t. } \ \mathbf{X}_l = \mathbf{\Theta}_l\mathbf{X}_{l-1}, \ \ \mathbf{X}_l \geq 0, \ \ \|\mathbf{x}_{lk}\|_2 \leq 1, \ \ \forall l = 1, 2, \ldots, m, \end{aligned} \tag{4.14}$$

where $\mathbf{X}$ is assigned to $\mathbf{X}_0$, and α, β, and γ are three penalty parameters corresponding to the different terms. In detail, the first term is a reconstruction loss term in order to effectively eliminate the spectral variability, since we found that such complex variability usually fails to be reconstructed linearly. The second term yields progressive prediction through multi-coupled projections. To facilitate structure learning, we also perform the local manifold regularization to each latent subspace, as shown in the third term. Figure 4.7 illustrates the holistic J-Play framework.

Table 3. Quantitative performance comparisons on two hyperspectral datasets with optimal dimensions determined by 10-fold cross-validation via three different classifiers – NN: nearest neighbor, KSVM: kernel SVM, and CCF: canonical correlation forests [Rainforth and Wood, 2015]. Note that J-Play$_l$ denotes the J-Play method with l number of layers. The best results for the different classifiers are shown in bold.

Methods	Optimal Dimensions	Indian Pines Dataset			Houston2013 Dataset		
		NN	KSVM	CCF	NN	KSVM	CCF
Baseline	(220/144)	65.89%	66.56%	81.71%	72.83%	80.19%	82.60%
PCA	(20/20)	65.40%	75.25%	79.26%	72.75%	79.54%	83.90%
LPP	(20/30)	64.86%	63.02%	68.48%	75.31%	78.43%	81.77%
LDA	(15/14)	64.14%	63.88%	65.61%	75.81%	76.66%	79.62%
LFDA	(15/14)	73.86%	74.25%	75.17%	75.52%	80.46%	82.27%
LSDR	(50/40)	73.67%	76.84%	77.38%	76.80%	80.39%	81.64%
LSQMID	(60/80)	66.94%	78.90%	79.32%	76.31%	80.23%	81.69%
J-Play$_1$	(20/30)	78.81%	82.04%	82.24%	78.22%	83.32%	85.09%
J-Play$_2$	(20/30)	80.87%	83.75%	83.23%	79.16%	**84.41**%	85.15%
J-Play$_3$	(20/30)	83.59%	85.08%	84.44%	**80.13**%	83.68%	**88.19**%
J-Play$_4$	(20/30)	**83.92**%	85.21%	**84.57**%	79.64%	83.25%	85.63%
J-Play$_5$	(20/30)	83.76%	**85.30**%	84.41%	80.00%	82.21%	85.81%
J-Play$_6$	(20/30)	83.56%	84.79%	83.82%	79.69%	82.45%	84.82%
J-Play$_7$	(20/30)	82.70%	83.82%	83.04%	77.81%	81.03%	83.23%

Moreover, the non-negativity constraint with respect to each learned dimension-reduced feature (e.g., $\{\mathbf{X}_l\}_{l=1}^m \geq 0$) is considered since we aim to obtain a meaningful low-dimensional feature representation similar to original image data acquired in a non-negative unit.

4.2.2 Model Learning Process

A. Auto-reconstructing Initialization

Obviously, the proposed model is complex and the non-convex, hence we pre-train our model to have an initial approximation of subspace projections $\{\mathbf{\Theta}_l\}_{l=1}^m$ as this can greatly reduce the model's training time and also help finding an optimal solution easier. For that, we propose a pre-training model with respect to $\mathbf{\Theta}_l, \forall l = 1, ..., m$ by simplifying Eq.(4.14) as

$$\min_{\mathbf{\Theta}_l} \frac{1}{2}\|\mathbf{X}_{l-1} - \mathbf{\Theta}_l^{\mathrm{T}}\mathbf{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\mathbf{\Theta}_l\mathbf{X}_{l-1}\mathbf{L}\mathbf{X}_{l-1}^{\mathrm{T}}\mathbf{\Theta}_l^{\mathrm{T}}) \ \text{ s.t. } \ \mathbf{X}_l \geq 0, \ \|\mathbf{x}_{lk}\|_2 \leq 1, \tag{4.15}$$

which is named as **auto**-**r**econstructing **u**nsupervised **le**arning (AutoRULe). It can be effectively solved via the ADMM-based framework, where the AutoRULe needs to be initialized by LPP as well. Please find the details in the **Appendix B**.

B. Global Algorithm of J-Play

Given the outputs of AutoRULe, the solution of Eq. (4.14) can be obtained by an alternatively minimizing strategy that separately solves two individual subproblems with respect to the variables $\{\mathbf{\Theta}_l\}_{l=1}^m$ and $\mathbf{P}$. The detailed procedures for the global algorithm of J-Play are given as follows:

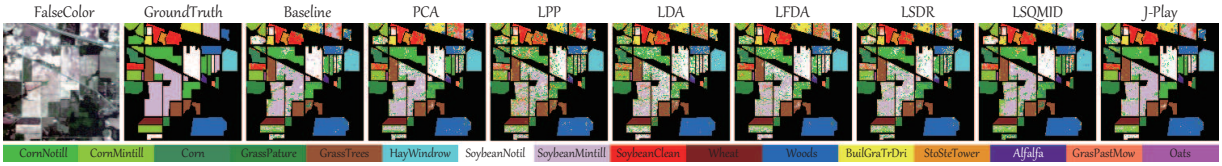


Fig. 4.8. A false-color image, groundtruth, and classification maps of the different algorithms obtained using CCF classifier on the Indine Pines dataset with the corresponding categories.

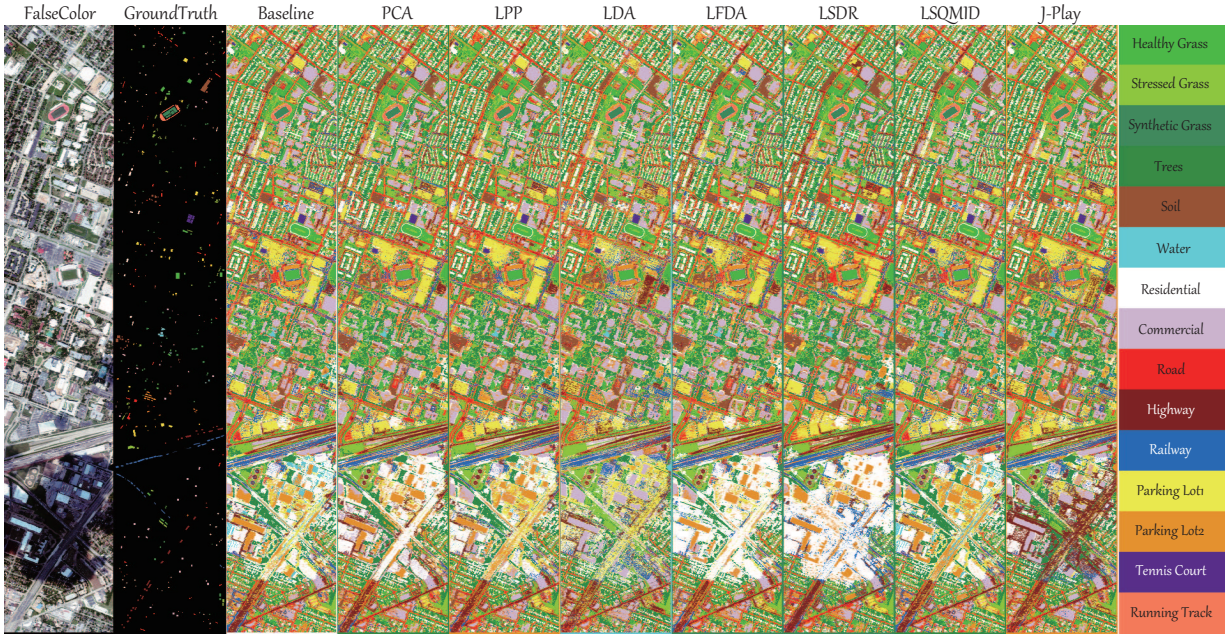


Fig. 4.9. A false-color image, groundtruth, and classification maps of the different algorithms obtained using CCF classifier on the Houston2013 dataset with the corresponding categories.

1) **Initialization Step:** Greedily initialize layer-wise $\mathbf{\Theta}_l$ corresponding to each latent subspace:
   ① $\mathbf{\Theta}_l^0 \leftarrow LPP(\mathbf{X}_{l-1})$
   ② $\mathbf{\Theta}_l \leftarrow AutoRULe(\mathbf{X}_{l-1}, \mathbf{\Theta}_l^0, \mathbf{L})$
   ③ $\mathbf{X}_l \leftarrow \mathbf{\Theta}_l \mathbf{X}_{l-1}$
   ④ repeat until l is equal to m
2) **Fine-tuning Step:**
   ① Fix other variables to update $\mathbf{P}$ by solving a subproblem of $\mathbf{P}$
   ② Fix other variables to update all $\{\mathbf{\Theta}_l\}_{l=1}^m$ by solving a subproblem of $\mathbf{\Theta}_l$
   ③ Repeat these optimization procedures until a stopping criterion is satisfied

4.2.3 Results and Analysis on Hyperspectral Data

Similarly in Section 4.1, the two same hyperspectral datasets are used for the performance assessment of HDR. Table 3 lists classification performances of the different methods with the optimal subspace dimensions obtained by cross-validation using three different classifiers. Correspondingly, the classification maps are given in Figures 4.8 and 4.9 to intuitively highlight the differences.

Remarkably, the performance of the proposed method (J-Play) is superior to the other methods on the two hyperspectral datasets. This indicates that J-Play is prone to learn a better feature representation and robust against noise. On the other hand, with the increase of m, the performance of J-Play steadily increases to the best with around 4 or 5 layers for the first dataset and 2 or 3 layers for the second one, and then gradually decreases with a slight

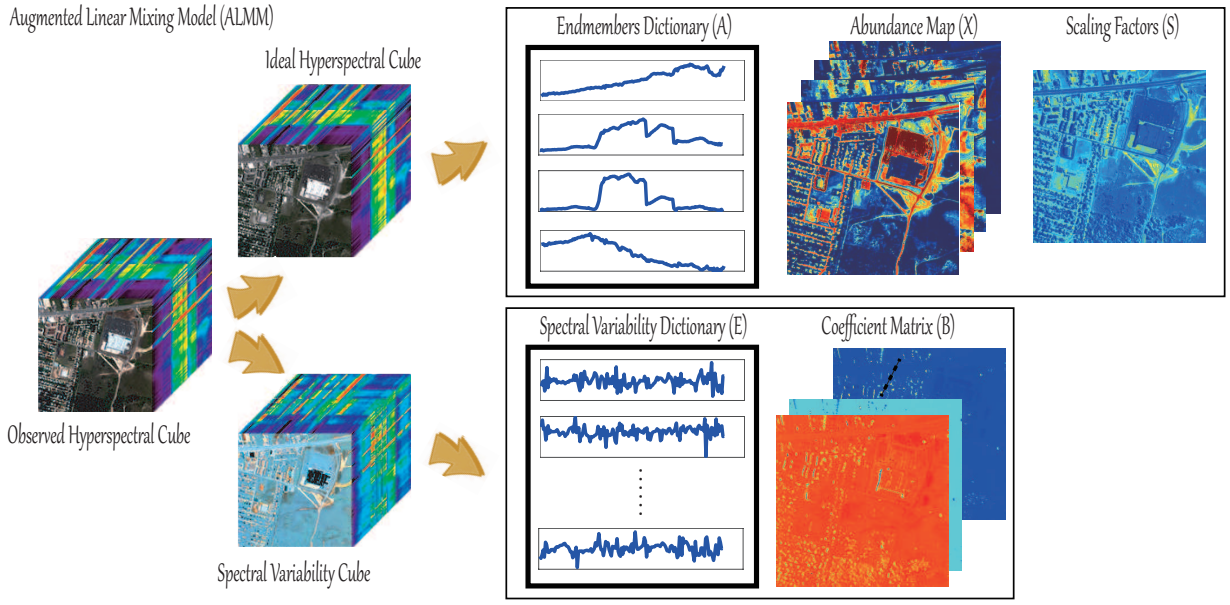


Fig. 4.10. The holistic diagram of spectral unmixing using the proposed ALMM.

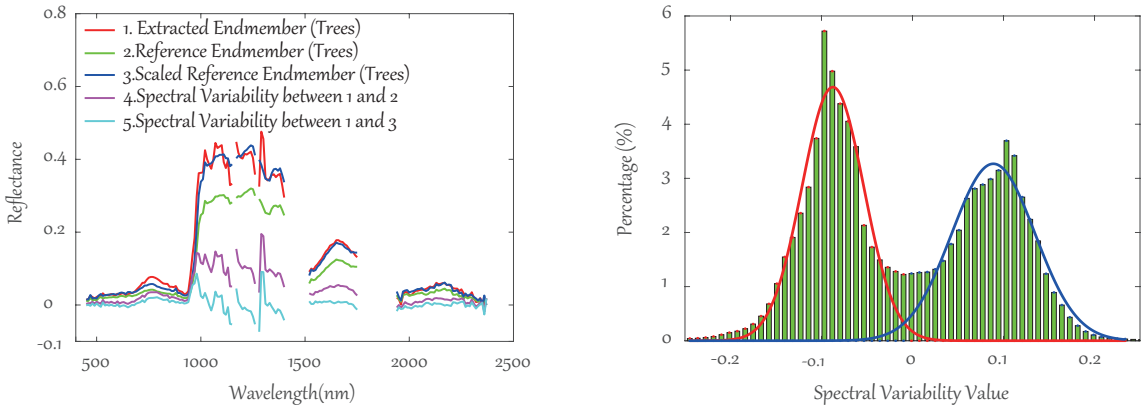
perturbation since our model is only trained on the training set.

4.3 Low-Coherence Learning for Hyperspectral Unmixing

In **Appendix C**, a novel augmented linear mixing model (ALMM) is proposed to address spectral variability for robust hyperspectral unmixing. The classical unmixing model, the linear mixing model (LMM), generally fails to accurately estimate abundance maps due to the existence of spectral variability. Thus, the proposed ALMM separately addresses the principle spectral variability (scaling factors) generated by variations in illumination or typography by means of the endmember dictionary and models other spectral variabilities caused by environmental conditions (e.g., atmospheric effects) and instrumental configurations (e.g., sensor noise) as well as material nonlinear mixing effects, by introducing an additional spectral variability dictionary in inverse problems of hyperspectral unmixing.

During the process, we found that the pure spectral signature should be low-coherent with the spectral variability, leading to the low-coherence learning strategy. Therefore, we formulate this property into our model so that the algorithm can jointly learn the spectral variability dictionary and estimate the abundance maps. An illustration for the ALMM is given in Figure 4.10. More specifically, the contributions of ALMM can be summarized as follows:

1) We propose a novel spectral mixture model, namely (ALMM), where scaling factors are modeled by the endmember dictionary and an additional dictionary is introduced to model the rest of spectral variabilities simultaneously;
2) A data-driven dictionary learning method is explored in the proposed framework of spectral unmixing in which a statistical prior is given, specifying that the spectral variability (except for scaling factors) be low-coherent with endmember spectral signatures, thereby achieving low-coherence learning for hyperspectral unmixing;
3) An optimization algorithm based on the alternating direction method of multipliers (ADMM) is designed to solve the proposed model.



(a) an example for spectral variability

(b) spectral variability distribution

Fig. 4.11. An explicit example to clarify the spectral variability. (a): The line (red) 1 denotes the endmember of the trees extracted using VCA from the Urban scene acquired from http://www.tec.army.mil/Hypercube, while the line (green) 2 is the corresponding reference endmember (Trees). The line (blue) 3 is estimated by multiplying a scaling factor on line 2. Line 4 (or 5) illustrates the differences between 1 and 2 (or 3) to clarify the existence of other spectral variabilities besides scaling factors. (b) gives a statistical distribution of spectral variability in the Urban scene that it is not a simple Gaussian distribution rather than more like a more complex Gaussian mixture distribution.

4.3.1 Spectral Variability Modeling

A. Existence and Complexity of Spectral Variability

Spectral variability refers to a variation of a spectral signature for a given material, due to illumination conditions and topography, atmospheric effects, or even the intrinsic variability of the material. A showcase of spectral variability is detailed in Figure 4.11(a). Its existence dramatically degrades the unmixing performance of traditional LMM-based methods, since the spectral variability is complex and does not strictly obey a Gaussian distribution in real scenarios. Direct evidence supporting this point is shown in Figure 4.11(b), which approximately satisfies a mixed Gaussian distribution.

B. Physical Significance of Spectral Variability Dictionary

Although most spectral variabilities coherent with endmembers (**A**) can be represented by scaling factors, yet the remaining spectral variabilities from either intra-class or inter-class can still hurt the unmixing performance in reality. An example is illustrated in Figure 4.12 to clarify that the spectral variability can not be fully explained by the scaled endmembers. Accordingly, we draw two points by reasoning as follows: 1) the scaled endmembers obtained by adding scaling factors on endmembers (**A**) fail to fully fit the gap in-between; 2) The errors marked in cyan of Figure 4.12(a) could be explained by spectral variabilities or a certain new material. We try to identify the errors by means of the USGS spectral library, generating the abundances with respect to the various materials as shown in Figure 4.12(d) where there is a rather high abundance in *Axinite* ranked as the second major component following *Actinolite*.

On the other hand, the physical significance of **E** could be also explained from the perspectives of intra-class and inter-class spectral variabilities. Without **E**, the intra-class spectral variability could be absorbed by endmembers (**A**), further leading an inaccurate estimation of abundance maps (**X**). If **E** is considered as inter-class spectral variability dictionary, and then the term (**EB**) might represent the spectral signatures of certain new materials that are not discovered by the LMM (see Figure 4.12 for example). The **E** used in the ALMM is therefore capable of calibrating the class-specific spectral variabilities into a unified or generalized spectral variability, which enables to simultaneously handle the intra- and inter-class variabilities. Figure 4.13 shows statistical evidence by collecting all cosine values between endmembers and spectral variabilities, where the cosine value is basically around 0,

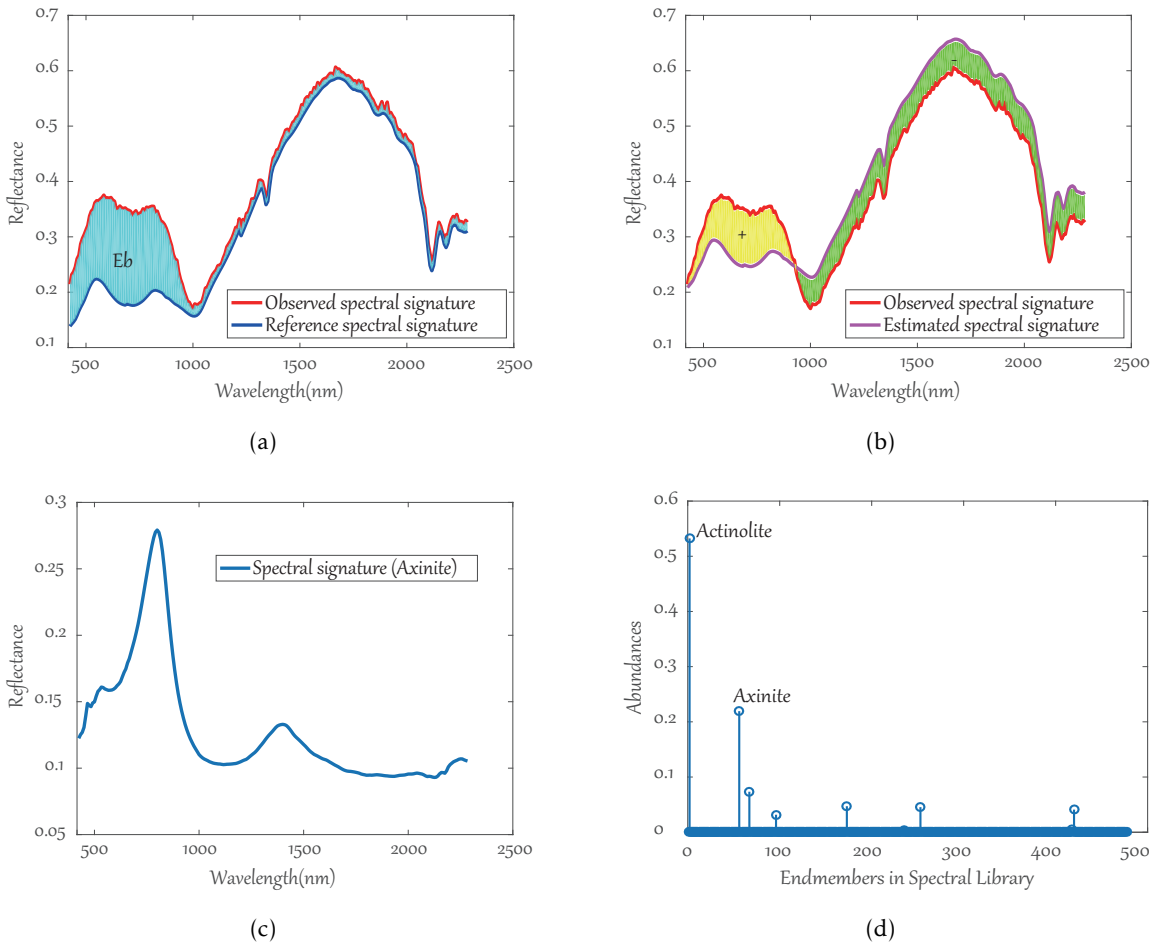


Fig. 4.12. An example in the real Cuprite scene to illustrate the physical meaning of **E**. (a) shows the differences (**Eb**) between the observed spectral signature and the real spectral signature that can not be explained by the endmember dictionary (**A**), but it can be represented well by an additional spectral variability dictionary (**E**). Correspondingly, if without **E**, the differences (spectral variability) could be absorbed by **A** as shown in (b), leading an inaccurate estimation of abundance maps (**X**). (c) gives a spectral signature of the material *Axinite* and (d) shows a real case of unmixing the observed spectral signature using USGS spectral library that except the *Actinolite*, the *Axinite* occupies the main abundances, which can well represents the **Eb** in (a).

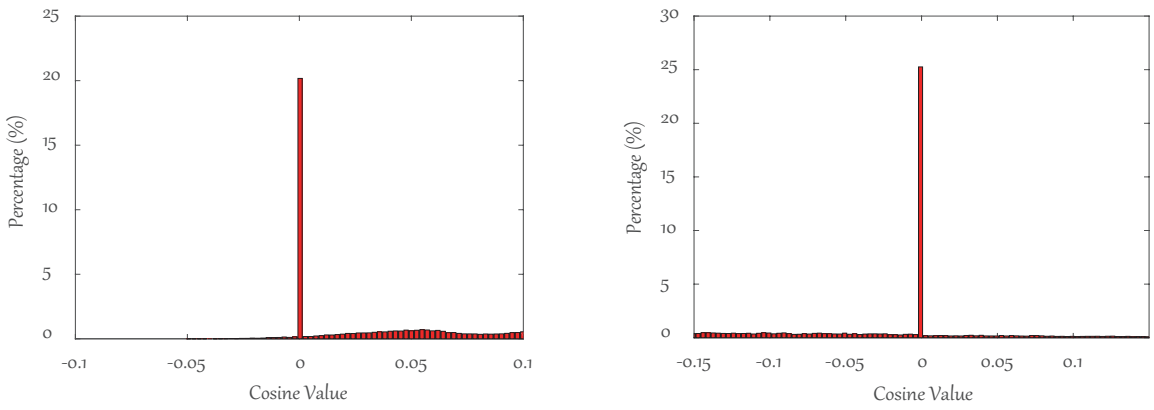


Fig. 4.13. Statistics of Cosine Value between endmembers and spectral variabilities on the first simulated dataset and real Urban scene, respectively, where the spectral variabilities are obtained by calculating the intra- and inter-class differences between the extracted endmembers and the given reference endmembers.

indicating that the spectral variability should, to a great extent, be low-coherent with the endmembers. This is basically consistent with the conclusion summarized above.

4.3.2 Augmented Linear Mixing Model

According to the aforementioned analysis and discussion of spectral variability, this augmented linear mixing model, or ALMM, is expressed by

$$\mathbf{Y} = \mathbf{AXS} + \mathbf{EB} + \mathbf{R}, \tag{4.16}$$

where $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_m, ..., \mathbf{e}_L] \in \mathbb{R}^{D \times L}$ denotes the spectral variability matrix (or dictionary), and L is the number of basis vectors in $\mathbf{E}$. The expression $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_k, ..., \mathbf{b}_N] \in \mathbb{R}^{L \times N}$ is the coefficient matrix corresponding to $\mathbf{E}$.

With the necessary non-negativity constraint, the problem (4.16) can be formulated as the following constrained optimization problem:

$$\min_{\mathbf{X},\mathbf{B},\mathbf{S},\mathbf{E}} \frac{1}{2}\|\mathbf{Y} - \mathbf{AXS} - \mathbf{EB}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{X}\|_{1,1} + \frac{\beta}{2}\|\mathbf{B}\|_{\mathrm{F}}^2 + \frac{\gamma}{2}\|\mathbf{A}^{\mathrm{T}}\mathbf{E}\|_{\mathrm{F}}^2 + \frac{\eta}{2}\|\mathbf{E}^{\mathrm{T}}\mathbf{E} - \mathbf{I}\|_{\mathrm{F}}^2 \ \text{ s.t. } \mathbf{X} \geq 0, \ \mathbf{S} \geq 0, \tag{4.17}$$

where the regularization terms successively correspond to *abundance regularization* ($\mathbf{X}$), *spectral variability coefficient regularization* ($\mathbf{B}$), and *spectral variability dictionary regularization* ($\mathbf{E}$), respectively.

Furthermore, non-negativity constraints ($\mathbf{X} \geq \mathbf{0}$ and $\mathbf{S} \geq \mathbf{0}$) usually have to be considered to satisfy the physical assumption. In addition to the non-negativity constraint, the sum-to-one also plays an important role in the abundance map. However, this constraint is relaxed by step-wise scaling operator, since the variables $\mathbf{X}$ and $\mathbf{S}$ are bundled together, leading to difficulty satisfying the sum-to-one constraint for $\mathbf{X}$. The solution and the specific details about the motivation of designing these terms of Eq. (4.17) can be thoroughly listed in **Appendix C**.

4.3.3 Visualization of Unmixing Results

Three datasets: a synthetic dataset presented in [Drumetz et al., 2016] and two real datasets over an urban area and the mining district in Cuprite, Nevada, are applied to visually evaluate the performance among the proposed ALMM and other state-of-the-art approaches, including FCLSU, PCLSU, SPCLSU, SUnSAL (ℓ_1-CLSU), SSUnSAL (scaled SUnSAL), as well as PLMM and ELMM.

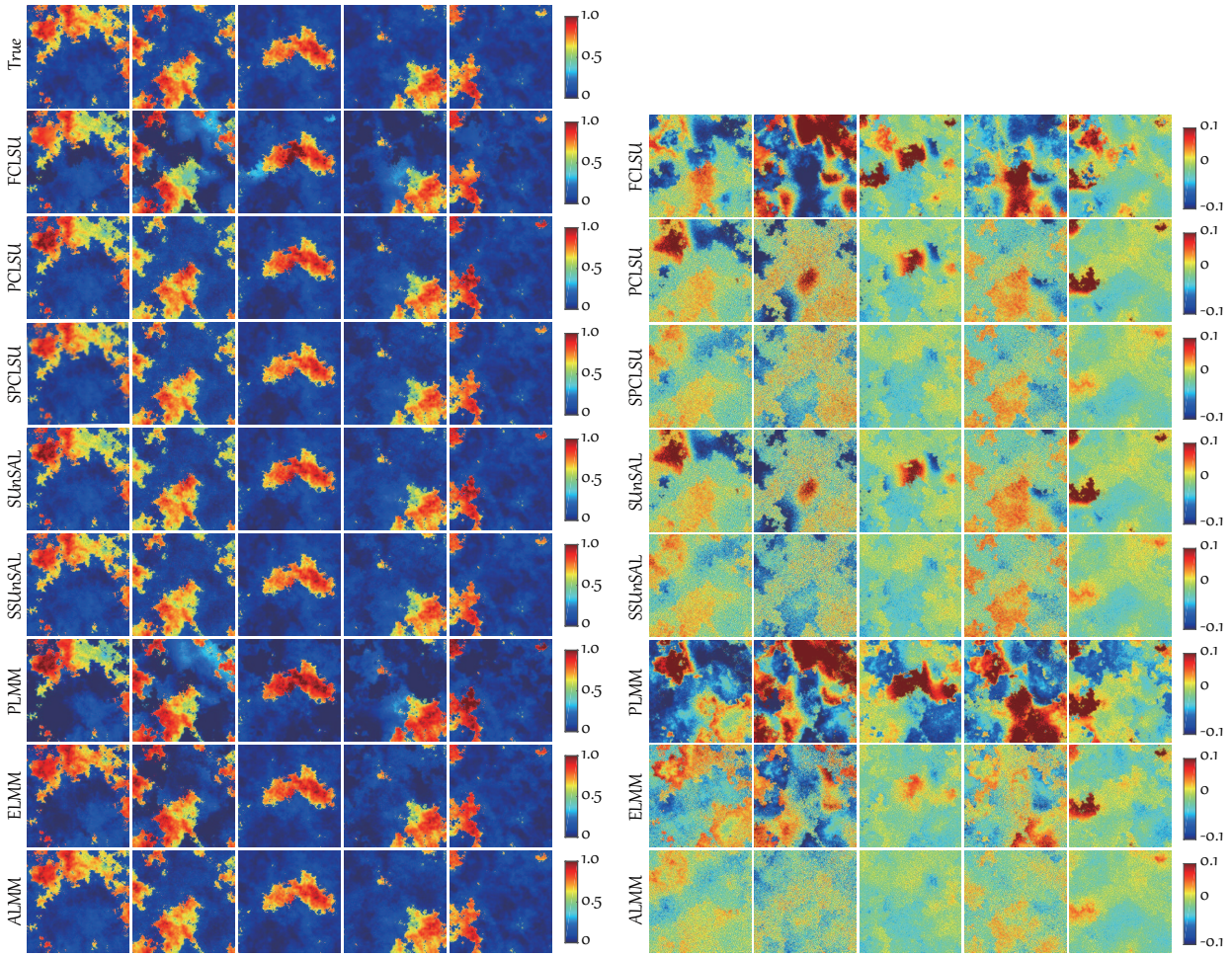
Simulated Hyperspectral Scene

Fig. 4.14(a) shows the estimated abundance maps for the aforementioned algorithms. Since the visual difference of the estimated abundance maps is not obvious among some of the algorithms, the abundance difference maps are also given in Fig. 4.14(b) to intuitively highlight the difference.

By comparison, the proposed method outperforms other algorithms, which suggests that this method can effectively learn the spectral variability, improving the accuracy of the abundance estimation. Figure 4.14(b) illustrates a more significant comparison by means of abundance difference maps between the ground truth and estimated abundance maps of the compared algorithms. The difference values obtained from ALMM are mostly close to zero, which indicates that the performance of ALMM is superior to that of the other methods.

Real Hyperspectral Scenes

For the second hyperspectral dataset – a real urban scene, we perform the SAM-based classification using the reference endmembers as reference spectra. The first row of Figure 4.15(a) shows the cosine similarity for the four classes, where negative samples are masked out with



(a) simulated dataset: abundance maps (b) simulated dataset: difference abundance maps

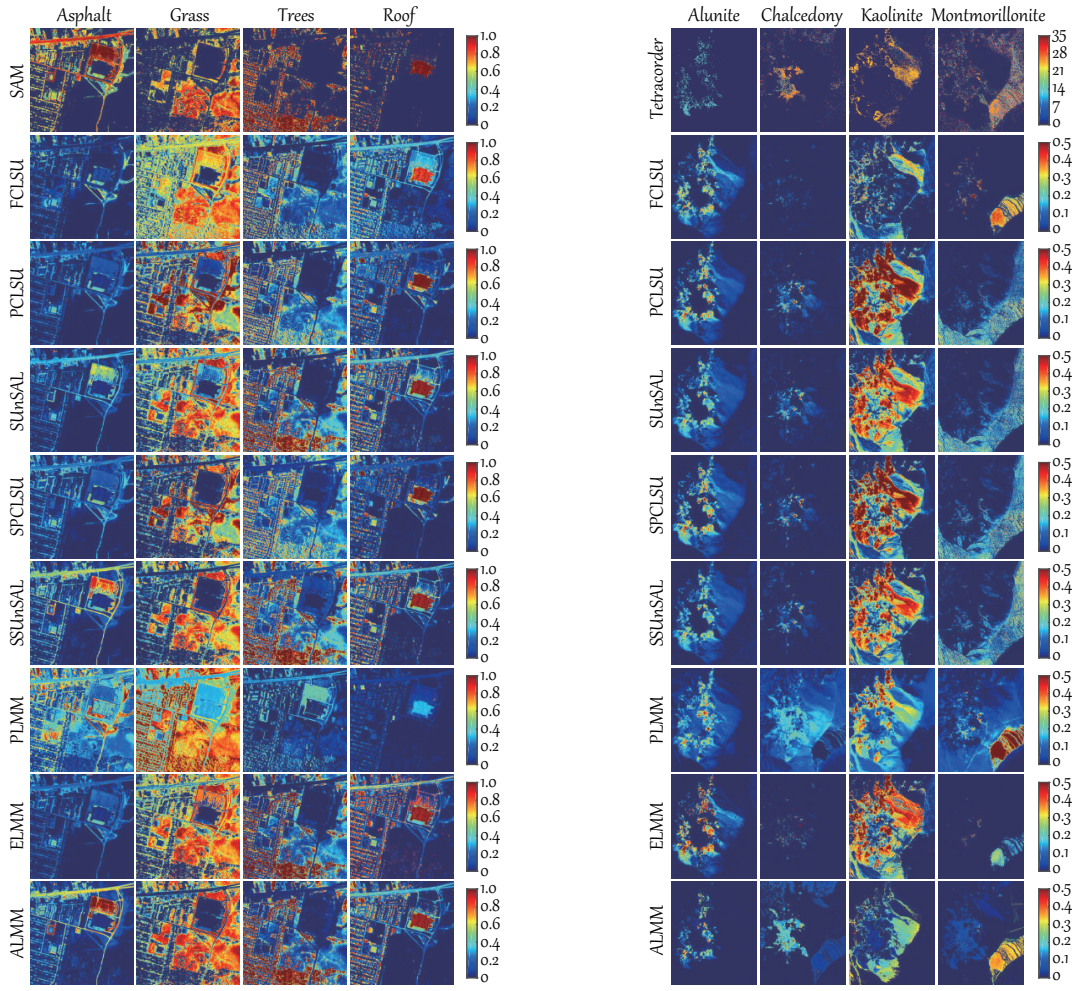
Fig. 4.14. Visualizing the unmixing results in the first simulated hyperspectral scene. (a) The abundances estimated by different spectral unmixing methods (each column corresponds to one endmember extracted by VCA) and the first row shows the ground truth. (b) The difference abundance maps using different spectral unmixing methods corresponding to Figure 4.14(a).

0. For the spectral unmixing results, we obtain classification maps by classifying each pixel into an endmember that has the maximum abundance value.

In this scene, there are many pure pixels, owing to high resolution; however, they are considered mixed pixels in the comparison of methods due to the existence of spectral variability. As shown in Fig. 4.15(a), the visual performance of the proposed ALMM method is superior to the other methods. More specifically, the asphalt is purely identified by ALMM, unlike the others; and a similar observation can be found in the grass as well. For the trees and the roof, the abundance maps estimated by ALMM show higher contrast than those estimated by other methods. This result implies that the proposed method successfully addresses spectral variability.

In the third Cuprite dataset, due to highly mixed effects, the data-driven endmember extraction is very challenging, hence the USGS spectral library is used to auxiliarily construct the endmember dictionary, where we only considered four principal minerals, i.e. alunite, chalcedony, kaolinite, and montmorillonite.

The estimated abundance maps of the four minerals are shown in Figure 4.15(b). The first row represents the reference classification maps generated by Tetracorder software [Clark et al., 2003]. The proposed method shows the best visual resemblance, compared with the results from the Tetracorder. The abundance maps generated by the proposed ALMM are more distinct and show greater contrast, and the distribution of each material is regional as



(a) real dataset (Urban) (b) real dataset (Cuprite)

Fig. 4.15. Visualization of the abundance maps of the proposed method and the state-of-art methods on two real hyperspectral datasets. (a) Urban scene: the groudtruth is given by the SAM-based measurement. (b) Cuprite scene: the first row shows the so-called ground truth generated by Tetracorder.

well, which implies that various spectral variabilities could be learned effectively.

4.4 Low-Rank Subspace Unmixing: A Novel Strategy

Unlike the previous approaches that unmix the spectral signatures directly in original space, in **Appendix D** a novel subspace-based unmixing strategy is introduced to robustly estimate the abundance maps in a low-dimensional latent subspace. With the low-rank attribute embedding, the original data is projected into a low-rank subspace where various spectral variabilities can be effectively addressed. This leads to a general subspace unmixing framework that jointly estimates subspace projections and abundance maps.

There is a trade-off between spectral information gain and the spectral variability in addressing the issue of spectral unmixing.

- ◇ On one hand, the spectrum are expected to be spectrally discriminative. This means, however, that more complex spectral variabilities might get involved in hyperspectral data.
- ◇ On the other hand, we also expect to robustify the unmixing process, that is, the spectral variability should be attained as little as possible.

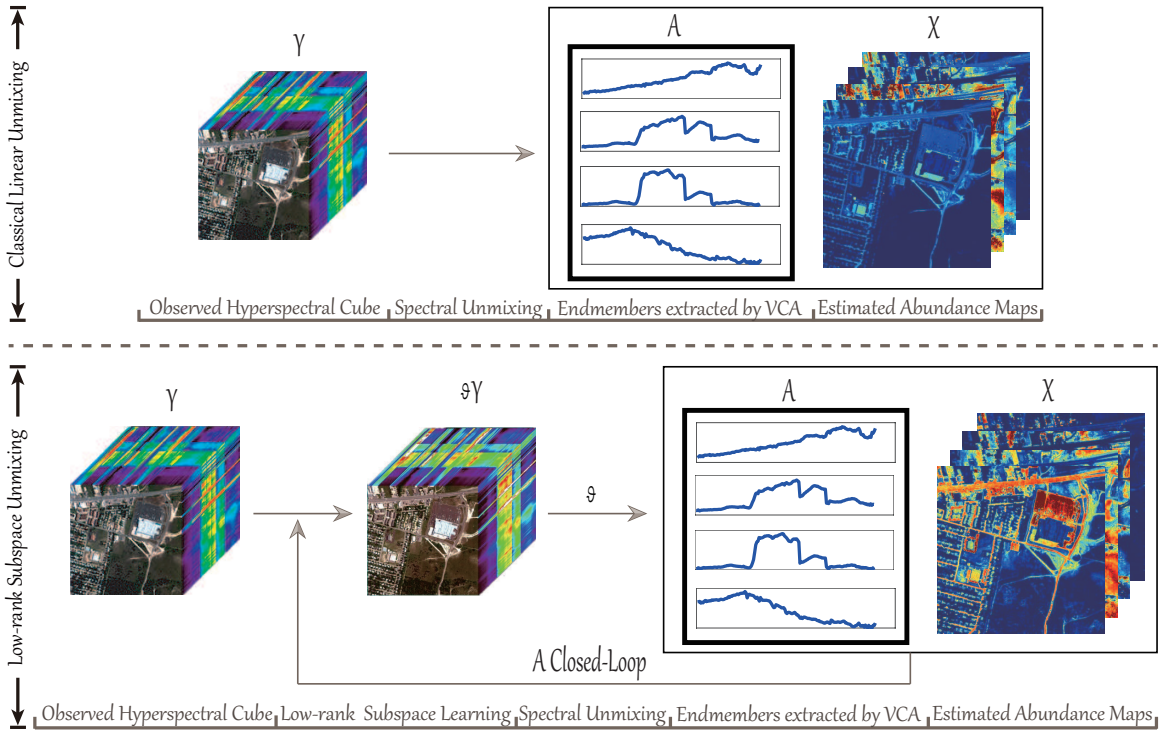


Fig. 4.16. An illustration to clarify the differences of the holistic workflow between the original-space-based spectral unmixing and subspace-based unmixing strategy.

4.4.1 General Remark in Subspace Unmixing

A feasible solution to meet the above two points is spectral unmixing in a 'raw' subspace rather than in the original space. In the learned subspace, the pixels belonging to the same class may be strongly correlated by using a low-rank attribute embedding. This naturally results in a general subspace unmixing, which can be mathematically modeled as

$$\mathbf{Y} = \mathbf{Y}^{'} + \mathbf{R}^{'} \ \text{ s.t. } \ \mathbf{Y}^{'} = \mathbf{\Theta Y}, \ \ \mathbf{Y}^{'} = \mathbf{\Theta AX} + \mathbf{R}^{''}, \tag{4.18}$$

where the variable $\mathbf{\Theta}$ is defined as the low-rank subspace projections, and $\mathbf{Y}^{'}$ is the subspace representation in the spectral domain after embedding the low-rank attribute.

Figure 4.16 clarifies the differences between the traditional spectral unmixing in the original space and the subspace-based strategy by the form of graphic analysis. More specifically, the proposed subspace-based unmixing method jointly performs subspace learning and unmixing in a closed-loop. With low-rank attribute embedding, the spectral variability can be effectively removed in the learned low-rank subspace, achieving a robust spectral unmixing. The main contributions can be unfolded as follows:

1) We propose a general subspace-based unmixing framework by jointly low-rank subspace learning and unmixing, called **s**ubspace **u**nmixing with **lo**w-**r**ank **a**ttribute embedding (SULoRA), to achieve a robust unmixing in a proper subspace rather than in the original space. Moreover, mostly linear unmixing models can be considered as special cases in this general framework.
2) With the low-rank attribute embedding, the proposed SULoRA can broadly mitigate the effects of various spectral variabilities by projecting the original data into a more representative low-rank subspace.
3) An ADMM-based optimization framework is designed to solve the resulting subspace unmixing model.

4.4.2 Low-rank Attribute Embedding

It is well-known that hyperspectral imagery inevitably suffers from various spectral variabilities in the process of imaging. These spectral variabilities, which are generated due to illumination conditions, topography change, atmospheric effects, and material nonlinear mixing, are complex and even hardly represented using a common model. Instead of directly modeling such changeable property, we hypothetically treat the spectral variability as *an unknown complex noise*. Therefore, modeling the complex spectral variability could be converted to a special denoising problem. Noises in the data can be generally removed through a projection transformation. During this process, one is expected to be the projected or denoised data as close as possible with the original data, resulting in a mathematical expression ($\mathbf{Y} \doteq \mathbf{\Theta Y}$). Besides, we also expect to structurally maintain consistency between noisy data ($\mathbf{Y}$) and processed data ($\mathbf{\Theta Y}$), which might be achieved by correlative or collaborative filtering to emphasize the correlation and structural property between the samples.

Not surprisingly, the low-rank assumption holds these characteristics well. As a result, the SULoRA model with low-rank attribute embedding can be formulated on the basis of Eq. (4.18) as

$$\min_{\mathbf{X},\mathbf{\Theta}} \frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{AX})\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{Y}-\mathbf{\Theta Y}\|_{\mathrm{F}}^2 + \beta\|\mathbf{\Theta}\|_* + \gamma\|\mathbf{X}\|_{1,1} \quad \text{s.t. } \mathbf{X} \geq \mathbf{0}. \tag{4.19}$$

The *subspace regularization* consists of second and third terms of Eq. (4.19) parameterized by α and β, respectively, which aims to find or learn a low-rank subspace projection so that the learned projection can play a correlative filtering-like role robustly against various spectral variabilities. The final term with the penlty parameter γ in Eq. (4.19) is nothing special but a commonly-used sparsity-promoting *abundance regularization*.

4.4.3 Visual Assessment of Abundance Maps

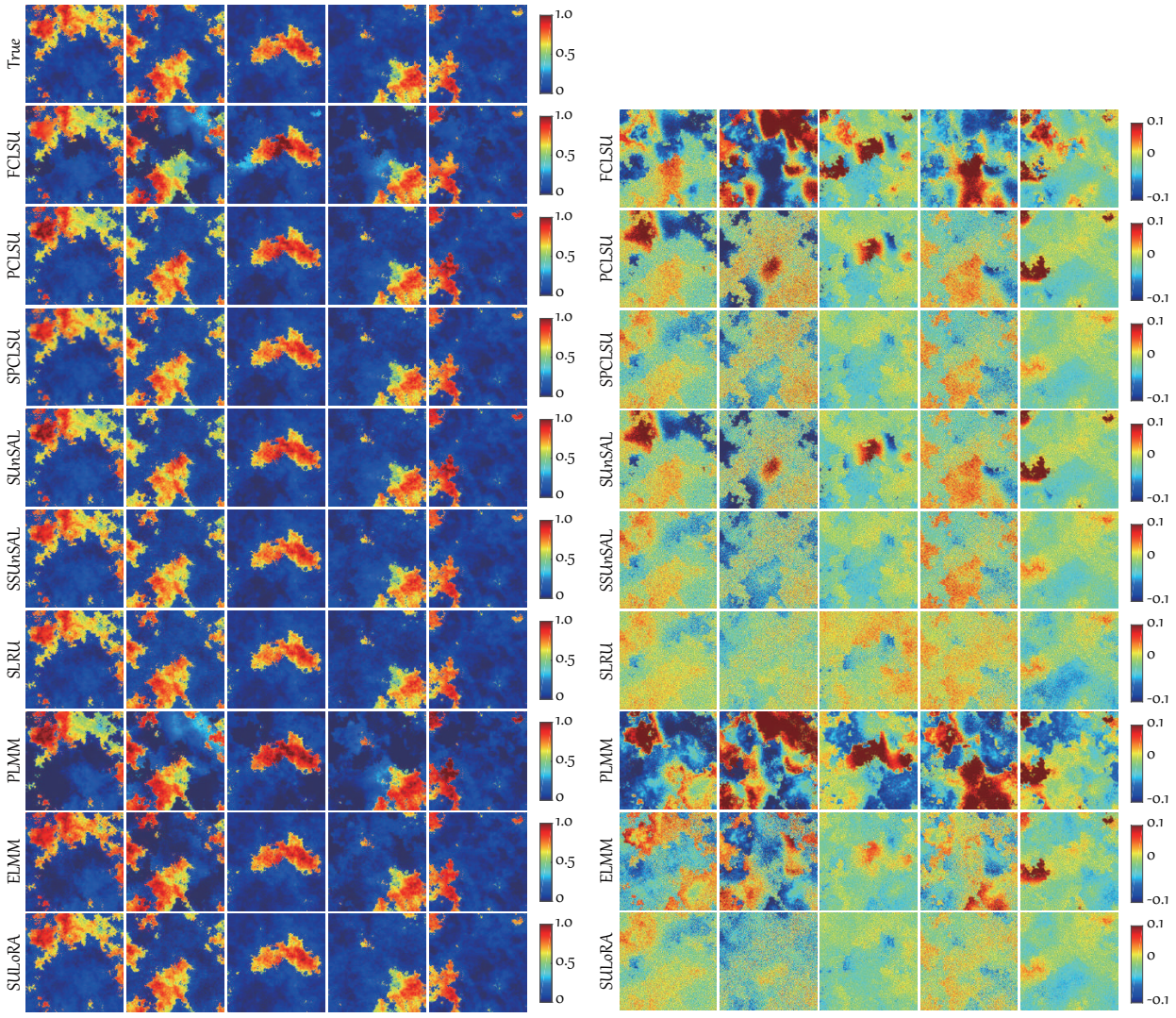
Similarly with **Appendix C**, we visually evaluate the unmixing performance of the SULoRA on a synthetic dataset and two real hyperspectral images over the areas of Urban and MUFFLE Gulfport Campus, in comparison with eight classical and state-of-the-art methods, including FCLSU, PCLSU, SPCLSU, SUnSAL, SSUnSAL, SLRU (sparse and low-rank unmixing) [Giampouras et al., 2016], PLMM and ELMM.

Simulated Scene

Fig. 4.17(a) shows the estimated abundance maps of the different algorithms. To highlight the visual differences, the abundance difference maps are displayed in Fig. 4.17(b). As expected, the performance of the subspace-based spectral unmixing (the proposed SULoRA) is superior to that of other algorithms unmixing in the original hyperspectral space, indicating its superiority and effectiveness in dealing with the spectral variability. Fig. 4.17(b) highlights a more significant comparison using abundance difference maps between the ground truth and the estimated abundance maps, where there are lower difference values in SULoRA than in others.

Parameter Sensitivity and Robustness Analysis

The performance of the proposed SULoRA algorithm in Eq. (4.19) is, to some extent, sensitive to the setting of three regularization parameters (α, β, and γ), it is, as a result, indispensable to search a set of optimal parameter combination. For this reason, the corresponding experiments are conducted to investigate the effects of the parameters on the performance of estimating abundance maps (measured by aRMSE), as specifically shown in Figure 4.18(a) where the optimal parameter combination in SULoRA is $\alpha = 0.1$, $\beta = 0.01$, and $\gamma = 8e-3$, respectively.



(a) simulated dataset: abundance maps (b) simulated dataset: difference abundance maps

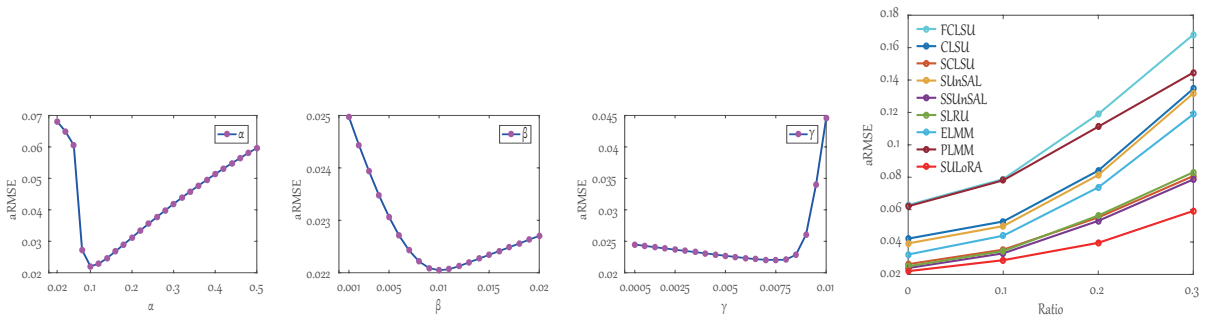
Fig. 4.17. Visual comparison of different spectral unmixing methods in the simulated hyperspectral scene. (a) The abundance maps with different spectral unmixing methods (each column corresponds to one endmember extracted by VCA) and the first row shows the groundtruth. (b) The difference abundance maps are given corresponding to Figure 4.17(a).

The robustness of the SULoRA against *sparse noise* is further investigated. For this purpose, the simulated data is corrupted by sparse noise with different corrupted levels, namely $ratio = \{0, 0.1, 0.2, 0.3\}$, where $ratio = 0$ denotes no additional sparse noise is added to the simulated data while $ratio = 0.1$, for instance, means that the 10% of total pixels are corrupted by additional sparse noise. As can be seen from Figure 4.18(b), with the increase of *sparse noise ratio*, the performance of most compared approaches dramatically degrades, yet SULoRA still holds a stable and robust performance.

Real Urban Scenes

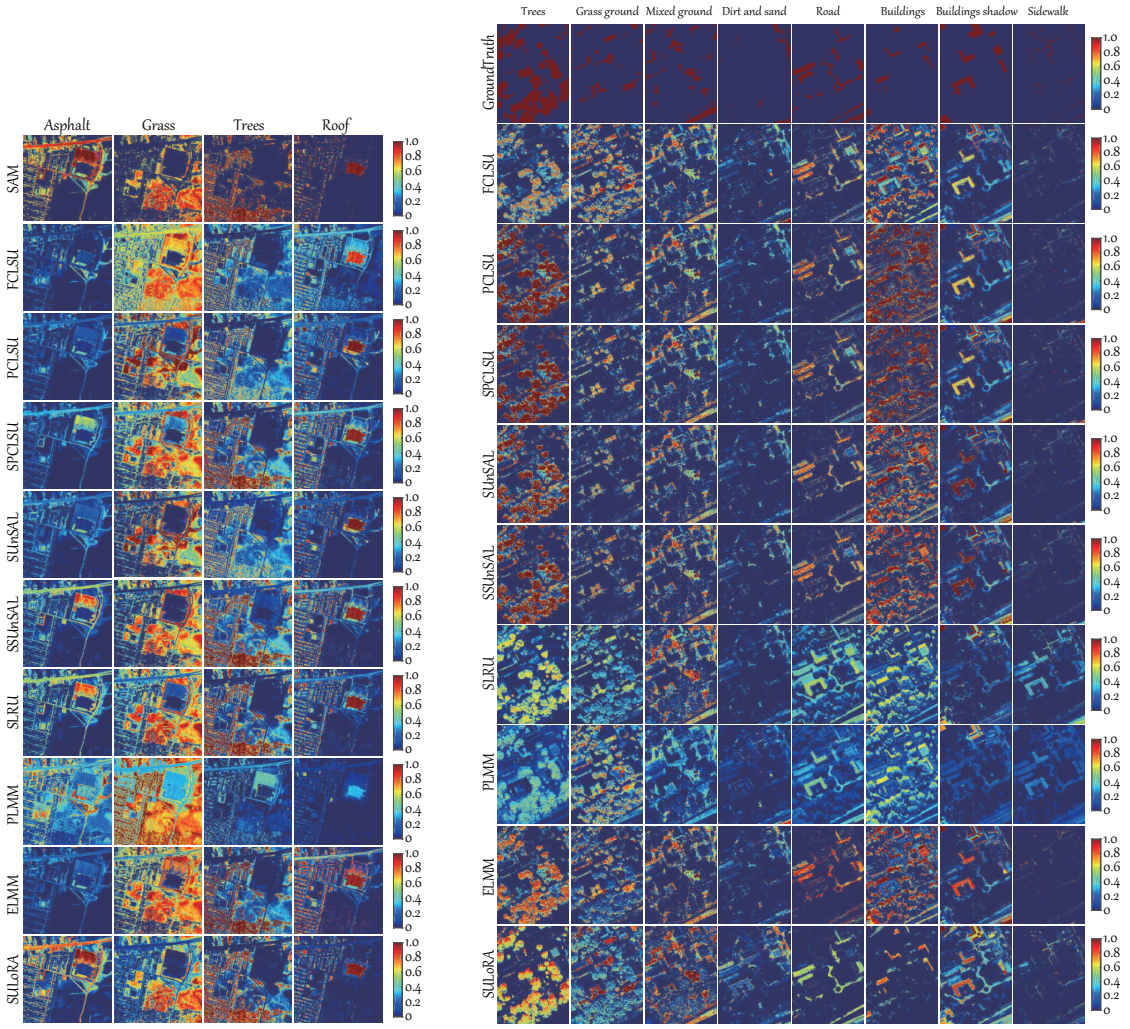
Figure 4.19 visualizes the abundance maps of several compared methods in the remaining two urban scenes.

Thanks to the high-resolution of the urban HSI, we can find many pure pixels, but they are mistaken as mixed pixels with the existence of spectral variability. This easily makes many pixels misclassified using those compared methods. Different from them, SULoRA can estimate the abundance maps in a robust subspace, so that its visual effect is superior to others', as shown in Figure 4.19(a). For instance, the asphalt and grass can be purely identified by SULoRA, unlike the others. The abundance maps of the tree and roof estimated by SULoRA show higher contrast as well.



(a) parameter sensitivity analysis (b) robustness analysis

Fig. 4.18. Parameter sensitivity and robustness analysis. (a) Sensitivity analysis of three regularization parameters (e.g., α, β, and γ) in SULoRA of Eq. (4.19) (b) Robustness evaluation of these compared algorithms using aRMSE at the different sparse noise ratio, where aRMSE is the acronym of abundance overall root mean square error.



(a) real dataset (Urban) (b) real dataset (MUFFLE Gulfport Campus)

Fig. 4.19. Visualization of the abundance maps for different methods on two real urban scenes. (a) Urban scene: the groudtruth is given by the SAM-based measurement. (b) MUFFLE Gulfport Campus scene: the first row shows the classification-based groundtruth.

By and large, these previously proposed methods basically pay more attentions on somewhat special spectral variability, lacking of generalization ability. Considering the complexity of the spectral variability in the real world, the proposed SULoRA accounts for spectral variability in a generalized fashion by embedding the low-rank attribute, resulting in more robust and effective unmixing results visually and quantitatively (see Figure 4.19(b)).

4.5 Learning Common Subspace across Multi- and Hyperspectral Modalities

With a large amount of open satellite multispectral (MS) imagery (e.g., Sentinel-2 and Landsat-8), considerable attention has been paid to global multispectral land-cover or land-use classification. However, its limited spectral information hinders further improving the classification performance. Hyperspectral (HS) imaging enables discrimination between spectrally similar classes but its swath width from space is narrow compared to multispectral ones. This challenge can be effectively transferred to model a joint learning framework to learn a shared subspace where multispectral features can be better represented by the guidance of the hyperspectral feature. To achieve accurate MS image classification over a larger coverage, we propose a novel cross-modality feature learning framework, called common subspace learning (CoSpace), detailed in **Appendix E**. CoSpace demonstrates its superiority mainly in the following aspects:

1) Subspace learning and classification are jointly considered in a unified framework by effectively bridging the learned features and label information, aiming at addressing the MS-HS cross-modal feature learning issue;
2) By locally aligning MS-HS data on the low-dimensional manifolds where the features of HS and MS images share the same dimension, CoSpace linearly learns a latent shared subspace from HS-MS correspondences, where samples are expected to be better classified. Owing to the subspace learned in a linear way, the out-of-samples data can be simply and smoothly embedded;
3) An optimization algorithm based on ADMM is properly designed to solve the proposed CoSpace model.

Figure 4.20 illustrates the holistic workflow of the CoSpace.

4.5.1 Cross-Modality Learning in Remote Sensing

Take the bi-modality as an example, the cross-modality learning refers to training a model on single modality and testing the model on bi-modality, or *vice versa* (training a model on bi-modality and testing the model on single modality). In remote sensing, particularly in MS-HS case, the cross-modality learning can be specified as a problem that given a large-scale MS image and a limited HS area partially overlapping with the MS data (see Figure 4.20 for example), we learn the low-dimensional embedding representation from the limited amount of MS-HS correspondences and transfer the learned features to the rest of MS data for improving the performance of larger-scale MS image classification and mapping. During the process, we expect to transfer the discrimination capability learned from the rich spectral information into MS data through the learned common subspace in order to more effectively identify some challenging classes that are hardly recognized by MS data due to its poor spectral information. Please note that we just start a preliminary investigation of cross-modality learning (MS-HS) in this section, that is, the MS and HS images share the same categories. An illustrative explanation is given in Figure 4.21 to distinguish the differences between multi-modality learning and cross-modality learning.

4.5.2 Learning to Align in the Latent Subspace

To fully take the benefits of HSI covering only a limited area of the MS image, and subsequently improve the classification results of the entire area covered by the MS image, our idea is to learn a HS-MS common subspace, in which the data from one domain can be adaptively transferred to another domain.

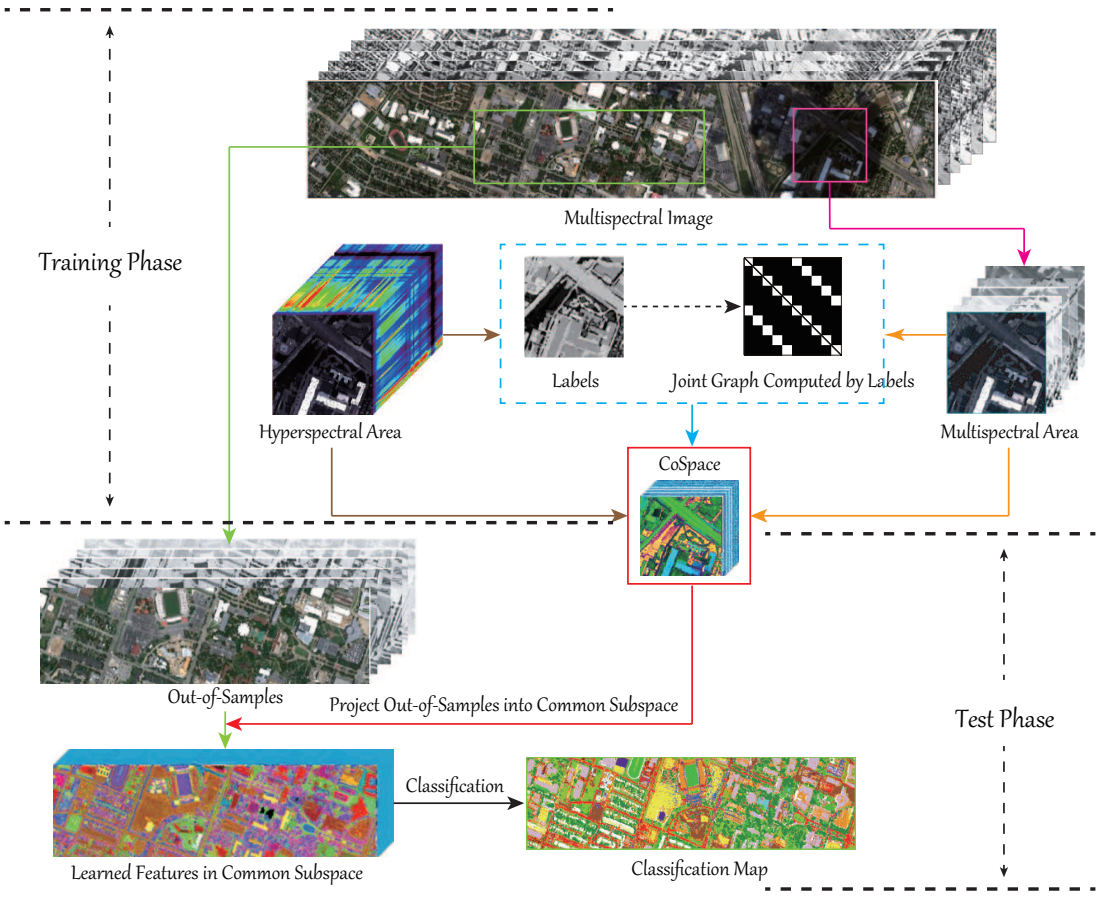


Fig. 4.20. The holistic workflow of the proposed CoSpace, including the training and test phases.

Let $\mathbf{X}_M \in \mathbb{R}^{d_M \times N}$ and $\mathbf{X}_H \in \mathbb{R}^{d_H \times N}$ be the observed MS image with d_M bands by N pixels and the HS image with d_H bands by N pixels, respectively. $\mathbf{Y} \in \mathbb{R}^{L \times N}$ is the label matrix represented by one-hot encoding. $\mathbf{\Theta}_M \in \mathbb{R}^{d \times d_M}$ ($\mathbf{\Theta}_H \in \mathbb{R}^{d \times d_H}$) is denoted as the projection matrix for connecting the MS (HS) data and the latent subspace. The variable $\mathbf{P} \in \mathbb{R}^{L \times d}$ is the weighted matrix specified by bridging the latent subspace and label information. Accordingly, the resulting model can be written as

$$\widetilde{\mathbf{Y}} = \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}} + \mathbf{E}, \tag{4.20}$$

where $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_H \end{bmatrix} \in \mathbb{R}^{(d_M+d_H)\times 2N}$ and $\widetilde{\mathbf{Y}} = [\mathbf{Y}, \mathbf{Y}] \in \mathbb{R}^{L \times 2N}$; $\mathbf{\Theta} = [\mathbf{\Theta}_M, \mathbf{\Theta}_H] \in \mathbb{R}^{d \times (d_M+d_H)}$. $\mathbf{E} \in \mathbb{R}^{L \times 2N}$ is the tolerated errors in the form of matrix.

Owing to more degrees of flexibility involved (e.g., latent subspace estimation), several assumptions (or prior knowledge) should be introduced into Eq. (4.20) using regularization technique, we then have the following constrained optimization problem:

$$\min_{\mathbf{P},\mathbf{\Theta}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\mathbf{\Theta}\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}) \ \text{ s.t. } \ \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}, \tag{4.21}$$

where the third term of Eq. (4.21) acts on a multi-modal manifold alignment that happens in the latent space. The joint Laplacian matrix $\mathbf{L}$ can be indirectly inferred by a LDA-like graph. Once the $\mathbf{\Theta}$ is obtained, the common subspace features are then represented as $\mathbf{\Theta}\widetilde{\mathbf{X}}$. In addition, please refer to the **Appendix E** for this model's solution with an ADMM solver.

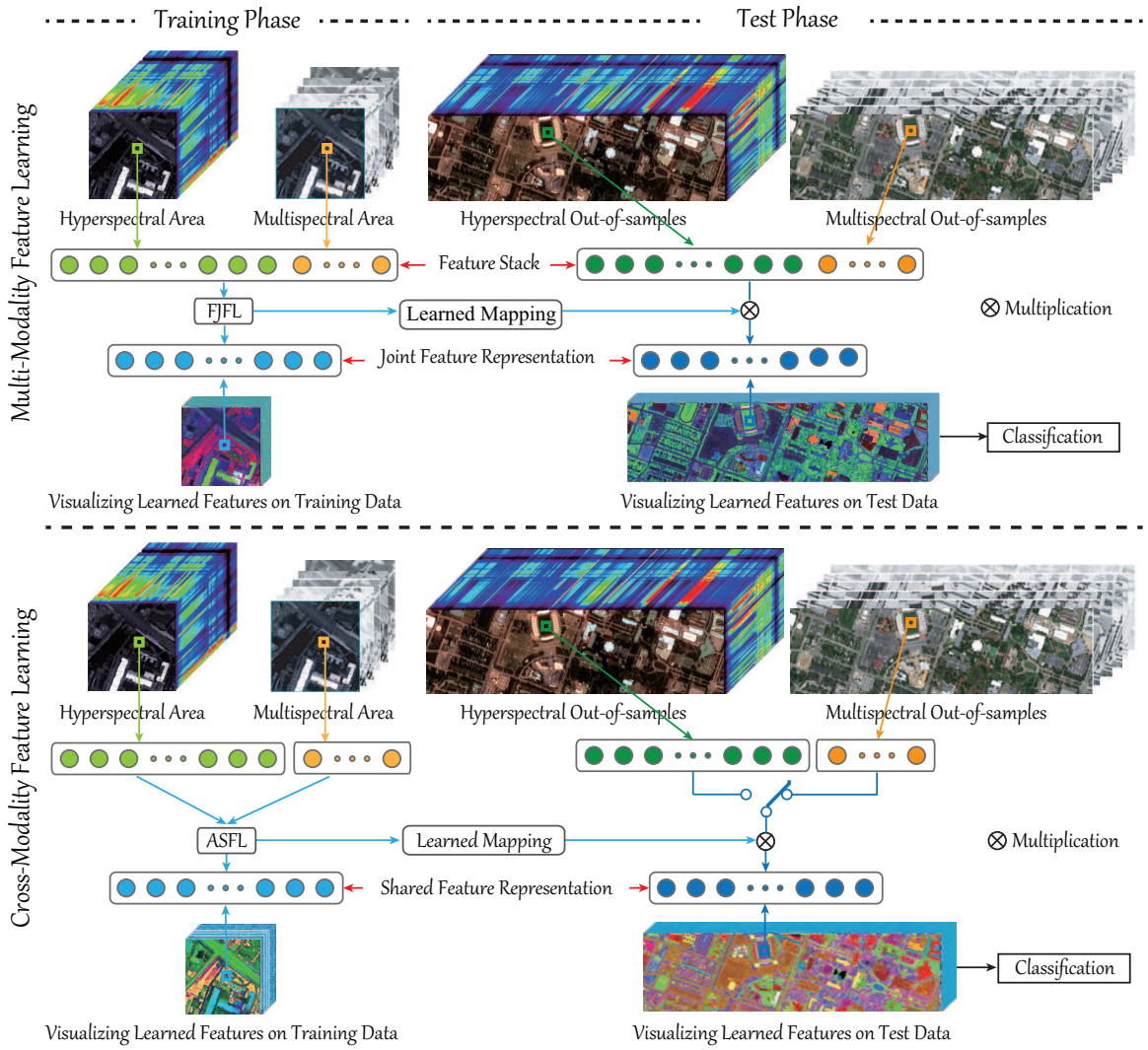
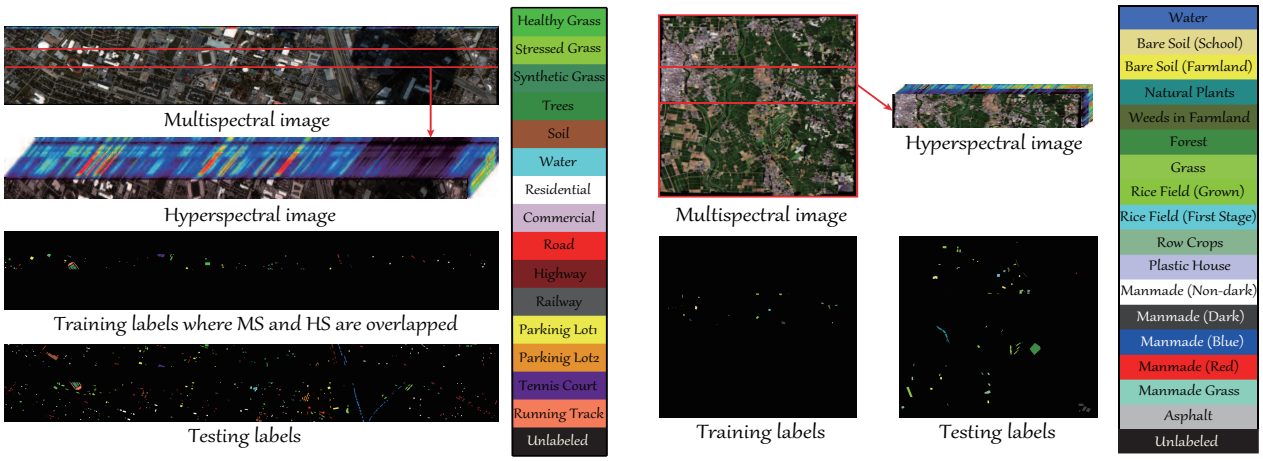


Fig. 4.21. An illustration to clarify the differences in training and test phases between the traditional multi-modality learning and cross-modality learning, where the switch (On-off) means that only one modality is involved as the test samples to meet the hypothesis of the cross-modality learning.



(a) University of Houston HS-MS Datasets (b) Chikusei HS-MS Datasets

Fig. 4.22. The multispectral image and its corresponding hyperspectral image that partially covers the same area, as well as the distributions and categories of training and test samples, for Houston2013 dataset (a) and Chikusei dataset (b), respectively.

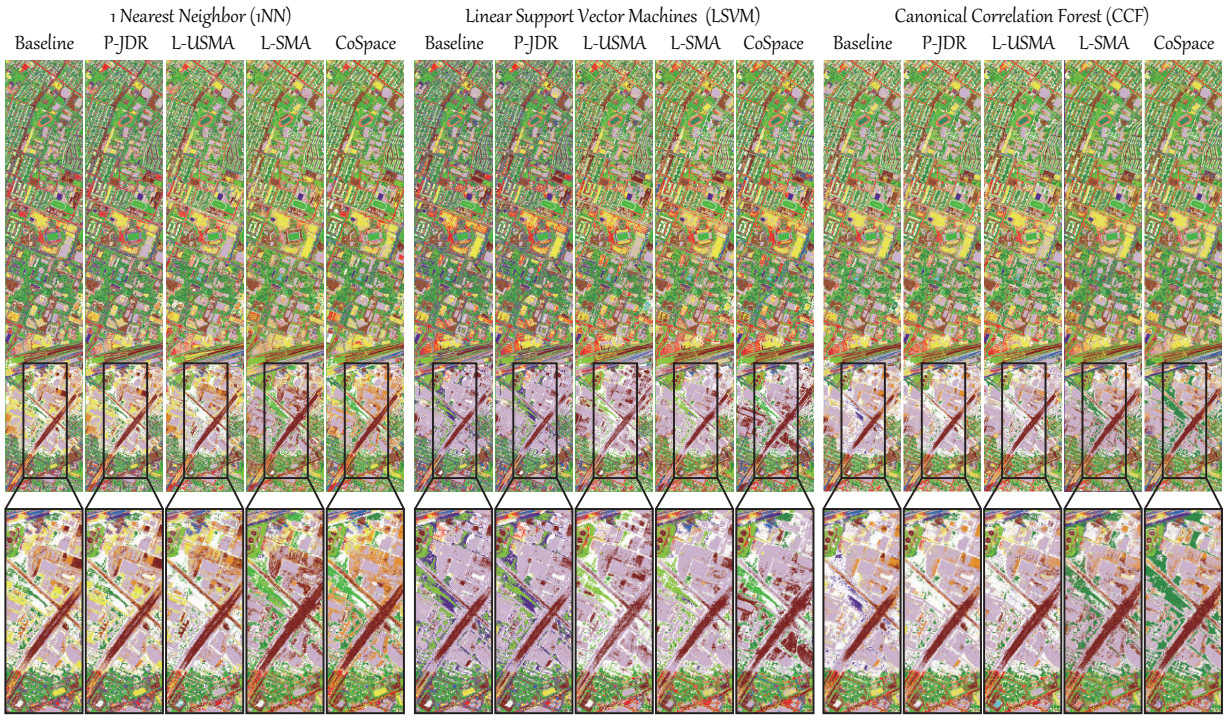


Fig. 4.23. Classification maps and a highlighted sub-area of different algorithms obtained using three classifiers on the Houston2013 dataset.

4.5.3 Larger Area Multispectral Classification with the Aids of HSI

To satisfy the problem setting of the cross-modality learning in remote sensing, two well-designed MS-HS datasets taken over the University of Houston and Chikusei are applied to quantitatively and qualitatively compare the classification performance of the proposed CoSpace and several compared methods, such as PCA-based on joint dimensionality reduction (P-JDR), LPP-based unsupervised manifold alignment (L-USMA), and LPP-based supervised manifold alignment (L-SMA) as well as the original MS features (Baseline). A showcase of cross-modality learning in remote sensing is figuratively shown in Figure 4.22.

A. Land Cover and Land Use Classification in a Large Area Multispectral Scene

Figures 4.23 and 4.24 show the classification maps of compared algorithms using three different classifiers: NN, linear SVM, and CCF.

As expected, CoSpace dramatically outperforms the others on the Houston2013 dataset, particularly for many small-scale classes, e.g., *Residential* and *Railway*. There is no denying, however, that CoSpace is superior to other algorithms to a larger extent, although it fails to effectively identify *Parking Lot2* as same with others.

To visually highlight the classification differences for the different methods, we enlarge the classification maps of a sub-area overshadowed by the cloud, as illustrated in Figure 4.23 in which it is clear to see that the methods with considering the HS information are able to generate the more discriminative features than the baseline, while the proposed CoSpace yields a better performance in identifying the materials in the shadow area, particularly for vegetation (e.g., *Grass*), *Residential* and *Commercial* that are easily misclassified by the traditional methods.

Similarly for the Chikusei hyperspectral dataset, a visual comparison of those compared algorithms is also made, as shown in Figure 4.24, where the CoSpace's superiority in classifying complex and similar land-cover classes is further shown as detailed in a salient region. Compared to other alignment-based methods, CoSpace is capable of better transferring HS information into MS data by means of joint subspace learning and classification, yielding a

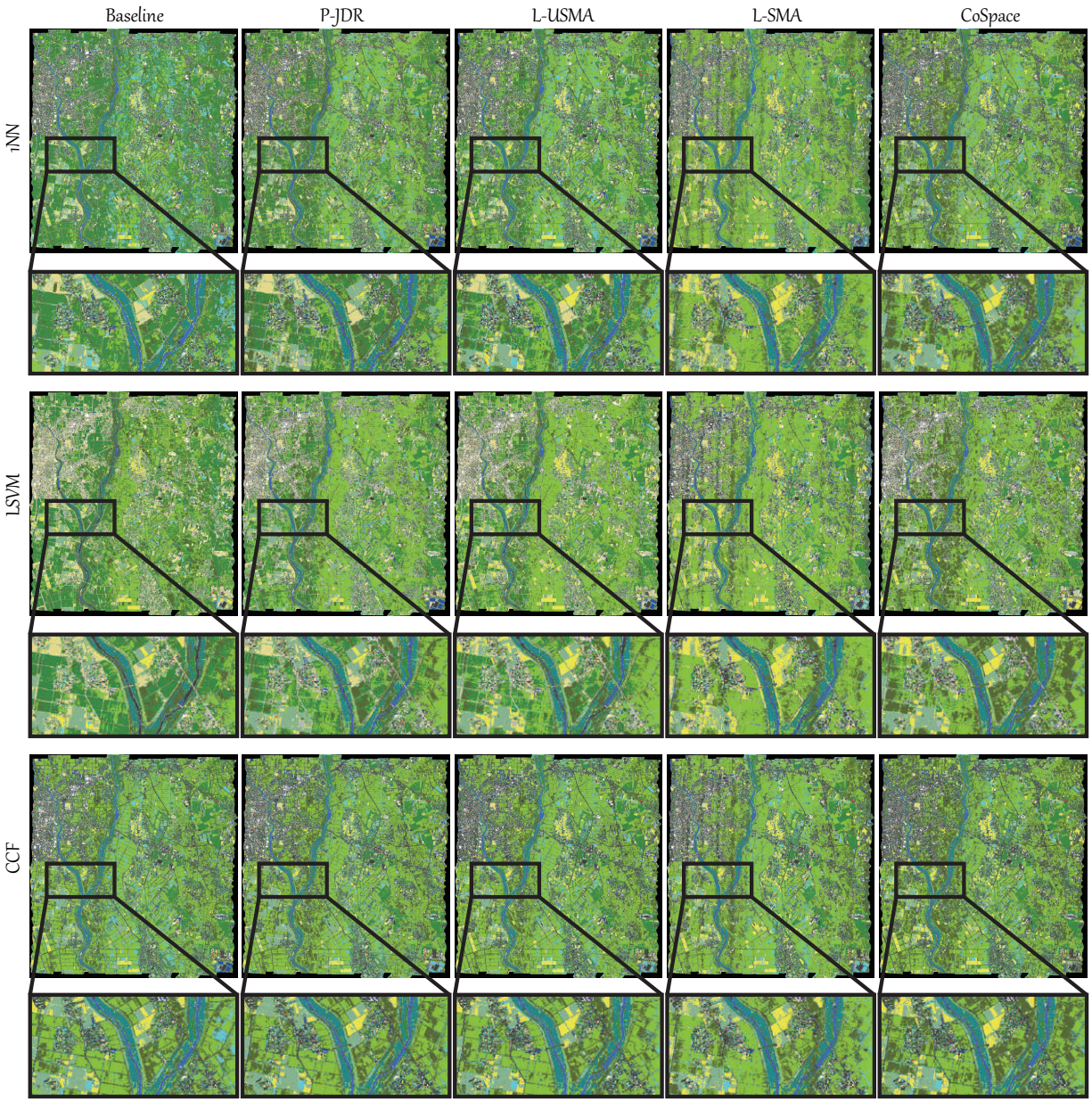


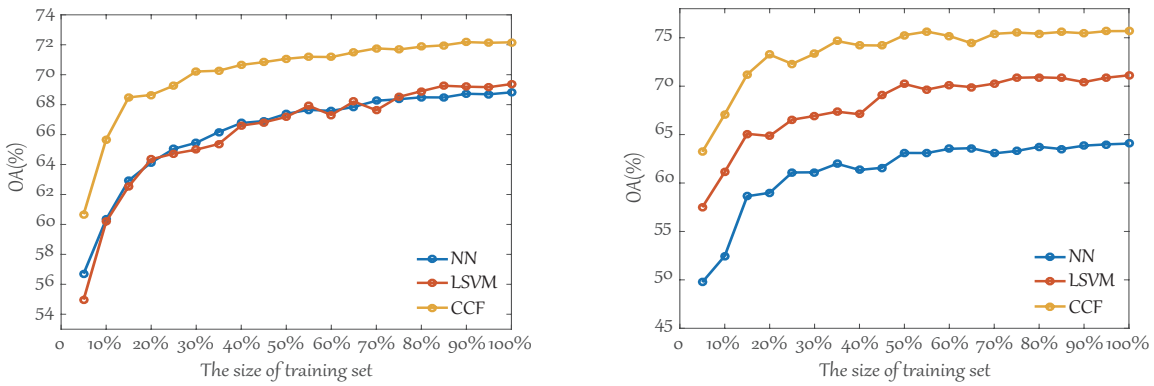
Fig. 4.24. Classification maps and a salient sub-area of different compared algorithms with three classifiers on the Chikusei hyperspectral scene.

more discriminative low-dimensional embedding. The learned features can recognize those classes of holding very similar features in MS data, such as *Bare Soil (Farmland)* and *Row Crops*, *Weeds in Farmland* and *Rice Field (Grown)*, more effectively. As shown in Figure 4.24, CoSpace performs more reasonable and competitive classification results, that is, on one hand the *Weeds in Farmland* and *Rice Field (Grown)* are most likely to be coexisted in a scene; on the other hand, the *Bare Soil (Farmland)* and *Row Crops* are separated more correctly. This can be explained by a powerful transferability of HS information in the proposed CoSpace.

B. Sensitivity Analysis to the Training Set Size

As the performance of the CoSpace largely depends on the number of training samples, it is, therefore, indispensable to investigate the sensitivity of the training set size.

In the Houston2013 datasets, the classification is conducted using the CoSpace by fixing the test set and setting a series of new training sets randomly selected from the original training set with the different percentages ranging from 5% to 100% at a 5% interval. As can be seen in Figure 4.25(a), there is a similar trend in OAs using different classifiers, that



(a) Houston2013 MS-HS datasets (b) Chikusei HS-MS datasets

Fig. 4.25. Sensitivity analysis to the sizes of training set using three different classifiers on the two MS-HS datasets.

is, the classification accuracy improves with the training set size, faster in the early, and later basically stabilized.

Similar to the former MS-HS datasets, we apply the same investigating strategy and observe the trend of classification performance using CoSpace with different sizes of training sets on the MS-HS Chikusei datasets in Figure 4.25(b). There is a very substantial change in classification accuracy with the increase of the training set size ranging from 5% to 40% of total training samples, while the performance tends to be stable after the training set size is over 50%.

4.6 Learnable Manifold Alignment in Cross-Modality: A Semi-Supervised Way

In the section, we start with revisiting the offered general but interesting cross-modality learning question in remote sensing, as mentioned in the section 4.5, – *can a limited amount of highly-discriminative (e.g., HS) training data improve the performance of a classification task using a large amount of poorly-discriminative (e.g., MS) data?* Beyond the supervised cross-modality learning, e.g., CoSpace, in **Appendix F** we propose a novel semi-supervised cross-modality learning framework, called learnable manifold alignment (LeMA). As the name suggests, LeMA learns a joint graph structure directly from the data instead of using a given fixed graph, i.e. defined by a Gaussian kernel function. With the learned graph, we can further capture the data distribution by graph-based label propagation (GLP) [Zhu et al., 2003], which enables finding a more accurate decision boundary. Figure 4.26 illustrates the workflow of the LeMA.

More specifically, the LeMA's contributions can be summarized as follows:

1) Unlike jointly feature learning in which the model is both trained and tested from completed HS-MS correspondences, LeMA learns an aligned feature subspace from the labeled HS-MS correspondences and partially unlabeled MS data, allowing to identify out-of-samples using either MS data or HS data;
2) Instead of directly computing graph structure with some pre-defined functions, e.g., RBF, a data-driven graph learning method is exploited behind LeMA to enhance the abilities to transfer and generalize;
3) An optimization framework based on ADMM-based operator is designed to fast and effectively solve the LeMA model.

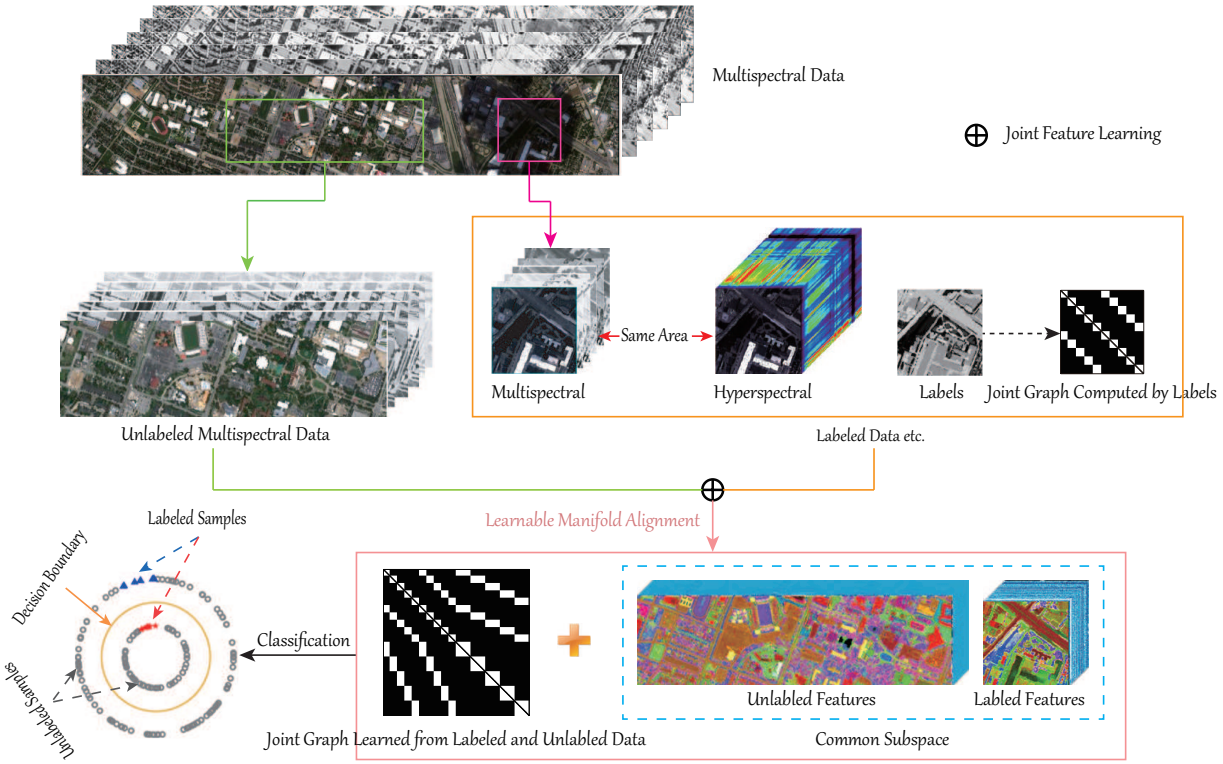


Fig. 4.26. An illustration of the proposed LeMA method.

4.6.1 Data-Driven VS Hand-Crafted Graph Construction

The adjacency matrix, also known as graph structure, is usually constructed by computing the pixel-wise similarity based on certain fixed functions (e.g., RBF) or label information (LDA-like graph) if the ground truth is partially available. The LDA-like and RBF-based graphs can be represented as

$$\mathbf{W}^{i,j}_{(\mathrm{LDA})} = \begin{cases} 1/N_k, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the } k\text{-}th \text{ class;} \\ 0, & \text{otherwise,} \end{cases} \tag{4.22}$$

and

$$\mathbf{W}^{i,j}_{(\mathrm{RBF})} = \begin{cases} \exp\dfrac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}, & \text{if } \mathbf{x}_j \text{ is one of } k \text{ neighbors of } \mathbf{x}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{4.23}$$

For example, the RBF-based and LDA-like graphs are applied in J-Play (see section 4.2) and CoSpace (see section 4.5), respectively.

Different from the two hand-crafted graphs, LeMA performs a data-driven graph learning directly from a common subspace so as to make the multimodal data comparable as well as improve the explainability of the learned common subspace, which further results in a better transferability. The graph learning problem can be generalized as a constrained optimization problem:

$$\min_{\mathbf{W}} \operatorname{tr}(\mathbf{WZ}) \;\text{ s.t. }\; 1/N_k \geq \mathbf{W}_{i,j} \geq 0, \;\; \|\mathbf{W}\|_{1,1} = s, \tag{4.24}$$

where $\mathbf{Z}$ is defined as *a pairwise Euclidean distance matrix*. Using the following equation:

$$\operatorname{tr}(\mathbf{HLH}^{\mathrm{T}}) = \frac{1}{2}\operatorname{tr}(\mathbf{WZ}) = \frac{1}{2}\|\mathbf{W} \odot \mathbf{Z}\|_{1,1}, \tag{4.25}$$

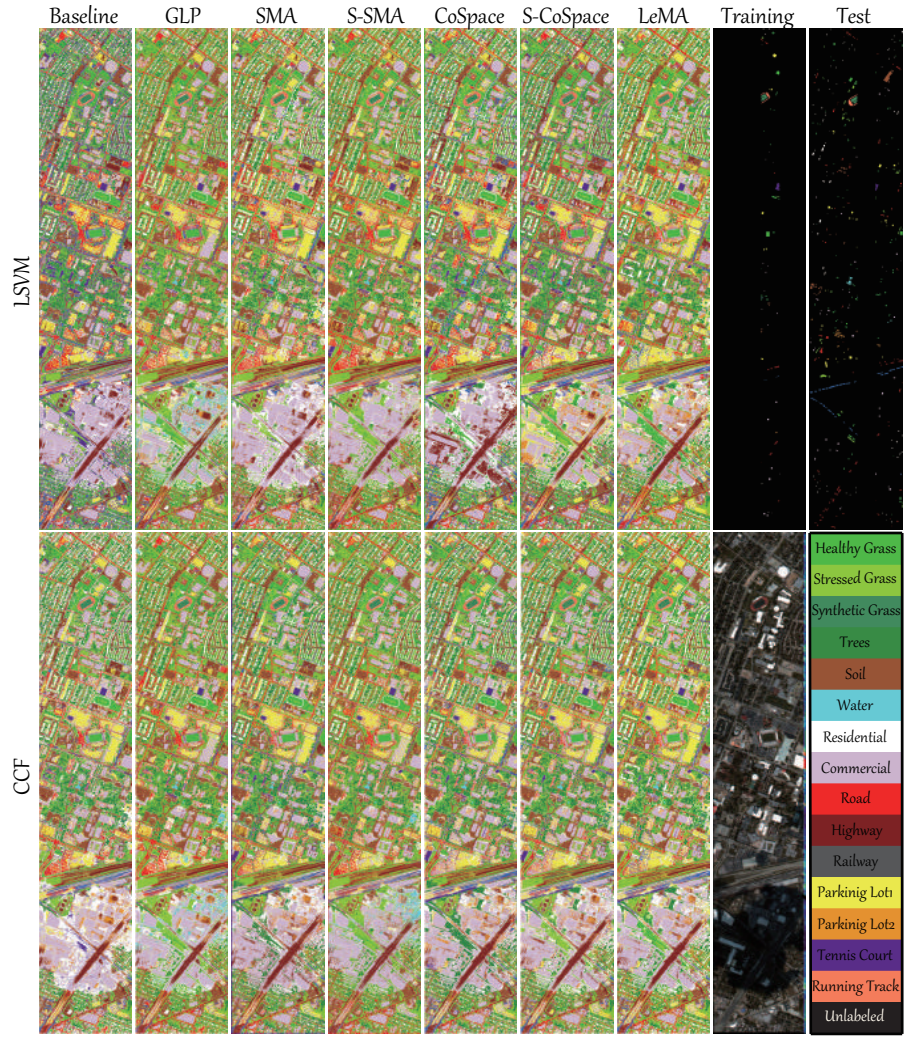


Fig. 4.27. Classification maps of the different algorithms obtained using two kinds of classifiers on the University of Houston dataset.

the Eq. (4.25) equivalently becomes

$$\min_{\mathbf{W}} \|\mathbf{W} \odot \mathbf{Z}\|_{1,1} \ \text{ s.t. } \ 1/N_k \geq \mathbf{W}_{i,j} \geq 0, \ \ \|\mathbf{W}\|_{1,1} = s. \tag{4.26}$$

Intuitively, the Eq. (4.26), which can be effectively optimized via the ADMM solver, is able to yield a data-driven graph construction.

4.6.2 Manifold Alignment Meets Graph Learning

Combined with the learnable graph (or manifold) structure, the multi-modal data are able to be adaptively aligned in a data-driven fashion. This also means such strategy not only can align the different modalities but also align the labeled and unlabeled samples, thereby yielding the proposed semi-supervised LeMA algorithm for the cross-modality learning.

Mathematically, the optimization problem of LeMA can be formulated with extra constraints related to necessary conditions with respect to the to-be-learned joint Laplacian matrix ($\mathbf{L}$) as follows:

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{\Theta},\mathbf{L}} \ \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\operatorname{tr}(\mathbf{\Theta}\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}) \\ &\text{s.t. } \mathbf{H} = \mathbf{\Theta}\widetilde{\mathbf{X}}', \ \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}, \ \mathbf{L} = \mathbf{L}^{\mathrm{T}}, \ \mathbf{L}_{i,j,i\neq j} \leq 0, \ \mathbf{L}_{i,j,i=j} \geq 0, \ \operatorname{tr}(\mathbf{L}) = s, \end{aligned} \tag{4.27}$$

Table 4. Quantitative performance comparison with the different algorithms in terms of three indices: overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) as well as the accuracy for each class on the first homogeneous MS-HS datasets (Houston2013). The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	d		(k,σ,d)		d		(k,σ,d)		(α,β,d)		(α,β,d)		(α,β,d)	
	10		(10,1,10)		30		(10,0.1,30)		(0.01,0.01,30)		(0.1,0.01,30)		(0.01,0.01,30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	62.12	68.21	64.71	70.01	68.01	69.59	69.29	70.10	69.38	72.17	70.41	73.75	73.42	**76.35**
AA	65.97	70.47	68.18	72.18	70.50	71.02	72.00	72.88	71.69	73.56	73.12	75.61	74.76	**77.18**
κ	0.5889	0.6543	0.6164	0.6728	0.6520	0.6695	0.6659	0.6754	0.6672	0.6975	0.6784	0.7146	0.7110	**0.7428**
Class1	76.39	67.95	77.83	77.97	75.25	68.53	74.25	73.53	75.54	69.96	**91.85**	87.98	89.56	85.84
Class2	80.59	78.08	93.85	**98.01**	97.57	77.9	97.57	93.67	73.74	77.99	90.12	91.59	93.67	93.85
Class3	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Class4	85.51	92.27	89.66	96.62	94.78	98.74	95.85	98.55	98.74	98.26	92.75	97.29	97.49	**99.61**
Class5	99.06	99.4	99.49	**99.66**	98.97	99.14	99.32	99.4	99.4	99.4	99.4	**99.66**	99.49	99.57
Class6	86.14	86.14	96.37	99.01	86.47	70.96	**99.67**	99.67	85.48	85.15	**99.67**	96.70	86.47	86.47
Class7	50.62	63.76	48.63	64.01	72.32	77.14	72.15	69.66	73.98	80.05	75.06	80.96	83.21	**88.03**
Class8	56.49	56.06	56.60	59.85	62.01	62.23	**64.61**	63.85	63.53	62.01	55.84	60.39	62.77	62.01
Class9	56.22	70.58	69.63	69.02	49.96	61.27	50.57	45.00	59.79	64.93	65.8	**71.54**	64.49	61.88
Class10	45.36	45.25	45.46	49.89	58.12	52.32	58.33	63.61	**64.14**	57.70	58.97	51.79	60.97	53.59
Class11	27.43	43.88	22.45	38.65	28.86	36.46	36.46	34.77	36.54	47.26	35.78	38.65	41.27	**49.96**
Class12	31.64	56.08	31.75	37.83	35.84	62.50	34.18	55.2	46.79	62.72	34.29	58.52	45.02	**76.88**
Class13	0.00	0.67	0.00	1.11	0.00	0.00	0.00	0.45	0.00	0.45	0.00	0.89	0.00	**1.78**
Class14	97.53	98.77	94.44	92.59	**100.00**	**100.00**	99.38	98.15	**100.00**	99.38	99.38	**100.00**	99.38	**100.00**
Class15	96.59	98.16	96.59	**98.43**	97.38	98.16	97.64	97.64	97.64	98.16	97.90	98.16	97.64	98.16

where $\widetilde{\mathbf{X}}' = \begin{bmatrix} \mathbf{X}_H & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_M & \mathbf{X}_U \end{bmatrix} \in \mathbb{R}^{(d_H+d_M)\times(2N+N_U)}$, $\widetilde{\mathbf{L}} \in \mathbb{R}^{(2N+N_U)\times(2N+N_U)}$, and $\mathbf{X}_U \in \mathbb{R}^{d_M\times N_U}$ represents the unlabeled MS samples and $s > 0$ controls the scale.

Using the Eq. (4.25), the optimization problem of smooth manifold in (4.27) can be equivalently converted to that of graph sparsity:

$$\min_{\mathbf{P},\mathbf{\Theta},\mathbf{W}} \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 + \frac{\beta}{4}\|\mathbf{W}\odot\mathbf{Z}\|_{1,1} \text{ s.t. } \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I},\ \mathbf{W} = \mathbf{W}^{\mathrm{T}},\ \mathbf{W}_{i,j} \geq 0,\ \|\mathbf{W}\|_{1,1} = s, \tag{4.28}$$

where $\mathbf{Z} \in \mathbb{R}^{(2N+N_U)\times(2N+N_U)}$ can be computed by $\mathbf{Z}_{i,j} = \|(\mathbf{\Theta}\widetilde{\mathbf{x}}_i) - (\mathbf{\Theta}\widetilde{\mathbf{x}}_j)\|^2$.

4.6.3 Application in Cross-Modality Data Analysis

In addition to the two homogeneous MS-HS datasets used in **Appendix E**, an additional real multispectral-lidar and hyperspectral dataset provided by 2018 IEEE GRSS data fusion contest (DFC2018) [Le Saux et al., 2018] is also considered to investigate the heterogeneous cross-modality data analysis. Moreover, we compare the performance of the proposed LeMA and several other state-of-art algorithms, i.e. GLP, SMA, Semi-supervised SMA (S-SMA), CoSpace and Semi-supervised CoSpace (S-CoSpace). The original data feature is used as a baseline. SMA constructs an LDA-like joint graph using label information. Besides label information, S-SMA method also uses unlabeled samples to generate the joint graph by computing the similarity based on Euclidean distance. The same strategy of graph construction is adopted for CoSpace and S-CoSpace.

A. The First Homogeneous MS-HS Dataset (Houston2013)

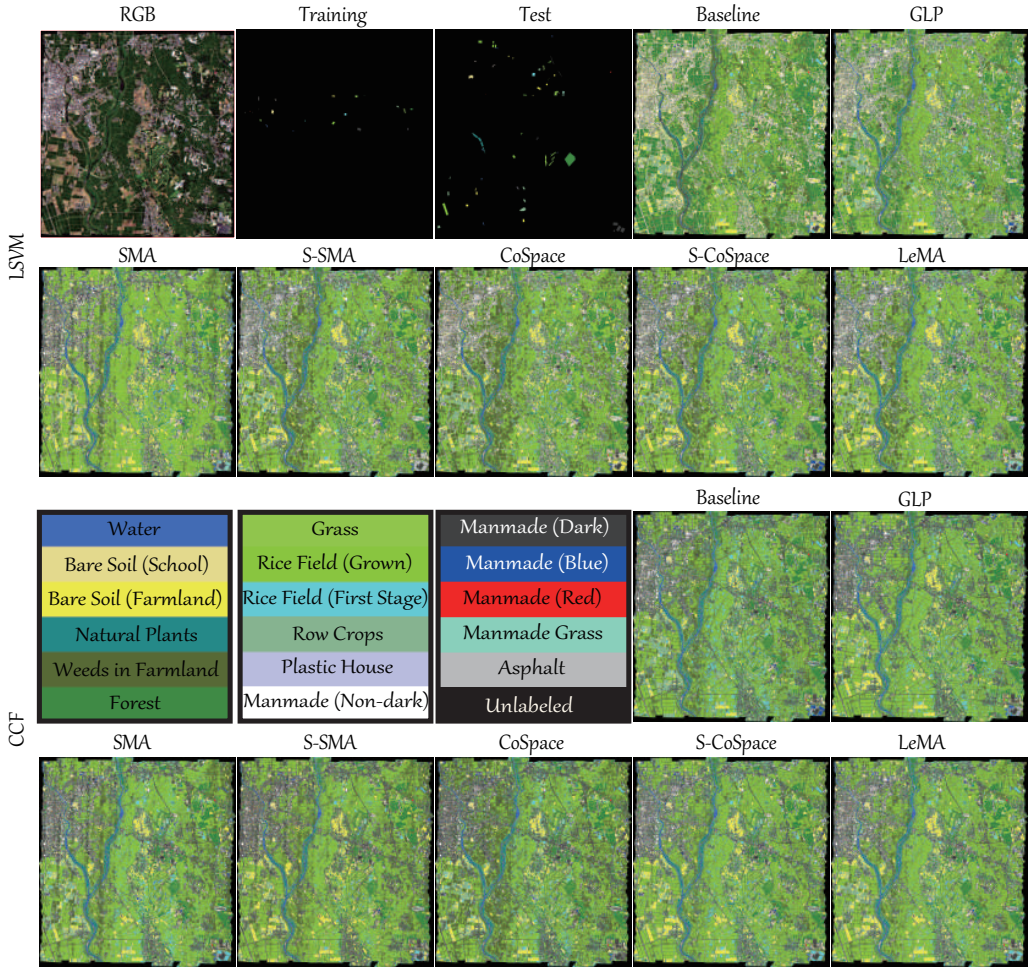


Fig. 4.28. Classification maps of the different algorithms obtained using two kinds of classifiers on the Chikusei dataset.

Figure 4.27 shows the classification maps of several compared algorithms using two classifiers of LSVM and CCF, while Table 4 lists the specific quantitative assessment results with optimal parameters obtained by 10-fold cross-validation.

By fully considering the connectivity of the common subspace, label information, and unlabeled information encoded by the learned graph structure, the performance of LeMA is much more superior to that of any other methods as can be observed in Table 4. This demonstrates that LeMA is likely to learn a more discriminative feature representation and to find a better decision boundary.

Unlike other methods, LeMA can adaptively learn a data-driven graph structure where the labels tend to spread more smoothly, which can result in a more effective material identification for those challenging classes (few training samples), such as *Trees*, *Residential*, *Railway*, *Parking Lot1*. In addition, we can also observe an easily overlooked phenomenon that the LeMA's ability in identifying certain classes still remains limited, such as *Parking Lot2*(only 1.78%) and *Railway* (49.96%). *Parking Lot2* is basically classified to *Commercial* and *Parking Lot1*, while *Railway* is largely identified as *Road* and *Commercial*. This might be explained by the limited number of training samples as well as fairly similar spectral properties between several classes.

B. The Second Homogeneous MS-HS Dataset (Chikusei)

We assess the classification performance of the different algorithms for the Chikusei MS-HS data both quantitatively and visually, as shown in Figure 4.28 and Table 5.

Similarly to the University of Houston MS-HS data, there is a basically consistent trend for

Table 5. Quantitative performance comparison with the different algorithms in terms of OA, AA, κ, and the accuracy of each class on the second homogeneous MS-HS datasets (Chikusei). The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	d		(k,σ,d)		d		(k,σ,d)		(α,β,d)		(α,β,d)		(α,β,d)	
	10		(10,1,10)		20		(10,0.1,20)		(0.1,0.01,30)		(0.1,0.01,30)		(0.1,0.01,30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	60.20	71.11	62.30	72.26	67.90	71.53	69.68	73.27	71.12	75.69	72.60	77.11	75.11	**81.71**
AA	69.42	70.40	69.80	70.71	70.79	66.47	72.27	70.01	73.96	71.46	71.64	71.33	75.29	**75.73**
κ	0.5523	0.6761	0.5784	0.6894	0.6391	0.6802	0.6602	0.6818	0.6746	0.7260	0.6911	0.7420	0.7194	**0.7933**
Class1	78.21	80.54	78.09	80.42	98.72	82.52	**99.53**	97.90	92.54	79.25	98.83	98.37	98.25	98.83
Class2	94.43	82.70	94.11	93.84	93.20	92.50	93.20	93.09	93.47	**94.91**	87.04	93.63	93.20	93.79
Class3	23.54	50.06	37.75	76.87	62.57	55.31	68.41	76.55	80.40	77.71	80.65	77.23	89.29	**89.90**
Class4	92.13	92.56	92.23	95.72	90.57	91.53	92.51	88.76	90.59	96.23	94.64	92.49	95.11	**96.96**
Class5	**97.65**	94.68	96.84	88.45	28.43	16.06	24.01	32.85	83.94	66.52	51.81	43.32	60.74	67.78
Class6	62.01	81.48	57.47	69.67	62.52	78.91	68.27	79.67	63.61	79.02	72.34	**88.48**	76.34	87.27
Class7	99.67	99.93	99.66	**100.00**	96.87	97.79	95.40	99.37	97.74	99.75	98.41	99.87	97.63	99.80
Class8	57.11	93.40	69.06	98.93	95.59	93.49	96.88	96.53	95.05	92.72	**99.48**	98.45	99.27	99.18
Class9	**100.00**	**100.00**	**100.00**	99.92	99.53	99.13	99.45	99.21	98.66	99.76	99.21	98.34	99.76	**100.00**
Class10	24.81	19.56	**26.64**	19.06	21.39	15.48	20.94	13.09	22.35	18.00	22.75	14.83	26.47	26.46
Class11	0.00	2.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	5.47	0.63	**5.68**
Class12	**90.32**	88.91	**90.32**	89.61	90.14	85.92	90.14	89.44	90.32	80.46	89.96	89.44	88.38	90.14
Class13	33.11	33.09	33.11	36.50	32.61	56.25	31.32	30.88	33.11	67.90	33.11	54.93	33.11	**68.73**
Class14	**94.20**	85.38	79.12	59.40	72.85	59.40	**94.20**	86.31	59.40	52.44	14.39	49.19	45.01	53.60
Class15	**100.00**	**100.00**	**100.00**	**100.00**	93.58	**100.00**	**100.00**	**100.00**	93.58	97.86	**100.00**	**100.00**	**100.00**	**100.00**
Class16	74.88	88.62	74.19	93.52	99.71	99.51	99.80	98.82	97.84	**100.00**	97.35	97.25	98.04	95.78
Class17	58.03	3.84	58.03	0.24	65.23	7.91	62.11	7.67	64.75	0.00	77.70	11.27	**78.66**	13.43

the different algorithms in the Chikusei MS-HS data. As expected, the performance of the LeMA is significantly superior to that of others, thanks to the great contributions of a common subspace learning from MS-HS data, a data-driven graph learning and semi-supervised learning strategy. Despite so, the LeMA still fails to recognize some challenging classes, such as *Weeds in Farmland*, *Row Crops*, *Plastic House*, and *Asphalt*. The reasons could be two-fold. On one hand, the performance of LeMA is limited, to some extent, by the unbalanced data sets. On the other hand, LeMA' transferring ability would sharply degrade when a great spectral variability between training and test samples exists.

C. The Final Heterogeneous MS-Lidar and HS Datasets (DFC2018)

Although we follow strict simulation procedures, yet the two MS-HS datasets used above (Houston2013 and Chikusei) essentially originate from a similar data source (homogeneous), which means there is a strong correlation in their spectral features. This makes the information of the different modalities transferred more effectively, but could limit the investigation in generalization ability. To this end, we make a quick shot to see a heterogeneous case by applying a real bi-modal dataset – multispectral-lidar and hyperspectral provided by the latest IEEE GRSS DFC2018.

We randomly assign 10% of total labeled samples as the training set and the rest of it as the test set in the experiment. Moreover, 16 main classes are selected out of 20 by removing several small classes with too few samples, e.g., *Artificial Turf*, *Water*, *Crosswalks*, and *Unpaved Parking Lots*.

The averaged results of the different algorithms out of 10 runs are reported for a relatively fair comparison, since the training and test sets are randomly generated from total samples in each round, as listed in Table 6. Correspondingly, Figure 4.29 visually shows the

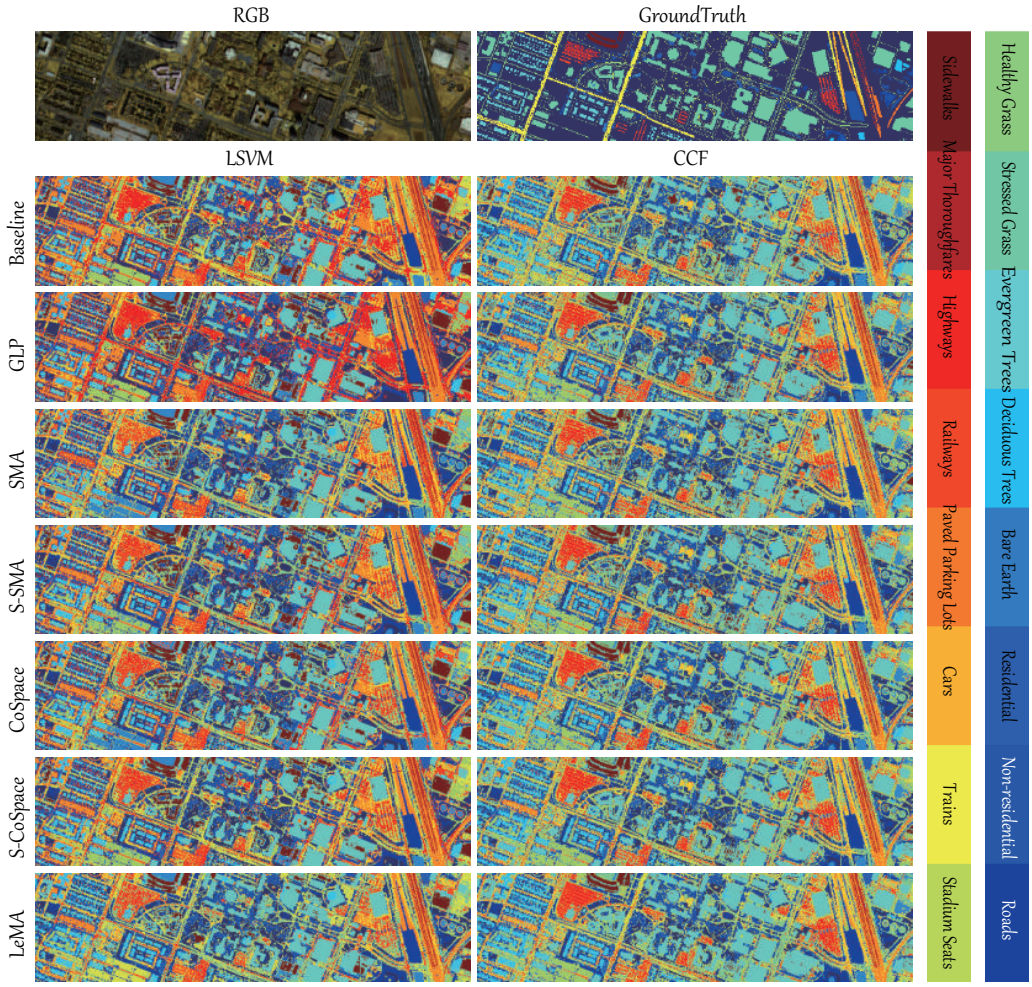


Fig. 4.29. Classification maps of the different algorithms obtained using two kinds of classifiers on the real dataset of DFC2018 (Multispectral-Lidar and Hyperspectral data).

Table 6. Comparison of classification accuracies (OA, AA, and κ) using different alignment algorithms on the DFC2018 dataset. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	d		(k,σ,d)		d		(k,σ,d)		(α,β,d)		(α,β,d)		(α,β,d)	
	7		(10,1,7)		30		(10,1,30)		(0.1,0.1,30)		(0.1,0.01,30)		(0.1,0.01,30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	51.35	72.84	52.28	73.15	52.73	70.37	54.69	72.13	55.56	74.04	58.65	76.59	61.69	**79.98**
AA	59.46	78.64	60.57	81.64	58.06	77.78	65.34	78.72	66.16	80.46	67.72	83.67	65.54	**88.82**
κ	0.4194	0.6534	0.4289	0.6587	0.4366	0.6256	0.4598	0.6441	0.4670	0.6682	0.4987	0.6990	0.5284	**0.7414**

differences of classification maps obtained using the different methods.

Generally speaking, hyperspectral information embedding can effectively improve the classification performance of the multispectral-lidar data, which implies that the models based common subspace learning (e.g., SMA, S-SMA, CoSpace, S-CoSpace, and LeMA) can transfer the knowledge from one modality to another modality to some extent, even though the input data are heterogeneous.

Not unexpectedly, the proposed LeMA integrating rich spectral information and unlabeled samples achieves a superior performance, which demonstrates that the learning-based graph structure is more applicable to capturing the data distribution and further find a potential optimal decision boundary.

One thing to be noted, however, is that compared to the performance of the different algorithms in the simulated MS-HS datasets from similar sources (homogeneous), the knowledge transferring ability of these algorithms in handling the real multispectral-lidar and hyperspectral datasets from different sources (heterogeneous) remains limited, since all listed methods including our LeMA are modeled in a linearized way. Unfortunately, a single linear transformation fails to fit the gap between heterogeneous modalities well, despite a limited performance improvement.

5 Conclusion and Outlook

5.1 Conclusion

This thesis aims at developing a novel paradigm – regression-induced representation learning – to revisit the remotely sensed hyperspectral imagery analysis, making it possible for HSI to reach a trade-off between robustness and discrimination, as well as to pinpoint its irreplaceable position in the next-generation large-scale earth observation tasks. As a result, a general objective is summarized as

"developing advanced algorithms to analyzing the hyperspectral data more robustly and efficiently with the potential contributions to improving the classification or mapping tasks in the regional and even global coverage".

In order to achieve this goal, three more specific challenges are detailed in Chapter 1, corresponding to the three sub-topics of HDR, spectral unmixing, cross-modality fusion and learning, respectively. Accordingly, the solutions are presented in the methodology of regression-induced representation learning with six main contributions, thus the conclusions can be drawn as follows:

- Redundant and noisy spectral bands inevitably degrade the performance of high-level hyperspectral data analysis. HDR should therefore give a priority to be made. However, classic graph embedding-based HDR methods are sensitive to complex noise and multicollinearity. To tackle the two problems, a robust local manifold representation (RLMR) is developed to effectively compress the spectral dimension, involving
  - A joint normalization is proposed by locally and globally mitigating the effects of various spectral variabilities;
  - After being coarsely selected, neighbors of each data point are further refined with a KLD measurement on locally constructed features;
  - Spatial contextual information is jointly embedded in the process of graph construction.

  Classification is explored as a potential application for validating the performance of HDR. Experimental results have demonstrated the effectiveness of the RLMR in HDR.
- There has been a trade-off between robustness and discrimination of dimension-reduced spectral features in the process of HDR. Interestingly, we found that those complex spectral variabilities are hardly reconstructed in a linear system. In light of the discovery, we reconsider the HDR as an issue of linear regression. To simultaneously maintain the spectral discriminant as high as possible, a multi-layered linearized regression model is proposed to learn the low-dimensional representation in a joint and progressive fashion. The obtained features using the proposed method yield the state-of-the-art classification results on two commonly-used benchmark datasets.
- The extremely complex spectral variability, such as environmental conditions (e.g., local temperature and humidity, atmospheric effects) and instrumental configurations (e.g., sensor noise), hinders the unmixing performance of most existing methods from being further improved. A feasible way to address the problem is to model a relatively general unmixing framework. For that, a novel spectral mixture model, called ALMM, is designed to consider not only the principal spectral variability (e.g., scaling factors) but also other various spectral variabilities to expand the scalability of the endmember dictionary.

  Significantly, a statistical trend is found, that is, the endmembers and spectral variabilities are highly uncorrelated. This motivates us to additionally learn a spectral variability dictionary , whose atoms are assumed to be low-coherent with spectral signatures

of endmembers.
It is worth while to note that the proposed method is able to obtain a more accurate abundance estimation when the spectral variabilities extensively exist, compared to other state-of-the-art algorithms.

- It is crucial to carefully consider the fact that the spectral signature in the original hyperspectral space inevitably suffers from largely and diversely spectral variabilities, possibly due to the high spectral dimension. For this reason, unmixing the HSI in a subspace might be a technically effective alternative. Such subspace unmixing model jointly learns a subspace projection and abundance maps.
  With the low-rank attribute embedding, the learned subspace is robust enough against a variety of spectral variabilities in a more general way, when facing the inverse problems of hyperspectral unmixing.
  Experimental results have demonstrated a higher unmixing performance both visually and quantitatively, in comparison with those methods without considering the subspace strategy.
- HSI will play an irreplaceable role in the coming high-performance and large-scale earth observation tasks. Owing to its narrower spectral sampling width than that of MS image, HSI is able to provide a great possibility to improve the performance of large-scale MS classification or mapping.
  The CoSpace model presented in this thesis locally aligns the manifold structure of MS and HS modalities in order to linearly learn a shared latent subspace. Through the subspace, the highly-discriminative spectral information is expected to be transferred into the MS data over a large coverage, yielding a better large-scale land cover classification. We have shown the superiority of the CoSpace on two MS-HS datasets, which have trade-offs between coverage and spectral resolution.
- In methodology, we further extended the supervised CoSpace model to a semi-supervised version (LeMA) for addressing the issue of the cross-modality learning. LeMA is not limited to the fixed graph structure and few training samples any more, as the graph structure can be adaptively learned from both labeled and unlabeled data.
  Apart from the homologous MS-HS datasets, we also investigated the performance of the proposed LeMA on the heterogeneous MS-Lidar and HS datasets. The best performance using LeMA in comparison with other state-of-the-art methods is obtained either homologous or heterogeneous datasets, demonstrating its effectiveness.

5.2 Outlook

According to the current requirements in various earth observation tasks presented in this dissertation, the future work regarding HSI is supposed to face towards more intellectualized and globalized applications of Earth Vision. There are, therefore, a few potential topics that need to be concerned in the further HSI-related study, which are outlined in the following.

5.2.1 High-Efficiency and Low-Loss Hyperspectral Data Compression

Currently, the data overload in data collection brings a serious challenge in storage capability, particularly for HSI that still holds spectral dimension expect the 2-D image structure. Hyperspectral data compression is an effective and feasible way to fix the issue, thus leading to more specific focuses in what follows:

- ◇ It is more promising to reduce the hyperspectral dimension and meanwhile compress the 2-D image structure, so as to develop the advanced 3-D data compression technique while preserving the spatial-spectral cube structure of HSI.

- ◇ How to reasonably explore few labeled and many unlabeled samples is a key in HDR to make a trade-off between the information loss and spectral discrimination (i.e. semi-supervised strategy).
- ◇ We also expect to perform the data compression and recovery both efficiently and effectively.

5.2.2 Weakly-Supervised Learning-based Hyperspectral Unmixing

It is well-known that spectral unmixing is a special case of blind source separation, that is, it belongs to a kind of unsupervised learning approach from the machine learning perspective. How to take the supervised information in the process of unmixing into consideration would be a promising direction. Two paths can be foreseen to achieve this goal, i.e.

- ◇ We may utilize the artificial abundance maps to train a model and transfer it into the real data.
- ◇ We may also collect the endmembers from the lab or extract them from the real hyperspectral scene, and then accurately estimate the abundance maps as the nearly groundtruth.

5.2.3 Evaluation of Spectral Unmixing: Build the Benchmark Datasets

Unlike the labeling in classification tasks, the groundtruth of abundance maps in the real scene is hardly obtained, consequently leading to the difficulty in quantitatively evaluating the performance of unmixing methods. This challenge can be naturally overcame by building the benchmark datasets. To my best knowledge, up to present, there are only some artificial so-called benchmark datasets in the unmixing assessment. A possible solution to the challenge is to make use of the multi-modal data. For example, given the high spatial resolution RBG or multispectral image, they have to share the same study scene with the HSI of low spatial resolution. Then the corresponding abundance maps could be generated by learning the spatial relationships across multi-modalities.

5.2.4 Time-Series Hyperspectral Data Analysis

The upcoming launch of hyperspectral satellites (e.g., EnMap) enable the time-series hyperspectral data freely available on a larger scale. This brings many meaningful research topics in the future, which can be unfolded with

- ◇ time-series or seasonal spectral unmixing, i.e. used for growth environment analysis of crops or ecological precaution;
- ◇ time-series data fusion, e.g., image super-resolution with spatiotemporal images;
- ◇ change detection for disaster responses, such as water-flood, earthquake, volcano eruption, and so on.

5.2.5 Geospatial Object Detection

HSI is characterized by very rich spectral information, which enables to detect the objects of interest easier from the bird's view. The hyperspectral data, as often as not, fail to individually perform the goespatial object detection, due to its much narrow coverage from the space (a large GSD, e.g., 1m). Nevertheless, the hyperspectral data can be still viewed as a complementary data source to improve the detection accuracy of other modalities (e.g., very high resolution RBG or multispectral image). In particular, HSI's high spectral resolution is capable of identifying the pixel-level or sub-pixel-level objects. This is never achievable for those data sources only with the limited number bands.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Proceedings of Asian Conference on Computer Vision (ACCV), Springer, 180–196.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140: 20–32.

Balasubramanian, M., Schwartz, E. L., 2002. The isomap algorithm and topological stability. Science 295 (5552): 7–7.

Baumgardner, M. F., Biehl, L. L., Landgrebe, D. A., 2015. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. [Online] : Available: https://purr.purdue.edu/ publications/1947/1.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (6): 1373–1396.

Bengio, Y., Paiement, J.-f., Vincent, P., Delalleau, O., Roux, N. L., Ouimet, M., 2004. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), 177–184.

Bioucas-Dias, J. M., Figueiredo, M. A., 2010. Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. In: 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J., 2012. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5 (2): 354–379.

Borel, C. C., Gerstl, S. A., 1994. Nonlinear spectral mixing models for vegetative and soil surfaces. Remote Sensing of Environment 47 (3): 403–416.

Broadwater, J., Banerjee, A., 2009. A comparison of kernel functions for intimate mixture models. In: 2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Broadwater, J., Banerjee, A., 2010. A generalized kernel for areal and intimate mixtures. In: 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Broadwater, J., Banerjee, A., 2011. Mapping intimate mixtures using an adaptive kernel-based technique. In: 2011 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Broadwater, J., Banerjee, A., Burlina, P., 2009. Kernel methods for unmixing hyperspectral imagery. Kernel Methods for Remote Sensing Data Analysis : 249–270.

Cai, D., He, X., Han, J., 2007. Semi-supervised discriminant analysis. In: Proceedings OF IEEE International Conference on Computer Vision (ICCV), 1–7.

Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J., 2006. Composite kernels for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters 3 (1): 93–97.

Chandrasekhar, S., 2013. Radiative transfer. Courier Corporation.

Chen, J., Wang, C., Wang, R., 2009. Using stacked generalization to combine svms in magnitude and shape feature spaces for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 47 (7): 2193–2205.

Chen, L., Huang, R., Huang, W., 2010. Graph-based semi-supervised weighted band selection for classification of hyperspectral data. In: Proceedings OF International Conference on Audio, Language and Image Processing, IEEE, 1123–1126.

Chen, S. S., Donoho, D. L., Saunders, M. A., 2001. Atomic decomposition by basis pursuit. SIAM Review 43 (1): 129–159.

Chen, Y., Nasrabadi, N. M., Tran, T. D., 2011. Hyperspectral image classification using dictionary-

based sparse representation. IEEE Transactions on Geoscience and Temote Sensing 49 (10): 3973–3985.

Clark, R. N., Swayze, G. A., Livo, K. E., Kokaly, R. F., Sutley, S. J., Dalton, J. B., McDougal, R. R., Gent, C. A., 2003. Imaging spectroscopy: Earth and planetary remote sensing with the usgs tetracorder and expert systems. Journal of Geophysical Research: Planets 108 (E12).

Comon, P., 1994. Independent component analysis, a new concept? Signal Processing 36 (3): 287–314.

Dalla Mura, M., Villa, A., Benediktsson, J. A., Chanussot, J., Bruzzone, L., 2011. Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis. IEEE Geoscience and Remote Sensing Letters 8 (3): 542–546.

Dobigeon, N., Tourneret, J.-Y., Richard, C., Bermudez, J. C. M., McLaughlin, S., Hero, A. O., 2014. Nonlinear unmixing of hyperspectral images: Models and algorithms. IEEE Signal Processing Magazine 31 (1): 82–94.

Drumetz, L., Meyer, T. R., Chanussot, J., Bertozzi, A. L., Jutten, C., 2019. Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms. IEEE Transactions on Image Processing .

Drumetz, L., Veganzones, M.-A., Henrot, S., Phlypo, R., Chanussot, J., Jutten, C., 2016. Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability. IEEE Transactions on Image Processing 25 (8): 3890–3905.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. Remote sensing of Environment 120: 25–36.

Du, B., Zhang, M., Zhang, L., Hu, R., Tao, D., 2017. Pltd: Patch-based low-rank tensor decomposition for hyperspectral images. IEEE Transactions on Multimedia 19 (1): 67–79.

Dube, T., Mutanga, O., 2016. The impact of integrating worldview-2 sensor and environmental variables in estimating plantation forest species aboveground biomass and carbon stocks in umgeni catchment, south africa. ISPRS Journal of Photogrammetry and Remote Sensing 119: 415–425.

Fan, W., Hu, B., Miller, J., Li, M., 2009. Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data. International Journal of Remote Sensing 30 (11): 2951–2962.

Fauvel, M., Chanussot, J., Benediktsson, J. A., 2009. Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. EURASIP Journal on Advances in Signal Processing 2009 (1): 783194.

Feng, Y.-Z., Sun, D.-W., 2012. Application of hyperspectral imaging in food safety inspection and control: a review. Critical reviews in food science and nutrition 52 (11): 1039–1058.

Fu, X., Ma, W.-K., Bioucas-Dias, J. M., Chan, T.-H., 2016. Semiblind hyperspectral unmixing in the presence of spectral library mismatches. IEEE Transactions on Geoscience and Remote Sensing 54 (9): 5171–5184.

Ghamisi, P., Benediktsson, J. A., Sveinsson, J. R., 2014. Automatic spectral–spatial classification framework based on attribute profiles and supervised feature extraction. IEEE Transactions on Geoscience and Remote Sensing 52 (9): 5771–5782.

Ghamisi, P., Höfle, B., Zhu, X. X., 2017. Hyperspectral and lidar data fusion using extinction profiles and deep convolutional neural network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (6): 3011–3024.

Ghiyamat, A., Shafri, H. Z., 2010. A review on hyperspectral remote sensing for homogeneous and heterogeneous forest biodiversity assessment. International Journal of Remote Sensing 31 (7): 1837–1856.

Giampouras, P. V., Themelis, K. E., Rontogiannis, A. A., Koutroumbas, K. D., 2016. Simultaneously sparse and low-rank abundance matrix estimation for hyperspectral image unmixing. IEEE Transactions on Geoscience and Remote Sensing 54 (8): 4775–4789.

Goetz, A. F., Vane, G., Solomon, J. E., Rock, B. N., 1985. Imaging spectrometry for earth remote sensing. Science 228 (4704): 1147–1153.

Gowen, A., O'Donnell, C., Cullen, P., Downey, G., Frias, J., 2007. Hyperspectral imaging–an emerging process analytical tool for food quality and safety control. Trends in food science & technology 18 (12): 590–598.

Guo, R., Wang, W., Qi, H., 2015. Hyperspectral image unmixing using autoencoder cascade. In: 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Halimi, A., Altmann, Y., Dobigeon, N., Tourneret, J.-Y., 2011. Nonlinear unmixing of hyperspectral images using a generalized bilinear model. IEEE Transactions on Geoscience and Remote Sensing 49 (11): 4153–4162.

Hall, M. A., 1999. Correlation-based feature selection for machine learning .

Hang, R., Liu, Q., Hong, D., Ghamisi, P., 2019. Cascaded recurrent neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 57 (8): 5384–5394.

Hapke, B., 1981. Bidirectional reflectance spectroscopy: 1. theory. Journal of Geophysical Research: Solid Earth 86 (B4): 3039–3054.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Proceedings of Asian Conference on Computer Vision (ACCV), Springer, 213–228.

He, X., Cai, D., Yan, S., Zhang, H.-J., 2005. Neighborhood preserving embedding. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Vol. 2, IEEE, 1208–1213.

He, X., Niyogi, P., 2004. Locality preserving projections. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), 153–160.

Heinz, D. C., et al., 2001. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 39 (3): 529–545.

Heylen, R., Burazerovic, D., Scheunders, P., 2011. Fully constrained least squares spectral unmixing by simplex projection. IEEE Transactions on Geoscience and Remote Sensing 49 (11): 4112–4122.

Hong, D., Chanussot, J., Yokoya, N., Heiden, U., Heldens, W., Zhu, X. X., 2019a. Wu-net: A weakly-supervised unmixing network for remotely sensed hyperspectral imagery. In: 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 1–4.

Hong, D., Liu, W., Su, J., Pan, Z., Wang, G., 2015. A novel hierarchical approach for multispectral palmprint recognition. Neurocomputing 151: 511–521.

Hong, D., Liu, W., Wu, X., Pan, Z., Su, J., 2016a. Robust palmprint recognition based on the fast variation vese–osher model. Neurocomputing 174: 999–1012.

Hong, D., Pan, Z., Wu, X., 2014a. Improved differential box counting with multi-scale and multi-direction: A new palmprint recognition method. Optik-International Journal for Light and Electron Optics 125 (15): 4154–4160.

Hong, D., Su, J., Hong, Q., Pan, Z., Wang, G., 2014b. Blurred palmprint recognition based on stable-feature extraction using a vese–osher decomposition model. PloS one 9 (7): e101866.

Hong, D., Yokoya, N., Chanussot, J., Xu, J., Zhu, X. X., 2019b. Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction. ISPRS Journal of Photogrammetry and Remote Sensing 158: 35–49.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X. X., 2017a. Learning a low-coherence dictionary to address spectral variability for hyperspectral unmixing. In: Proceedings of IEEE International Conference on Image Processing (ICIP), IEEE, 235–239.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X. X., 2019c. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. IEEE Transactions on Image Processing 28 (4): 1923–1938.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X. X., 2019d. Cospace: Common subspace learning from hyperspectral-multispectral correspondences. IEEE Transactions on Geoscience and Remote Sensing 57 (7): 4349–4359.

Hong, D., Yokoya, N., Ge, N., Chanussot, J., Zhu, X. X., 2019e. Learnable manifold alignment (lema): A semi-supervised cross-modality learning framework for land cover and land use classification.

ISPRS Journal of Photogrammetry and Remote Sensing 147: 193–205.

Hong, D., Yokoya, N., Xu, J., Zhu, X., 2018. Joint & progressive learning from high-dimensional data for multi-label classification. In: Proceedings of European Conference on Computer Vision (ECCV), 469–484.

Hong, D., Yokoya, N., Zhu, X. X., 2016b. The k-lle algorithm for nonlinear dimensionality ruduction of large-scale hyperspectral data. In: 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–5.

Hong, D., Yokoya, N., Zhu, X. X., 2017b. Learning a robust local manifold representation for hyperspectral dimensionality reduction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (6): 2960–2975.

Hong, D., Zhu, X. X., 2018. Sulora: Subspace unmixing with low-rank attribute embedding for hyperspectral data analysis. IEEE Journal of Selected Topics in Signal Processing 12 (6): 1351–1363.

Hong, D. F., Yokoya, N., Zhu, X. X., 2016c. Local manifold learning with robust neighbors selection for hyperspectral dimensionality reduction. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 40–43.

Hu, J., Hong, D., Wang, Y., Zhu, X. X., 2019. A comparative review of manifold learning techniques for hyperspectral and polarimetric sar image fusion. Remote Sensing 11 (6): 681.

Huang, H., Luo, F., Liu, J., Yang, Y., 2015. Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding. ISPRS Journal of Photogrammetry and Remote Sensing 106: 42–54.

Huang, R., Hong, D., Xu, Y., Yao, W., Stilla, U., 2019. Multi-scale local context embedding for lidar point cloud classification. IEEE Geoscience and Remote Sensing Letters .

Imani, M., Ghassemian, H., 2015. Feature space discriminant analysis for hyperspectral data feature reduction. ISPRS Journal of Photogrammetry and Remote Sensing 102: 1–13.

Indyk, P., Motwani, R., 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM, 604–613.

Iordache, M.-D., Bioucas-Dias, J. M., Plaza, A., 2012. Total variation spatial regularization for sparse hyperspectral unmixing. IEEE Transactions on Geoscience and Remote Sensing 50 (11): 4484–4502.

Iordache, M.-D., Bioucas-Dias, J. M., Plaza, A., 2014. Collaborative sparse regression for hyperspectral unmixing. IEEE Transactions on Geoscience and Remote Sensing 52 (1): 341–354.

Jayaprakash, C., Damodaran, B. B., Soman, K., et al., 2018. Randomized ica and lda dimensionality reduction methods for hyperspectral image classification. arXiv preprint arXiv:1804.07347 .

Ji, S., Ye, J., 2009. Linear dimensionality reduction for multi-label classification. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI).

Jiang, X., Lai, J., 2015. Sparse and dense hybrid representation via dictionary decomposition for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (5): 1067–1079.

Jolliffe, I., 2011. Principal component analysis. Springer.

Kang, X., Li, C., Li, S., Lin, H., 2018. Classification of hyperspectral images by gabor filtering based deep network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (4): 1166–1178.

Karami, A., Yazdi, M., Mercier, G., 2012. Compression of hyperspectral images using discerete wavelet transform and tucker decomposition. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5 (2): 444–450.

Karila, K., Yu, X., Vastaranta, M., Karjalainen, M., Puttonen, E., Hyyppä, J., 2019. Tandem-x digital surface models in boreal forest above-ground biomass change detection. ISPRS Journal of Photogrammetry and Remote Sensing 148: 174–183.

Khan, A., Kim, I., Kong, S. G., 2009. Dimensionality reduction of hyperspectral images using kernel ica. In: Sensing for Agriculture and Food Quality and Safety, Vol. 7315, International Society for Optics and Photonics, 731510.

Kim, H., Park, H., 2008. Nonnegative matrix factorization based on alternating nonnegativity con-

strained least squares and active set method. SIAM Journal on Matrix Analysis and Applications 30 (2): 713–730.

Koller, D., Sahami, M., 1996. Toward optimal feature selection. Tech. rep., Stanford InfoLab.

Kubelka, P., Munk, F., 1931. Reflection characteristics of paints. Zeitschrift fur Technische Physik 12: 593–601.

Le Saux, B., Yokoya, N., Hänsch, R., Prasad, S., 2018. 2018 ieee grss data fusion contest: Multimodal land use classification [technical committees]. IEEE Geoscience and Remote Sensing Magazine 6 (1): 52–54.

Li, K.-C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86 (414): 316–327.

Li, W., Du, Q., 2016. Laplacian regularized collaborative graph for discriminant analysis of hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 54 (12): 7066–7076.

Li, W., Liu, J., Du, Q., 2016. Sparse and low-rank graph for discriminant analysis of hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 54 (7): 4094–4105.

Liao, D., Qian, Y., Tang, Y. Y., 2018. Constrained manifold learning for hyperspectral imagery visualization. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (4): 1213–1226.

Liao, W., Bellens, R., Pizurica, A., Philips, W., Pi, Y., 2012. Classification of hyperspectral data over urban areas using directional morphological profiles and semi-supervised feature extraction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5 (4): 1177–1190.

Liao, W., Pižurica, A., Bellens, R., Gautama, S., Philips, W., 2015. Generalized graph-based fusion of hyperspectral and lidar data using morphological features. IEEE Geoscience and Remote Sensing Letters 12 (3): 552–556.

Liao, W., Pizurica, A., Scheunders, P., Philips, W., Pi, Y., 2013. Semisupervised local discriminant analysis for feature extraction in hyperspectral images. IEEE Transactions on Geoscience and Remote Sensing 51 (1): 184–198.

Licciardi, G., Marpu, P. R., Chanussot, J., Benediktsson, J. A., 2012. Linear versus nonlinear pca for the classification of hyperspectral data based on the extended morphological profiles. IEEE Geoscience and Remote Sensing Letters 9 (3): 447–451.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y., 2013. Robust recovery of subspace structures by low-rank representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1): 171–184.

Liu, X., Deng, C., Chanussot, J., Hong, D., Zhao, B., 2019. Stfnet: A two-stream convolutional neural network for spatiotemporal image fusion. IEEE Transactions on Geoscience and Remote Sensing 57 (9): 6552–6564.

Lunga, D., Prasad, S., Crawford, M. M., Ersoy, O., 2014. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. IEEE Signal Processing Magazine 31 (1): 55–66.

Ly, N. H., Du, Q., Fowler, J. E., 2014a. Collaborative graph-based discriminant analysis for hyperspectral imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (6): 2688–2696.

Ly, N. H., Du, Q., Fowler, J. E., 2014b. Sparse graph-based discriminant analysis for hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 52 (7): 3872–3884.

Ma, L., Crawford, M. M., Tian, J., 2010. Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 48 (11): 4099–4109.

Ma, L., Crawford, M. M., Yang, X., Guo, Y., 2015. Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 53 (5): 2832–2844.

Ma, L., Ma, A., Ju, C., Li, X., 2016a. Graph-based semi-supervised learning for spectral-spatial hyperspectral image classification. Pattern Recognition Letters 83: 133–142.

Ma, L., Zhang, X., Yu, X., Luo, D., 2016b. Spatial regularized local manifold learning for classifica-

tion of hyperspectral images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (2): 609–624.

Matasci, G., Volpi, M., Kanevski, M., Bruzzone, L., Tuia, D., 2015. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 53 (7): 3550–3564.

Matasci, G., Volpi, M., Tuia, D., Kanevski, M., 2011. Transfer component analysis for domain adaptation in image classification. In: Image and Signal Processing for Remote Sensing XVII, Vol. 8180, International Society for Optics and Photonics, 81800F.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks 12 (2).

Nascimento, J. M., Dias, J. M., 2005. Vertex component analysis: A fast algorithm to unmix hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43 (4): 898–910.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., 2011. Multimodal deep learning. In: Proceedings of International Conference onMachine Learning (ICML), 689–696.

Ortenberg, F., Thenkabail, P., Lyon, J., Huete, A., 2011. Hyperspectral sensor characteristics: airborne, spaceborne, hand-held, and truck-mounted; integration of hyperspectral data with lidar.

Pacifici, F., Du, Q., Prasad, S., 2013. Report on the 2013 ieee grss data fusion contest: Fusion of hyperspectral and lidar data [technical committees]. IEEE Geoscience and Remote Sensing Magazine 1 (3): 36–38.

Palsson, B., Sigurdsson, J., Sveinsson, J. R., Ulfarsson, M. O., 2018. Hyperspectral unmixing using a neural network autoencoder. IEEE Access 6: 25646–25656.

Pan, L., Li, H.-C., Deng, Y.-J., Zhang, F., Chen, X.-D., Du, Q., 2017. Hyperspectral dimensionality reduction by tensor sparse and low-rank graph-based discriminant analysis. Remote Sensing 9 (5): 452.

Pan, S. J., Tsang, I. W., Kwok, J. T., Yang, Q., 2011. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks 22 (2): 199–210.

Patra, S., Modi, P., Bruzzone, L., 2015. Hyperspectral band selection based on rough set. IEEE Transactions on Geoscience and Remote Sensing 53 (10): 5495–5503.

Qu, Y., Guo, R., Qi, H., 2017. Spectral unmixing through part-based non-negative constraint denoising autoencoder. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 209–212.

Rainforth, T., Wood, F., 2015. Canonical correlation forests. arXiv preprint arXiv:1507.05444 .

Renard, N., Bourennane, S., 2009. Dimensionality reduction based on tensor modeling for classification methods. IEEE Transactions on Geoscience and Remote Sensing 47 (4): 1123–1131.

Riedmann, M., Milton, E., 2003. Supervised band selection for optimal use of data from airborne hyperspectral sensors. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Vol. 3, IEEE, 1770–1772.

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

Roweis, S. T., Saul, L. K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500): 2323–2326.

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., et al., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment 145: 154–172.

Sainui, J., Sugiyama, M., 2013. Direct approximation of quadratic mutual information and its application to dependence-maximization clustering. IEICE Transactions on Information and Systems 96 (10): 2282–2285.

Shao, Z., Zhang, L., 2014. Sparse dimensionality reduction of hyperspectral image based on semi-supervised local fisher discriminant analysis. International Journal of Applied Earth Observation and Geoinformation 31: 122–129.

Shkuratov, Y., Starukhina, L., Hoffmann, H., Arnold, G., 1999. A model of spectral albedo of particulate surfaces: Implications for optical properties of the moon. Icarus 137 (2): 235–246.

Somers, B., Cools, K., Delalieux, S., Stuckens, J., Van der Zande, D., Verstraeten, W. W., Coppin, P., 2009. Nonlinear hyperspectral mixture analysis for tree cover estimates in orchards. Remote Sensing of Environment 113 (6): 1183–1193.

Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L. S., Gao, W., 2015. Multi-task learning with low rank attribute embedding for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR), 3739–3747.

Su, H., Du, P., Du, Q., 2012a. Semi-supervised dimensionality reduction using orthogonal projection divergence-based clustering for hyperspectral imagery. Optical Engineering 51 (11): 111715.

Su, H., Sheng, Y., Du, P., Liu, K., 2012b. Adaptive affinity propagation with spectral angle mapper for semi-supervised hyperspectral band selection. Applied Optics 51 (14): 2656–2663.

Su, Y., Li, J., Plaza, A., Marinoni, A., Gamba, P., Chakravortty, S., 2019. Daen: Deep autoencoder networks for hyperspectral unmixing. IEEE Transactions on Geoscience and Remote Sensing .

Su, Y., Marinoni, A., Li, J., Plaza, J., Gamba, P., 2018. Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing. IEEE Geoscience and Remote Sensing Letters 15 (9): 1427–1431.

Sugiyama, M., 2007. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. Journal of Machine Learning Research 8 (May): 1027–1061.

Sun, W., Du, Q., 2018. Graph-regularized fast and robust principal component analysis for hyperspectral band selection. IEEE Transactions on Geoscience and Remote Sensing 56 (6): 3185–3195.

Suzuki, T., Sugiyama, M., 2013. Sufficient dimension reduction via squared-loss mutual information estimation. Neural Computation 25 (3): 725–758.

Tangkaratt, V., Sasaki, H., Sugiyama, M., 2017. Direct estimation of the derivative of quadratic mutual information with application in supervised dimension reduction. Neural Computation 29 (8): 2076–2122.

Thompson, B., 2007. Factor analysis. The Blackwell Encyclopedia of Sociology .

Thouvenin, P.-A., Dobigeon, N., Tourneret, J.-Y., 2016. Hyperspectral unmixing with spectral variability using a perturbed linear mixing model. IEEE Transactions on Signal Processing 64 (2): 525–538.

Tsang, L., Kong, J. A., Shin, R. T., 1985. Theory of microwave remote sensing .

Tuia, D., Camps-Valls, G., 2016. Kernel manifold alignment for domain adaptation. PloS One 11 (2): e0148655.

Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G., 2014. Semisupervised manifold alignment of multimodal remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 52 (12): 7708–7720.

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E., Steininger, M., 2003. Remote sensing for biodiversity science and conservation. Trends in ecology & evolution 18 (6): 306–314.

Uezato, T., Fauvel, M., Dobigeon, N., 2019. Hyperspectral unmixing with spectral variability using adaptive bundles and double sparsity. IEEE Transactions on Geoscience and Remote Sensing .

Ulaby, F. T., Moore, R. K., Fung, A. K., 1986. Microwave remote sensing: Active and passive. volume 3-from theory to applications .

Veganzones, M. A., Drumetz, L., Tochon, G., Dalla Mura, M., Plaza, A., Bioucas-Dias, J., Chanussot, J., 2014. A new extended linear mixing model to address spectral variability. In: 2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, 1–4.

Villa, A., Benediktsson, J. A., Chanussot, J., Jutten, C., 2011. Hyperspectral image classification with independent component discriminant analysis. IEEE Transactions on Geoscience and Remote Sensing 49 (12): 4865–4876.

Wang, C., Mahadevan, S., 2009. A general framework for manifold alignment. In: 2009 AAAI Fall Symposium Series.

Wang, C., Mahadevan, S., 2011. Heterogeneous domain adaptation using manifold alignment. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI).

Wang, J., Chang, C.-I., 2006. Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis. IEEE Transactions on Geoscience and Remote Sens-

ing 44 (6): 1586–1600.

Wang, L., Hao, S., Wang, Q., Wang, Y., 2014. Semi-supervised classification for hyperspectral imagery based on spatial-spectral label propagation. ISPRS Journal of Photogrammetry and Remote Sensing 97: 123–137.

Wang, X., Gao, Y., Cheng, Y., 2016. A non-negative sparse semi-supervised dimensionality reduction algorithm for hyperspectral data. Neurocomputing 188: 275–283.

Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y., 2009. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2080–2088.

Wu, H., Prasad, S., 2018. Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels. Pattern Recognition 74: 212–224.

Wu, X., Hong, D., Chanussot, J., Xu, Y., Tao, R., Wang, Y., 2019a. Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection. arXiv preprint arXiv:1905.11074 .

Wu, X., Hong, D., Ghamisi, P., Li, W., Tao, R., 2018. Msri-ccf: Multi-scale and rotation-insensitive convolutional channel features for geospatial object detection. Remote Sensing 10 (12): 1990.

Wu, X., Hong, D., Tian, J., Chanussot, J., Li, W., Tao, R., 2019b. Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. IEEE Transactions on Geoscience and Remote Sensing .

Xia, J., Bombrun, L., Adali, T., Berthoumieu, Y., Germain, C., 2016. Classification of hyperspectral data with ensemble of subspace ica and edge-preserving filtering. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1422–1426.

Xia, J., Chanussot, J., Du, P., He, X., 2014. (semi-) supervised probabilistic principal component analysis for hyperspectral remote sensing image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (6): 2224–2236.

Xiong, F., Qian, Y., Zhou, J., Tang, Y. Y., 2018. Hyperspectral unmixing via total variation regularized nonnegative tensor factorization. IEEE Transactions on Geoscience and Remote Sensing .

Xue, Z., Du, P., Li, J., Su, H., 2015. Simultaneous sparse graph embedding for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 53 (11): 6114–6133.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S., 2007. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence (1): 40–51.

Yang, H. L., Crawford, M. M., 2016. Domain adaptation with preservation of manifold geometry for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (2): 543–555.

Yang, J., Yang, J.-y., 2003. Why can lda be performed in pca transformed space? Pattern Recognition 36 (2): 563–566.

Yao, J., Meng, D., Zhao, Q., Cao, W., Xu, Z., 2019. Nonconvex-sparsity and nonlocal-smoothness based blind hyperspectral unmixing. IEEE Transactions on Image Processing .

Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Le Saux, B., Moser, G., Tuia, D., 2018. Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (5): 1363–1377.

Zabalza, J., Ren, J., Yang, M., Zhang, Y., Wang, J., Marshall, S., Han, J., 2014. Novel folded-pca for improved feature extraction and data reduction with hyperspectral imaging and sar in remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 93: 112–122.

Zhang, L., Zhang, L., Du, B., You, J., Tao, D., 2019. Hyperspectral image unsupervised classification by robust manifold matrix factorization. Information Sciences 485: 154–169.

Zhang, L., Zhang, L., Tao, D., Huang, X., 2013a. Tensor discriminative locality alignment for hyperspectral image spectral–spatial feature extraction. IEEE Transactions on Geoscience and Remote Sensing 51 (1): 242–256.

Zhang, X., He, Y., Zhou, N., Zheng, Y., 2013b. Semisupervised dimensionality reduction of hyper-

spectral images via local scaling cut criterion. IEEE Geoscience and Remote Sensing Letters 10 (6): 1547–1551.

Zhu, X., Ghahramani, Z., Lafferty, J. D., 2003. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of International Conference on Machine Learning (ICML), 912–919.

List of Figures

List of Tables

Acknowledgement

Over the past nearly four years, I devoted to the development of hyperspectral data analysis in remote sensing with the great expectancy to make it applicable to the next-generation earth observation tasks. During the period, the achievement I made could not have been fulfilled without the great support of many people.

My deepest gratitude goes first and foremost to Prof. Xiaoxiang Zhu, my first supervisor, who gives me the chance to purse my doctor degree in her group. Otherwise, it is hardly possible to have the follow-up offers in a variety of amazing resources and cooperation from the famous institution – DLR and university – TUM. I still remembered that day when I was interviewed by her with Prof. Richard Bamler and Dr. Naoto Yokoya. They were carefully listening to my answers spoken in unskillful English and discussed with me patiently, which deeply touched me. I would also like to thank Prof. Zhu for providing a great Ph.D topic: hyperspectral data analysis, which is promising and cutting-edge, and multiple scientific discussions to make my research smoothly be in progress. She has spent the time not only offering me valuable suggestions in the academic studies but also getting to know the difficulties in my life and helping me fix them to a great extent. Without her consistent and illuminating instruction, this thesis could not reach its present form.

Especially, I am greatly indebted to my second supervisor Dr. Naoto Yokoya, who has walked me through the first two most difficult years with his patient instruction, insightful criticism, and expert guidance. His professionalism does not lie in pointing you to a specific topic but guiding you to find a leading-edge research problem, his conscientiousness does not lie in revising your draft word-for-word but teaching you how to organize the structure of your paper logically, and his enthusiasm does not lie in giving you a hand whenever and wherever possible but assisting you in a proper or key time to gradually improve the adaptivity and resistance in handling the emergencies. After he left our group to be a leader in RIKEN AIP, he still spent much time to discuss with me and point out the mistakes in my drafted papers, as well as give me tremendous encouragement and unwavering support when I encountered difficulties both for work and life. A very touching moment is that he often utilized non-working hours to answer me questions via email or Skype at dead of night (due to the time difference between Tokyo (night) and Munich (afternoon)) in order to timely give me feedback. His these excellent qualities render me to approach him to be a real researcher like him in the future. I believe, it must be so!

The same heartfelt gratitude goes to my another supervisor Prof. Jocelyn Chanussot, who is full professor from a world-famous GIPSA-lab of University of Grenoble Aples, Grenoble, France. He is a very knowledgeable expert in remote sensing. In particular, he is one of the most distinguished contributors in the HSI-related fields. Until today, I still vividly recall that moment that it was the first time for me to have an online meeting with Jocelyn via Skype. At that time, I was kind of nervous to report the work to him using loosely professional terms with my broken English, which puts me to shame and could be a behavior of wasting time. Conversely, I was deeply moved by his incentive and positive feedback to my work. This experience reignites the passion and strengthens the conviction to my research topics. Also, he is very friendly, easy-going, versatile and responsible person. I have had the privileges of visiting Jocelyn's group and working together with him. What impressed me most was that it is unshakeable to have our routine meeting once a week, although he was nearly fully occupied by tons of things. That is something that I can not imagine before, especially for big professors like him. The only thing that I can do is to work harder to pay him back, keeping the latest ideas and progress reported in each routine meeting.

Besides my supervisors, my sincere thanks also go to all my colleagues and friends of TUM and DLR for their scholarly advice and generous help during my stay in Germany. In par-

ticular, I am most grateful to several people who play a vital and indispensable role in the period of pursing the doctor degree. Dr. Jian Xu, who is a senior and more like an intimate friend, is not only a spiritual mentor to help me overcome the culture shock period of encountering the new cultures at the very beginning, but also like an unofficially academic mentor to broaden my horizon and enrich the ways of thinking. Mr. Jingliang Hu, who is one of my great friends and also my colleague in our group, has provided me countless help in various aspects, especially in the niggling but important matters of everyday life. Mr. Yuanxin Xia, who is a second-year junior Ph.D student and my good friend, has witnessed my own ups and downs in the years and encouraged me to look ahead. As my another good friend, although Mr. Bo Zhang is just a one-year visiting student, yet I gratefully acknowledge his instruction in the use of GSI-software and Google cloud as well as the applications of temperature inversion appeared in our collaborative paper. Often being invited by Jian, Jingling, Yuanxin, Bo, and Ms. Song Liu to hang out or have the party makes me happy and forget the pain of homesickness temporarily. With the aid of Mr. Nan Ge, who is very good at modeling the inverse problem and its optimization, I successfully solved many challenges in my topic, and the collision of ideas between us boosted the progress of our work as well.

I would also like to extend my gratitude to Dr. Yusheng Xu, Miss. Rong Huang, and Dr. Zhen Ye, who are from Chair of Photogrammetry and Remote Sensing, TUM. Their specialism in 3-D point cloud data processing and analysis provides me a new insight to see the remote sensing techniques, and they have also put considerable time and effort into the valuable comments and suggestions for my thesis and future career.

This thesis could not be completed without the great support from the other colleagues of our group SiPEO led by my supervisor Prof. Zhu and EOC of DLR. Their willingness to do me a favor so generously has been much appreciated, including Dr. Michael Schmitt as the deputy of SiPEO for dealing with various paper work in silence in order to create a stable research environment to us, Dr. Martin Werner for willingly helping us solve the IT-related difficulties, Dr. Pedram Ghamisi for voluntarily providing us technical support and research guidance, Dr. Gerald Baier for registering the access to DLR for me with warmth and affection many times and sharing his experience not for the purpose of return, Dr. Yuanyuan Wang as the first graduated Ph.D in SiPEO for sharing his precious experience with us and also providing us a lot of convenience, Dr. Rong Liu and Dr. Wei Yao for kindly organizing the round-table meeting, Dr. Claas Grohnfeldt for systematically summarizing his work to me, Mr. Jian Kang for frequently making academic discussion with me and assisting me to solve massive knotty problems, Mr. Sina Montazeri for kindly providing me the information regarding the reimbursements from the TUM graduate school, sorting out and sharing the defense process to us, and Ms. Chunping Qiu for bringing the stuff from the Asian supermarkets to me so kindly, as well as Dr. Muhammad Shahzad, Mr. Hossein Bagheri, Mr. Lichao Mou, Mrs. Yao Sun, Mr. Lloya Hughes, Mr. Matthias Häberle, Mr. Eike Hoffmann, Mr. Yuansheng Hua, Mr. Yilei Shi, Mr. Hao Li, Ms. Guicheng Zhang, Mr. Yu Li, Mr. Zhuoru Wang, Miss. Jun Zhang, Prof. Kai Qin, and so on.

I am also deeply grateful to Dr. Uta Heiden, Dr. Wieke Heldens, and Miss. Chaonan Ji from EOC-DFD of DLR for organizing an amazing discussion meeting and sharing the processed data to me.

I also owe a special debt of gratitude to all the professors and colleagues from other institutions or universities, who helped me directly or indirectly in my studies, such as Prof. Pierre Common who offered me two very important discussions regrading the tensor when I was visiting the GIPSA-lab, Prof. Lucas Drumetz for providing me the simulated datasets for my topic, Prof. Yang Xu and Mr. Xun Liu enthusiastically for serving me while I was living in Grenoble for academic exchange, etc.

Last but not least, my gratitude would go to my beloved family: my parents, for giving birth

to me, for continuously educating me, for never giving up me, and for their loving considerations and great confidence in me all through these years; my grandfather and grandmother, for unconditionally supporting me, for cheering me up, and for paying close attention to my latest situations all the time. Emphatically for my wife (Xin Wu), her great care and tolerance in life deserve more thanks than I can find words to express. I have to say, however, that properly managing the long-distance relationship between us is challenging, yet mutual understanding and tolerance make us more mature and more cherish what we have and where we are on our path right now.

Appendices

A Hong D., Yokoya N., Zhu X. X., 2017. Learning a Robust Local Manifold Representation for Hyperspectral Dimensionality Reduction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), 10(6): 2960-2975.

https://ieeexplore.ieee.org/document/7985008

Learning a Robust Local Manifold Representation for Hyperspectral Dimensionality Reduction

Danfeng Hong, *Student Member, IEEE*, Naoto Yokoya, *Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*—Local manifold learning has been successfully applied to hyperspectral dimensionality reduction in order to embed nonlinear and nonconvex manifolds in the data. Local manifold learning is mainly characterized by affinity matrix construction, which is composed of two steps: neighbor selection and computation of affinity weights. There is a challenge in each step: First, the neighbor selection is sensitive to complex spectral variability due to nonuniform data distribution, illumination variations, and sensor noise; second, the computation of affinity weights is challenging due to highly correlated spectral signatures in the neighborhood. To address the two issues, in this paper, a novel manifold learning methodology based on locally linear embedding is proposed through learning a robust local manifold representation. More specifically, a hierarchical neighbor selection is designed to progressively eliminate the effects of complex spectral variability using joint normalization and to robustly compute affinity (or reconstruction) weights reducing multicollinearity via the refined neighbor selection. Additionally, an idea that combines spatial–spectral information is introduced into the proposed manifold learning methodology to further improve the robustness of affinity calculations. Classification is explored as a potential application for validating the proposed algorithm. The classification accuracy in the use of different dimensionality reduction methods is evaluated and compared, while two kinds of strategies are applied in selecting the training and test samples: random sampling and region-based sampling. Experimental results show the classification accuracy obtained by the proposed method is superior to those state-of-the-art dimensionality reduction methods.**

***Index Terms*—Dimensionality reduction (DR), hyperspectral image, local manifold learning (LML), multicollinearity, nonuniform data distribution.**

I. Introduction

HYPERSPECTRAL data are characterized by very rich spectral information, which enables us to detect targets of interest and analyze data attributes more easily, but also introduces drawbacks caused by its high dimensionality. As a result, the dimensionality reduction (DR) is a necessary and essential ingredient to address the aforementioned issue. A large number of DR techniques have been developed for a wide range of applications, including image segmentation [1], biometric [2], large-scale data classification [3], image/video analysis [4], and visualization [5]. Generally, these DR approaches can be categorized into linear and nonlinear methods.

Classical linear methods, such as principal component analysis (PCA) [6], easily fail to excavate the underlying data structure that lies in the complex real world. Comparatively, many nonlinear techniques, such as manifold learning (Isomap [7], locally linear embedding (LLE) [8], Laplacian eigenmaps (LE) [9], and local tangent space alignment (LTSA) [10]), exhibit unique advantages in DR and obtain state-of-the-art results in many fields. These examples of successful use of manifold learning mentioned above have widely attracted the attention of researchers working in the field of hyperspectral data analysis. Owing to merits of manifold learning, which can effectively map nonlinear and nonconvex manifolds in low-dimensional space, massive related approaches are introduced into hyperspectral image processing and successfully applied to various tasks, e.g., feature extraction [11], [12], classification [13]–[16], detection [17], [18], and multitemporal analysis [19]. In addition, it has been proven in [3] that the algorithm performance with global manifold methods is inferior to that with local manifold methods. As a typical and benchmark local manifold learning (LML) method, LLE explores locally linear and globally nonlinear assumptions to effectively capture the underlying intrinsic structure of data. LLE has been successfully applied to hyperspectral classification. Ma *et al.* [13] integrated LML with improved k-nearest neighbor for hyperspectral classification tasks. In [14], Ma *et al.* extended their work and proposed a kind of semisupervised hyperspectral image classification method based on LML. Tang *et al.* [16] proposed manifold based on sparse representation for hyperspectral classification, and they embedded the local geometric property using the local manifold representation into classification framework based on sparse representation in order to enforcedly keep consistent from sparse code to local manifold representation.

Current research on manifold learning methods in hyperspectral data processing mostly focuses on their potential for classification or detection tasks and frequently neglects the representation capability of the manifold structure, leading to difficulty in improving the classification accuracy. In other words,

Manuscript received September 30, 2016; revised December 11, 2016 and January 26, 2017; accepted February 25, 2017. Date of publication July 18, 2017; date of current version July 17, 2017. This work was supported in part by the European Research Council (ERC) under the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement ERC-2016-StG-714087 (Acronym: So2Sat), and in part by the Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de). *(Corresponding author: Xiao Xiang Zhu.)*

D. Hong and X. X. Zhu are with the Remote Sensing Technology Institute (IMF), German Aerospace Center, Weßling 82234, Germany, and also with the Signal Processing in Earth Observation, Technical University of Munich, Munich 80333, Germany (e-mail: danfeng.hong@dlr.de; xiao.zhu@dlr.de).

N. Yokoya is with the Remote Sensing Technology Institute (IMF), German Aerospace Center, Weßling 82234, Germany, with the Signal Processing in Earth Observation, Technical University of Munich, Munich 80333, Germany, and also with the Department of Advanced Interdisciplinary Studies, University of Tokyo, Tokyo 153-8904, Japan (e-mail: yokoya@sal.rcast.u-tokyo.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2017.2682189

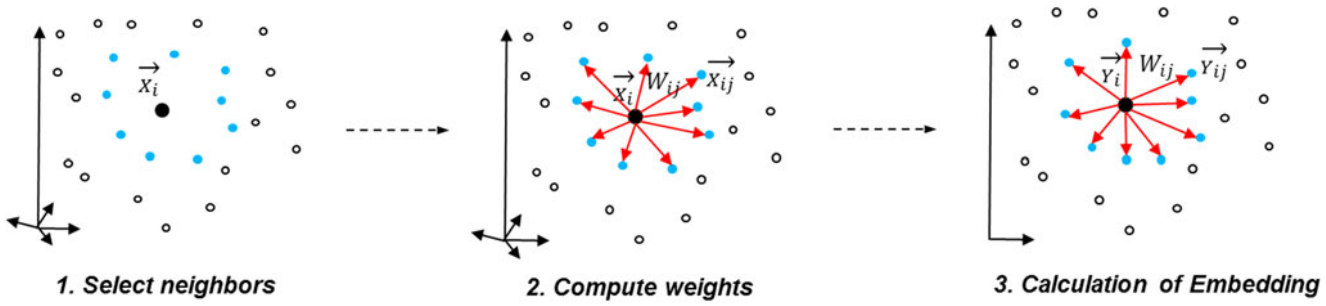


Fig. 1. Unified framework of the LML algorithm.

considerable attention has been paid to feature fusion and classifier design on manifold-based hyperspectral data processing; however, studies on manifold representation are still lacking. Consequently, the classification accuracy can be limited by bottlenecks in manifold learning, where a breakthrough in the level of the classifier is hardly made. To this end, a better manifold representation can break the stalemate.

In general, LML can be regarded as local graph embedding [20], while the most important part of the graph-embedding framework is the calculation of affinities (or similarities) of vertex pairs in a graph, i.e., the affinity matrix. The construction of the affinity matrix comprises two steps: neighbor selection (NS) and computation of affinity weights. There is a challenge in each step: 1) The NS is sensitive to the complex spectral variability due to environmental conditions (e.g., illumination and atmospheric conditions) and instrumental configurations (e.g., sensor noise) as well as data inherent structure (e.g., data distribution); 2) the computation of affinity weights is challenging due to highly correlated spectral signatures in the neighborhood. The latter issue is called *multicollinearity* when multiple regression analysis is used to obtain affinity weights. More specifically, multicollinearity refers to a phenomenon where multiple explanatory variables (spectral signatures in our case) are highly correlated in a linear regression model. This phenomenon in LML easily results in an inaccurate estimation of the affinity matrix.

To tackle these challenges, it is important to develop a robust and effective local manifold representation approach. In this paper, we mainly focus on improving LLE, which is one of the benchmark LML methods in many fields. A novel LML methodology on the basis of LLE is proposed, which aims at learning a robust local manifold representation (RLMR). Two main contributions of this paper are as follows: First, the hierarchical NS (HNS), which comprises joint normalization (JN) and refined NS (RNS), has been embedded into the original LLE framework to robustly select neighbors and mitigate multicollinearity in calculating affinity weights at the same time; Second, inspired by successful applications of spatial information in the hyperspectral classification, we model the spatial information into the proposed DR methodology in order to further improve the robustness of affinity calculations.

The remainder of this paper is described as follows: In Section II, we begin with a brief review of LML with three representative LML methods and provide comparative analysis. Section III introduces our methodology. Experimental results on classification are presented in Section IV. Finally, we provide conclusions and future outlook in Section V.

II. Local Manifold Learning

In this section, three representative LML methods, i.e., LE, LLE, and LTSA, are introduced in the graph-embedding framework, focusing on their advantages and disadvantages.

Generally, LML methods attempt to capture the underlying local manifold structure of the original data and preserve it in a low-dimensional space, which enables nonlinear DR. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ denotes N data samples that have D-dimensional features and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^{d\times N}$ denotes their low-dimensional representations, where $d \ll D$. LML comprised mainly three steps:

1) neighbor selection;
2) computation of affinity weights; and
3) calculation of embedding.

The above-mentioned steps are illustrated in Fig. 1. Pairwise similarity measurements are performed to selected k neighbors for each data sample. Euclidean distance is commonly used for similarity measurement. Let $\mathbf{W} \in \mathbb{R}^{N\times N}$ be a sparse affinity matrix with the (i, j)th entry of the matrix representing the affinity weight from the ith sample and jth sample, where $j \in \phi_i$ and ϕ_i is a set of neighbors of the ith sample. The calculation of embedding coordinates is generally formulated as [20]

$$\begin{aligned}\hat{\mathbf{Y}} &= \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \sum_{j\in\phi_i} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \mathbf{W}_{ij} \right\}, \text{s.t. } \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \\ &= \arg\min_{\mathbf{Y}} \left\{ \mathrm{tr}\left(\mathbf{Y}\mathbf{L}\mathbf{Y}^{\mathrm{T}}\right) \right\}, \text{ s.t. } \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \end{aligned} \quad (1)$$

where $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the Laplacian matrix defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and $\mathbf{D}$ is a diagonal matrix defined by $\forall i\ \mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. $\mathbf{B}$ is a constant matrix defined by the formulation of each manifold learning method. LML methods can be mainly characterized by the construction of the affinity matrix $\mathbf{W}$, as described below.

In the following, three popular LML methods—namely LE, LLE, and LTSA—are introduced in details according to the aforementioned unified framework of the LML algorithm.

LE: The basic principle is to compute the affinity matrix for each data point in the original high-dimensional space using the

Gaussian function as [9]

$$\mathbf{W}_{ij}^{\mathrm{LE}} = \begin{cases} \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 \Big/ 2\sigma^2\right) & \text{if } j \in \phi_i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The constant matrix $\mathbf{B}$ is defined as $\mathbf{B} = \mathbf{D}$. The low-dimensional representations can be obtained by solving the optimization equation (1).

LE is a very typical graph-based embedding method, which has been proven in [9] to be simple to implement and robust against outliers and noise. However, its limitation is also obvious [21], namely a local manifold structure is artificially designed by exploiting approximately pairwise distances with heat kernel, which brings relatively weak representation of local manifold without considering the property of local neighbors.

LLE: It represents the underlying local manifold structure by exploiting the local symmetries of linear reconstructions [5] between each data point and its neighbors in the high-dimensional space and then computes the low-dimensional embedding coordinates that preserve the reconstruction coefficients. The reconstruction coefficients, denoted as $\mathbf{A} \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}}$, are obtained by the minimization

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{j\in\phi_i} \mathbf{A}_{ij}\mathbf{x}_j \right\|_2^2 \right\}$$
$$\text{s.t.} \sum_{j\in\phi_i} \mathbf{A}_{ij} = 1 \tag{3}$$

where $\mathbf{A}_{ij}$ denotes the reconstruction weight between $\mathbf{x}_i$ and $\mathbf{x}_j$, if the *j*th data point is not one of the k neighbors of the *i*th data point ($j \in \phi_j$); otherwise $\mathbf{A}_{ij} = 0$. The reconstruction weights obey an important symmetry of being invariant to rotations, rescalings, and translations of any target data point and its neighbors [5]. The low-dimensional coordinates are obtained by minimizing the embedding cost function as

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \left\| \mathbf{y}_i - \sum_{j\in\phi_i} \mathbf{A}_{ij}\mathbf{y}_j \right\|_2^2 \right\}$$
$$\text{s.t.} \sum_{i=1}^{N} \mathbf{y}_i = 0, \quad \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i\mathbf{y}_i^{\mathrm{T}} = \mathbf{I}. \tag{4}$$

From the viewpoint of the graph-embedding framework, LLE can also be induced as the graph-embedding problem; therefore, (4) can be rewritten in the form of (1) as

$$\begin{aligned} \hat{\mathbf{Y}} &= \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \left\| \mathbf{y}_i - \sum_{j\in\phi_i} \mathbf{A}_{ij}\mathbf{y}_j \right\|_2^2 \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \\ &= \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \sum_{j\in\phi_i} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2 \mathbf{W}_{ij}^{\mathrm{LLE}} \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \\ &= \arg\min_{\mathbf{Y}} \left\{ \mathrm{tr}\left(\mathbf{Y}\mathbf{L}^{\mathrm{LLE}}\mathbf{Y}^{\mathrm{T}}\right) \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \end{aligned} \tag{5}$$

where the affinity matrix ($\mathbf{W}^{\mathrm{LLE}}$) can be computed by the following equation [20]:

$$\mathbf{W}_{ij}^{\mathrm{LLE}} = \begin{cases} \mathbf{A}_{ij} + \mathbf{A}_{ji} - \mathbf{A}_{ij}\mathbf{A}_{ji} & \text{if } j \in \phi_i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and Laplacian matrix of LLE can be given by $\mathbf{L}^{\mathrm{LLE}} = \mathbf{D} - \mathbf{W}^{\mathrm{LLE}} = (\mathbf{I} - \mathbf{A})^{\mathrm{T}}(\mathbf{I} - \mathbf{A})$ [5].$\mathbf{B}$ is defined as $\mathbf{B} = \mathbf{I}$.

With a local regression technique [22], the property of local data is fully taken into consideration in LLE, which means that a local manifold structure can be effectively learned from local data. It is natural that it is able to improve the representation ability of the local manifold. That is not to say, however, that the RLMR can be obtained using LLE, since LLE is very sensitive to data distribution [23], variability [24], as well as multicollinearity.

LTSA: Similar to LLE, LTSA attempts to mine the underlying local manifold structure assuming local linearity. The core idea of LTSA is to utilize a local tangent space to represent a local manifold structure via a linear mapping, such as PCA. Therefore, it can be solved naturally as a graph-embedding problem, and the affinity matrix can be defined as $\mathbf{W}^{\mathrm{LTSA}} = \mathbf{D} - \mathbf{L}^{\mathrm{LTSA}}$, more specifically formulated as follows [14]:

$$\mathbf{W}_{ij}^{\mathrm{LTSA}} = \begin{cases} \frac{1}{k} + \frac{1}{k-1}\theta_i^{\mathrm{T}}\mathbf{\Lambda}^{-1}\theta_j & \text{if } j \in \phi_i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where θ_i and θ_j are the local tangent coordinates of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively, and $\mathbf{\Lambda}$ stands for the leading d eigenvalues of the covariance matrix of ϕ_i, and k is the number of neighbors for $\mathbf{x}_i$. The low-dimensional embedding is calculated by the following minimization:

$$\begin{aligned} \hat{\mathbf{Y}} &= \arg\min_{\mathbf{Y}} \left\{ \mathrm{tr}\left(\mathbf{Y}\mathbf{L}^{\mathrm{LTSA}}\mathbf{Y}^{\mathrm{T}}\right) \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \\ &= \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \sum_{j\in\phi_i} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2 \mathbf{W}_{ij}^{\mathrm{LTSA}} \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \\ &= \arg\min_{\mathbf{Y}} \left\{ \sum_{i=1}^{N} \left\|\mathbf{y}_i\mathbf{H} - \mathbf{T}_i\theta_i\right\|_2^2 \right\}, \text{s.t.} \mathbf{Y}\mathbf{B}\mathbf{Y}^{\mathrm{T}} = \mathbf{I} \end{aligned} \tag{8}$$

where $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^{\mathrm{T}}/k$ is the centering matrix, and $\mathbf{e}$ is a uniform vector with the size of $k \times 1$. $\mathbf{T}_i$ is a local transformation matrix with linearity, and $\mathbf{B}$ is defined as $\mathbf{B} = \mathbf{I}$.

Typically, a concept of local tangent space is proposed in LTSA to linearly and approximately estimate the local manifold structure, which is able to better capture the intrinsic structure of the underlying manifold [10]. However, such approximated estimation of the local manifold structure is possibly inaccurate, particularly in nonuniform distributed data [25], due to those data in the local manifold space without lying in, or closing to, a linear subspace. Also, although the performance of LTSA can improve the local manifold representation compared to LLE to some extent, it still fails when taking the data variability (e.g., noise) into consideration [26]. Furthermore, unlike LLE, LTSA explores a linear mapping (e.g., PCA) to find the principle information to depict the local manifold structure, accordingly

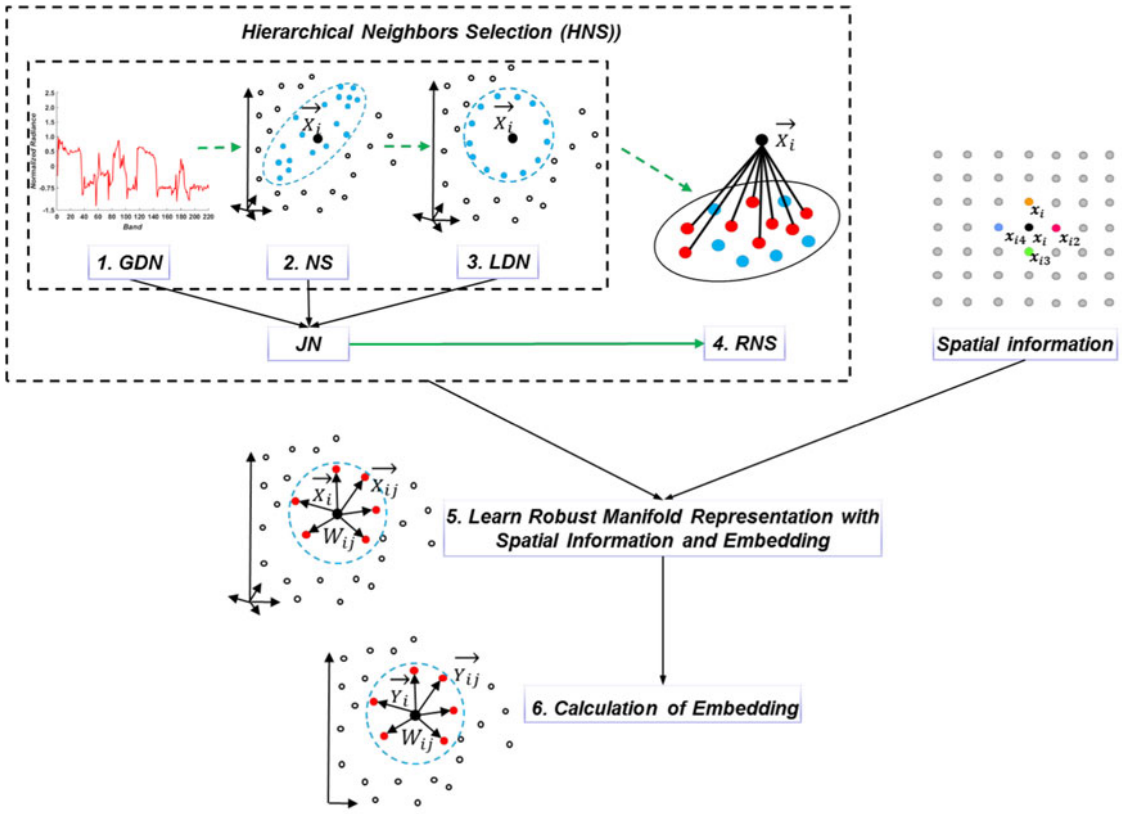


Fig. 2. Holistic diagram of the proposed method.

resulting in inevitable loss of useful information (e.g., geometric structure and local minutiae).

In summary, among the three LML methods, one advantage of LLE and LTSA over LE is that by using LLE or LTSA we can obtain a potentially better performance in DR due to their reasonably linear representation in the local manifold space. But the drawback of LLE and LTSA is that neither is highly robust against complex data variability, e.g., caused by noise, illumination, or nonuniform data distribution. Therefore, how to robustly learn the local manifold representation is an unsolved problem in LML. As a promising LML framework, LLE has been successfully applied in many fields and has obtained some amazing experimental results due to effectively and reasonably local linear assumption, for example, in hyperspectral data processing [3], [13], [14], [16], [17], [22]. However, sensitivity to variability and multicollinearity when calculating the local linear representation are hindering the advancement of LLE toward robustness and high performance. Therefore, in the next section, we emphatically introduce the proposed novel methodology based on LLE in an attempt to address the two issues mentioned above.

III. Robust Local Manifold Representation

In this section, a novel LML methodology is introduced in detail in order to learn an RLMR, mainly including the design of HNS and the integration of spatial contextual information. Fig. 2 shows the holistic diagram of the proposed methodology that mainly comprises the six steps given below, where the first four correspond to HNS and the fifth is the integration of spatial information.

Step 1. *Global data normalization (GDN)* is performed to deal with the spectral variability modeled by scaling and shifting.

Step 2. *NS* coarsely selects local neighbors of the target pixel.

Step 3. *Local data normalization (LDN)* is applied to make local data distribution more uniform and isotropic and further eliminate locally spectral variability.

Step 4. *RNS* aims at mitigating multicollinearity in the local manifold space, making it possible to obtain a relatively accurate and intrinsic structure of underlying manifold.

Step 5. *Computation of reconstruction weights with contextual information* jointly embeds spectral and spatial information for a robust calculation of the reconstruction weights.

Step 6. *Calculation of embedding* obtains the low-dimensional feature representation by embedding robust local manifold properties into the low-dimensional space.

A. Hierarchical Neighbors Selection

Fig. 3 shows the detailed diagram of HNS, which is composed of JN and RNS.

1) Joint Normalization: Data normalization is widely used in data preprocessing procedure, including hyperspectral data analysis [27], [28]. It aims at reducing the effect of numerous variations and improving the performance of subsequent algorithms. Generally, data normalization includes GDN and

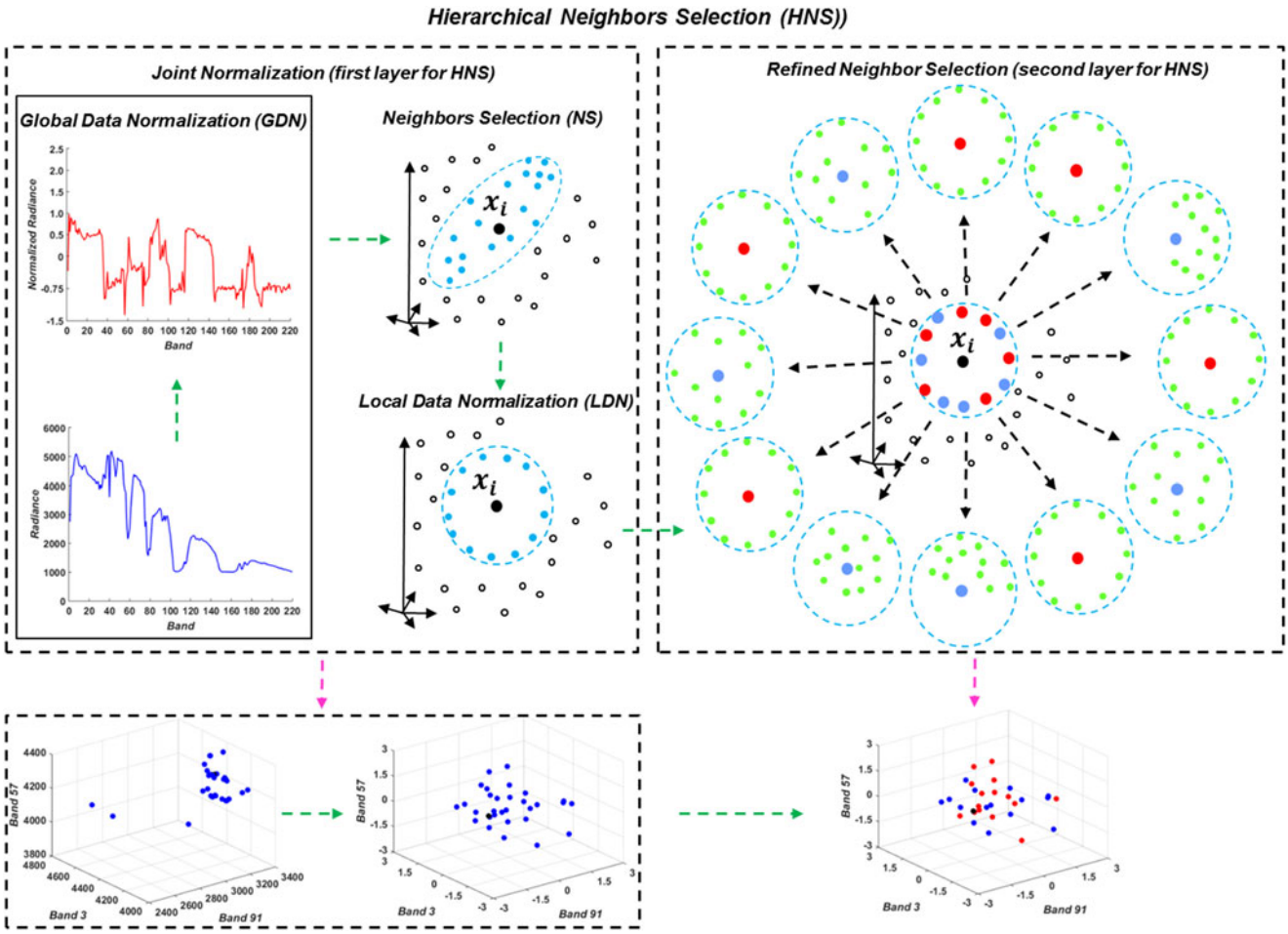


Fig. 3. Detailed diagram of HNS.

LDN [29]. The purpose of GDN is to mitigate illumination variations and modify the global data distribution so that it is more uniform and isotropic [30], [31], enabling them to be measured in the same, or similar, level or unit. Therefore, GND should be performed on the whole hyperspectral image. Unlike GDN, LDN tends to uniformize the mean and variance of the local neighborhood, which is particularly useful for nonuniform distributed data [32], [33]. Owing to the merits of GDN and LDN, JN is an appropriate approach to effectively address the issues of spectral variability and nonuniform data distribution, which can be implemented step-by-step via the following formulations

1) Global data normalization: It performs the following computations:

$$\mathbf{x}_i^{ns} = \frac{\mathbf{x}_i^o - c_i^o}{s_i^o} \tag{9}$$

$$\mathbf{x}_i^g = (\mathbf{x}_i^{ns} - \mathbf{c}^{ns})./\mathbf{s}^{ns} \tag{10}$$

where "./" means the elementwise division, $\mathbf{x}_i^o \in \mathbb{R}^{D\times 1}$ is the ith original spectral signature, and c_i^o and s_i^o are the mean value and variance corresponding to $\mathbf{x}_i^o$, respectively. $\mathbf{x}_i^{ns} \in \mathbb{R}^{D\times 1}$ stands for the normalized spectral signature. $\mathbf{X}^{ns} \in \mathbb{R}^{D\times N}$ represents all normalized spectral signatures made up of $\mathbf{x}_i^{ns}$, and $\mathbf{c}^{ns} \in \mathbb{R}^{D\times 1}$ and $\mathbf{s}^{ns} \in \mathbb{R}^{D\times 1}$ correspond to the mean value and variance of $\mathbf{X}^{ns}$, respectively. $\mathbf{x}_i^g \in \mathbb{R}^{D\times 1}$ stands for the normalized spectral signature of GDN. The normalization obtained by performing (9) can mitigate the effects of spectral variability that can be explained by scaling and shifting, whereas (10) makes the global data distribution more uniform and isotropic and puts the same weight on all the spectral bands, as shown in Fig. 3(Top-left).

2) Local data normalization: After selecting coarse neighbors for each data point using the Euclidean distance, LDN is exploited to make data distribution more uniform and isotropic in the local manifold space, which can be formulated as

$$\mathbf{x}_{ij}^l = \begin{cases} (\mathbf{x}_i^g - \mathbf{c}_i^g)./\mathbf{s}_i^g & j = 0 \\ (\mathbf{x}_{ij}^g - \mathbf{c}_i^g)./\mathbf{s}_i^g & j = 1, 2, \ldots, K \end{cases} \tag{11}$$

where "./" means the elementwise division, $\mathbf{X}_i^g = [\mathbf{x}_i^g, \mathbf{x}_{i1}^g, ..., \mathbf{x}_{ij}^g, ..., \mathbf{x}_{iK}^g] \in \mathbb{R}^{D\times(K+1)}$ consists of the globally normalized spectral features of ith data point and its K neighbors. $\mathbf{c}_i^g \in \mathbb{R}^{D\times 1}$ and $\mathbf{s}_i^g \in \mathbb{R}^{D\times 1}$ represent the mean value and variance of $\mathbf{X}_i^g$, respectively. $\mathbf{X}_i^l = [\mathbf{x}_i^l, \mathbf{x}_{i1}^l, ..., \mathbf{x}_{ij}^l, ..., \mathbf{x}_{iK}^l] \in \mathbb{R}^{D\times(K+1)}$ represents the final normalized spectral features for ith data point and its neighbors by JN. An example of local data distribution is shown in Fig. 3(Bottom-left). We can see that the data distribution becomes more uniform and isotropic

by means of LDN reducing the effects of nonuniform data distribution.

2) Refined Neighbor Selection: After JN, we obtain the rough results of NS where the influence of spectral variability has been mitigated, but multicollinearity still exists among neighbors. Multicollinearity leads to an inaccurate estimation of the affinity matrix, thereby degrade the quality of the local manifold structure. To address this issue, RNS is performed as the second layer of HNS. RNS, which is inspired by the local manifold alignment, is proposed to reduce the information redundancy [34] in the coarse neighborhood, as illustrated in Fig. 3(Right). RNS can mitigate the effects of multicollinearity in the next step, i.e., the calculation of reconstruction weights, while preserving local manifold properties. In detail, LFS is divided into two parts.

First, inspired by [35] and [36], we construct the local structure feature $\mathbf{F}_p^{\text{local}}$ for the data point p in the feature space using its neighbor's information $\mathbf{X}_p^l = [\mathbf{x}_{p1}^l, ..., \mathbf{x}_{pj}^l, ..., \mathbf{x}_{pK}^l] \in \mathbb{R}^{D\times K}$. $\mathbf{F}_p^{\text{local}}$ can be formed by the distance property between the feature of p with those of its neighbors using a Gaussian function:

$$F_{pj}^{\text{local}} = \exp\left(-\left\|\mathbf{x}_p^l - \mathbf{x}_{pj}^l\right\|_2^2\right) \tag{12}$$

$$\mathbf{F}_p^{\text{local}} = \left[F_{p1}^{\text{local}}, \ldots, F_{pj}^{\text{local}}, \ldots, F_{pK}^{\text{local}}\right]. \tag{13}$$

The second part is to screen out new local neighbors that have similar data distribution using the Kullback–Leibler divergence (KLD). The KLD has been justified to effectively measure the similarity of hyperspectral data distribution [37]. The difference of local features $\mathbf{d}^f = [d_1^f, ..., d_q^f, ..., d_K^f] \in \mathbb{R}^{1\times K}$ between the point p and its neighbor q can be measured as

$$d_q^f = \text{KLD}\left(\mathbf{F}_p^{local}||\mathbf{F}_q^{\text{local}}\right) + \alpha\text{KLD}\left(\mathbf{F}_q^{\text{local}}||\mathbf{F}_p^{\text{local}}\right) \tag{14}$$

$$\text{KLD}\left(\mathbf{F}_p^{\text{local}}||\mathbf{F}_q^{\text{local}}\right) = \sum_{j=1}^{K} F_{pj}^{\text{local}} \times \log_2\left(\frac{F_{pj}^{\text{local}}}{F_{qj}^{\text{local}}}\right) \tag{15}$$

$$\text{KLD}\left(\mathbf{F}_q^{\text{local}}||\mathbf{F}_p^{\text{local}}\right) = \sum_{j=1}^{K} F_{qj}^{\text{local}} \times \log_2\left(\frac{F_{qj}^{\text{local}}}{F_{pj}^{\text{local}}}\right) \tag{16}$$

where $\mathbf{F}_p^{\text{local}} \in \mathbb{R}^{1\times K}$ and $\mathbf{F}_q^{\text{local}} \in \mathbb{R}^{1\times K}$ stand for the local structure features of p and q in the spectral domain, respectively, and α is a penalty parameter balancing the two terms described in (15) and (16). Neighbors with the k smallest $\mathbf{d}^f$ value are chosen from the coarse neighbors as the new neighbors of the data point p, namely $\mathbf{X}_p^{nl} = [\mathbf{x}_{p1}^{nl}, ..., \mathbf{x}_{pj}^{nl}, ..., \mathbf{x}_{pk}^{nl}] \in \mathbb{R}^{D\times k}$. k is the final number of neighbors for each point, and we make the value of K equal to twofold k.

An example showing the effect of RNS is given in Fig. 4, where correlations between the target pixel and its neighbors are shown with and without using RNS. To be specific, given any target pixel, k neighbors were selected without RNS, whereas for RNS, $2k$ were selected at first and then k neighbors are refined from $2k$ neighbors. Therefore, the same number of neighbors k was obtained without RNS and with RNS. Fig. 4(Left) shows spectral signatures of neighbors from two different strategies (without RNS and with RNS). Although it is not so obvious, it still

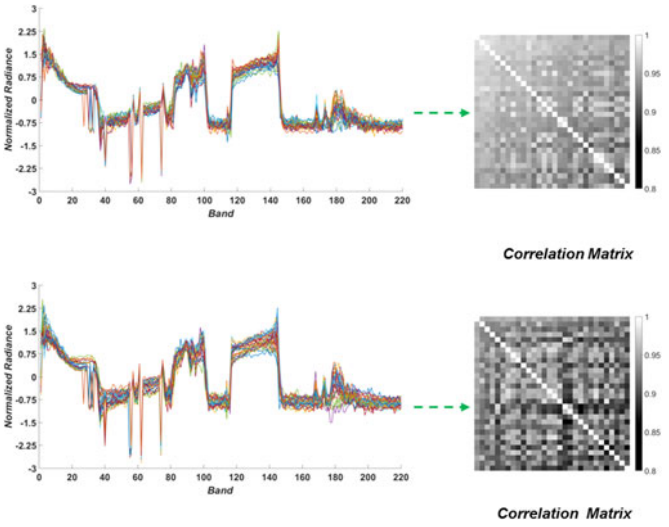


Fig. 4. (Left) Spectral signatures of local neighbors for an exemplar data point and (right) their correlations (top) without RNS and (bottom) with RNS.

emerges the slight difference that spectral signatures without RNS are more intensive than those with RNS, which means that those without RNS are likely to generate multicollinearity when computing the affine matrix (weight matrix). Fig. 4(Right) shows relatively obvious results regarding the reduction of multicollinearity. We can see that the values of correlation matrix with RNS are lower than those without RNS, which demonstrates that the linear correlations observed in the correlation matrix are effectively reduced after using RNS.

B. Local Manifold Representation With Spatial Contextual Information

To further improve the robustness of the calculation of reconstruction weights, the spatial information is incorporated into linear reconstructions. We assume that spatially neighboring spectral pixels can be explained by the same or similar reconstruction weights [38], if spatially neighboring pixels include similar spectral components. The calculation of reconstruction weights with spatial contextual information can be formulated based on (1) by adding the constraint that the reconstruction weights of the target pixel are approximately equal to the average of those of its neighboring pixels, as shown in the following:

$$\begin{aligned}\mathbf{a}_i^0 = &\arg\min_{\mathbf{w}_i^0}\left\{\sum_{s=0}^{4}\left\|\mathbf{x}_{is}^{nl} - \mathbf{X}_i^{nl}\mathbf{a}_i^s\right\|_2^2\right\},\\ &\text{s.t.}\left\|\mathbf{X}_i^{nl}\left(4\mathbf{a}_i^0 - \sum_{s=1}^{4}\mathbf{a}_i^s\right)\right\|_2^2 \le \eta,\ (\mathbf{a}_i^s)^{\mathrm{T}}\mathbf{a}_i^s = 1,\\ &s = 0, 1, \ldots, 4\end{aligned} \tag{17}$$

where $\mathbf{X}_i^{nl} = [\mathbf{x}_{i1}^{nl}, ..., \mathbf{x}_{ij}^{nl}, ..., \mathbf{x}_{ik}^{nl}] \in \mathbb{R}^{D\times k}$ is the k-nearest neighbors selected by HNS. $\mathbf{x}_{is}^{nl}$, $s = 0, 1, ..., 4$ are the target spectral pixel and its four spatial neighbors, respectively, as an example shown in Fig. 5. Correspondingly, $\mathbf{a}_i^s \in \mathbb{R}^{k\times 1}$, $s = 0, 1, ..., 4$ are their reconstruction weights. η is a tiny real number

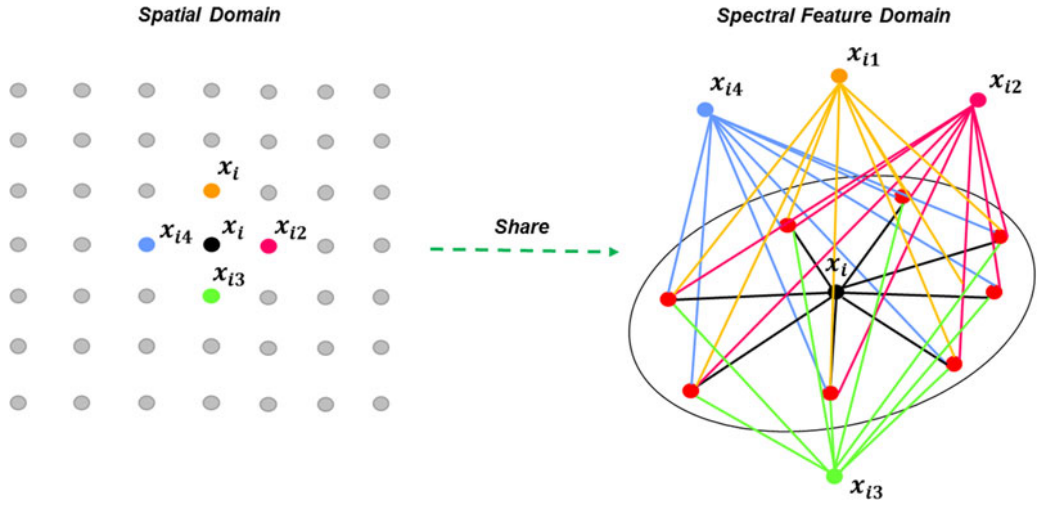


Fig. 5. Diagram for spatial–spectral combination in hyperspectral DR.

(here $\eta = 10^{-3}$) that represents the limit of error. Note that LDN should be conducted on this dataset composed of target spectral pixel and its spatial and spectral neighbors before calculating reconstruction weights.

We can regard (17) as a joint optimization problem. In this case, the objective function of (17) can be rewritten as

$$\mathbf{a}_i^0 = \arg\min_{\mathbf{a}_i^0} \left\{ \left\| \hat{\mathbf{X}}_i^{nl} - \mathbf{L}\hat{\mathbf{A}}_i \right\|_F^2 \right\}, \text{s.t.} \mathbf{C}\hat{\mathbf{A}}_i = [1\ 1\ 1\ 1\ 1]^{\mathrm{T}}$$

$$\mathbf{L} = \begin{bmatrix} 4\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} & -\beta\mathbf{X}_i^{nl} \\ \mathbf{X}_i^{nl} & & & & \\ & \mathbf{X}_i^{nl} & & & \\ & & \mathbf{X}_i^{nl} & & \\ & & & \mathbf{X}_i^{nl} & \\ & & & & \mathbf{X}_i^{nl} \end{bmatrix},$$

$$\hat{\mathbf{A}}_i = \begin{bmatrix} \mathbf{a}_i^0 \\ \mathbf{a}_i^1 \\ \mathbf{a}_i^2 \\ \mathbf{a}_i^3 \\ \mathbf{a}_i^4 \end{bmatrix}, \quad \hat{\mathbf{X}}_i^{nl} = \begin{bmatrix} \mathbf{0} \\ \mathbf{x}_{i0}^{nl} \\ \mathbf{x}_{i1}^{nl} \\ \mathbf{x}_{i2}^{nl} \\ \mathbf{x}_{i3}^{nl} \\ \mathbf{x}_{i4}^{nl} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{e} & & & & \\ & \mathbf{e} & & & \\ & & \mathbf{e} & & \\ & & & \mathbf{e} & \\ & & & & \mathbf{e} \end{bmatrix} \tag{18}$$

where the sizes of $\mathbf{L}$, $\hat{\mathbf{X}}_i^{nl}$, $\hat{\mathbf{A}}_i$,$\mathbf{C}$ are $6D \times 5k$, $6D \times 1$, $5k \times 1$ and $5 \times 5k$, respectively. And $\mathbf{e} \in \mathbb{R}^{1\times k}$ is the unit vector with a size of $1 \times k$, and β is a penalty parameter to balance the importance between error item and constraint item in (18).

In order to solve (18), it can be further relaxed by means of Lagrange multipliers as represented by

$$\mathbf{a}_i^0 = \arg\min_{\mathbf{a}_i^0} \left\{ \left\| \hat{\mathbf{X}}_i^{nl} - \mathbf{L}\hat{\mathbf{A}}_i \right\|_F^2 + \lambda \left\| \mathbf{C}\hat{\mathbf{A}}_i - \hat{\mathbf{e}} \right\|_2^2 \right\} \tag{19}$$

where λ is also a penalty parameter, and here let it be 1 for simplicity as well as $\hat{\mathbf{e}} = [1\ 1\ 1\ 1\ 1]^{\mathrm{T}} \in \mathbb{R}^{5\times 1}$. The solution in (19) can be analytically derived [39] by matrix derivation operation as

$$\mathbf{a}_i^0 = \left(\mathbf{L}^{\mathrm{T}}\mathbf{L} + \lambda\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\left(\mathbf{L}^{\mathrm{T}}\hat{\mathbf{X}}_i^{nl} + \lambda\mathbf{C}^{\mathrm{T}}\hat{\mathbf{e}}\right). \tag{20}$$

Therefore, $\mathbf{a}_i^0$ is the weight vector for ith pixel by using RLMR. Following the framework shown in Fig. 2, the result of DR can be obtained by calculating the embedding using (1).

IV. Experiment

In this section, we explore the classification as a potential application and quantitatively evaluate the performance of DR algorithms using overall classification accuracy. The main focus of this paper is to learn a more robust and discriminative feature representation, rather than how to develop a more advanced classifier. Therefore, we use two common classifiers, namely the nearest neighbor (NN) algorithm based on the Euclidean distance and linear support vector machines (SVMs).

A. Hyperspectral Datasets

The experiments are carried out using two benchmark hyperspectral datasets.

1) Indian Pines AVIRIS Image: The first dataset was acquired by NASA's AVIRIS sensor over the Indian Pines test site in Northwest Indiana with the size of $145 \times 145 \times 220$ and 10 nm spectral resolutions over the range of 400–2500 nm, mainly including several kinds of vegetation. More specific classes and the number of samples can be found in Table I.
2) 2013 IEEE GRSS Data Fusion Contest (DFC) image: The second dataset was provided for the 2013 IEEE GRSS DFC acquired by the ITRES-CASI 1500 sensor with the size of $349 \times 1905 \times 144$ in the range of 380–1050 nm, which includes more varied categories.

B. Results of Indian Pines AVIRIS Data

For the first dataset, we adopted two sampling strategies to select training samples and test samples: random sampling and region-based sampling. Random sampling is a common way

TABLE I
NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR EACH CLASS

No.	Class Name	Total	Cross Validation	Training	Testing
1	Corn-Notill	1434	50	50	1334
2	Corn-Mintill	834	50	50	734
3	Corn	234	50	50	134
4	Grass-Pasture	497	50	50	397
5	Grass–Trees	747	50	50	647
6	Hay-Windrowed	489	50	50	389
7	Soybean-Notill	968	50	50	868
8	Soybean-Mintill	2468	50	50	2368
9	Soybean-Clean	614	50	50	514
10	Wheat	212	50	50	112
11	Woods	1294	50	50	1194
12	Bldg-Gra-Tr-Driv	380	50	50	280
13	Stone–Stel–Tower	95	15	15	65
14	Alfalfa	54	10	10	34
15	Grass-Past-Mowed	26	5	5	16
16	Oats	20	5	5	10

for the validation of the hyperspectral classification. In contrast, classification using region-based sampling is more practical and challenging due to high correlation and limited variability of training samples, and thus an effective way to investigate the performance of the proposed method. We randomly assigned around 5% of total samples as cross-validation samples and then divided the rest into two parts: training samples (5% of total samples), by random sampling or region-based sampling, and test samples (90% of total samples). Moreover, ten replications were performed for selecting training and test samples based on the two aforementioned sampling strategies. The specific number of cross validation, training, and test samples is listed in Table I [40]. We compare the classification results on dimensionality-reduced data using the proposed method with those using some benchmark DR methods (PCA, KPCA [41], LLE, LE, and LTSA) and original spectral features (OSF). Three step-by-step methods, i.e., JN, HNS, and RLMR, are used for the proposed methods to investigate the effects of JN, LFS, and the integration of spatial information.

1) Performance Comparison and Analysis Between RLMR and Classical DR Methods: Initially, we conducted a fivefold cross validation on training samples in order to select the optimal parameter combination. Table II gives the classification accuracies obtained by using the nine methods with optimal parameters (*d, k*). It should be noted that two kinds of classification accuracy are applied here, including overall accuracy (total classification accuracy of all classes) and average accuracy (the average of the classification accuracy of each class), to evaluate the performance of the listed methods.

The proposed methods outperform the other methods both with random sampling and region-based sampling. Compared to OSF, JN, HNS, and RLMR increase the overall accuracy by 8.25%, 12.71%, and 21.1%, respectively, with random sampling, and 7.42%, 8.83%, and 10.46%, respectively, with region-based sampling. For the average accuracy, on the other hand, the corresponding increases are, respectively, 10.2%, 12.89%, 18.11% with random sampling, and 9.68%, 10.95%, 11.54% with region-based sampling.

The classification maps are shown in Figs. 6 and 7. It can be seen that the classification maps of JN, HNS, and RLMR include less salt-and-pepper errors. In particular, those of RLMR are smoother in the local spatial region, resulting from the embedding of spatial information. These results demonstrate the effectiveness of all three technical components of the RLMR, i.e., JN, RNS, and the integration of spatial information, and imply that they successfully contribute to extracting robust and discriminative low-dimensional feature representations. In contrast, the classification accuracies of the classical LML methods (e.g., LLE, LTSA) are holistically higher than those obtained by using OSF and PCA, and yet lower than the results of our proposed methods due to the sensitivity of variability with respect to LLE and the unavoidable loss of information with respect to LTSA. As for the performance of LE, it is even inferior to the performances of OSF and PCA, and considerably lower than LLE and LTSA, as discussed in Section II. This indicates that the performance of these methods is unstable in DR due to challenges involved in NS and affinity calculations.

To effectively support the conclusion obtained by the NN classifier, an advanced and common classifier—SVM [44] is also applied for classification under the same condition. In this paper, a linear version of SVM is selected for the classifier rather than nonlinear versions to investigate the capability of handing nonlinear structure in the data for all DR methods under comparison. Classification accuracies obtained via SVM and corresponding optimal parameters for nine methods are listed in Table III. Figs. 8 and 9 show classification maps for the different methods using the random sampling and region-based sampling strategies, respectively.

In addition, we can observe from Tables II and III that the performance of JN, HNS, and RLMR is progressively increased, which can be contributed by the used of normalization, RNS, and spatial information, respectively. To investigate the effectiveness of RNS, we compare the performance with RNS and without RNS via the NN classifier, listed in Table IV. We can clearly see that the classification accuracies of those methods with RNS are stably higher than those without RNS while the proposed method JN+RNS (HNS) shows the best performance.

2) Sensitivity Analysis of Parameters and Robustness Against Noise

a) Sensitivity analysis of parameters: The sensitivity of parameters is examined by varying the number of neighbors (k) and the size of reduced dimensionality (d) for LML methods, and the variance (v) of kernel for KPCA. As shown in Figs. 10 and 11, the performance of the LML methods is less sensitive to the parameters. In general, as observed from the data dimensionality point of view, the classification accuracy increases with increasing dimensionality, to a certain extent, and then holds steady. When the reduced dimensionality d reaches approximately 50, the results are basically stable for those ML-based methods, while the number of neighbors k is around 60 when accuracy reaches the nearly optimum level. As the number of neighbors gradually increases, the corresponding classification accuracy progressively increases to a peak (e.g., k is equal to around 50) and then dramatically drops. A large number of neighbors may obscure the local structure, whereas a small number of

TABLE II
CLASSIFICATION ACCURACIES USING OPTIMAL PARAMETERS VIA NN FOR DIFFERENT DR METHODS IN INDIAN PINE DATASET

Method	Optimal Parameters	Classification Accuracy			
		Random Sampling		Region-based sampling	
		Overall Accuracy	Average Accuracy	Overall Accuracy	Average Accuracy
OSF	/	64.74%	72.72%	44.78%	56.67%
PCA	$d = 50$	64.62%	72.66%	44.74%	56.64%
KPCA	$d = 50, v = 10$	66.95%	76.03%	48.79%	61.25%
LLE	$d = 60, k = 40$	68.49%	75.51%	47.45%	59.55%
LE	$d = 60, k = 7$	59.57%	68.19%	40.92%	52.73%
LTSA	$d = 60, k = 70$	71.22%	81.12%	51.63%	66.09%
JN	$d = 70, k = 40$	72.99%	82.92%	52.20%	66.35%
HNS	$d = 70, k = 40$	77.45%	85.61%	53.61%	67.62%
RLMR	$d = 50, k = 80$	**85.84%**	**90.83%**	**55.24%**	**68.21%**

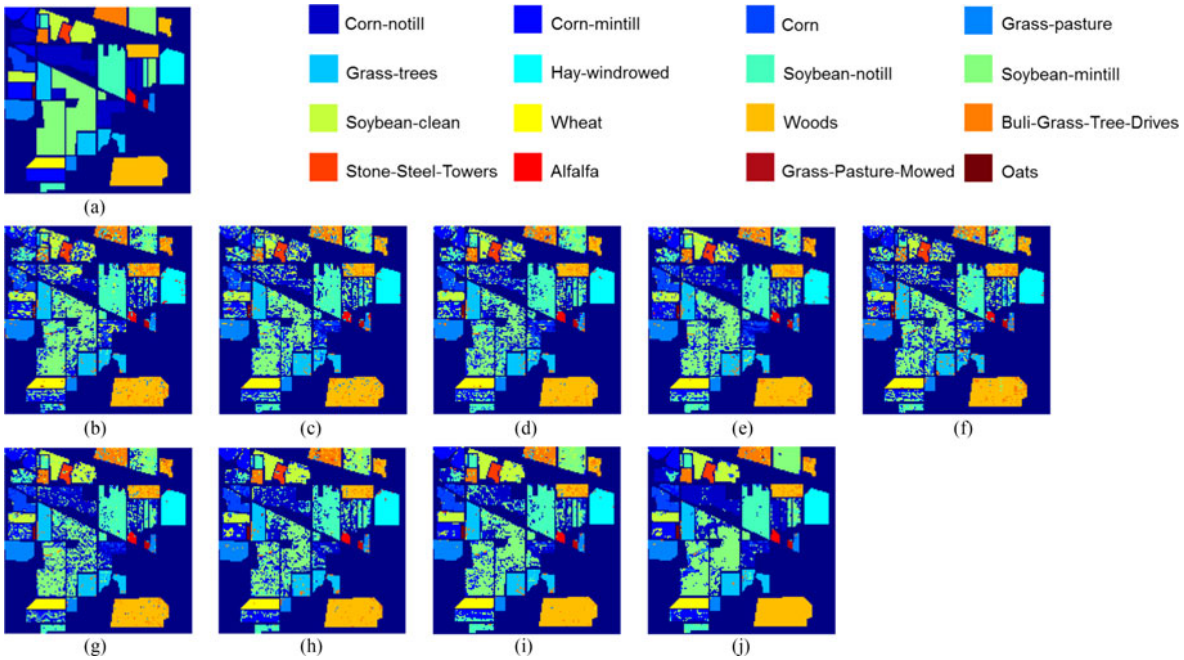


Fig. 6. NN classification maps for the Indian Pines dataset using all DR methods under comparison with the optimal parameters in Table II based on random sampling. (a) Ground truth and (b)–(j) results for OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

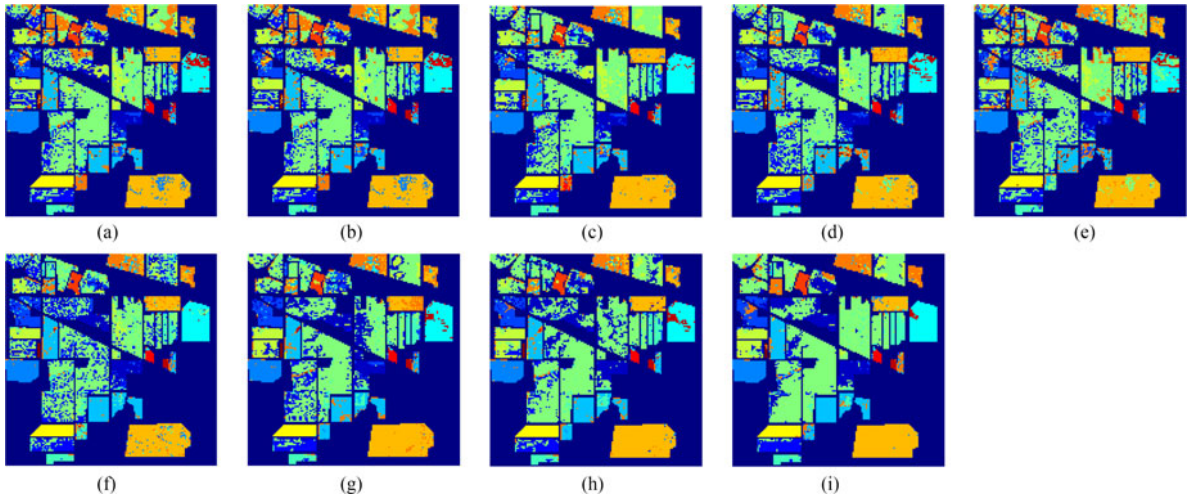


Fig. 7. NN classification maps for the Indian Pines dataset using all DR methods under comparison with the optimal parameters in Table II based on region-based sampling. (a)–(i) Results for OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

TABLE III
CLASSIFICATION ACCURACIES USING OPTIMAL PARAMETERS VIA SVM FOR DIFFERENT DR METHODS IN INDIAN PINES DATASET

Method	Optimal Parameters	Classification Accuracy			
		Random Sampling		Region-Based Sampling	
		Overall Accuracy	Average Accuracy	Overall Accuracy	Average Accuracy
OSF	/	73.86%	76.04%	47.39%	61.87%
PCA	$d = 30$	70.60%	79.50%	47.82%	58.38%
KPCA	$d = 60$, $v = 10$	72.16%	80.88%	50.36%	63.52%
LLE	$d = 40$, $k = 50$	71.47%	72.51%	47.23%	62.49%
LE	$d = 80$, $k = 3$	56.93%	65.06%	36.59%	52.85%
LTSA	$d = 40$, $k = 70$	75.49%	84.93%	52.79%	64.51%
JN	$d = 90$, $k = 60$	76.52%	83.03%	52.83%	66.95%
HNS	$d = 100$, $k = 50$	78.75%	85.04%	54.73%	68.03%
RLMR	$d = 40$, $k = 90$	**87.06%**	**90.93%**	**56.92%**	**69.24%**

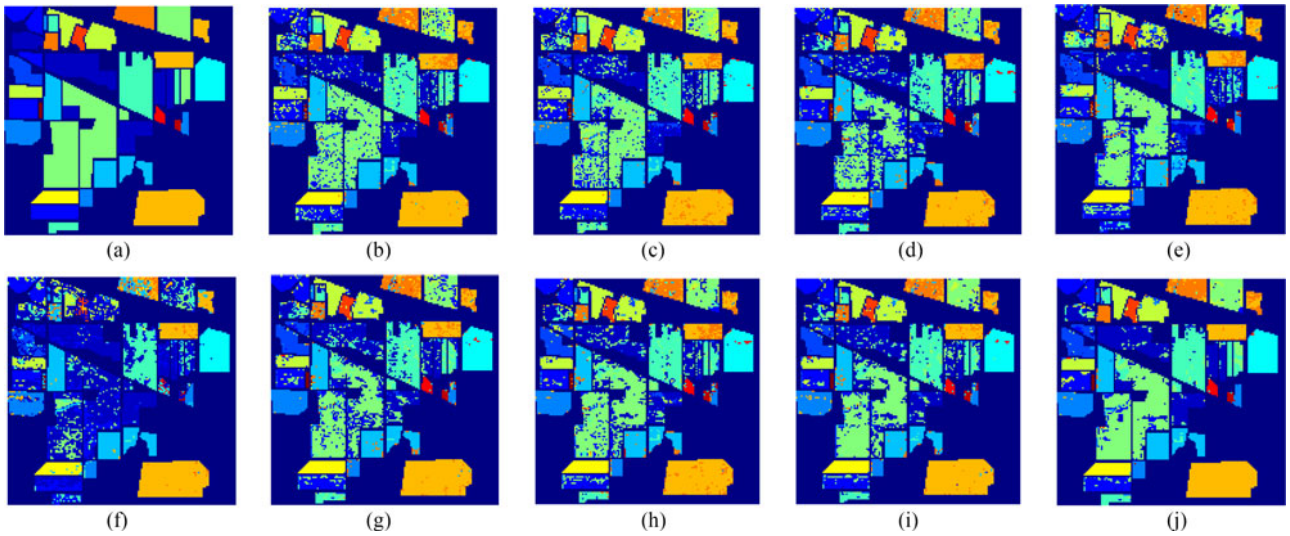


Fig. 8. SVM classification maps for the Indian Pines dataset using all DR methods under comparison with the optimal parameters in Table III based on random sampling. (a) Ground truth and (b)–(j) results for OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

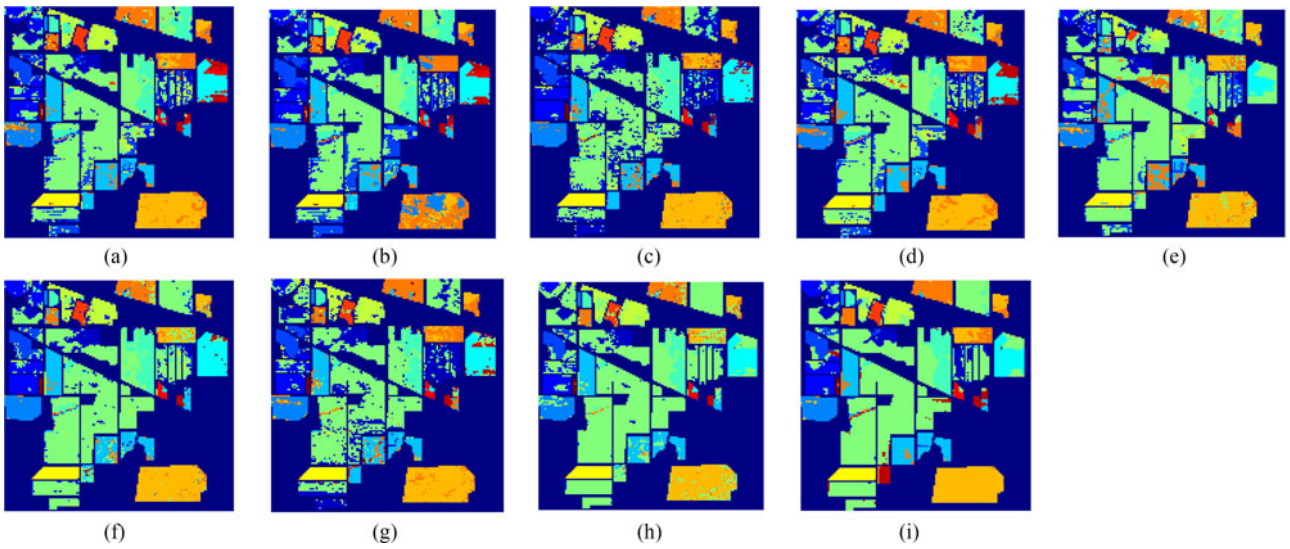


Fig. 9. SVM classification maps for the Indian Pines dataset using all DR methods under comparison with the optimal parameters in Table III based on region-based sampling. (a)–(i) Results for OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

TABLE IV
CLASSIFICATION ACCURACIES OBTAINED WITH NN CLASSIFIER FOR THE INDIAN PINES DATASET USING LLE WITH DIFFERENT NS METHODS

NS Method	Optimal Parameters	Classification Accuracy	
		Random Sampling	Region-Based Sampling
Euclidean	$d = 60, k = 40$	68.49%	47.45%
Euclidean+RNS	$d = 90, k = 50$	70.24%	48.85%
SAM	$d = 60, k = 80$	70.85%	48.97%
SAM+RNS	$d = 70, k = 50$	72.67%	49.50%
JN	$d = 70, k = 40$	72.99%	52.20%
JN+RNS (HNS)	$d = 70, k = 40$	**77.45%**	**53.61%**

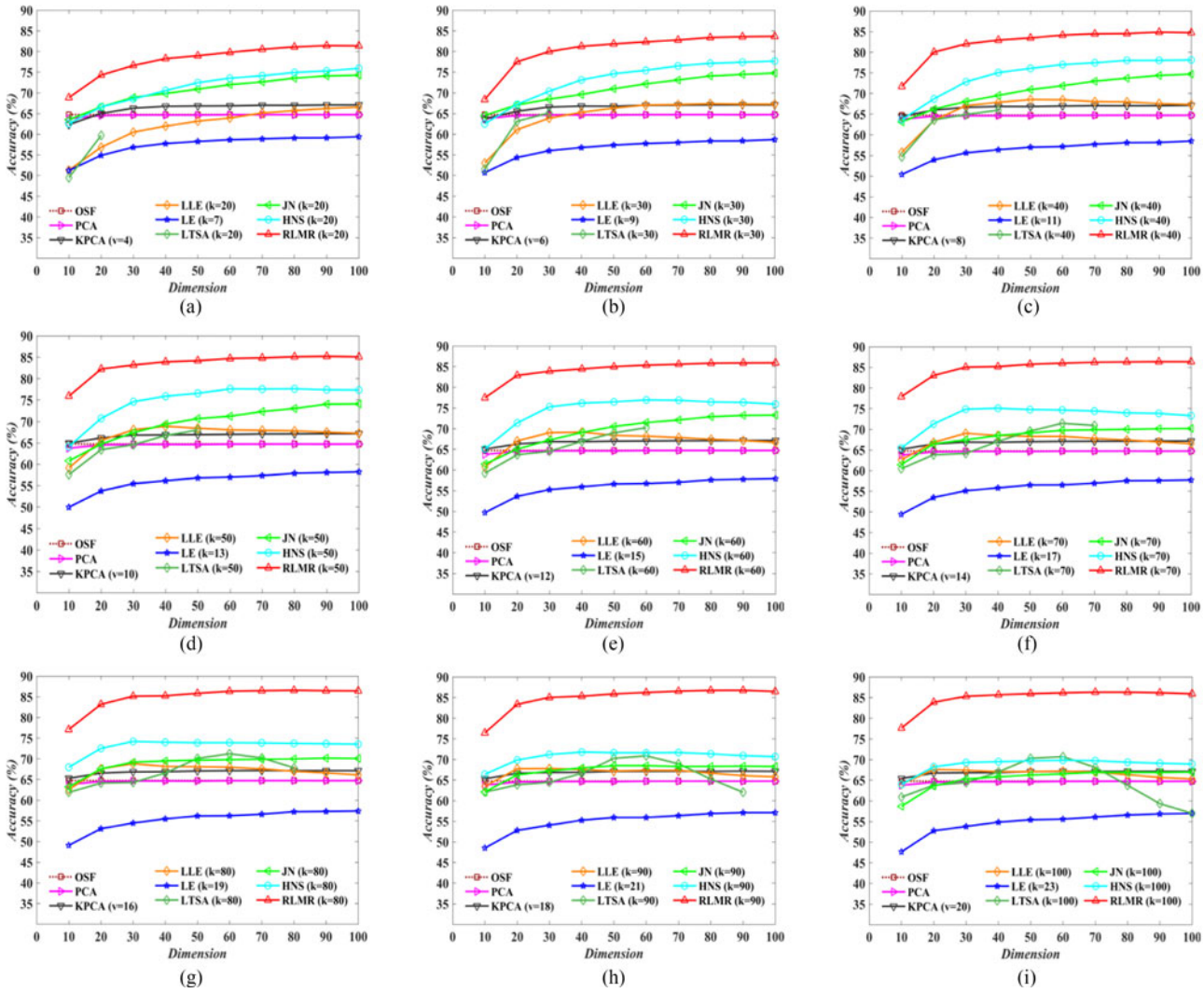


Fig. 10. Performance comparison: Classification accuracy as a function of data dimension using random sampling for the Indian Pines dataset. (a)–(i) Results using different numbers of neighbors, respectively.

neighbors may not sufficiently represent the local structure, causing the degradation of the DR performance. Proper parameters are determined from Figs. 10 and 11, which are basically consistent with parameter selection defined via cross validation given in Table II, where the LML methods are used for classification. However, it is worth noting that due to robustness of our proposed method (RLMR), its results remain stable with the increase in the number of neighbors k and reduced dimensionality d. Conversely, the performances of JN and HNS are progressively degrading with the change of parameters; particularly in a situation with a large k, the classification accuracies even degrade to a level similar to classical LML methods.

Unlike manifold learning methods, the size of reduced dimensionality (d) is the only parameter for PCA, and a limited number of d, around 30, is sufficient to obtain the best classification accuracy. Compared to PCA, KPCA shows a better

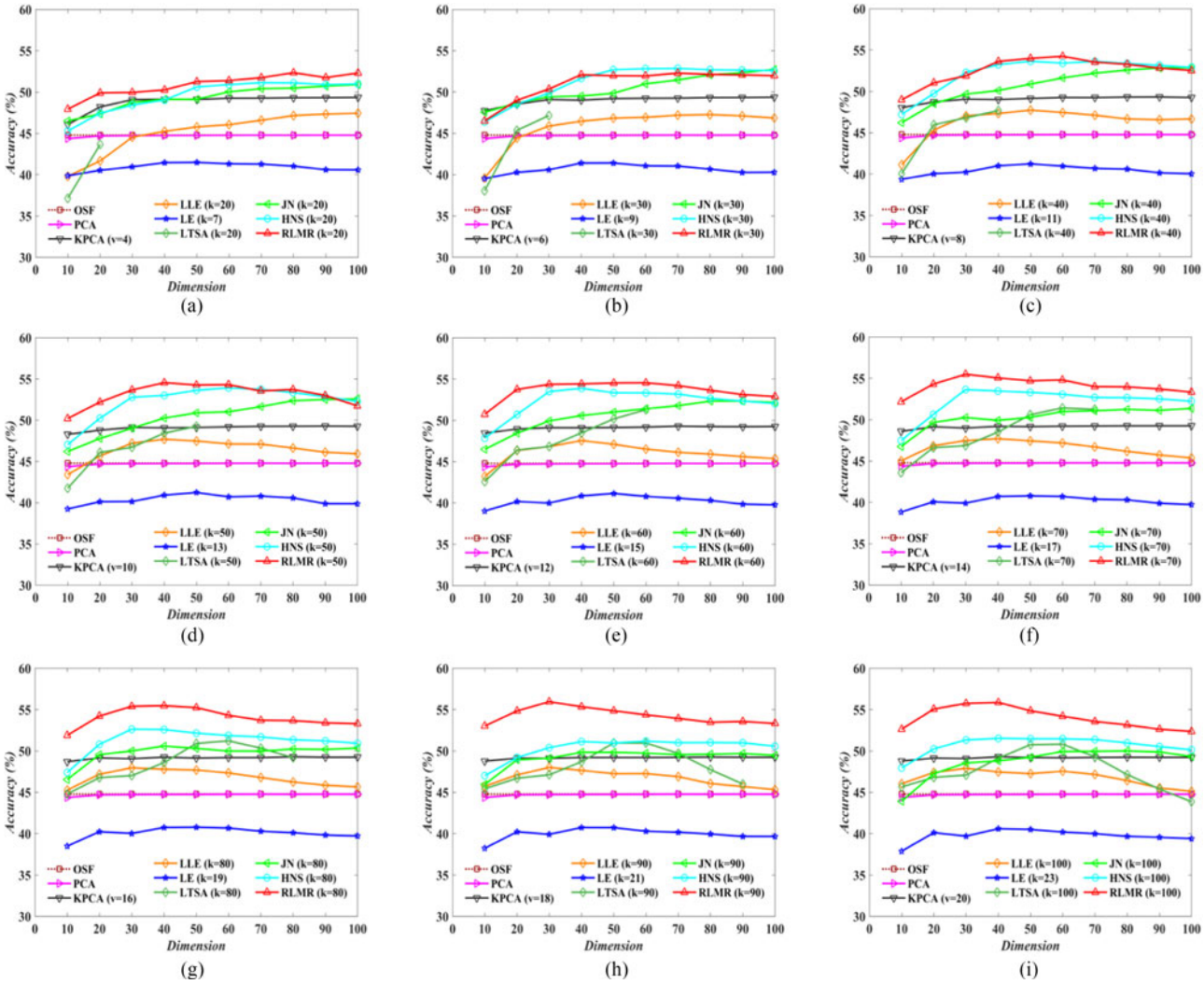


Fig. 11. Performance comparison: Classification accuracy as a function of data dimension using region-based sampling for the Indian Pines dataset. (a)–(i) Results using different numbers of neighbors, respectively.

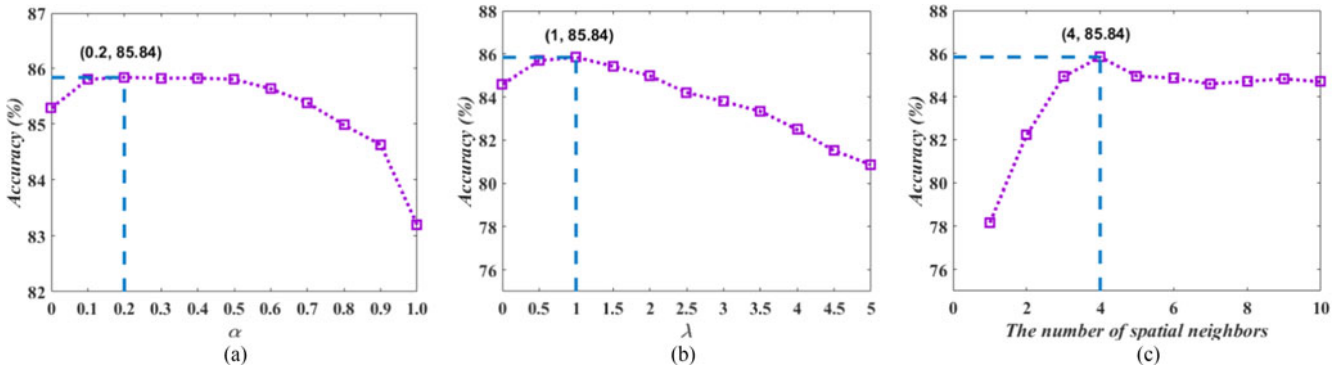


Fig. 12. Sensitivity of the proposed method to the change of three parameters: (a) $\boldsymbol{\alpha}$, (b) $\boldsymbol{\lambda}$, and (c) number of spatial neighbors.

performance owing to its advantage to capture nonlinear properties of the data; however, the parameter selection of kernel is important.

Except for the two parameters, the number of neighbors (k) and the size of reduced dimensionality (d), there are still several parameters in the proposed method, including α in RNS (14), the penalty parameter λ (19), and the number of spatial neighbors (17). With the change of these parameters, the best classification accuracies can be found on the Indian Pines dataset via the NN classifier, and the optimal parameters can be obtained accordingly, as shown in Fig. 12. More specifically, the parameter α in (14) balances similarities generated by KLD between

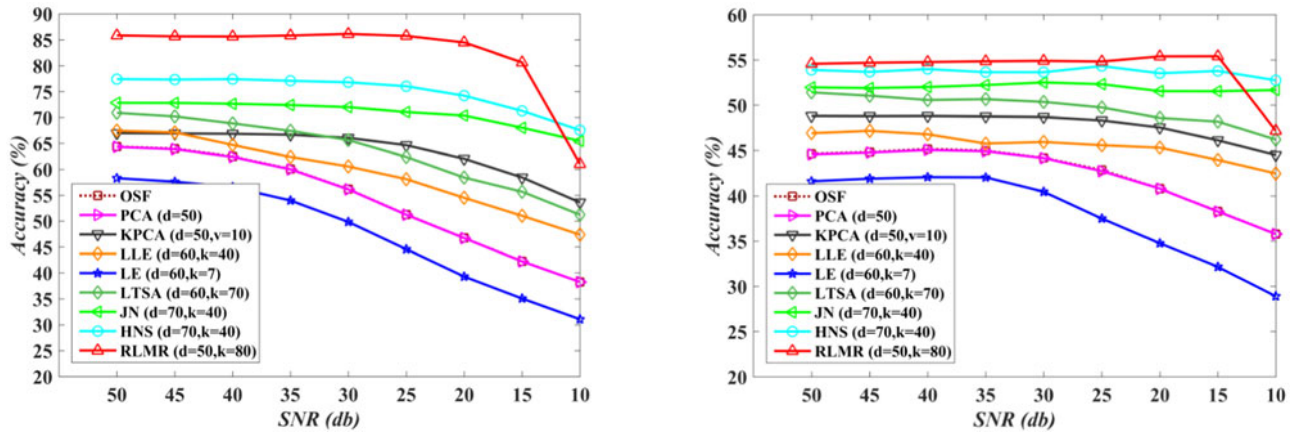


Fig. 13. Classification accuracy for the Indian Pines dataset with different SNRs using all DR methods under comparison based on (left) random sampling and (right) region-based sampling.

the target point and its neighbors. Equation (14) consists of two parts: One is the similarity of data distribution from the target point to its neighbors, and the other is the similarity of data distribution from neighbors to the target point. Obviously, the similarity of the former should be more important than that of latter, which means the parameter α should be less than 1. The optimal value of α is 0.2 corresponding to the best classification accuracy. Regarding the parameter λ, it is used to strike a balance between the error and constraint terms in (19). A proper value ($\lambda = 1$) is set according to experimental results shown in Fig. 12(b). If the number of spatial neighbors is too large or too small, spatial information can be overused or underused, as indicated by degraded classification performance in Fig. 12(c). The value of this parameter should be selected eclectically, and it is set as 4 in terms of the best classification accuracy observed in Fig. 12(c).

b) Robustness analysis: In order to validate the robustness of RLMR, a further experiment is performed, which adds noise with a different signal-to-noise ratio (SNR) into the AVIRIS Indian Pines image. The Gaussian noises are added to the image band by band with the same SNR. Classification was performed with various SNRs to investigate the robustness of the DR algorithms against noise. Fig. 13 shows the classification accuracies under the two sampling strategies. As the SNR decreases, the performance of JN, HNS, and RLMR are comparatively stable and superior compared to those of classical ML methods, PCA, KPCA, and OSF. This demonstrates the robustness of the proposed method against noise and implies its effectiveness for low SRN hyperspectral images.

TABLE V
CLASSIFICATION ACCURACIES FOR THE DFC DATASET USING NN AND DIFFERENT DR METHODS WITH OPTIMAL PARAMETERS

Method	Optimal Parameters	Classification Accuracy	
		Overall Accuracy	Average Accuracy
OSF	/	72.83%	76.16%
PCA	$d = 50$	72.85%	76.19%
KPCA	$d = 50, v = 10$	73.80%	77.79%
LLE	$d = 40, k = 50$	74.23%	77.49%
LE	$d = 60, k = 20$	66.70%	70.66%
LTSA	$d = 40, k = 50$	75.40%	78.75%
JN	$d = 60, k = 50$	77.45%	80.69%
HNS	$d = 80, k = 70$	78.52%	81.75%
RLMR	$d = 70, k = 50$	**80.87%**	**82.77%**

TABLE VI
CLASSIFICATION ACCURACIES FOR THE DFC DATASET USING SVM AND DIFFERENT DR METHODS WITH OPTIMAL PARAMETERS

Method	Optimal Parameters	Classification Accuracy	
		Overall Accuracy	Average Accuracy
OSF	/	74.68%	77.84%
PCA	$d = 30$	74.78%	77.79%
KPCA	$d = 30, v = 10$	75.12%	78,14%
LLE	$d = 60, k = 40$	75.33%	78.03%
LE	$d = 20, k = 30$	70.71%	72.98%
LTSA	$d = 30, k = 50$	76.04%	79.18%
JN	$d = 70, k = 60$	77.86%	80.12%
HNS	$d = 90, k = 60$	78.98%	82.01%
RLMR	$d = 90, k = 100$	**81.13%**	**82.79%**

C. Results of 2013 IEEE GRSS DFC Data

Similarly, we obtained the classification accuracies for the nine methods under the optimal parameters tuned by fivefold cross validation via NN and SVM classifiers using the given training samples in DFC, as listed in Tables V and VI. As can be seen in Tables V and VI, RLMR outperforms the other methods in DFC dataset. This demonstrates that the proposed novel ML method can indeed obtain the good feature representation, thereby further improving the classification accuracy. To be specific, similar results from the different classifiers listed in Tables V and VII also demonstrate the effectiveness and stability of the proposed method.

For simplicity, a general framework for the out-of-samples extension of ML proposed by Bengio [42], [43] is used in this paper in order to obtain the full classification map. The out-of-samples extension can be separated into two parts: first, an appropriate kernel function should be constructed (here, a Gaussian kernel is chosen); next, the Nystrom formulation should be applied for the generalization of a new data point. Classification

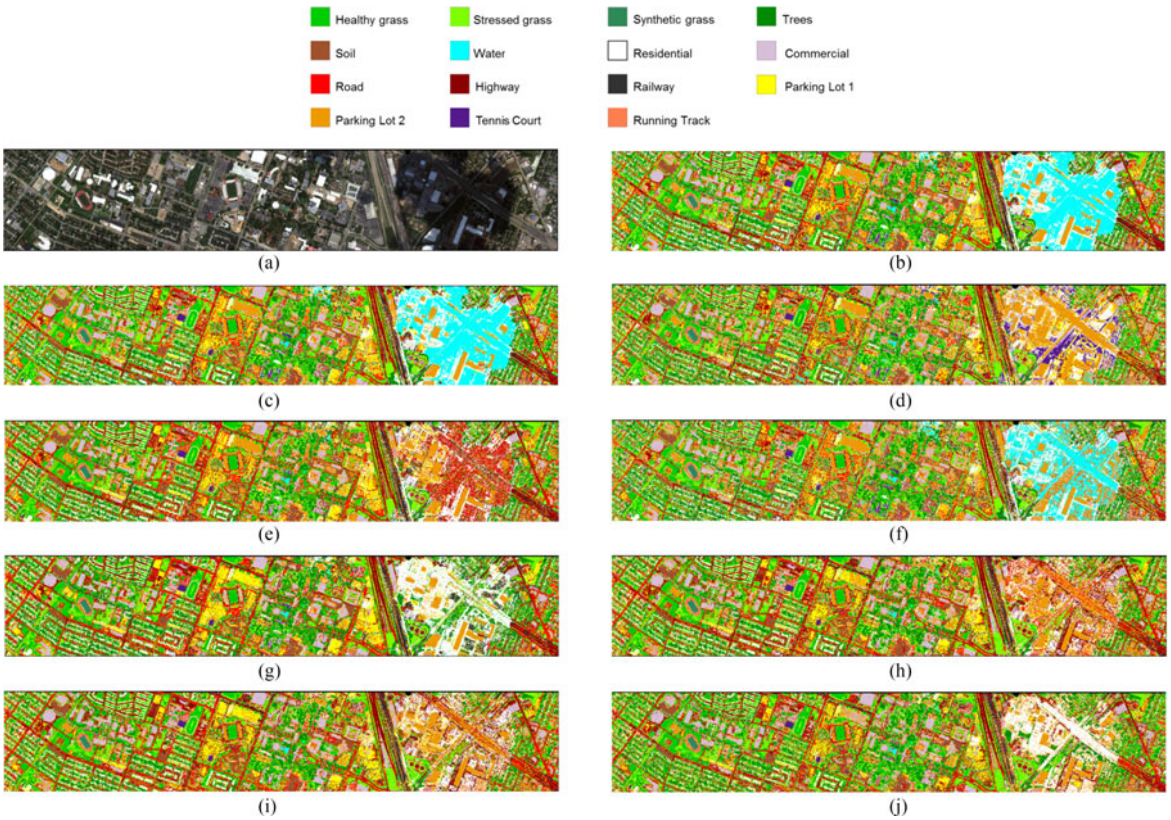


Fig. 14. NN classification maps of the DFC dataset using all DR methods under comparison with optimal parameters in Table V. (a) RGB image from the original hyperspectral image. (b)–(j) Results using OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

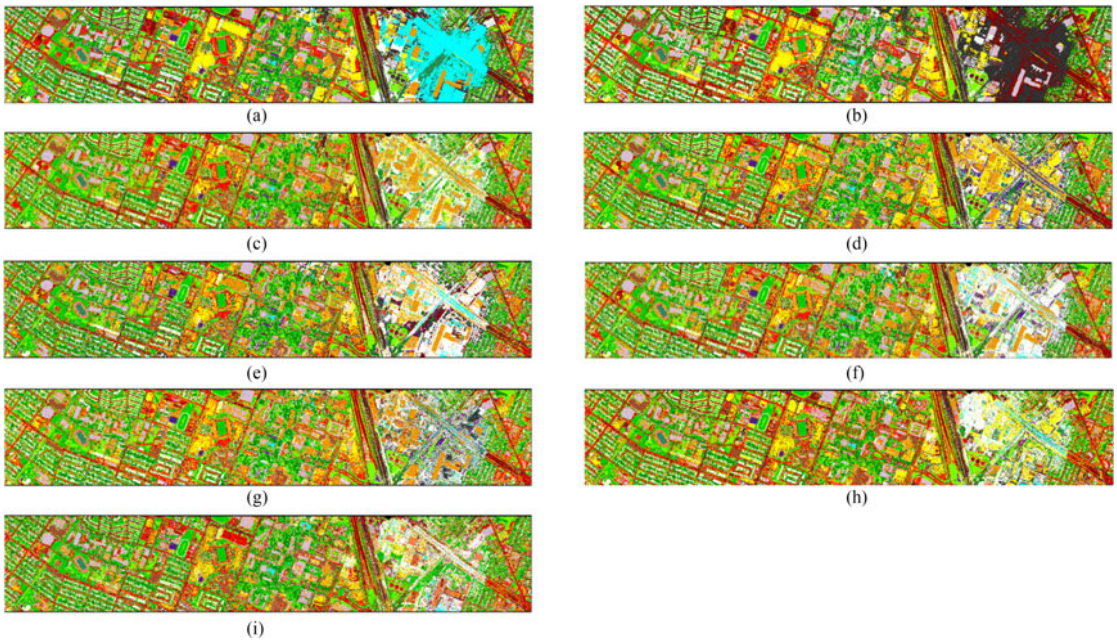


Fig. 15. SVM classification maps of the DFC dataset using all DR methods under comparison with optimal parameters in Table VI. (a)–(i) Results using OSF, PCA, KPCA, LLE, LE, LTSA, JN, HNS, and RLMR, respectively.

maps for different DR methods using the aforementioned optimal parameters are given in Figs. 14 and 15, respectively, corresponding to NN and SVM classifiers. As shown in Fig. 14(a), the east side of the scene is covered with shadows of clouds, resulting in the performance degradation of those previous DR methods—such as in Fig. 14(b)–(g) and Fig. 15(a)–(f)—while our proposed methods are rather robust against this variability observed in Figs. 14(h)–(j) and Fig. 15(g)–(j).

V. Conclusion

In this paper, a novel LML methodology—RLMR—is developed for hyperspectral DR in order to tackle two challenges of LML, involving: 1) NS due to complex spectral variability (e.g., noise, illumination, nonuniform data distribution), and 2) the computation of affinity weights due to multicollinearity. The proposed method is based on JN, RNS, and the integration of spatial information. It was validated via the classification using two benchmark hyperspectral datasets. Compared to other state-of-the-art methods, the proposed method achieves better performance in terms of the classification accuracy. RLMR has a more robust and stable performance than the other methods due to JN, RNS, and the embedding of spatial information, as shown in a series of experiments. In the future, we will further focus on how to more effectively embed the spatial information into DR framework. Additionally, the application of manifold learning methods to large-scale data should be given more attention in the future.

References

[1] Q. Zhang, R. Souvenir, and R. Pless, "On manifold structure of cardiac MRI data: Application to segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1092–1098.

[2] J. Yang, D. Zhang, J. Yang, and B. Niu, "Globally maximizing, locally minimization: Unsupervised discriminant projection with application to face and palm biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 650–664, Apr. 2007.

[3] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning, " *IEEE Signal Process.*, vol. 31, no. 1, pp. 55–66, Jan. 2014.

[4] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani, "Multi-class classification on Riemannian manifolds for video surveillance," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 378–391.

[5] L. K. Saul and S. T. Roweis, "Thing globally, fit locally: Unsupervised learning of low dimensional manifolds," *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, Jun. 2003.

[6] I. T. Jolliffe, *Principle Component Analysis*. New York, NY, USA: Springer-Verleg, 1986.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, Dec. 2000.

[8] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5550, pp. 2323–2326, 2000.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Mar. 2003.

[10] Z. Y. Zhang and H. Y. L. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, Dec. 2004.

[11] C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina, "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 441–454, Mar. 2005.

[12] J. He, L. Zhang, Q. Wang, and Z. Li, "Using diffusion geometric coordinates for hyperspectral imagery representation, " *IEEE Trans. Geosci. Remote Sens.*, vol. 6, no. 4, pp. 767–771, Jan. 2009.

[13] L. Ma, M. M. Crawford, and J. W. Tian, "Local manifold learning-based-k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

[14] L. Ma, M. M. Crawfor, X. Yang, and Y. Guo, "Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2832–2844, May. 2015.

[15] H. Huang, H. Huo, and T. Fang, "Hierarchical manifold learning with application to supervised classification for high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1677–1692, Mar. 2013.

[16] Y. Tan, H. Yuan, and L. Li, "Manifold-based sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7606–7618, Dec. 2014.

[17] L. Ma, M. M. Crawford, and J. W. Tian, "Anomaly detection for hyperspectral images based on robust locally linear embedding, "*J. Infrared Millim. THz Waves*, vol. 31, no. 6, pp. 753–763, 2010.

[18] L. Zhang, L. Zhang, D. Tao, and X. Huang, "Sparse transfer manifold embedding for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1030–1042, Feb. 2014.

[19] H. L. Yang and M. M. Crawford, "Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 51–64, Jan. 2016.

[20] S. Yan, X. Dong, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 40–51, Jan. 2007.

[21] S. Gerver, T. Tasdizen, and R. Whitaker, "Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 281–288.

[22] H. Chang and D. Y. Yeung, "Robust locally linear embedding," *Pattern Recognit.*, vol. 39, no. 6, pp. 1053–1065, Jun. 2006.

[23] Y. Goldberg and Y. A. Ritov, "LDR-LLE: LLE with low-dimensional neighborhood representation," in *Advances in Visual Computing*, vol. 5359. Berlin, Germany: Springer, Dec. 2008, pp. 43–54.

[24] J. A. Lee and M. Verleysen, "*Nonlinear Dimensionality Reduction*. New York, NY, USA: Springer, Oct. 2007.

[25] P. Zhang, H. Qiao, and B. Zhang, "An improved local tangent space alignment method for manifold learning," *Pattern Lett.*, vol. 32, no. 2, pp. 181–189, Jan. 2011.

[26] Y. Zhan and J. Yin, "Robust local tangent space alignment," *Neural Inf. Process.*, vol. 5863. Berlin, Germany: Springer, Dec. 2009, pp. 293–301.

[27] S. T. Monteiro, K. Uto, Y. Kosugi, K. Oda, Y. Lino, and G. Saito, "Hyperspectral image classification of grass species in northeast Japan," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, Jul. 2008, pp. IV-399–IV-402.

[28] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Jan. 2007.

[29] S. Lyu and E. P. Simoncelli, "Nonlinear image representation using divisive normalization, " in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[30] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2007, pp. 85–86.

[31] S. O. Los, P. R. J. North, W. M. F. Gery, and M. J. Barnsley, "A method to convert AVHRR normalized difference vegetation index time series to a standard viewing and illumination geometry," *Remote Sens. Environ.*, vol. 99, no. 4, pp. 400–411, Dec. 2005.

[32] D. Sage, "Local normalization filter to reduce the effect of non-uniform illumination," 2011. [Online]. Available: http://bigwww.epfl.ch/sage/soft/localnormalization/

[33] S. Azadi, J. Maitin-Shepard, and P. Abbeel, "Optimization-based artifact correction for electron microscopy image stacks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 219–235.

[34] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Sep. 2015.

[35] Y. Pei, F. Huang, F. Shi, and H. Zhi, "Unsupervised image matching based on manifold alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1658–1664, Jun. 2012.

[36] C. Wang, J. Lai, and J. Zhu, "Graph-based multiprototype competitive learning and its applications," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 934–946, Dec. 2012.

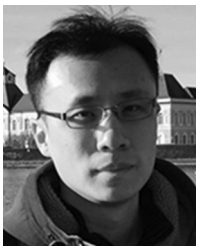
[37] M. Jia, M. Gong, E. Zhang, Y. Li, and L. Jiao, "Hyperspectral image classification based on nonlocal means with a novel class-relativity measurement," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 7, pp. 1300–1304, Jul. 2014.

[38] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation, " *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[39] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 471–478, Nov. 2011.

[40] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden Markov random fields, " *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2565–2574, Feb. 2014.

[41] B. Scholkopf, A. J. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural. Comput.*, vol. 10, no. 5, pp. 583–588, 1998.
[42] Y. Bengio, O. Delalleau, and N. Le Roux, "Learning eigenfunctions links spectral embedding and kernel PCA," *Neural Comput.*, vol. 16, no. 10, pp. 2197–2219, 2004.
[43] Y. Bengio, J. F. Paiement, and P. Vincent., "Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, 2003, pp. 177–184.
[44] J. A. Benediktsson and P. Ghamisi, *Spectral-Spatial Classification of Hyperspectral Remote Sensing Images*. Boston, MA, USA: Artech House.

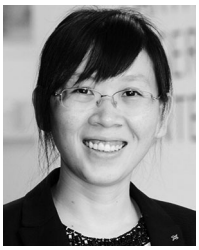


Danfeng Hong (S'16) received the B.Sc. degree in computer science and technology from the Neusoft College of Information, Northeastern University, Dalian, China, in 2012, the M. Sc. degree in computer vision from Qingdao University, Qingdao, China, in 2015. He has been working toward the Ph.D. degree in the hyperspectral data analysis including dimensionality reduction and nonlinear spectral unmixing, Technical University of Munich, Munich, Germany, and the Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany, since September 2015.

His research interests include image processing, pattern recognition, and machine learning and their applications in hyperspectral data analysis.

Naoto Yokoya (S'10–M'13) received the M.Sc. and Ph.D. degrees in aerospace engineering from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

From 2012 to 2013, he was a Research Fellow with the Japan Society for the Promotion of Science, Tokyo, Japan. Since 2013, he has been an Assistant Professor with the University of Tokyo. Since 2015, he has also been an Alexander von Humboldt Research Fellow with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and Technical University of Munich (TUM), Munich, Germany. His research interests include image analysis and data fusion in remote sensing.



Xiao Xiang Zhu (S'10–M'12–SM'14) received the Bachelor's degree in space engineering from the National University of Defense Technology, Changsha, China, in 2006. She received the M.Sc., Dr.-Ing., and Habilitation degrees in the field of signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

Since 2011, she has been a Scientist with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany, where she is also the Head of the Team Signal Analysis. Since 2013, she has also been a Helmholtz Young Investigator Group Leader and appointed as a TUM Junior Fellow. In 2015, she was appointed as a Professor in the Signal Processing in Earth Observation, TUM. She was a Guest Scientist or Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and the University of California, Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests include advanced InSAR techniques, such as high-dimensional tomographic SAR imaging and SqueeSAR; computer vision in remote sensing including object reconstruction and multidimensional data visualization; big data analysis in remote sensing; and modern signal processing, including innovative algorithms, such as sparse reconstruction, nonlocal means filter, robust estimation, and deep learning, with applications in the field of remote sensing, such as multi/hyperspectral image analysis.

Dr. Zhu is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Appendices

B Hong D., Yokoya N., Xu J., Zhu X. X., 2018. Joint & Progressive Learning from High-Dimensional Data for Multi-Label Classification. European Conference on Computer Vision (ECCV), Munich, Germany, September, pp. 469-484.

https://eccv2018.org/.../Danfeng_Hong_Joint__Progressive_ECCV_2018_paper.pdf

Joint & Progressive Learning from High-Dimensional Data for Multi-Label Classification

Danfeng Hong[1,2][0000−0002−3212−9584], Naoto Yokoya[3][0000−0002−7321−4590], Jian Xu[1][0000−0003−2348−125X], and Xiaoxiang Zhu[1,2][0000−0001−5530−3613]

[1] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
{danfeng.hong,jian.xu,xiao.zhu}@dlr.de
[2] Signal Processing in Earth Observation (SiPEO), Technical University of Munich, Munich, Germany
[3] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
{naoto.yokoya}@riken.jp

Abstract. Despite the fact that nonlinear subspace learning techniques (e.g. manifold learning) have successfully applied to data representation, there is still room for improvement in explainability (explicit mapping), generalization (out-of-samples), and cost-effectiveness (linearization). To this end, a novel linearized subspace learning technique is developed in a joint and progressive way, called **j**oint and **p**rogressive **l**earning str**a**teg**y** (J-Play), with its application to multi-label classification. The J-Play learns high-level and semantically meaningful feature representation from high-dimensional data by 1) jointly performing multiple subspace learning and classification to find a latent subspace where samples are expected to be better classified; 2) progressively learning multi-coupled projections to linearly approach the optimal mapping bridging the original space with the most discriminative subspace; 3) locally embedding manifold structure in each learnable latent subspace. Extensive experiments are performed to demonstrate the superiority and effectiveness of the proposed method in comparison with previous state-of-the-art methods.

Keywords: Alternating direction method of multipliers · High-dimensional data · Manifold regularization · Multi-label classification · Joint learning · Progressive learning

1 Introduction

High-dimensional data are often characterized by very rich and diverse information, which enables us to classify or recognize the targets more effectively and analyze data attributes more easily, but inevitably introduces some drawbacks (e.g. information redundancy, complex noise effects, high storage-consuming, etc.) due to *the curve of dimensionality*. A general way to address this problem

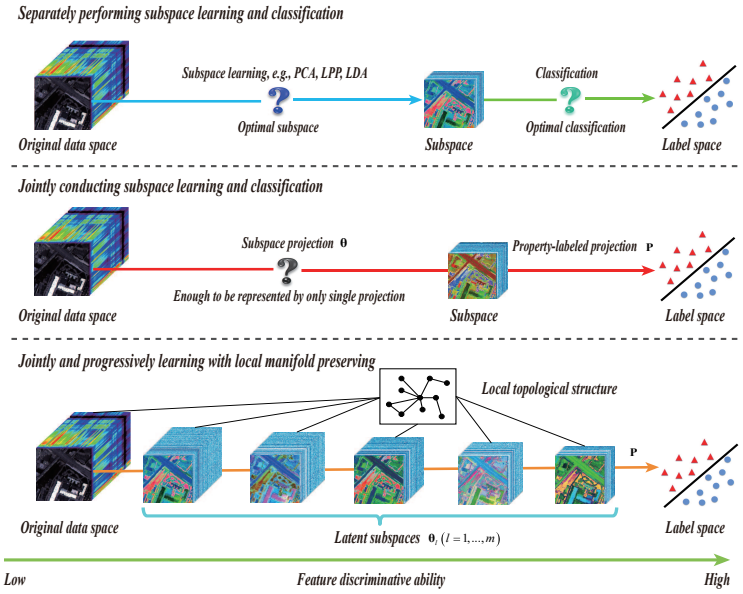


Fig. 1. The motivation interpolation from separately performing subspace learning and classification to joint learning to joint & progressive learning again. The subspaces learned from our model indicates the higher feature discriminative ability as explained by the green bottom line.

is to learn a low-dimensional and high-discriminative feature representation. In general, it is also called as dimensionality reduction or subspace learning. In the past decades, a large number of subspace learning techniques have been developed in the machine learning community, with successful applications to biometrics [20][5][9][10], image/video analysis [26], visualization [22], hyperspectral data analysis (e.g., dimensionality reduction and unmixing) [12][13][14]. These subspace learning techniques are generally categorized into linear or nonlinear methods. Theoretically, nonlinear approaches are capable of curving the data structure in a more effective way. There is, however, no explicit mapping function (poor explainability), and meanwhile it is relatively hard to embed the out-of-samples into the learned subspace (weak generalization) as well as high computational cost (lack of cost-effectiveness). Additionally, for a task of multi-label classification, these classic subspace learning techniques, such as principal component analysis (PCA) [29], local discriminant analysis (LDA) [20], local fisher discriminant analysis (LFDA) [23], manifold learning (e.g. Laplacian eigenmaps (LE) [1], locally linear embedding (LLE) [21]) and their linearized methods (e.g. locality preserving projection (LPP)[6], neighborhood preserving embedding (NPE)[4]), are commonly applied as a disjunct feature learning step before classification, whose limitation mainly lies in a weak connection between features

by subspace learning and label space (see the top panel of Fig. 1). It is unknown which learned features (or subspace) can improve the classification.

Recently, a feasible solution to the above problems can be generalized as a joint learning framework [17] that simultaneously considers linearized subspace learning and classification, as illustrated in the middle panel of Fig. 1. Following it, more advanced methods have been proposed and applied in various fields, including supervised dimensionality reduction (e.g. least-squares dimensionality reduction (LSDR) [24] and its variants: least-squares quadratic mutual information derivative (LSQMID) [25]), multi-modal data matching and retrieval [28, 27], and heterogeneous features learning for activity recognition [15, 16]. In these work, the learned features (or subspace) and label information are effectively connected by regression techniques (e.g. linear regression) to adaptively estimate a latent and discriminative subspace. Despite this, they still fail to find an optimal subspace, as single linear projection is hardly enough to represent the complex transformation from the original data space to the potential optimal subspace.

Motivated by the aforementioned studies, we propose a novel **j**oint and **p**rogressive **l**earning str**a**teg**y** (J-Play) to linearly find an optimal subspace for general multi-label classification, illustrated in the bottom panel of Fig. 1. We practically extend the existing joint learning framework by learning a series of subspaces instead of single subspace, aiming at progressively converting the original data space to a potentially optimal subspace through multi-coupled intermediate transformations [18]. Theoretically, by increasing the number of subspaces, coupled subspace variations are gradually narrowed down to a very small range that can be represented effectively via a *linear transformation*. This renders us to find a good solution easier, especially when the model is complex and non-convex. We also contribute to structure learning in each latent subspace by locally embedding manifold structure.

The main highlights of our work can be summarized as follows:

- A linearized progressive learning strategy is proposed to describe the variations from the original data space to potentially optimal subspace, tending to find a better solution. A joint learning framework that simultaneously estimates subspace projections (connect the original space and the latent subspaces) and a property-labeled projection (connect the learned latent subspaces and label space) is considered to find a discriminative subspace where samples are expected to be better classified.
- Structure learning with local manifold regularization is performed in each latent subspace.
- Based on the above techniques, a novel joint and progressive learning strategy (J-Play) is developed for multi-label classification.
- An iterative optimization algorithm based on the alternating direction method of multipliers (ADMM) is designed to solve the proposed model.

2 Joint & Progressive Learning Strategy (J-Play)

2.1 Notations

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_k, ..., \mathbf{x}_N] \in \mathbb{R}^{d_0 \times N}$ be a data matrix with d_0 dimensions and N samples, and the matrix of corresponding class labels be $\mathbf{Y} \in \{0, 1\}^{L \times N}$. The kth column of $\mathbf{Y}$ is $\mathbf{y}_k = [\mathbf{y}_{k1}, ..., \mathbf{y}_{kt}, ..., \mathbf{y}_{kL}]^T \in \mathbb{R}^{L \times 1}$ whose each element can be defined as follows:

$$\mathbf{y}_{kt} = \begin{cases} 1, & \text{if } \mathbf{y}_k \text{ belongs to the } t\text{-th class;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In our task, we aim to learn a set of coupled projections $\{\mathbf{\Theta}_l\}_{l=1}^m \in \mathbb{R}^{d_l \times d_{l-1}}$ and a property-labeled projection $\mathbf{P} \in \mathbb{R}^{L \times d_m}$, where m stands for the number of subspace projections and $\{d_l\}_{l=1}^m$ are defined as the dimensions of those latent subspaces respectively, while d_0 is specified as the dimension of $\mathbf{X}$.

2.2 Basic Framework of J-Play from the View of Subspace Learning

Subspace learning is to find a low-dimensional space where we expect to maximize certain properties of the original data, e.g. variance (PCA), discriminative ability (LDA), and graph structure (manifold learning). Yan et al. [30] summarized these subspace learning methods in a general graph embedding framework.

Given an undirected similarity graph $G = \{\mathbf{X}, \mathbf{W}\}$ with the vertices $\mathbf{X} \in \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and the adjacency matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$, we can intuitively measure the similarities among the data. By preserving the similarities relationship, the high-dimensional data can be well embedded into the low-dimensional space, which can be formulated by denoting the low-dimensional data representation as $\mathbf{Z} \in \mathbb{R}^{d \times N}$ ($d \ll d_0$) in the following

$$\min_{\mathbf{Z}} \operatorname{tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}), \quad \text{s.t.} \quad \mathbf{Z}\mathbf{D}\mathbf{Z}^{\mathrm{T}} = \mathbf{I}, \tag{2}$$

where $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ is a diagonal matrix, $\mathbf{L}$ is a Laplacian matrix defined by $\mathbf{L} = \mathbf{D} - \mathbf{W}$ [3], and $\mathbf{I}$ is the identity matrix. In our case, we aim at learning multi-coupled linear projections to find optimal mapping, therefore a linearized subspace learning problem can be reformulated on the basis of Eq. (2) by substituting $\mathbf{\Theta}\mathbf{X}$ for $\mathbf{Z}$

$$\min_{\mathbf{\Theta}} \operatorname{tr}(\mathbf{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}), \quad \text{s.t.} \quad \mathbf{\Theta}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}, \tag{3}$$

which can be solved by generalized eigenvalue decomposition.

Different from the previously mentioned subspace learning methods, a regression-based joint learning model [17] can explicitly bridge the learned latent subspace and labels, which can be formulated in a general form:

$$\min_{\mathbf{P}, \mathbf{\Theta}} \frac{1}{2}\mathbf{E}(\mathbf{P}, \mathbf{\Theta}) + \frac{\beta}{2}\mathbf{\Phi}(\mathbf{\Theta}) + \frac{\gamma}{2}\mathbf{\Psi}(\mathbf{P}), \tag{4}$$

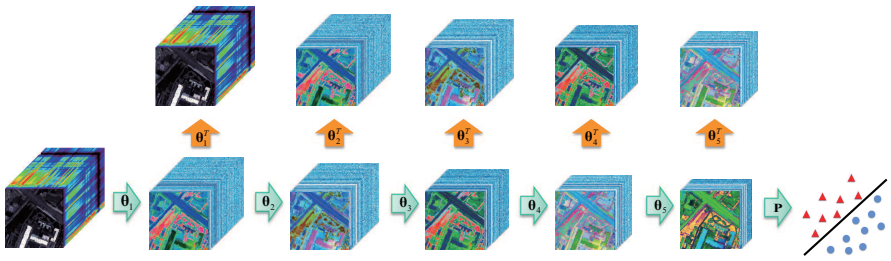


Fig. 2. The illustration of the proposed J-Play framework.

where $\mathbf{E}(\mathbf{P}, \mathbf{\Theta})$ is the error term defined as $\|\mathbf{Y} - \mathbf{P}\mathbf{\Theta}\mathbf{X}\|_{\mathrm{F}}^2$, $\|\bullet\|_{\mathrm{F}}$ represents a Frobenius norm, β and γ are the corresponding penalty parameters. $\mathbf{\Phi}$ and $\mathbf{\Psi}$ denote regularization functions, which might be l_1 norm, l_2 norm, $l_{2,1}$ norm or manifold regularization. Herein, the variable $\mathbf{\Theta}$ is called intermediate transformation and the corresponding subspace generated by $\mathbf{\Theta}$ is called latent subspace where the feature can be further structurally learned and represented in a more suitable way [16].

On the basis of Eq. (5), we further extend the framework by following a progressive learning strategy:

$$\min_{\mathbf{P}, \{\mathbf{\Theta}_l\}_{l=1}^m} \frac{1}{2}\mathbf{E}(\mathbf{P}, \{\mathbf{\Theta}_l\}_{l=1}^m) + \frac{\beta}{2}\mathbf{\Phi}(\{\mathbf{\Theta}_l\}_{l=1}^m) + \frac{\gamma}{2}\mathbf{\Psi}(\mathbf{P}), \tag{5}$$

where $\mathbf{E}(\mathbf{P}, \{\mathbf{\Theta}_l\}_{l=1}^m)$ is specified as $\|\mathbf{Y} - \mathbf{P}\mathbf{\Theta}_m ... \mathbf{\Theta}_l ... \mathbf{\Theta}_1 \mathbf{X}\|_{\mathrm{F}}^2$ and $\{\mathbf{\Theta}_l\}_{l=1}^m$ represent a set of intermediate transformations.

2.3 Problem Formulation

Following the general framework given in Eq.(6), the proposed J-Play can be formulated as the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{P}, \{\mathbf{\Theta}_l\}_{l=1}^m} & \frac{1}{2}\mathbf{\Upsilon}(\{\mathbf{\Theta}_l\}_{l=1}^m) + \frac{\alpha}{2}\mathbf{E}(\mathbf{P}, \{\mathbf{\Theta}_l\}_{l=1}^m) + \frac{\beta}{2}\mathbf{\Phi}(\{\mathbf{\Theta}_l\}_{l=1}^m) + \frac{\gamma}{2}\mathbf{\Psi}(\mathbf{P}) \\ & \text{s.t.} \quad \mathbf{X}_l = \mathbf{\Theta}_l \mathbf{X}_{l-1}, \quad \mathbf{X}_l \succeq 0, \quad \|\mathbf{x}_{lk}\|_2 \preceq 1, \quad \forall l = 1, 2, ..., m, \end{aligned} \tag{6}$$

where $\mathbf{X}$ is assigned to $\mathbf{X}_0$, while α, β, and γ are three penalty parameters corresponding to the different terms, which aim at balancing the importance between the terms. Fig. 2 illustrates the J-Play framework. Since Eq. (7) is a typically ill-posed problem, reasonable assumptions or priors need to be introduced to search a solution in a narrowed range effectively. More specifically, we cast Eq.(7) as a least-square regression problem with reconstruction loss term ($\mathbf{\Upsilon}(\bullet)$), prediction loss term ($\mathbf{E}(\bullet)$) and two regularization terms ($\mathbf{\Phi}(\bullet)$ and $\mathbf{\Psi}(\bullet)$). We detail these terms one by one as follows.

1) Reconstruction Loss Term $\mathbf{\Upsilon}(\{\mathbf{\Theta}_l\}_{l=1}^m)$: Without any constraints or prior, directly estimating multi-coupled projections in J-Play is hardly performed with

the increase of the number of estimated projections. This can be reasonably explained by gradient missing between the two neighboring variables estimated in the process of optimization. That is, the variations between these neighboring projections are made to be tiny and even zero. In particular, when the number of projections increases to a certain extent, most of learned projections tend to be zero and become meaningless. To this end, we adopt a kind of autoencoder-like scheme to make the learned subspace projected back to the original space as much as possible. The benefits of the scheme are, on one hand, to prevent the data over-fitting to some extent, especially avoiding overmuch noises from being considered; on the other hand, to establish an effective link between the original space and the subspace, making the learned subspace more meaningful. Therefore, the resulting expression is

$$\boldsymbol{\Upsilon}(\{\boldsymbol{\Theta}_l\}_{l=1}^m) = \sum\nolimits_{l=1}^{m} \|\mathbf{X}_{l-1} - \boldsymbol{\Theta}_l^T \boldsymbol{\Theta}_l \mathbf{X}_{l-1}\|_{\mathrm{F}}^2. \tag{7}$$

In our case, to fully utilize the advantages of this term, we consider it in each latent subspace as shown in Eq.(8).

2) Predication Loss Term $\mathbf{E}(\mathbf{P}, \{\boldsymbol{\Theta}_l\}_{l=1}^m)$: This term is to minimize the empirical risk between the original data and the corresponding labels through multi-coupled projections in a progressive way, which can be formulated as

$$\mathbf{E}(\mathbf{P}, \{\boldsymbol{\Theta}_l\}_{l=1}^m) = \|\mathbf{Y} - \mathbf{P}\boldsymbol{\Theta}_m ... \boldsymbol{\Theta}_l ... \boldsymbol{\Theta}_1 \mathbf{X}\|_{\mathrm{F}}^2. \tag{8}$$

3) Local Manifold Regularization $\boldsymbol{\Phi}(\{\boldsymbol{\Theta}_l\}_{l=1}^m)$: As introduced in [27], a manifold structure is an important prior for subspace learning. Superior to vector-based feature learning, such as artificial neural network (ANN), a manifold structure can effectively capture the intrinsic structure between samples. To facilitate structure learning in J-Play, we perform the local manifold regularization to each latent subspace. Specifically, this term can be expressed by

$$\boldsymbol{\Phi}(\{\boldsymbol{\Theta}_l\}_{l=1}^m) = \sum\nolimits_{l=1}^{m} \mathrm{tr}(\boldsymbol{\Theta}_l \mathbf{X}_{l-1} \mathbf{L} \mathbf{X}_{l-1}^{\mathrm{T}} \boldsymbol{\Theta}_l^{\mathrm{T}}). \tag{9}$$

4) Regression Coefficient Regularization $\boldsymbol{\Psi}(\mathbf{P})$: The regularization term can promote us to derive a more reasonable solution with a reliable generalization to our model, which can be written as

$$\boldsymbol{\Psi}(\mathbf{P}) = \|\mathbf{P}\|_{\mathrm{F}}^2. \tag{10}$$

Moreover, the non-negativity constraint with respect to each learned dimension-reduced feature (e.g. $\{\mathbf{X}_l\}_{l=1}^m \succeq 0$) is considered since we aim to obtain a meaningful low-dimensional feature representation similar to original image data acquired in a non-negative unit. In addition to the non-negativity constraint, we also impose a norm constraint [4] for sample-based of each subspace: $\|\mathbf{x}_{lk}\|_2 \preceq 1, \forall k = 1, ..., N$ and $l = 1, ..., m$.

[4] Regarding this constraint,please refer to [19] for more details.

Algorithm 1: Joint & Progressive Learning Strategy (J-Play)

Input: $\mathbf{Y}, \mathbf{X}, \mathbf{L}$, and parameters α, β, γ and $maxIter$.
Output: $\{\mathbf{\Theta}_l\}_{l=1}^m$.

Initialization Step:
Greedily initialize $\mathbf{\Theta}_l$ corresponding to each latent subspace:
for $l = 1 : m$ **do**
  $\mathbf{\Theta}_l^0 \leftarrow LPP(\mathbf{X}_{l-1})$
  $\mathbf{\Theta}_l \leftarrow AutoRULe(\mathbf{X}_{l-1}, \mathbf{\Theta}_l^0, \mathbf{L})$
  $\mathbf{X}_l \leftarrow \mathbf{\Theta}_l \mathbf{X}_{l-1}$
end
Fine-tuning Step:
$t = 0, \zeta = 1e-4$;
while *not converged* or $t > maxIter$ **do**
  Fix other variables to update $\mathbf{P}$ by solving a subproblem of $\mathbf{P}$;
  for $i = 1 : m$ **do**
    Fix other variables to update $\mathbf{\Theta}_l^{t+1}$ by solving a subproblem of $\mathbf{\Theta}_l$;
  end
  Compute the objective function value Obj^{t+1} and check the convergence condition: **if** $|\frac{Obj^{t+1} - Obj^t}{Obj^t}| < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
  end
end

2.4 Model Optimization

Considering the complexity and the non-convexity of our model, we pretrain our model to have an initial approximation of subspace projections $\{\mathbf{\Theta}_l\}_{l=1}^m$ as this can greatly reduce the model's training time and also help finding an optimal solution easier. This is a common tactic that has been successfully employed in deep autoencoders [8]. Inspired by this trick, we propose a pre-training model with respect to $\mathbf{\Theta}_l, \forall l = 1, ..., m$ by simplifying Eq.(7) as

$$\min_{\mathbf{\Theta}_l} \frac{1}{2}\mathbf{\Upsilon}(\mathbf{\Theta}_l) + \frac{\eta}{2}\mathbf{\Phi}(\mathbf{\Theta}_l) \quad \text{s.t.} \quad \mathbf{X}_l \succeq 0, \quad \|\mathbf{x}_{lk}\|_2 \preceq 1, \tag{11}$$

which is named as **auto-r**econstructing **u**nsupervised **le**arning (AutoRULe). Given the outputs of AutoRULe, the problem of Eq. (7) can be more effectively solved by an alternatively minimizing strategy that separately solves two subproblems with respect to $\{\mathbf{\Theta}_l\}_{l=1}^m$ and $\mathbf{P}$. Therefore, the global algorithm of J-Play can be summarized in **Algorithm 1**,where AutoRULe is initialized by LPP.

The pre-training method (AutoRULe) can be effectively solved via the ADMM-based framework. Following this, we consider an equivalent form of Eq. (12) by introducing multiple auxiliary variables $\mathbf{H}$, $\mathbf{G}$, $\mathbf{Q}$ and $\mathbf{S}$ to replace $\mathbf{X}_l$, $\mathbf{\Theta}_l$, $\mathbf{X}_l^+$ and $\mathbf{X}_l^\sim$, respectively, where $()^+$ denotes an operator that converts each component of the matrix to its absolute value and $()^\sim$ is a proximal operator for

solving the constraint of $\|\mathbf{x}_{lk}\|_2 \preceq 1$ [7], written as follows

$$\begin{aligned}\min_{\boldsymbol{\Theta}_l,\mathbf{H},\mathbf{G},\mathbf{Q},\mathbf{S}} \frac{1}{2}\boldsymbol{\Upsilon}(\mathbf{G},\mathbf{H}) + \frac{\eta}{2}\boldsymbol{\Phi}(\boldsymbol{\Theta}_l) &= \frac{1}{2}\|\mathbf{X}_{l-1} - \mathbf{G}^{\mathrm{T}}\mathbf{H}\|_{\mathrm{F}}^2 + \frac{\eta}{2}\operatorname{tr}(\mathbf{X}_l\mathbf{L}\mathbf{X}_l^{\mathrm{T}}) \\ \text{s.t.} \quad \mathbf{Q} \succeq 0, \quad \|\mathbf{s}_k\|_2 \preceq 1, \quad \mathbf{X}_l &= \boldsymbol{\Theta}_l\mathbf{X}_{l-1}, \\ \mathbf{X}_l = \mathbf{H}, \quad \boldsymbol{\Theta}_l = \mathbf{G}, \quad \mathbf{X}_l &= \mathbf{Q}, \quad \mathbf{X}_l = \mathbf{S}.\end{aligned} \tag{12}$$

The augmented Lagrangian version of Eq. (13) is

$$\begin{aligned}&\mathscr{L}_\mu\left(\boldsymbol{\Theta}_l,\mathbf{H},\mathbf{G},\mathbf{Q},\mathbf{S},\{\boldsymbol{\Lambda}_n\}_{n=1}^4\right) \\ &= \frac{1}{2}\|\mathbf{X}_{l-1} - \mathbf{G}^{\mathrm{T}}\mathbf{H}\|_{\mathrm{F}}^2 + \frac{\eta}{2}\operatorname{tr}(\boldsymbol{\Theta}_l\mathbf{X}_{l-1}\mathbf{L}\mathbf{X}_{l-1}^{\mathrm{T}}\boldsymbol{\Theta}_l^{\mathrm{T}}) + \boldsymbol{\Lambda}_1^{\mathrm{T}}(\mathbf{H} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}) \\ &+ \boldsymbol{\Lambda}_2^{\mathrm{T}}(\mathbf{G} - \boldsymbol{\Theta}_l) + \boldsymbol{\Lambda}_3^{\mathrm{T}}(\mathbf{Q} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}) + \boldsymbol{\Lambda}_4^{\mathrm{T}}(\mathbf{S} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}) + \frac{\mu}{2}\|\mathbf{H} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 \\ &+ \frac{\mu}{2}\|\mathbf{G} - \boldsymbol{\Theta}_l\|_{\mathrm{F}}^2 + \frac{\mu}{2}\|\mathbf{Q} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 + \frac{\mu}{2}\|\mathbf{S} - \boldsymbol{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 + l_R^+(\mathbf{Q}) + l_R^{\sim}(\mathbf{S}),\end{aligned} \tag{13}$$

where $\{\boldsymbol{\Lambda}_n\}_{n=1}^4$ are Lagrange multipliers and μ is the penalty parameter. The two terms $l_R^+(\bullet)$ and $l_R^{\sim}(\bullet)$ represent two kinds of projection operators, respectively. That is, $l_R^+(\bullet)$ is defined as

$$max(\bullet) = \begin{cases} \bullet\,, & \bullet \succ 0 \\ 0\,, & \bullet \preceq 0, \end{cases} \tag{14}$$

while $l_R^{\sim}(\bullet_k)$ is a vector-based operator defined by

$$prox_f(\bullet_k) = \begin{cases} \frac{\bullet_k}{\|\bullet_k\|_2}\,, & \|\bullet_k\|_2 \succ 1 \\ \bullet_k\,, & \|\bullet_k\|_2 \preceq 1, \end{cases} \tag{15}$$

where $\bullet_k$ is the kth column of matrix $\bullet$. **Algorithm 2** details the procedures of AutoRULe.

The two subproblems in **Algorithm 1** can be optimized alternatively as follows:

Optimization with respect to $\mathbf{P}$: This is a typical least square regression problem, which can be written as

$$\min_{\mathbf{P}} \frac{\alpha}{2}\mathbf{E}(\mathbf{P}) + \frac{\gamma}{2}\boldsymbol{\Psi}(\mathbf{P}) = \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{P}\boldsymbol{\Theta}_m...\boldsymbol{\Theta}_l...\boldsymbol{\Theta}_1\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\gamma}{2}\|\mathbf{P}\|_{\mathrm{F}}^2, \tag{16}$$

which has a closed-form solution

$$\mathbf{P} \leftarrow (\alpha\mathbf{Y}\mathbf{V}^{\mathrm{T}})(\alpha\mathbf{V}\mathbf{V}^{\mathrm{T}} + \gamma\mathbf{I})^{-1}, \tag{17}$$

where $\mathbf{V} = \boldsymbol{\Theta}_m...\boldsymbol{\Theta}_l...\boldsymbol{\Theta}_1, \forall l = 1, ..., m$.

Optimization with respect to $\{\boldsymbol{\Theta}_l\}_{l=1}^m$: The variables $\{\boldsymbol{\Theta}_l\}_{l=1}^m$ can be individually optimized, and hence the optimization problem of each $\boldsymbol{\Theta}_l$ can be generally

Algorithm 2: Auto-reconstructing unsupervised learning (AutoRULe)

Input: $\mathbf{X}_{l-1}, \mathbf{\Theta}_l^0, \mathbf{L}$, and parameters η and $maxIter$.
Output: $\mathbf{\Theta}_l$.
Initialization: $\mathbf{H}^0 = \mathbf{\Theta}_l^0\mathbf{X}_{l-1}, \mathbf{G}^0 = \mathbf{0}, \mathbf{Q}^0 = \mathbf{P}^0 = \mathbf{0}, \mathbf{\Lambda}_2^0 = \mathbf{0}, \mathbf{\Lambda}_1^0 = \mathbf{\Lambda}_3^0 = \mathbf{\Lambda}_4^0 = \mathbf{0}, \mu^0 = 1e-3, \mu_{max} = 1e6, \rho = 2, \varepsilon = 1e-6, t = 0$.
while *not converged* or $t > maxIter$ **do**
  Fix $\mathbf{H}^t, \mathbf{G}^t, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{\Theta}_l^{t+1}$ by
$$\mathbf{\Theta}_l = (\mu\mathbf{H}\mathbf{X}_{l-1}^T + \mathbf{\Lambda}_1\mathbf{X}_{l-1}^T + \mu\mathbf{G} + \mathbf{\Lambda}_2 + \mu\mathbf{Q}\mathbf{X}_{l-1}^T + \mathbf{\Lambda}_3\mathbf{X}_{l-1}^T + \mu\mathbf{P}\mathbf{X}_{l-1}^T + \mathbf{\Lambda}_4\mathbf{X}_{l-1}^T)(\eta(\mathbf{X}_{l-1}\mathbf{L}\mathbf{X}_{l-1}^T) + 3\mu(\mathbf{X}_{l-1}\mathbf{X}_{l-1}^T) + \mu\mathbf{I})^{-1}.$$
  Fix $\mathbf{\Theta}_l^{t+1}, \mathbf{G}^t, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{H}^{t+1}$ by
$$\mathbf{H} = (\mathbf{G}\mathbf{G}^T + \mu\mathbf{I})^{-1}(\mathbf{G}\mathbf{X}_{l-1} + \mu\mathbf{\Theta}_l\mathbf{X}_{l-1} - \mathbf{\Lambda}_1).$$
  Fix $\mathbf{H}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{G}^{t+1}$ by
$$\mathbf{G} = (\mathbf{H}\mathbf{H}^T + \mu\mathbf{I})^{-1}(\mathbf{H}\mathbf{X}_i + \mu\mathbf{\Theta}_l - \mathbf{\Lambda}_2).$$
  Fix $\mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{P}^t$ to update $\mathbf{Q}^{t+1}$ by
$$\mathbf{Q} = max(\mathbf{\Theta}_l\mathbf{X}_{l-1} - \mathbf{\Lambda}_3/\mu, 0).$$
  Fix $\mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{Q}^{t+1}$ to update $\mathbf{P}^{t+1}$ by
$$\mathbf{P} = prox_f(\mathbf{\Theta}_l\mathbf{X}_{l-1} - \mathbf{\Lambda}_4/\mu).$$
  Update Lagrange multipliers by
$$\mathbf{\Lambda}_1^{t+1} = \mathbf{\Lambda}_1^t + \mu^t(\mathbf{H}^{t+1} - \mathbf{\Theta}_i^{t+1}\mathbf{X}_{l-1}), \mathbf{\Lambda}_2^{t+1} = \mathbf{\Lambda}_2^t + \mu^t(\mathbf{G}^{t+1} - \mathbf{\Theta}_i^{t+1}),$$
$$\mathbf{\Lambda}_3^{t+1} = \mathbf{\Lambda}_3^t + \mu^t(\mathbf{Q}^{t+1} - \mathbf{\Theta}_i^{t+1}\mathbf{X}_{l-1}), \mathbf{\Lambda}_4^{t+1} = \mathbf{\Lambda}_4^t + \mu^t(\mathbf{P}^{t+1} - \mathbf{\Theta}_i^{t+1}\mathbf{X}_{l-1}).$$
  Update penalty parameter by
$$\mu^{t+1} = min(\rho\mu^t, \mu_{max}).$$
  Check the convergence conditions: **if** $\|\mathbf{H}^{t+1} - \mathbf{\Theta}_l^{t+1}\mathbf{X}_{l-1}\|_F < \varepsilon$ *and* $\|\mathbf{G}^{t+1} - \mathbf{\Theta}_l^{t+1}\|_F < \varepsilon$ *and* $\|\mathbf{Q}^{t+1} - \mathbf{\Theta}_l^{t+1}\mathbf{X}_{l-1}\|_F < \varepsilon$ *and* $\|\mathbf{P}^{t+1} - \mathbf{\Theta}_l^{t+1}\mathbf{X}_{l-1}\|_F < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
  end
end

formulated by

$$\begin{aligned}\min_{\mathbf{\Theta}_l} \frac{1}{2}\mathbf{\Upsilon}(\mathbf{\Theta}_l) + \frac{\alpha}{2}\mathbf{E}(\mathbf{\Theta}_l) + \frac{\beta}{2}\mathbf{\Phi}(\mathbf{\Theta}_l) &= \frac{1}{2}\|\mathbf{X}_{l-1} - \mathbf{\Theta}_l^{\mathrm{T}}\mathbf{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 \\ &+ \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{P}\mathbf{\Theta}_m...\mathbf{\Theta}_l...\mathbf{\Theta}_1\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\operatorname{tr}(\mathbf{\Theta}_l\mathbf{X}_{l-1}\mathbf{L}\mathbf{X}_{l-1}^{\mathrm{T}}\mathbf{\Theta}_l^{\mathrm{T}}) \\ \text{s.t.} \quad \mathbf{X}_l = \mathbf{\Theta}_l\mathbf{X}_{l-1}, \quad \mathbf{X}_l \succeq 0, \quad \|\mathbf{x}_{lk}\|_2 &\preceq 1,\end{aligned} \tag{18}$$

which can be basically deduced by following the framework of **Algorithm 2**. The only difference lies in the optimization subproblem with respect to $\mathbf{H}$ whose solution can be collected by solving the following problem:

$$\begin{aligned}\min_{\mathbf{H}} \frac{1}{2}\|\mathbf{X}_{l-1} - \mathbf{G}^{\mathrm{T}}\mathbf{H}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{Y} - \mathbf{P}_l\mathbf{H}\|_{\mathrm{F}}^2 + \mathbf{\Lambda}_1^{\mathrm{T}}(\mathbf{H} - \mathbf{\Theta}_l\mathbf{X}_{l-1}) \\ + \frac{\mu}{2}\|\mathbf{H} - \mathbf{\Theta}_l\mathbf{X}_{l-1}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathbf{P}_l = \mathbf{P}_{l-1}\mathbf{\Theta}_{l+1}, \quad \mathbf{P}_0 = \mathbf{P}.\end{aligned} \tag{19}$$

The analytical solution of Eq. (20) is given by

$$\mathbf{H} \leftarrow (\alpha\mathbf{P}_l^{\mathrm{T}}\mathbf{P}_l + \mathbf{G}\mathbf{G}^{\mathrm{T}} + \mu\mathbf{I})^{-1}(\alpha\mathbf{P}_l^{\mathrm{T}}\mathbf{Y} + \mathbf{G}\mathbf{X}_{l-1} + \mu\mathbf{\Theta}_l\mathbf{X}_{l-1} - \mathbf{\Lambda}_1). \tag{20}$$

Finally, we repeat these optimization procedures until a stopping criterion is satisfied. Please refer to **Algorithm 1** and **Algorithm 2** for more explicit steps.

3 Experiments

In this section, we conduct the classification to quantitatively evaluate the performance of the proposed method (J-Play) using three popular and advanced classifiers, namely the nearest neighbor (NN) based on the Euclidean distance, kernel support vector machines (KSVM) and canonical correlation forest (CCF), in comparison with previous state-of-the-art methods. Overall accuracy (OA) is given to quantify the classification performance.

3.1 Data Description

The experiments are performed on two different types of datasets: hyperspectral datasets and face datasets, as both of them easily suffer from the information redundancy and need to improve the representative ability of features. We have used the following two hyperspectral datasets and two face datasets:

1) *Indian Pines AVIRIS Image:* The first hyperspectral cube was acquired by the AVIRIS sensor with the size of $145 \times 145 \times 220$, which consists of 16 class of vegetation. More specific classes and the arrangement of training and test samples can be found in [11]. The first image of Fig. 3 shows a false color image of Indian Pines data.

2) *University of Houston Image:* The second hyperspectral cube was provided for the 2013 IEEE GRSS data fusion contest acquired by ITRES-CASI sensor with size of $349 \times 1905 \times 144$. The information regarding classes and corresponding train and test samples can be found in [13]. A false color image of the study scene is shown in the first image of Fig. 4.

3) *Extended Yale-B Dataset:* We only choose a subset of the mentioned dataset with the frontal pose and the different illuminations of 38 subjects (2414 images in total), which can widely used in evaluating the performance of subspace learning [32][2]. These images were aligned and cropped to the size of 32×32, that is, 1024-dimensional vector-based representation. Each individual has 64 near frontal images under different illuminations.

4) *AR Dataset:* Similar to [31], we choose a subset of AR under the conditions of illumination and expressions, which comprises of 100 subjects. Each person has 14 images with seven ones from Session 1 as training set and others from Session 2 as testing samples. The images are resized to 60×43.

3.2 Experimental Steup

As the fixed training and testing samples are given for the hyperspectral datasets, subspace learning techniques can directly be performed on training set to learn an optimal subspace where the testing set can be simply classified by NN, KSVM, and CCF. For the face datasets, since there is no standard training and testing

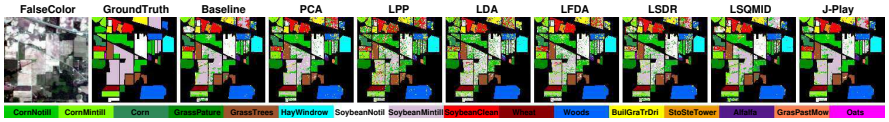


Fig. 3. A false color image, ground truth and classification maps of the different algorithms obtained using CCF on the Indian Pines dataset.

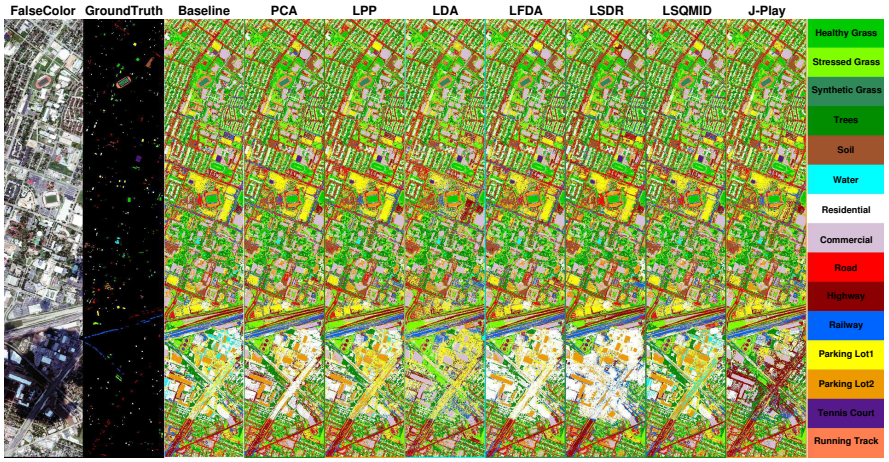


Fig. 4. A false color image, ground truth and classification maps of the different algorithms obtained using CCF on the Houston dataset.

sets, ten replications are performed for randomly selecting training and testing samples. A random subset with 10 facial images per individual is chosen with labels as the training set and the rest of it is considered to be the testing set. Furthermore, we compare the performance of the proposed method (J-Play) with the baseline (original features without dimensionality reduction) and six popular and advanced methods (PCA, LPP, LDA, LFDA, LSDR, and LSQMID). With learning the different number of coupled projections, the proposed method can be successively specified as J-Play$_1$,...,J-Play$_l$,...,J-Play$_m$, $\forall l = 1, ..., m$. To investigate the trend of OAs, m are uniformly set up to 7 on the four datasets.

3.3 Results of Hyperspectral Data

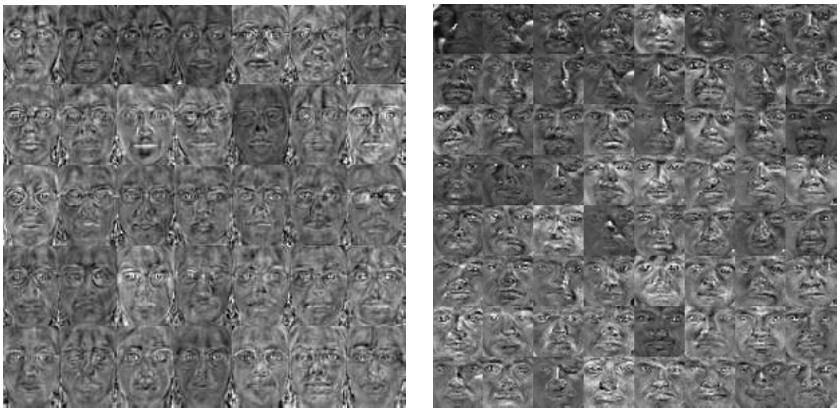
Initially, we conduct a 10-fold cross-validation for the different algorithms on the training set in order to estimate the optimal parameters which can be selected from $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. Table 1 lists classification performances of the different methods with the optimal subspace dimensions obtained by cross-validation using three different classifiers. Correspondingly, the classification maps are given in Figs. 3 and 4 to intuitively highlight the difference.

Table 1. Quantitative performance comparisons on two hyperspectral datasets. The best results for the different classifiers are shown in red.

Methods	Indian Pines dataset			Houston dataset		
	NN	KSVM	CCF	NN	KSVM	CCF
Baseline (220/144)	65.89%	66.56%	81.71%	72.83%	80.19%	82.60%
PCA (20/20)	65.40%	75.25%	79.26%	72.75%	79.54%	83.90%
LPP (20/30)	64.86%	63.02%	68.48%	75.31%	78.43%	81.77%
LDA (15/14)	64.14%	63.88%	65.61%	75.81%	76.66%	79.62%
LFDA (15/14)	73.86%	74.25%	75.17%	75.52%	80.46%	82.27%
LSDR (50/40)	73.67%	76.84%	77.38%	76.80%	80.39%	81.64%
LSQMID (60/80)	66.94%	78.90%	79.32%	76.31%	80.23%	81.69%
J-Play$_1$ (20/30)	78.81%	82.04%	82.24%	78.22%	83.32%	85.09%
J-Play$_2$ (20/30)	80.87%	83.75%	83.23%	79.16%	84.41%	85.15%
J-Play$_3$ (20/30)	83.59%	85.08%	84.44%	80.13%	83.68%	88.19%
J-Play$_4$ (20/30)	83.92%	85.21%	84.57%	79.64%	83.25%	85.63%
J-Play$_5$ (20/30)	83.76%	85.30%	84.41%	80.00%	82.21%	85.81%
J-Play$_6$ (20/30)	83.56%	84.79%	83.82%	79.69%	82.45%	84.82%
J-Play$_7$ (20/30)	82.70%	83.82%	83.04%	77.81%	81.03%	83.23%

Overall, PCA performs basically similar performance with the baseline using the three different classifiers on the two datasets. For LPP, due to its sensitivity to noise, it yields a poor performance on the first dataset, while on the relatively high-quality second dataset, LPP steadily outperforms the baseline and PCA. In the supervised algorithms, owing to the limitation of training samples and discriminative power, the classification accuracies of classic LDA is holistically lower than those previously mentioned. With a more powerful discriminative criterion, LFDA obtains more competitive results by locally focusing on discriminative information, which are generally better than those of the baseline, PCA, LPP, and LDA. However, the features learned by LFDA is sensitive to noise and the number of neighbors, resulting in the unstable performance particularly for the different classifiers. For LSDR and LSQMID, they aim to find a linear projection by maximizing the mutual information between input and output from the view of statistics. With fully considering the mutual information, they achieve the good performance on the two given hyperspectral datasets.

Remarkably, the performance of the proposed method (J-Play) is superior to the other methods on the two hyperspectral datasets. This indicates that J-Play is prone to learn a better feature representation and robust against noise. On the other hand, with the increase of m, the performance of J-Play steadily increases to the best with around 4 or 5 layers for the first dataset and 2 or 3 layers for



(a) Extended Yale-B dataset (b) AR dataset

Fig. 5. Visualization of partial facial features learned by the proposed J-Play on two face datasets.

Table 2. Quantitative performance comparisons on two face datasets. The best results for the different classifiers are shown in red.

Methods	Extended Yale-B dataset			AR dataset		
	NN	KSVM	CCF	NN	KSVM	CCF
Baseline (1024/2580)	45.77%	45.87%	76.99%	71.71%	72.29%	80.29%
PCA (120/80)	41.05%	81.47%	83.53%	68.43%	80.29%	81.43%
LPP (170/70)	70.75%	76.55%	77.48%	70.86%	74.00%	79.86%
LDA (37/99)	80.88%	78.37%	83.68%	81.43%	82.29%	85.38%
LFDA (37/99)	81.02%	80.88%	83.58%	71.29%	75.71%	80.38%
LSDR (60/80)	71.29%	76.40%	78.66%	75.14%	79.00%	80.14%
LSQMID (60/80)	71.48%	77.09%	78.37%	73.29%	74.29%	79.29%
J-Play$_1$ (170/210)	73.01%	79.30%	80.29%	73.57%	79.86%	77.86%
J-Play$_2$ (170/210)	81.17%	84.27%	85.22%	82.29%	86.00%	84.57%
J-Play$_3$ (170/210)	83.43%	85.50%	85.76%	85.43%	88.71%	87.43%
J-Play$_4$ (170/210)	84.07%	86.09%	86.55%	85.29%	87.71%	87.71%
J-Play$_5$ (170/210)	84.56%	86.14%	86.20%	85.71%	87.29%	88.86%
J-Play$_6$ (170/210)	85.35%	85.64%	86.53%	85.14%	87.29%	88.29%
J-Play$_7$ (170/210)	85.74%	85.45%	86.20%	86.57%	86.86%	88.71%

the second one, and then gradually decreases with a slight perturbation since our model is only trained on the training set.

3.4 Results of Face Images

As J-Play is proposed as a general subspace learning framework for multi-label classiciation, we additionally used two popular face datasets to further assess its generalization capability. Similarly, cross-validation on training set is conducted for estimating the optimal parameter combination on the extended Yale-B and AR datasets. Considering the high-dimensional vector-based face images, we first perform the PCA for face images in order to roughly reduce the feature redundancy, whose results are further explored to the dimensionality reduction methods by following the previous work on face recognition (e.g. LDA (Fisherfaces) [20] and LPP (Laplacianfaces) [5]). Table 2 gives the corresponding OAs using the different methods on the two face datasets respectively.

By comparison, the performance of PCA and LPP is steadily superior to that of baseline, while PCA is even better than LPP. For supervised approaches, LDA performs better than baseline, PCA, LPP and even LFDA, showing an impressive result. Due to the less number of training samples from face datasets, LSDR and LSQMID are limited to effectively estimate the mutual information between the training samples and labels, resulting in the performance degradation compared to the hyperspectral data. The proposed method outperforms other algorithms, which indicates that this method can effectively learn an optimal mapping from original space to label space, further improving the classification accuracy. Likewise, there is a similar trend for the proposed method with the increase of m that J-Play can basically obtain the optimal OAs with around 4 or 5 layers and more layers would lead to the performance degradation. We also characterize and visualize each column of the learned projection, as shown in Fig. 5 where those high-level or semantically meaningful features, i.e. face features under the different pose and illumination, can be learned well, making the faces identified easier.

4 Conclusions

To effectively find an optimal subspace where the samples can be semantically represented and thereby be better classified or recognized, we proposed a novel linearized subspace learning framework (J-Play) which aims at learning the feature representation from the high-dimensional data in a joint and progressive way. Extensive experiments of multi-label classification are conducted on two types of datasets: hyperspectral images and face images, in comparison with some previously proposed state-of-the-art methods. The promising results using J-Play demonstrate its superiority and effectiveness. In the future, we will further build an unified framework based on J-Play by extending it to semi-supervised learning, transfer learning, or multi-task learning.

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6), 1373–1396 (2003)

2. Cai, D., He, X., Han, J.: Spectral regression: A unified approach for sparse subspace learning. In: International Conference on Data Mining (ICDM). pp. 73–82 (2007)
3. Chung, F.R.K.: Spectral graph theory. American Mathematical Society (1997)
4. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood preserving embedding. In: International Conference on Computer Vision (ICCV). vol. 2, pp. 1208–1213 (2005)
5. He, X., Hu, S., Niyogi, P., Zhang, H.J.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **27**(3), 328–340 (2005)
6. He, X., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems (NIPS). pp. 153–160 (2004)
7. Heide, F., Heidrich, W., Wetzstein, G.: Fast and flexible convolutional sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5135–5143 (2015)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
9. Hong, D., Liu, W., Su, J., Z.Pan, Wang, G.: A novel hierarchical approach for multispectral palmprint recognition. Neurocomputing **151**, 511–521 (2015)
10. Hong, D., Liu, W., Wu, X., Pan, Z., Su, J.: Robust palmprint recognition based on the fast variation vese–osher model. Neurocomputing **174**, 999–1012 (2016)
11. Hong, D., Yokoya, N., Zhu, X.: The k-lle algorithm for nonlinear dimensionality ruduction of large-scale hyperspectral data. In: IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS). pp. 1–5. IEEE (2016)
12. Hong, D., Yokoya, N., Zhu, X.: Local manifold learning with robust neighbors selection for hyperspectral dimensionality reduction. In: IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS). pp. 40–43. IEEE (2016)
13. Hong, D., Yokoya, N., Zhu, X.: Learning a robust local manifold representation for hyperspectral dimensionality reduction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) **10**(6), 2960–2975 (2017)
14. Hong, D., Yokoya, N., Chanussot, J., Zhu, X.X.: Learning a low-coherence dictionary to address spectral variability for hyperspectral unmixing. In: Image Processing (ICIP), 2017 IEEE International Conference on. pp. 235–239. IEEE (2017)
15. Hu, J., Zheng, W., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5344–5352 (2015)
16. Hu, J., Zheng, W., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition (2016)
17. Ji, S., Ye, J.: Linear dimensionality reduction for multi-label classification. In: International Joint Conference on Artifical Intelligence (IJCAI). vol. 9, pp. 1077–1082 (2009)
18. Kan, M., Shan, S., Chang, H., Chen, X.: Stacked progressive auto-encoders (spae) for face recognition across poses. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1883–1890 (2014)
19. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems (NIPS). pp. 801–808 (2007)
20. Martnez, A.M., Avinash, C.K.: Pca versus lda. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **23**(2), 228–233 (2001)
21. Roweis, S.T., Lawrence, K.S.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)

22. Saul, S.L., Roweis, S.T.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research (JMLR) **4**, 119–155 (2003)
23. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. Journal of Machine Learning Research (JMLR) **8**, 1027–1061 (2007)
24. Suzuki, T., Sugiyama, M.: Sufficient dimension reduction via squared-loss mutual information estimation. Neural Computation **25**(3), 725–758 (2013)
25. Tangkaratt, V., Sasaki, H., Sugiyama, M.: Direct estimation of the derivative of quadratic mutual information with application in supervised dimension reduction. Neural Computation **29**(8), 2076–2122 (2017)
26. Tosato, D., Farenzena, M., Spera, M., Murino, V., Cristani, M.: Multi-class classification on riemannian manifolds for video surveillance. In: Europe Conference on Computer Vision (ECCV). pp. 378–391 (2010)
27. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **38**(10), 2010–2023 (2016)
28. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: International Conference on Computer Vision (ICCV). pp. 2088–2095 (2013)
29. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems **2**(1), 37–52 (1987)
30. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **29**(1), 40–51 (2007)
31. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 625–632 (2011)
32. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: which helps face recognition?. In: International Conference on Computer Vision (ICCV). pp. 471–478 (2011)

Appendices

C Hong D., Yokoya N., Chanussot J., Zhu X. X., 2019. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. IEEE Transactions on Image Processing (TIP), 28(4): 1923-1938.

https://ieeexplore.ieee.org/document/8528557

An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing

Danfeng Hong, *Student Member, IEEE*, Naoto Yokoya, *Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— Hyperspectral imagery collected from airborne or satellite sources inevitably suffers from spectral variability, making it difficult for spectral unmixing to accurately estimate abundance maps. The classical unmixing model, the linear mixing model (LMM), generally fails to handle this sticky issue effectively. To this end, we propose a novel spectral mixture model, called the augmented LMM, to address spectral variability by applying a data-driven learning strategy in inverse problems of hyperspectral unmixing. The proposed approach models the main spectral variability (i.e., scaling factors) generated by variations in illumination or typography separately by means of the endmember dictionary. It then models other spectral variabilities caused by environmental conditions (e.g., local temperature and humidity and atmospheric effects) and instrumental configurations (e.g., sensor noise), and material nonlinear mixing effects, by introducing a spectral variability dictionary. To effectively run the data-driven learning strategy, we also propose a reasonable prior knowledge for the spectral variability dictionary, whose atoms are assumed to be low-coherent with spectral signatures of endmembers, which leads to a well-known low-coherence dictionary learning problem. Thus, a dictionary learning technique is embedded in the framework of spectral unmixing so that the algorithm can learn the spectral variability dictionary and estimate the abundance maps simultaneously. Extensive experiments on synthetic and real datasets are performed to demonstrate the superiority and effectiveness of the proposed method in comparison with the previous state-of-the-art methods.**

***Index Terms*— Alternating direction method of multipliers, low-coherent dictionary learning, remote sensing, spectral unmixing, spectral variability.**

Manuscript received May 22, 2017; revised February 1, 2018 and September 15, 2018; accepted October 26, 2018. Date of publication November 9, 2018; date of current version December 12, 2018. This work was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program under Grant [ERC-2016-StG-714087], Acronym: *So2Sat*, in part by the Helmholtz Association under the framework of the Young Investigators Group "SiPEO" under Grant VH-NG-1018, in part by the Bavarian Academy of Sciences and Humanities in the Framework of Junges Kolleg, and in part by ANR ASTRID (Project APHYPIS) under Grant ANR-16-ASTR-0027-01. This work of N. Yokoya was supported by the Japan Society for the Promotion of Science (KAKENHI) under Grant 18K18067. This paper was presented at the ICIP2017 [1]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter Tay. *(Corresponding author: Xiao Xiang Zhu.)*

D. Hong and X. X. Zhu are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany, and also with the Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: danfeng.hong@dlr.de; xiaoxiang.zhu@dlr.de).

N. Yokoya is with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: naoto.yokoya@riken.jp).

J. Chanussot is with the Université Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France, and also with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland (e-mail: jocelyn@hi.is).

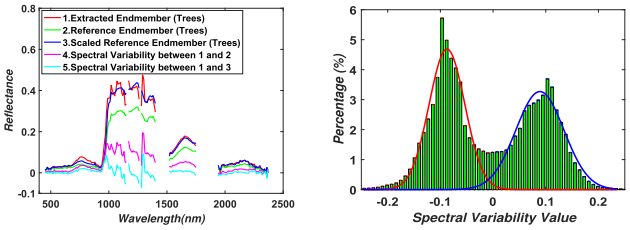
Digital Object Identifier 10.1109/TIP.2018.2878958

I. Introduction

With the rapid development of imaging spectrometers, considerable attention has been paid to spectral-based data processing and analysis, including dimensionality reduction [2]–[4], spectral unmixing [5], [6], segmentation [7], classification [8], [9], and object detection and recognition [10], [11]. Many pixels in hyperspectral data suffer from the effect of material mixtures due to a lower spatial resolution than that of color or multispectral imaging and so on. Mixed pixels inevitably degrade the performance of high-level data analysis. Therefore, spectral unmixing has been gaining importance for hyperspectral image analysis. Hyperspectral unmixing is a procedure that decomposes the measured pixel spectrum of hyperspectral data into a collection of constituent spectral signatures (or *endmembers*) and a set of corresponding fractional abundances. Hyperspectral unmixing techniques have been widely used for a variety of applications [12], such as mineral mapping [13] and land-cover change detection [14].

The linear mixing model (LMM) is a simple but effective model that is extensively used for spectral unmixing. However, two main factors, nonlinearity and spectral variability, still hinder the LMM's ability to yield high performance. In hyperspectral imaging, nonlinearity - i.e., nonlinearly mixed spectral signatures - is the result of multiple scattering and intimate mixing. Spectral variability refers to a variation of a spectral signature for a given material, due to illumination conditions and topography, atmospheric effects, or even the intrinsic variability of the material [15], [16]. Quite recently, considerable attention has been paid to dealing with spectral variability in hyperspectral unmixing [16]–[19]. Variations of spectral signatures for a material can result in significant errors in hyperspectral unmixing.

In the literature, several theories have been proposed to model spectral variability. In [20] and [21], the normal compositional model and the beta compositional model were designed by assuming that spectral variability follows a given probability distribution. Fu *et al.* proposed a spectral-library-based spectral unmixing approach, called the dictionary-adjusted nonconvex sparsity-encouraging regression (DANSER), to model the mismatch between the spectral library and the observed spectral signatures [22]. This kind of mismatch can be also treated as spectral variability in general. Obviously, the spectral variability in a certain scene can hardly be modeled by giving an explicit distribution in reality. Thouvenin *et al.* [23] indicated that spectral variability can be represented using a perturbed linear mixing model (PLMM),



(a) (b)

Fig. 1. An explicit example to clarify the spectral variability. (a): The line (red) 1 denotes the endmember of the trees extracted using VCA from the Urban scene (see Section IV), while the line (green) 2 is the corresponding reference endmember (Trees) by referring to [29] and [30]. The line (blue) 3 is estimated by multiplying a scaling factor on the line 2. The line 4 (or 5) illustrates the differences between 1 and 2 (or 3) in order to clarify the existence of other spectral variabilities besides scaling factors. (b) gives a statistical distribution of spectral variability in Urban scene that it is not a simple Gaussian distribution rather than more like a more complex Gaussian mixture distribution (please refer to the Section II.D for more details).

where the variability is explained by an additive perturbation term for each endmember. One drawback of this model is a lack of physical meaning. For instance, as a principal spectral variability, scaling factors should be coherent with endmember spectral signatures, while other variabilities are often incoherent with endmember spectral signatures. Intuitively, such attributed spectral variability can not be represented by an additional term. In contrast, an interesting approach, called an extended linear mixing mode (ELMM), has been proposed in [24] and [25]. This work mainly focuses on modeling the scaling factors on the endmembers, but is a slight deficiency in that other spectral variabilities cannot be considered correspondingly. Only taking the scaling factors into account is incomplete due to those innegligible spectral variabilities (e.g. atmospheric effects or nonlinear spectral mixing) that are restrictively represented only using scaling factors. Figs. 1(a) and 3(a) show the intuitive examples to clarify the significance of considering other spectral variabilities.

To address the limitations of the PLMM and the ELMM, the purpose of this paper is to model the scaling factors and other spectral variability simultaneously, according to their distinctive properties. More specifically, our contributions can be summarized as follows:

- We propose a novel spectral mixture model, called an augmented linear mixing model (ALMM), where scaling factors are modeled by the endmember dictionary and an additional dictionary is introduced to model the rest of spectral variabilities simultaneously;
- A data-driven dictionary learning method is explored in the proposed framework of spectral unmixing in which a statistical prior is given, specifying that the spectral variability (except for scaling factors) be low-coherent with endmember spectral signatures;
- An optimization algorithm based on the alternating direction method of multipliers (ADMM) is designed to solve the proposed model.

The remainder of this paper is organized as follows. Section II describes the classical LMM and its variations, particularly analyzing their advantages and disadvantages. In Section III, we elaborate on our motivation and propose the methodology for the novel spectral mixture model (ALMM) and the corresponding optimization algorithm. Section IV presents the experimental results using three different datasets and discusses the qualitative and quantitative analysis. Finally, Section V concludes with a summary.

II. The Linear Mixing Model and Its Variations

In this section, we introduce the LMM and discuss its variations, the ELMM and the PLMM. Their respective motivations to address spectral variability are presented and analyzed in detail, with a focus on their advantages and disadvantages.

A. Linear Mixing Model

Let $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_k, ..., \mathbf{y}_N] \in \mathbb{R}^{D\times N}$ be an observed hyperspectral image with D bands and N pixels, and $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_P] \in \mathbb{R}^{D\times P}$ be the endmember matrix (or dictionary), where P is the number of endmembers. $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_k, ..., \mathbf{x}_N] \in \mathbb{R}^{P\times N}$ is the abundance map, with each column vector representing the abundance vector at each pixel. $\mathbf{R} = [\mathbf{r}_1, ..., \mathbf{r}_k, ..., \mathbf{r}_N] \in \mathbb{R}^{D\times N}$ is the corresponding residual matrix containing the additive noise and other errors.

With these notations, the LMM can be modeled, based on pixel-wise $\mathbf{y}_k \in \mathbb{R}^{D\times 1}$, as

$$\mathbf{y}_k = \mathbf{A}\mathbf{x}_k + \mathbf{r}_k, \tag{1}$$

with the two reasonable constraints adapting to reality [26] as follows: 1) the abundance non-negative constraint (ANC), namely $\mathbf{x}_k \succeq 0$; and 2) the abundance sum-to-one constraint (ASC), namely $\mathbf{1}_P^T\mathbf{x}_k = 1$ ($\mathbf{1}_P = [1, 1, ..., 1]^T \in \mathbb{R}^P$). Considering all pixels, a compact matrix form for the LMM can be written as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{R}. \tag{2}$$

The LMM is an approximation of reality and the linearity assumption can hold in most real cases, but it is limited to handle the problem with spectral variability. Further, these changes or effects, as often as not, lead to more specific spectral variabilities, such as scaling factors, offset, or complex noise. Although the LMM coupled with the spectral bundles technique has provided a consideration for spectral variability, spectral bundles rely heavily on establishing a good dictionary. Coincidentally, it is barely possible to prepare a good dictionary in a real case, resulting in the failure of the LMM against spectral variability. Fig. 1(a) shows an example of spectral variability. We extracted the pure endmember (trees) via the vertex component analysis (VCA) [27] algorithm from an urban dataset (see Section IV), which can be simply identified by the reference endmembers [28], [29]. The differences between the two endmembers, (spectral variability) can be visually observed as shown in Fig. 1(a) (*curve 4*). As shown in Fig. 1, the extracted endmember (*curve 1*) can be better approximated by a scaled version of the reference endmember (*curve 3*) than directly by the reference endmember without a scaling factor (*curve 2*).

B. Extended Linear Mixing Model

Incorporating that estimation approach into the algorithm, Drumetz *et al.* [25] proposed the ELMM to fully consider the scaling factors in order to allow a pixel-wise variation of each endmember:

$$\mathbf{y}_k = \mathbf{A}\mathbf{S}_k\mathbf{x}_k + \mathbf{r}_k, \tag{3}$$

where $\mathbf{S}_k \in \mathbb{R}^{P\times P}$ is a diagonal matrix with the constraint that diagonal elements are nonnegative. Eq. (3) can be extended to a compact matrix form:

$$\mathbf{Y} = \mathbf{A}(\mathbf{S} \odot \mathbf{X}) + \mathbf{R}, \tag{4}$$

where $\mathbf{S} \in \mathbb{R}^{P\times N}$ aims at representing all scaling factors for all pixels whose k^{th} column is $\mathbf{S}_k$. The mathematical symbol $\odot$ denotes the Schur-Hadamard (termwise) product.

Typically, Eqs. (3) and (4) are non-convex optimization problems, which difficultly provide the analytic solutions. In [25], Drumetz *et al.* relaxed Eqs. (3) and (4) by employing a strategy of splitting variables, thereby obtaining the following objective function:

$$\{\hat{\mathbf{X}}, \hat{\mathbf{S}}, \hat{\underline{\mathbf{A}}}\} = arg \min_{\mathbf{X},\mathbf{S},\underline{\mathbf{A}}} \sum_{k=1}^{N} (\|\mathbf{y}_k - \mathbf{A}_k\mathbf{x}_k\|_2^2 + \lambda_S \|\mathbf{A}_k - \mathbf{A}_0\mathbf{S}_k\|_F^2) \tag{5}$$

where $\mathbf{A}_0$ is the reference endmember matrix, $\underline{\mathbf{A}} = \{\mathbf{A}_k\}$ is the collection of pixel-dependent endmember matrices, and λ_S denotes the penalty parameter to balance the two separated terms. Therefore, we can iteratively optimize individual variables by alternating nonnegative least squares (ANLS) [31].

C. Perturbed Linear Mixing Model

Inspired by a model proposed in [23] and [32] modeled spectral variability simply and flexibly through an additive perturbation information. This model, the PLMM, is formulated by

$$\mathbf{y}_k = (\mathbf{A} + \Delta_k)\mathbf{x}_k + \mathbf{r}_k, \tag{6}$$

where $\Delta_k \in \mathbb{R}^{D\times P}$ denotes the perturbation of the endmember matrix $\mathbf{A}$ in the k^{th} pixel, whose columns are the perturbation vectors associated with each endmember in $\mathbf{A}$. The matrix form of Eq. (6) can be expressed as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \underbrace{[\Delta_1\mathbf{x}_1|...|\Delta_k\mathbf{x}_k|...|\Delta_N\mathbf{x}_N]}_{\Delta} + \mathbf{R}, \tag{7}$$

where Δ is $[\Delta_1\mathbf{x}_1|...|\Delta_k\mathbf{x}_k|...|\Delta_N\mathbf{x}_N]$. $\mathbf{X}$ can be estimated by adopting an alternating minimization strategy based on an ADMM optimization framework [33]. Readers are referred to [23] for more details.

D. Discussion and Summary

In summary, according to the different prior assumptions, the LMM and its variations give the corresponding spectral mixing models for unmixing respectively. Unfortunately, in real scenarios they are not capable of effectively dealing with spectral variability (in the case of the LMM) or can only considering a special type of spectral variability (in the case of the ELMM for scaling factors). Although the PLMM tried to create a general model that incorporates spectral variabilities, the model does not consider the properties of spectral variability (e.g., the variation of illumination conditions). Furthermore, perturbation information may explain offset variability effectively but it ignores other important spectral variabilities, like scaling factors, which results in performance degradation in the unmixing process. Fig. 1 shows more evidence regarding the spectral variabilities. As can be clearly seen, the spectral variability *curve 4* generated by the difference between *curve 1* and *curve 2* can be largely explained by the scaling factor (shown in *curve 3*), but spectral variability other than scaling factors still remain (refer to *curve 5*). This is a self-evident example that demonstrates that individually considering scaling factors or perturbed information to model spectral variability is not adequate for modeling. Nevertheless, although DANSER and PLMM attempt model spectral variability in a generalized way, in [22] and [23] they both assume that spectral variability follows a Gaussian distribution and thus is constrained using the Frobenius norm in [22] and [23]. We have to point out, however, that spectral variability does not strictly obey a Gaussian distribution in real scenarios. Direct evidence supporting this point is shown in Fig. 1(b), which approximately satisfies a mixed Gaussian distribution. More specifically, we selected the potential pure endmembers from the urban data based on the reference endmembers provided and then were able to calculate the spectral variabilities between the reference endmembers and the extracted endmembers using a subtraction operation. In the end, we collected all scalars from the obtained spectral variabilities and displayed them in the form of statistics, as shown in Fig. 1(b).

III. AUGMENTED LINEAR MIXING MODEL

Scaling factors and other spectral variabilities are simultaneously considered in our model. Also, the reasonable prior assumptions are introduced as regularization terms into our model. Finally, an ADMM-based optimization algorithm is explored to solve the proposed model.

A. Motivation

In hyperspectral imaging, a local region in the real world, which is presented as a mixed pixel in an image, usually presents a similar scaling variability due to a similar illumination condition and topography. Considering another fact that more degrees of flexibility (e.g. endmember-wise scaling factors) are considered in the ELMM, this makes the model too ill-posed. The two facts motivated us to further slightly re-constrain the ELMM model by using a shared scaling factor on each endmember. In practice, this implementation is basically reasonable and useful as the scaling factors are strongly related to topography, which can indeed be assumed in most situations to be constant for all the endmembers of a given pixel at the scale of observation [24]. While those small induced remaining errors that can not be represented by the shared scaling factors, should be able to be corrected with

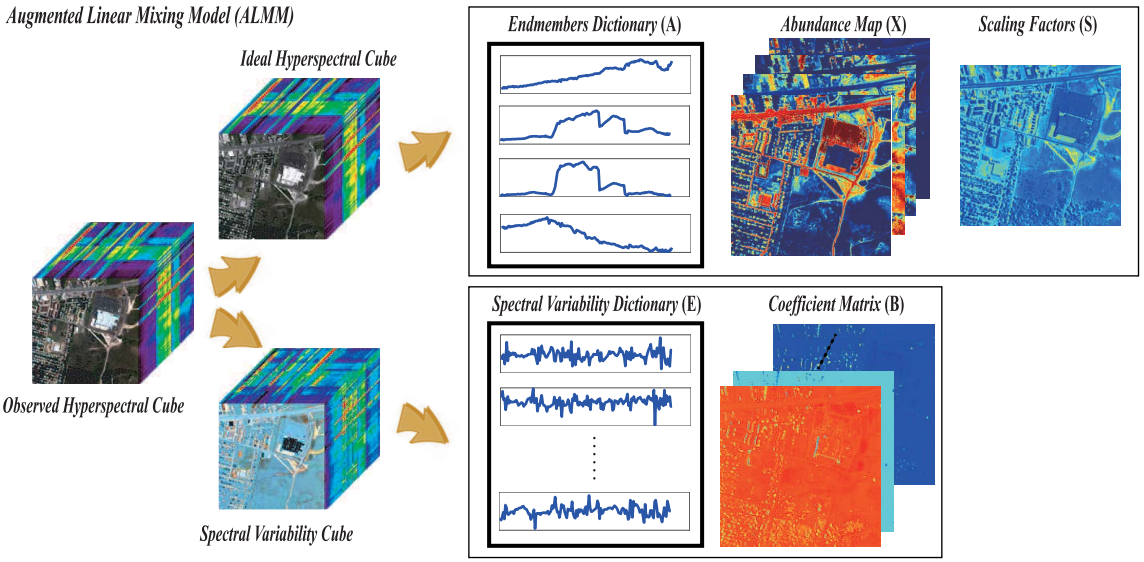


Fig. 2. The holistic diagram of spectral unmixing using the proposed ALMM.

the additional degrees of liberty[1] (several corrected examples can be found in Fig. 9). Additionally, more discussions and explanations have been done in [25] and [34]. The special case of the ELMM can be formulated as

$$\mathbf{y}_k = S_k(\mathbf{A}\mathbf{x}_k) + \mathbf{r}_k, \tag{8}$$

where S_k is a scalar in the k^{th} pixel that can be simply estimated using the regression between $\mathbf{y}_k$ and $\mathbf{A}\mathbf{x}_k$. Likewise, the matrix form of Eq. (8) can be written as

$$\mathbf{Y} = \mathbf{AXS} + \mathbf{R}, \tag{9}$$

where $\mathbf{S} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with its diagonal values $\mathbf{S}_k \succeq \mathbf{0}$.

In order to further overcome the shortcomings of the ELMM, which ignores the effects of other spectral variabilities, we extend the simplified ELMM to an augmented linear mixing model. This augmented linear mixing model, or ALMM, is expressed by

$$\mathbf{Y} = \mathbf{AXS} + \mathbf{EB} + \mathbf{R}, \tag{10}$$

where $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_m, ..., \mathbf{e}_L] \in \mathbb{R}^{D\times L}$ denotes the spectral variability matrix (or dictionary), and L is the number of basis vectors in $\mathbf{E}$. The expression $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_k, ..., \mathbf{b}_N] \in \mathbb{R}^{L\times N}$ is the coefficient matrix corresponding to $\mathbf{E}$.

Unlike the ELMM, where the spectral variability is modeled by endmember-wise scaling at each pixel, the ALMM represents the spectral signature by the endmember dictionary (i.e., **AXS**) with pixel-wise scaling and also spectral variabilities that cannot be explained using scaling by the spectral variability term (i.e., **EB**). On the other hand, unlike the PLMM, the ALMM gives an explicit physical consideration to scaling factors by inheriting concepts behind the ELMM, simultaneously modeling other variabilities by reasonable physical assumptions (see Subsection III-B for more details).

[1]For example, **EB** term which will be introduced in Eq. (10).

Fig. 2 gives the macro diagram of spectral unmixing using the proposed ALMM.

B. Problem Formulation

As introduced in Subsection III-A, the ALMM shown in Eq. (10) with a non-negativity constraint can be formulated as the following constrained optimization problem:

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{B},\mathbf{S},\mathbf{E}}{arg\,\min}\ \frac{1}{2}\|\mathbf{Y} - \mathbf{AXS} - \mathbf{EB}\|_{\mathrm{F}}^2 + \Phi(\mathbf{X}) + \Psi(\mathbf{B}) + \Upsilon(\mathbf{E}) \\ &\text{s.t. } \mathbf{X} \succeq 0, \quad \mathbf{S} \succeq 0, \end{aligned} \tag{11}$$

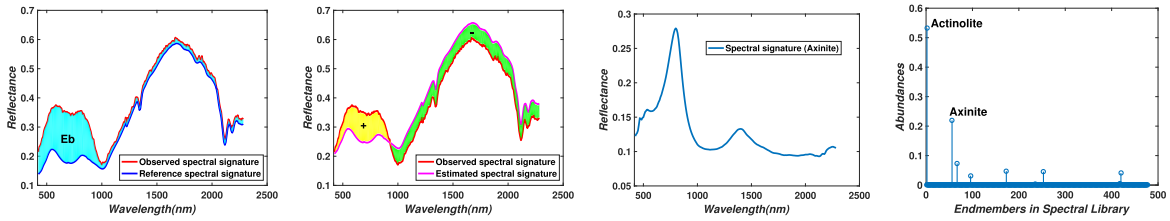
where the intent is to estimate the variables **X**, **S**, **E**, and **B**, while **A** is given. Since Eq. (11) is a typically ill-posed problem, several reasonable assumptions (or prior knowledge) should be introduced into the ALMM using regularization. Specifically, we defined three regularization functions Φ, Ψ, and Υ with respect to variables **X**, **B**, and **E**, respectively. The three regularization terms are described below.

1) Abundance Regularization $\Phi(\mathbf{X})$: In reality, a given spectral signature is usually composed of a limited number of materials in a hyperspectral scene, and hence the abundance regularization should be selected to be sparsity-prompting. In this paper, we applied $\|\mathbf{X}\|_{1,1} \equiv \sum_{k=1}^{N}\|\mathbf{x}_k\|_1$ to approximately estimate the sparsity-prompting term, which can be expressed with the penalty parameter α as

$$\Phi(\mathbf{X}) = \alpha\,\|\mathbf{X}\|_{1,1}. \tag{12}$$

2) Spectral Variability Coefficient Regularization $\Psi(\mathbf{B})$: Spectral variability is generally generated from various factors in a given hyperspectral scene. Except for scaling factors that can be modeled well by the endmember dictionary, the rest are diverse. To achieve a reliable generalization of our model, **E** should be regularized by a Frobenius Norm parameterized by β:

$$\Psi(\mathbf{B}) = \frac{\beta}{2}\|\mathbf{B}\|_{\mathrm{F}}^2. \tag{13}$$



(a) (b) (c) (d)

Fig. 3. An example in the real Cuprite scene to illustrate the physical meaning of **E**. (a) shows the differences (**Eb**) between the observed spectral signature and the real spectral signature that can not be explained by the endmember dictionary (**A**), but it can be represented well by an additional spectral variability dictionary (**E**). Correspondingly, if without **E**, the differences (spectral variability) could be absorbed by **A** as shown in (b), leading an inaccurate estimation of abundance maps (**X**). (c) gives a spectral signature of the material *Axinite* and (d) shows a real case of unmixing the observed spectral signature using USGS spectral library that except the *Actinolite*, the *Axinite* occupies the main abundances, which can well represents the **Eb** in (a).

3) Spectral Variability Dictionary Regularization $\Upsilon(\mathbf{E})$*:* To effectively find a better local optimal solution in our optimization problem, we acquire the variable **E** to be bounded by two prior knowledge assumptions: 1) the spectral variability dictionary (**E**) should be low-coherent with the endmember dictionary (**A**), formulated by $\frac{1}{2}\left\|\mathbf{A}^T\mathbf{E}\right\|_F^2$. 2) **E** should possess another property, making it possible for the basis vectors of **E** to be orthogonal, since such a dictionary can adequately represent various potential spectral variabilities. This makes the second prior assumption written by $\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1,j\neq i}^{L}\left\|\mathbf{e}_i^T\mathbf{e}_j\right\|_2^2$. Also, the constraint $\|\mathbf{e}_m\|_2^2 = 1(m = 1,...,L)$ should be satisfied in order to eliminate the trivial solution effectively; this second regularization term can be summarized as $\frac{1}{2}\left\|\mathbf{E}^T\mathbf{E}-\mathbf{I}\right\|_F^2$ (refer to [35] for more details regarding this term). The resulting expression of regularization with respect to **E** is

$$\Upsilon(\mathbf{E}) = \frac{\gamma}{2}\left\|\mathbf{A}^{\mathrm{T}}\mathbf{E}\right\|_{\mathrm{F}}^2 + \frac{\eta}{2}\left\|\mathbf{E}^{\mathrm{T}}\mathbf{E}-\mathbf{I}\right\|_{\mathrm{F}}^2, \tag{14}$$

where γ and η are the corresponding penalty parameters.

Moreover, non-negativity constraints ($\mathbf{X} \succeq \mathbf{0}$ and $\mathbf{S} \succeq \mathbf{0}$) usually have to be considered to satisfy the physical assumption. In addition to the non-negativity constraint, the sum-to-one also plays an important role in the abundance map. However, this constraint is not considered in our original problem [eq. (11)], since the variables **X** and **S** are bundled together, leading to difficultly satisfying the sum-to-one constraint for **X**. In the following section, we adopt the scaled constrained least squares unmixing (SCLSU) [24] technique to force **X** to follow the sum-to-one constraint.

C. Discussion on the Physical Significance of **E**

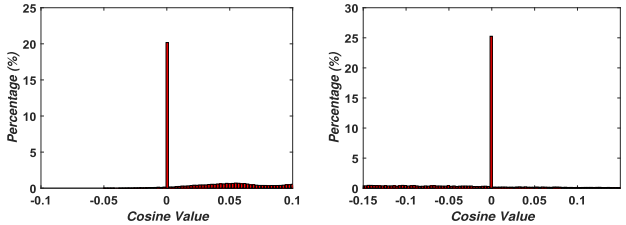
Followed by the instruction of $\Upsilon(\mathbf{E})$ shown in Eq. (14), we attempt to further discuss and explain the physical meaning of **E**, unfolded as follows:

On the one hand, although most spectral variabilities coherent with endmembers (**A**) can be represented by scaling factors, yet the remaining spectral variabilities, either intra-class or inter-class, can still hurt the performance of the spectral unmixing in reality. An example is illustrated in Fig. 3 to clarify that the spectral variability can not be fully explained by the scaled endmembers. The red curve in Fig. 3(a) is the observed spectral signature (*Actinolite*) extracted from the Cuprite scene and the blue one is the corresponding reference spectral signature (*Actinolite*) obtained from the USGS spectral library. Obviously, the differences (or spectral variabilities) between the two curves can not be well fit by the magenta curve, as shown in 3(b). Accordingly, we draw two points by reasoning as follows: 1) the scaled endmembers obtained by adding scaling factors on endmembers (**A**) fail to fully fit the gap in-between; 2) The errors marked in cyan of Fig. 3(a) could be explained by spectral variabilities or a certain new material. We try to identify the errors by means of the USGS spectral library, generating the abundances with respect to the various materials as shown in Fig. 3(d) where there is a rather high abundance in *Axinite* ranked as the second major component following *Actinolite*. In our model (ALMM), we represent the errors (or spectral variabilities) by an additional spectral (variability) dictionary (e.g., **E**). With the naked eye in Fig. 3(c), the spectral signature of the *Actinolite* yields a low-coherence with that of *Axinite* (A statistic will be given below.).

On the other hand, the physical significance of **E** could be also explained from the perspectives of intra-class and inter-class spectral variabilities. For instance, suppose only the existence of intra-class spectral variability that can be modeled by **E**, and thus the abundance maps **X** can be more accurately estimated by getting rid of the effects for the spectral variability (**EB**) that can not be explained by scaling factors, as shown in Figs. 1 and 3. Without **E**, the intra-class spectral variability could be absorbed by endmembers (**A**), further leading an inaccurate estimation of abundance maps (**X**). If **E** is considered as inter-class spectral variability dictionary, and then the term (**EB**) might represent the spectral signatures of certain new materials that are not discovered by the LMM (see Fig. 3 for example). The **E** used in the ALMM is therefore capable of calibrating the class-specific spectral variabilities into an unified or generalized spectral variability, which enables to simultaneously handle the intra- and inter-class variabilities. Fig. 4 shows a statistical evidence by collecting all cosine values between endmembers and spectral variabilities, where the cosine value is basically around 0, indicating that the spectral variability should, to a great extent, be low-coherent with the endmembers. This is basically consistent with the conclusion summarized above.

D. ADMM-Based Optimization Algorithm

In our case, the proposed ALMM framework can be roughly divided into two parts: *ALMM-based spectral unmix-*



(a) (b)

Fig. 4. Statistics of Cosine Value between endmembers and spectral variabilities on the first simulated dataset and real Urban scene, respectively, where the spectral variabilities are obtained by calculating the intra- and inter-class differences between the extracted endmembers and the given reference endmembers.

ing (SU) and *ALMM-based spectral variability dictionary learning (SVDL).*

1) ALMM-Based Spectral Unmixing: When $\mathbf{A}$ and $\mathbf{E}$ are given, Eq. (11) is naturally converted into a problem of spectral unmixing as

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{S},\mathbf{B}}{arg\min}\ \frac{1}{2}\|\mathbf{Y}-\mathbf{AXS}-\mathbf{EB}\|_{\mathrm{F}}^2+\Phi(\mathbf{X})+\Psi(\mathbf{B})\\ &\text{s.t. } \mathbf{X}\succeq 0,\quad \mathbf{S}\succeq 0. \end{aligned} \tag{15}$$

In order to conveniently and effectively collect all abundance vectors $\mathbf{X}$, we separately consider the problem (15) over the N pixels as pixel-wise spectral unmixing:

$$\begin{aligned} &\underset{\mathbf{x}_k,S_k,\mathbf{b}_k}{arg\min}\ \frac{1}{2}\|\mathbf{y}_k-(S_k\mathbf{A})\mathbf{x}_k-\mathbf{E}\mathbf{b}_k\|_2^2+\Phi(\mathbf{x}_k)+\Psi(\mathbf{b}_k)\\ &\text{s.t. } \mathbf{x}_k\succeq 0,\quad S_k\succeq 0. \end{aligned} \tag{16}$$

In [24], SCLSU is proposed to effectively solve the problem of scaled spectral unmixing. Equivalently, we formulate and solve the following nonnegative least squares (NNLS) problem [25], [36]:

$$\underset{\mathbf{x}_k\succeq 0}{arg\min}\ \frac{1}{2}\|\mathbf{y}_k-\mathbf{A}\mathbf{x}_k\|_2^2. \tag{17}$$

Once $\mathbf{x}_k$ is estimated by solving Eq. (17), then $\mathbf{x}_k$ and S_k can be simply derived, while satisfying the sum-to-one constraint with respect to $\mathbf{x}_k$ by

$$\hat{S}_k=\mathbf{1}^T\mathbf{x}_k,\quad \hat{\mathbf{x}}_k=\mathbf{x}_k/\mathbf{1}^T\mathbf{x}_k. \tag{18}$$

In the following, we will effectively embed this idea into our framework to satisfy the sum-to-one constraint with respect to $\mathbf{x}_k$ and update S_k simultaneously. Generally, the problem in (16) can be seen as a Constrained Bilinear Regression Problem (CBRP). A similar CBRP has been effectively solved by the ADMM optimization algorithm [37].

To facilitate an effective use of ADMM, we consider an equivalent form of (16) by introducing multiple auxiliary variables $\mathbf{g}_k$, $\mathbf{h}_k$ to replace $\mathbf{x}_k$, $\mathbf{x}_k^+$, respectively, where $()^+$ denotes an operator that converts each component of the matrix to its absolute value, and $l_R^+(\mathbf{x})$ is defined as $\mathbf{x}\succeq\mathbf{0}$.

$$\begin{aligned} &\underset{\mathbf{x}_k,S_k,\mathbf{b}_k,\mathbf{g}_k,\mathbf{h}_k}{arg\min}\ \frac{1}{2}\|\mathbf{y}_k-(S_k\mathbf{A})\mathbf{x}_k-\mathbf{E}\mathbf{b}_k\|_2^2\\ &\qquad+\Phi(\mathbf{g}_k)+\Psi(\mathbf{b}_k)+l_R^+(\mathbf{h}_k)\\ &\text{s.t. } S_k\succeq 0,\quad \mathbf{x}_k=\mathbf{g}_k,\ \mathbf{x}_k^+=\mathbf{h}_k,\ \mathbf{h}_k\succeq\mathbf{0}. \end{aligned} \tag{19}$$

Algorithm 1 ALMM-Based Pixel-Wise SU

Input: $\mathbf{y}_k, \mathbf{A}, \mathbf{E}$, and parameters $\alpha, \beta, maxIter$.
Output: $\mathbf{x}_k, S_k, \mathbf{b}_k$.
Initialization: $\mathbf{g}_k^0=\mathbf{h}_k^0=\mathbf{0}, S_k^0=1, \mathbf{b}_k^0=\mathbf{0}, \boldsymbol{\lambda}_k^0=\boldsymbol{\nu}_k^0=\mathbf{0}$, $\mu^0=1e-3, \mu_{max}=1e6, \rho=1.5, \varepsilon=1e-6, t=0$.
while *not converged* or $t>maxIter$ **do**
  Fix $\mathbf{b}_k^t, \mathbf{g}_k^t, \mathbf{h}_k^t$ to update $\mathbf{x}_k^{t+1}$ and S_k^{t+1} by (22-24);
  Fix $\mathbf{x}_k^{t+1}, S_k^{t+1}, \mathbf{g}_k^t, \mathbf{h}_k^t$ to update $\mathbf{b}_k^{t+1}$ by (26);
  Fix $\mathbf{x}_k^{t+1}, S_k^{t+1}, \mathbf{b}_k^{t+1}, \mathbf{h}_k^t$ to update $\mathbf{g}_k^{t+1}$ by (28);
  Fix $\mathbf{x}_k^{t+1}, S_k^{t+1}, \mathbf{b}_k^{t+1}, \mathbf{g}_k^{t+1}$ to update $\mathbf{h}_k^{t+1}$ by (31);
  Update Lagrange multipliers by $\boldsymbol{\lambda}_k^{t+1}=\boldsymbol{\lambda}_k^t+\mu^t(\mathbf{g}_k^{t+1}-\mathbf{x}_k^{t+1})$, $\boldsymbol{\nu}_k^{t+1}=\boldsymbol{\nu}_k^t+\mu^t(\mathbf{h}_k^{t+1}-\mathbf{x}_k^{t+1})$;
  Update penalty parameter by $\mu^{t+1}=\min(\rho\mu^t,\mu_{max})$;
  Check the convergence conditions:
  if $\|\mathbf{g}_k^{t+1}-\mathbf{x}_k^{t+1}\|_2<\varepsilon$ *and* $\|\mathbf{h}_k^{t+1}-\mathbf{x}_k^{t+1}\|_2<\varepsilon$ *and* $\|\mathbf{x}_k^{t+1}-\mathbf{x}_k^t\|_2<\varepsilon$ **then**
    Stop iteration;
  else
    $t\leftarrow t+1$;
  end
end

The augmented Lagrangian version of Eq. (19) is

$$\begin{aligned} &\mathscr{L}_U(\mathbf{x}_k,S_k,\mathbf{b}_k,\mathbf{g}_k,\mathbf{h}_k,\boldsymbol{\lambda}_k,\boldsymbol{\nu}_k)\\ &=\frac{1}{2}\|\mathbf{y}_k-(S_k\mathbf{A})\mathbf{x}_k-\mathbf{E}\mathbf{b}_k\|_2^2+\Phi(\mathbf{g}_k)+\Psi(\mathbf{b}_k)\\ &\quad+\boldsymbol{\lambda}_k^T(\mathbf{g}_k-\mathbf{x}_k)+\boldsymbol{\nu}_k^T(\mathbf{h}_k-\mathbf{x}_k)+\frac{\mu}{2}\|\mathbf{g}_k-\mathbf{x}_k\|_2^2\\ &\quad+\frac{\mu}{2}\|\mathbf{h}_k-\mathbf{x}_k\|_2^2+l_R^+(\mathbf{h}_k), \end{aligned} \tag{20}$$

where $\boldsymbol{\lambda}_k$, $\boldsymbol{\nu}_k$, and $\boldsymbol{\pi}_k$ are Lagrange multipliers and μ is the penalty parameter.The resulting algorithm of ALMM-based pixel-wise SU is detailed in **Algorithm 1**, and the solution to each subproblem is given in Appendix A. Correspondingly, variables $\mathbf{X}, \mathbf{B}$, and $\mathbf{S}$ for all pixels can be collected using ALMM-based pixel-wise SU in turn.

2) ALMM-Based Spectral Variability Dictionary Learning: If and only when $\mathbf{E}$ is unknown in Eq. (11), we have to simultaneously perform spectral unmixing and dictionary learning using *ALMM-based SVDL*, resulting in alternately updating variables $\mathbf{X}$, $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{S}$. This task essentially guides us to solve the optimization problem in Eq. (11). Facing such a multi-variable optimization problem, we once again explore the ADMM algorithm for a fast and effective solution.It is noteworthy that concurrently estimating variables $\mathbf{X}$, $\mathbf{E}$, $\mathbf{B}$, and $\mathbf{S}$ in each iteration is of benefit to provide us a broader solution space and further find a better local minimum close to the global one easier.

By introducing multiple auxiliary variables $\mathbf{G}$, $\mathbf{H}$, $\mathbf{M}$, $\mathbf{T}$, and $\mathbf{Q}$ to replace $\mathbf{X}$, $\mathbf{X}^+$, $\mathbf{XS}$, $\mathbf{S}^+$, and $\mathbf{E}$, respectively, the augmented Lagrangian function of Eq. (11) can be written as

$$\begin{aligned} &\mathscr{L}_D(\mathbf{X},\mathbf{S},\mathbf{E},\mathbf{B},\mathbf{G},\mathbf{H},\mathbf{M},\mathbf{T},\mathbf{Q},\boldsymbol{\Lambda},\boldsymbol{V},\boldsymbol{\Omega},\boldsymbol{\Pi},\boldsymbol{\Delta})\\ &=\frac{1}{2}\|\mathbf{Y}-\mathbf{AM}-\mathbf{EB}\|_{\mathrm{F}}^2+\Phi(\mathbf{G})+\Psi(\mathbf{B})+\Upsilon(\mathbf{Q})\\ &\quad+\boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{G}-\mathbf{X})+\boldsymbol{V}^{\mathrm{T}}(\mathbf{H}-\mathbf{X})+\boldsymbol{\Pi}^{\mathrm{T}}(\mathbf{Q}-\mathbf{E})\\ &\quad+\boldsymbol{\Omega}^{\mathrm{T}}(\mathbf{M}-\mathbf{XS})+\boldsymbol{\Delta}^{\mathrm{T}}(\mathrm{T}-\mathbf{S})+\frac{\xi}{2}\|\mathbf{G}-\mathbf{X}\|_{\mathrm{F}}^2\\ &\quad+\frac{\xi}{2}\|\mathbf{H}-\mathbf{X}\|_{\mathrm{F}}^2+\frac{\xi}{2}\|\mathbf{Q}-\mathbf{E}\|_{\mathrm{F}}^2+\frac{\xi}{2}\|\mathbf{M}-\mathbf{XS}\|_{\mathrm{F}}^2\\ &\quad+\frac{\xi}{2}\|\mathbf{T}-\mathbf{S}\|_{\mathrm{F}}^2+l_R^+(\mathbf{H})+l_R^+(\mathbf{T}), \end{aligned} \tag{21}$$

Algorithm 2 ALMM-Based SVDL

Input: $\mathbf{Y}$, $\mathbf{A}$ and parameters $\alpha, \beta, \gamma, \eta, maxIter$.
Output: $\mathbf{E}, \mathbf{X}, \mathbf{S}, \mathbf{B}$.
Initialization: $\mathbf{G}^0 = \mathbf{H}^0 = \mathbf{M}^0 = \mathbf{0}, \mathbf{S}^0 = \mathbf{I}, \mathbf{B}^0 = \mathbf{0}, \mathbf{\Delta}^0 = \mathbf{0}$, $\mathbf{\Lambda}^0 = \boldsymbol{V}^0 = \mathbf{\Omega}^0 = \mathbf{0}, \mathbf{Q}^0 = \mathbf{0}, \mathbf{T}^0 = \mathbf{0}, \mathbf{\Pi}^0 = \mathbf{0}, \mathbf{X}^0, \mathbf{E}^0, t = 0$, $\xi^0 = 1e-3, \xi_{max} = 1e6, \rho = 1.5, \varepsilon = 1e-6$.
while *not converged* or $t > maxIter$ **do**
  Fix $\mathbf{E}^t, \mathbf{B}^t, \mathbf{X}^t, \mathbf{S}^t$ to update $\mathbf{M}^{t+1}$ by (35);
  Fix $\mathbf{E}^t, \mathbf{M}^{t+1}$ to update $\mathbf{B}^{t+1}$ by (36);
  Fix $\mathbf{G}^t, \mathbf{H}^t, \mathbf{S}^t, \mathbf{M}^{t+1}, \mathbf{\Lambda}^t, \boldsymbol{V}^t$ to update $\mathbf{X}^{t+1}$ by (38-39);
  Fix $\mathbf{M}^{t+1}, \mathbf{X}^{t+1}, \mathbf{T}^t, \mathbf{\Pi}^t, \mathbf{\Delta}^t$ to update $\mathbf{S}^{t+1}$ by (41);
  Fix $\mathbf{M}^{t+1}, \mathbf{B}^{t+1}, \mathbf{Q}^t, \mathbf{\Pi}^t$ to update $\mathbf{E}^{t+1}$ by (43);
  Fix $\mathbf{E}^t, \mathbf{E}^{t+1}, \mathbf{\Pi}^t$ to update $\mathbf{Q}^{t+1}$ by (45);
  Fix $\mathbf{X}^{t+1}, \mathbf{\Lambda}^t$ to update $\mathbf{G}^{(t+1)}$ by (46);
  Fix $\mathbf{X}^{t+1}, \boldsymbol{V}^t$ to update $\mathbf{H}^{(t+1)}$ by (47);
  Fix $\mathbf{S}^{t+1}, \mathbf{\Delta}^t$ to update $\mathbf{T}^{(t+1)}$ by (48);
  Update Lagrange multipliers by
    $\mathbf{\Lambda}^{t+1} \leftarrow \mathbf{\Lambda}^t + \xi^t(\mathbf{G}^{t+1} - \mathbf{X}^{t+1})$
    $\boldsymbol{V}^{t+1} \leftarrow \boldsymbol{V}^t + \xi^t(\mathbf{H}^{t+1} - \mathbf{X}^{t+1})$
    $\mathbf{\Omega}^{t+1} \leftarrow \mathbf{\Omega}^t + \xi^t(\mathbf{M}^{t+1} - \mathbf{X}^{t+1}\mathbf{S}^{t+1})$
    $\mathbf{\Pi}^{t+1} \leftarrow \mathbf{\Pi}^t + \xi^t(\mathbf{Q}^{t+1} - \mathbf{E}^{t+1})$
    $\mathbf{\Delta}^{t+1} \leftarrow \mathbf{\Delta}^t + \xi^t(\mathbf{T}^{t+1} - \mathbf{S}^{t+1})$
  Update penalty parameter by $\xi^{(t+1)} = \min(\rho\xi^{(t)}, \xi_{max})$;
  Check the convergence conditions:
  if $\|\mathbf{G}^{t+1} - \mathbf{X}^{t+1}\|_F < \varepsilon$ *and* $\|\mathbf{H}^{t+1} - \mathbf{X}^{t+1}\|_F < \varepsilon$ *and* $\|\mathbf{M}^{t+1} - \mathbf{X}^{t+1}\mathbf{S}^{t+1}\|_F < \varepsilon$ *and* $\|\mathbf{Q}^{(t+1)} - \mathbf{E}^{(t+1)}\|_F < \varepsilon$ *and* $\|\mathbf{T}^{t+1} - \mathbf{S}^{t+1}\|_F < \varepsilon$ *and* $\|\mathbf{E}^{t+1} - \mathbf{E}^t\|_F < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
  end
end

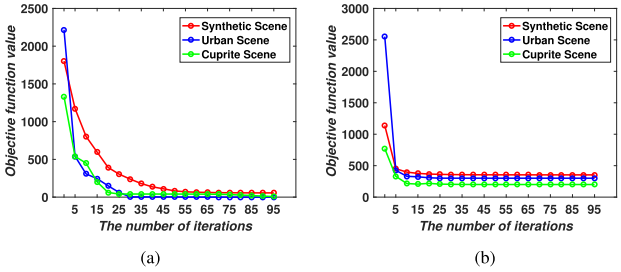


Fig. 5. Convergence analysis of ALMM are experimentally performed on three different datasets. (a) ALMM-based SU. (b) ALMM-based SVDL.

where $\mathbf{\Lambda}$, $\boldsymbol{V}$, $\mathbf{\Omega}$, $\mathbf{\Pi}$, and $\mathbf{\Delta}$ are Lagrange multipliers and ξ is the penalty parameter.

The proposed algorithm for dictionary learning is summarized in **Algorithm 2** (see Appendix B for more details), where careful initialization is necessary in our case since the optimization problem of dictionary learning is not convex. The abundances generated by the SCLSU algorithm are set as the initial value ($\mathbf{X}^0$) and a random orthogonal matrix is produced to the initialization of the spectral variability dictionary ($\mathbf{E}^0$).

E. Convergence and Computational Cost

The alternating scheme used in **Algorithm 1** and **Algorithm 2** is a typical multi-block ADMM optimization problem, whose convergence is theoretically supported in [38] and [39]. Moreover, similar work for solving this sort of multi-block ADMM-based optimization problem has been successfully applied in [40]–[42]. We experimentally visualize the convergence results for *ALMM-based SU* and *ALMM-based SVDL* on the three datasets, where the objective function value is recorded in each iteration (see Fig. 5). Notably, we collect the objective function values of all pixels computed by *ALMM-based pixel-wise SU* for obtaining *ALMM-based SU*'s.

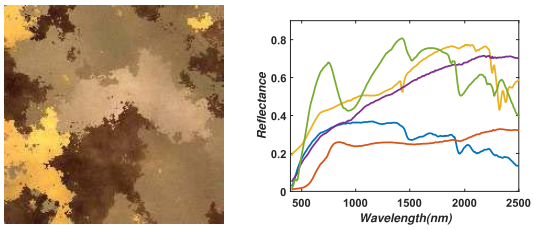


Fig. 6. A false color image of the synthetic data and five endmembers extracted by VCA. (a) A false color image. (b) Endmembers.

We can clearly observe from Appendix A and B that the computational cost of our method is dominated by matrix products, yielding an overall $\mathcal{O}(DLN)$ w.r.t. *ALMM-based SU* and $\mathcal{O}(DL^2N)$ w.r.t. *ALMM-based SVDL*, respectively.

IV. Experiments

In this section, we quantitatively and visually evaluate the performance of the proposed method on three datasets: a synthetic dataset presented in [25] and two real datasets over an urban area and the mining district in Cuprite, Nevada. We compare the proposed method (ALMM) with conventional and state-of-the-art approaches, including fully constrained least squares unmixing (FCLSU), constrained least squares unmixing (CLSU), scaled constrained least squares unmixing (SCLSU),[2] SUnSAL (ℓ_1-CLSU), SSUnSAL (scaled SUnSAL), as well as PLMM and ELMM. Since the different regularization parameters lead to different results for each algorithm, we empirically and experimentally set up them to maximize performance. Specifically, we set the penalty parameter of the sparsity-promoting term to be $6e-3$ in both SUnSAL and SSUnSAL, while three regularization parameters for abundances, endmembers, and perturbation in the PLMM are set to be $1e-2$, $1e-2$, and 1, respectively. The regularization parameter λ_S in the ELMM is set to be 0.5. In the following experiments, we fix a display range for the abundance maps, e.g., [0, 1] for Fig. **??** and [0, 0.5] for Fig. 14, in the interest of making fair visual comparisons. It should be noted that there are some abundances that show the maximum of the display range but actually exceed it, since they are generated by those algorithms without considering the scaling factors.

A. Synthetic Data

1) Data Description: The synthetic data was simulated using five reference endmembers with 224 spectral bands randomly selected from the United States Geological Survey (USGS) spectral library and 200×200 abundance maps generated using Gaussian fields, which were designed to satisfy the ANC and the ASC. Fig. 6 shows a false color

[2]Without the term of **EB**, our model (ALMM) is equivalent to SCLSU.

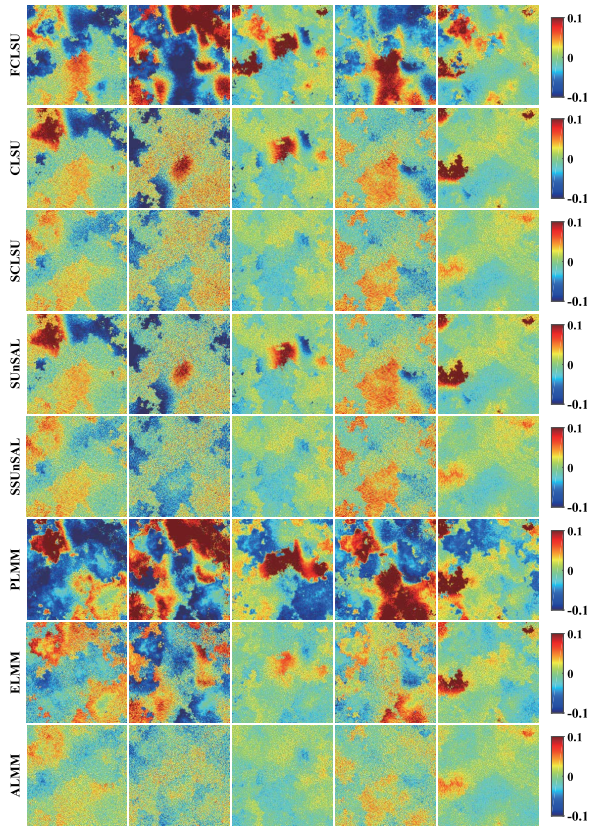


Fig. 7. The difference abundance maps using different spectral unmixing methods.

image of the synthetic data and five selected endmembers. It should be noted that a spectral signature of each pixel in this dataset includes spectral variability due to endmember-dependent scaling factors and complex noise. Specifically, given five reference endmembers, we respectively multiply those spectral signatures by randomly-generated scaling factors ranging in [0.75, 1.25] and then a 25dB white Gaussian noise was added to these scaled reference endmembers. Next, we follow the LMM to mix them by means of generated abundance maps, finally a 25dB white Gaussian was added to these mixed pixels again. Following this simulation process, the generated spectral variabilities can be explained - without considering scaling factors - using a special mixtures of Gaussian distributions. Therefore, this simulated data with such spectral variability will give us a proper scenario to validate the proposed approach.

2) Experimental Setup: For a fair comparison, we adopt VCA to construct the endmember dictionary for all algorithms (including the proposed ALMM) under comparison, while Hysime [43] is used to estimate the number of endmembers. Endmember identification can be effectively performed with the spectral angle and five reference endmembers. Note that we show the averaged results for the different algorithms out of 10 runs, because VCA cannot always guarantee the same estimations in each round.

Importantly, a good initialization leads to a reasonable solution in our optimization problem due to nonconvexity. We hereafter initialize the abundance maps ($\mathbf{X}^0$) using the result of SCLSU. For the setting of other parameters, please refer to **Algorithm 1** and **Algorithm 2** for details.

For the performance assessment of the algorithms, we introduce three criteria to quantify experimental results: abundance overall root mean square error (aRMSE), reconstruction overall root mean square error (rRMSE), and average spectral angle mapper (aSAM). When the groundtruth of abundance maps is given, aRMSE can be defined as

$$aRMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{P}\sum_{p=1}^{P}(\mathbf{x}_{kp} - \hat{\mathbf{x}}_{kp})^2}. \tag{22}$$

Without the groundtruth of abundance maps, we can also give the two measures for assessing the performance of the algorithms from the point of view of data reconstruction. One of these measurements is rRMSE, defined by

$$rRMSE = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{D}\sum_{l=1}^{D}(\mathbf{y}_{kd} - \hat{\mathbf{y}}_{kd})^2}, \tag{23}$$

while the other is aSAM, expressed as

$$aSAM = \frac{1}{N}\sum_{k=1}^{N} arccos\left(\frac{\mathbf{y}_k^T\hat{\mathbf{y}}_k}{\|\mathbf{y}_k\|\,\|\hat{\mathbf{y}}_k\|}\right). \tag{24}$$

Parameters Setting. As the performance of the proposed ALMM model is fairly sensitive to the setting of four regularization parameters α, β, γ, and η as well as the number of basis vectors (L) of $\mathbf{E}$, it is, therefore, indispensable to investigate the parameters setting in a proper range. For this reason, we attempt to find a group of stable and effective parameters by conducting several experiments on three different datasets (synthetic scene, urban scene, and Cuprite scene), as specifically shown in Fig. 8 where we can empirically observe that α plays a dominant role in estimating the abundance maps ($\mathbf{X}$), while for other parameters (β, γ, η, and L) the importance of L is visibly higher than that of the rest ones. With the increase of L, the algorithm performance is gradually improved until to around middle and then reaches a relatively stable state. An optimal performance can be obtained by setting these parameters as $\alpha = \beta = 2e-3$, $\gamma = \eta = 5e-3$, and $L = 100$ in the synthetic scene, $\alpha = \beta = 5e-2$, $\gamma = \eta = 1e-2$, and $L = 80$ in the urban scene, and $\alpha = 1e-2$, $\beta = 5e-2$, $\gamma = 5e-2$, $\eta = 1e-2$, and $L = 90$ in the Cuprite scene. Accordingly, we can empirically summarize a general trend for the parameter-setting, that is, for regularization parameters (α, β, γ, and η), they can be basically chosen in the range from $1e-3$ to $1e-2$, while L tends to be approximately assigned to one half of the spectral length.

3) Results and Analysis: Table I details the corresponding quantitative assessment results. Since the visual difference of the estimated abundance maps is not obvious among some of the algorithms, the abundance difference maps are also given in Fig. 7 to intuitively highlight the difference.

As can be seen in Table I, FCLSU yields a poor performance due to the presence of spectral variability. CLSU performs better than FCLSU since the abundances can be reasonably estimated in a cone not in a simplex by dropping the ASC. However, the spectral variability is not actually eliminated

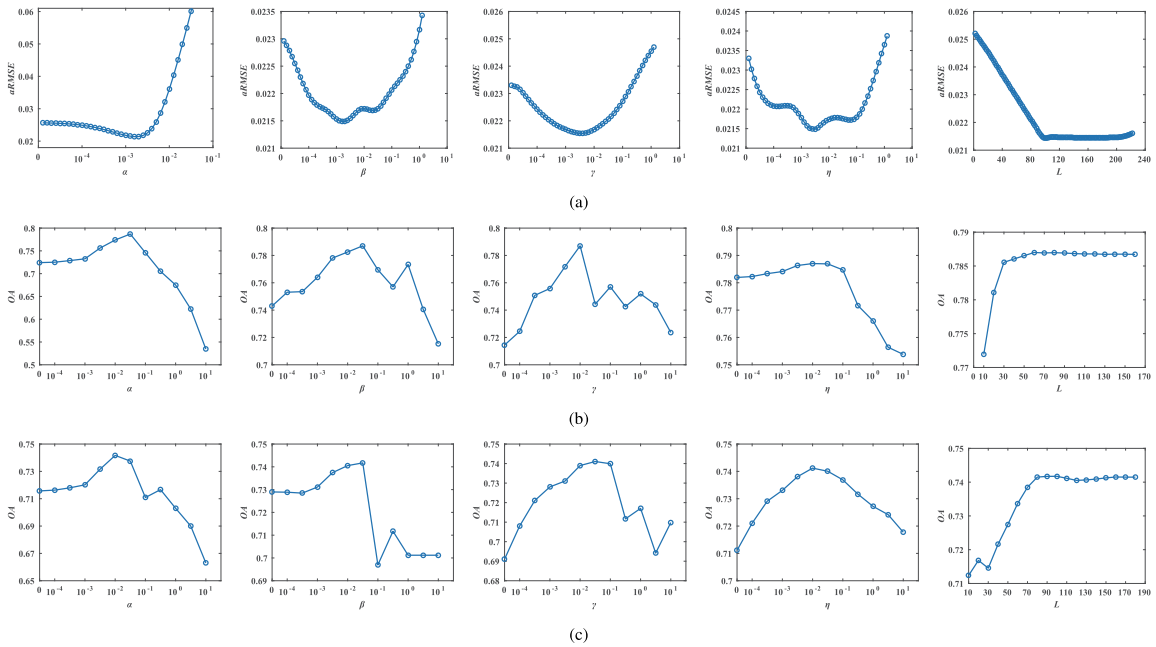


Fig. 8. Parameter sensitivity analysis of the proposed ALMM algorithm on three different study scenes for four regularization parameters: α, β, γ, and η as well as the number of basis vectors (L) of $\mathbf{E}$. (a) Synthetic Scene. (b) Urban Scene. (c) Cuprite Scene

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON THE SYNTHETIC DATA. THE BEST ONE IS SHOWN IN BOLD

Algorithm	FCLSU	CLSU	SUnSAL	SCLSU	SSUnSAL	PLMM	ELMM	ALMM
aRMSE (10^{-2})	6.30±0.70	4.21±0.20	3.99±0.33	2.63±0.23	2.43±0.16	6.21±0.43	3.23±0.28	**2.15±0.17**
rRMSE (10^{-2})	1.50±0.024	1.23±0.0012	1.23±0.0012	1.23±0.0012	1.23±0.0012	1.29±0.0037	0.58±0.0005	**0.018±0.00004**
aSAM (10^{-2})	198.36±2.54	177.17±0.016	177.26±0.018	177.17±0.016	177.26±0.018	184.27±1.09	83.92±0.007	**1.04±0.0002**

by CLSU, but rather absorbed by the abundances. With the consideration of the ASC, SCLSU performs better than CLSU, particularly being robust against scaling factors. By adding the sparsity term, SUnSAL and SSUnSAL can further improve the performance compared to CLSU and SCLSU without sparsity term, which experimentally explains that each pixel in the studied hyperspectral scene consists of a few materials.

Although the ELMM approach can model scaling variability with reasonable physical consideration, the difficult parameter estimation results in its limited performance. More specifically, the objective function of the ELMM is an obvious non-convex problem, since the scaling factors and abundance maps need to be estimated simultaneously, which easily drops to a local minimum and leads to the inaccurate estimation of the abundances and scaling factors. Generally, the scaling factors among different endmembers are highly correlated in reality, because the endmember variability is often dominated by the geometry effect; this is another factor that hinders the improvement of the performance of the ELMM. PLMM fails to specify the spectral variabilities (e.g., scaling factors) according to their properties.

By comparison, the proposed method outperforms other algorithms, which suggests that this method can effectively learn the spectral variability, improving the accuracy of the abundance estimation. Fig. 7 illustrates a more significant comparison by means of abundance difference maps between the groundtruth and estimated abundance maps of the compared algorithms. The difference values obtained from ALMM are mostly close to zero, which indicates that the performance of ALMM is superior to that of the other methods.

For the purpose of highlighting the learnability for spectral variabilities, we emphatically investigate several typical examples of learned spectral variabilities by giving the four spectral signatures under different conditions, as shown in Fig. 9, including

- the observed spectral signature ($\mathbf{y}_k$) and the reconstructed spectral signature ($\mathbf{A}\hat{\mathbf{x}}_k\hat{S}_k + \hat{\mathbf{E}}\hat{\mathbf{b}}_k$),
- the truth spectral signature ($\mathbf{A}\mathbf{x}_k$) and the estimated spectral signature ($\mathbf{A}\hat{\mathbf{x}}_k$),
- the truth scaled spectral signature ($\mathbf{A}\mathbf{x}_k S_k$) and the estimated scaled spectral signature ($\mathbf{A}\hat{\mathbf{x}}_k\hat{S}_k$),
- spectral variability ($|1 - S_k|\mathbf{A}\mathbf{x}_k + \mathbf{E}\mathbf{b}_k$) and learned spectral variability ($|1 - \hat{S}_k|\hat{\mathbf{A}}\hat{\mathbf{x}}_k + \hat{\mathbf{E}}\hat{\mathbf{b}}_k$) and
- spectral variability without scaling ($\mathbf{E}\mathbf{b}_k$) and learned spectral variability without scaling ($\hat{\mathbf{E}}\hat{\mathbf{b}}_k$).

Using ALMM, scaling factors can be approximately fit by S_k, as shown in Fig. 9(a), while for those spectral variabilities that cannot be fully explained by scaling factors, $\mathbf{E}\mathbf{b}_k$ can correct them (see Fig. 9(b)). More discussion can be detailed as follows

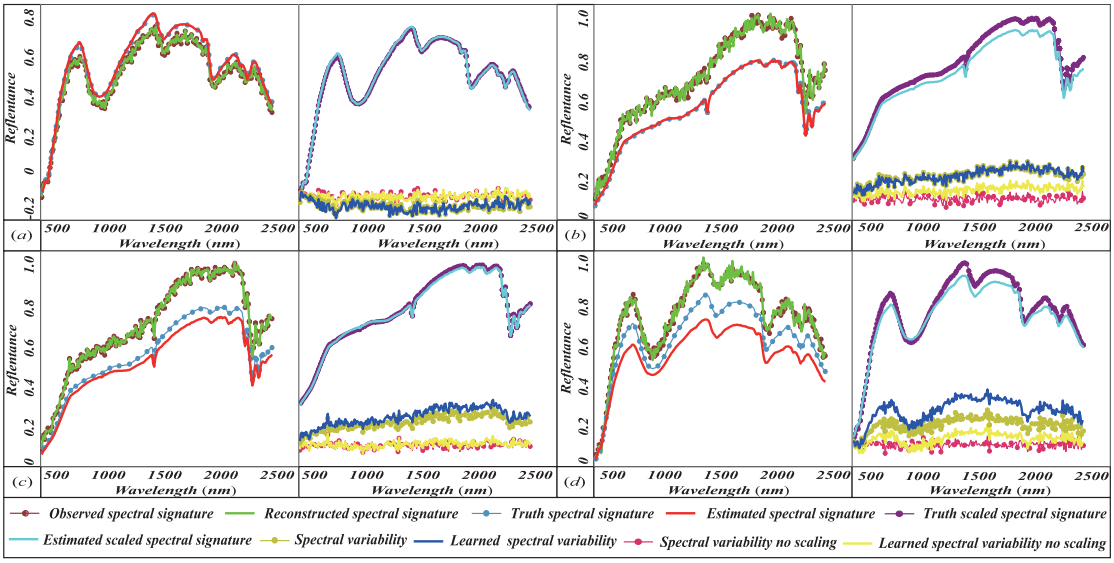
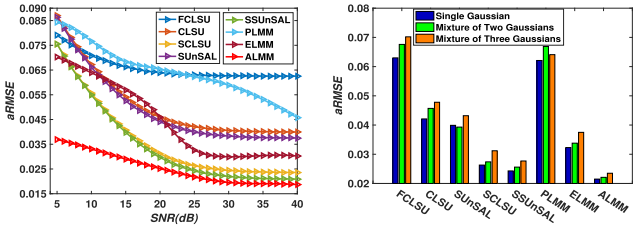


Fig. 9. Highlight some typical cases with respect to learned spectral variabilities. (a): ALMM not only can reconstruct the observed spectral signature well but also learn the various spectral variabilities (e.g., scaling factors and other complex variabilities) effectively. (b): In this case, although the scaling factors fail to be estimated well, yet the term of **EB** effectively fix the errors, still leading to a desirable abundances estimation. (c) shows a bad example in estimating scaling factors, while (d) gives a failure case of learning spectral variabilities. Please refer to the fifth paragraph in Section IV.A(3) for more analysis and discussion.

- Fig. 9(a) shows the expected competitive result: the spectral signature and learned spectral variabilities basically match the real ones. We have to point out that most of the pixels in this simulated data follow the expected results.
- Fig. 9(b) shows another case where each endmember in the given mixed pixel is not sharing similar scalar, so it would fail to estimate the scaling factors accurately. In such a case, however, the abundance map can be estimated well, since the spectral variability term ($\mathbf{Eb}_k$) can effectively represent the rest of the spectral variabilities that cannot be explained by a shared scaling factor, as displayed in the curves of the second figure of Fig. 9(b).
- As can be seen in Fig. 9(c), although our model can learn spectral variability without scaling factors well, it fails to effectively estimate the truth spectral signature, further causing inaccurate abundance maps. This probably results from the inaccurate estimation of scaling factors.
- A counterexample that cannot handle the spectral variability is given in Fig. 9(d). Such a negative example is unexpected but reasonable due to non-convexity, which masks it difficult for our model to precisely estimate all variables. Although proper prior assumptions and the endmember dictionary extracted by VCA are used in our model, they can only shrink the range of solutions rather than giving the globally optimal solution directly.

4) Robustness Study: To quantitatively validate the robustness of our method, we investigate the performances (aRMSEs) of the different algorithms on simulated data by adding Gaussian white noises with the different signal-to-noise-ratio (SNR) ranging from 5dB to 40dB at a 1dB interval. As can be clearly seen from Fig. 10(a) that ALMM is more robust and effective against noises with the different SNRs, compared to others. Also, we experimentally discuss another



(a) (b)

Fig. 10. Robustness analysis using single Gaussian noise with the different SNRs and mixtures of multiple Gaussian noises. (a) Single Gaussian noise (different SNRs). (b) Mixtures of Multiple Gaussian noises.

case of mixed Gaussian distributions as the noise input. More specifically, mixtures of Gaussian distributions can be generated by assembling several single Gaussian distributions with the different mean and variance randomly selected from 0 to 0.01. Using them (single Gaussian, mixtures of two Gaussian, and mixtures of three Gaussian), we horizontally compare the performance of different algorithms by the averaged aRMSEs out of 20 runs to achieve the reliable results. It is clear in Fig. 10(b) that the ALMM performs better and more robust against the Gaussian mixture noises than other comparative algorithms.

B. First Real Data (Urban)

1) Data Description: This dataset was collected by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) over an urban area at Copperas Cove, Texas, USA. The dataset has been widely used in the field of hyperspectral unmixing [28]–[30]. The latest data version was issued by Geospatial Research Laboratory (USA) and Engineer Research and Development Center (USA) in 2015.[3] The image consists

[3]http://www.tec.army.mil/Hypercube

TABLE II

QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON THE URBAN DATA. THE BEST ONE IS SHOWN IN BOLD

Algorithm	FCLSU	CLSU	SUnSAL	SCLSU	SSUnSAL	PLMM	ELMM	ALMM
OA (%)	54.66±7.59	68.08±5.24	71.26±5.50	68.08±5.24	71.26±5.50	58.55±7.21	62.41±6.87	**78.70±2.83**
rRMSE (10^{-2})	3.97±0.32	0.86±0.085	0.86±0.085	0.86±0.085	0.86±0.085	1.11±0.20	0.72±0.048	**0.27±0.0003**
aSAM (10^{-2})	867.34±50.14	295.69±26.19	295.82±32.03	295.69±26.19	295.82±32.03	362.40±60.04	203.78±19.95	**58.02±1.81**

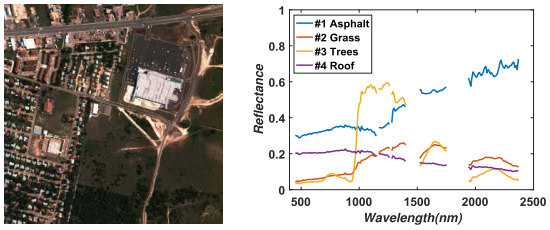


Fig. 11. A false color image of the Urban data and four endmembers used in spectral unmixing. (a) A false color image. (b) Endmembers.

of 307×307 pixels with 210 spectral bands in the wavelength from 400 nm to 2500 nm with spectral resolution of 10 nm at a ground sampling distance (GSD) of 2 m. Fig. 11(a) shows a false color image of the study scene. Due to water absorption and atmospheric effects, we reduced 210 bands to 162 bands by removing bands 1-4, 76, 87, 101-111, 136-153, and 198-210.

2) Experimental Setup: Four main endmembers can be observed in the scene: asphalt (road and parking lot), grass, trees, and roof. For more discussion and analysis regarding these endmembers, refer to [28] and [30]. Likewise, VCA and HySime are adopted to build the endmember dictionary and determine the number of endmembers for all algorithms (including ALMM), respectively. Fig. 11(b) shows the endmembers used in spectral unmixing. Furthermore, the material identification step is performed through comparison with the reference endmembers.[4]

3) Results and Analysis: For the quantitative assessment of the experimental results, we calculate the two indices, rRMSE and aSAM. Since there is no groundtruth of the abundance maps for the real data and meanwhile the metrics based on reconstruction errors (rRMSE and aSAM) are not suitable to assess the performance of spectral unmixing. For these reasons, we propose a classification-based evaluation strategy for assessing the abundance maps using the overall accuracy (OA). Firstly, we perform the spectral angle mapper (SAM) classification using the reference endmembers as reference spectra. The first row of Fig. 12 shows the cosine similarity for the four classes, where negative samples are masked out with 0. For the spectral unmixing results, we obtain classification maps by classifying each pixel into an endmember that has the maximum abundance value. By using the SAM classification result as the groundtruth, OA can be calculated for the different methods, as listed in Table II.

We also perform a visual examination to evaluate the performance of the algorithms for the estimation of abundance

[4]The reference endmembers are manually extracted from the original image. Please refer to [29] and [30] for details.

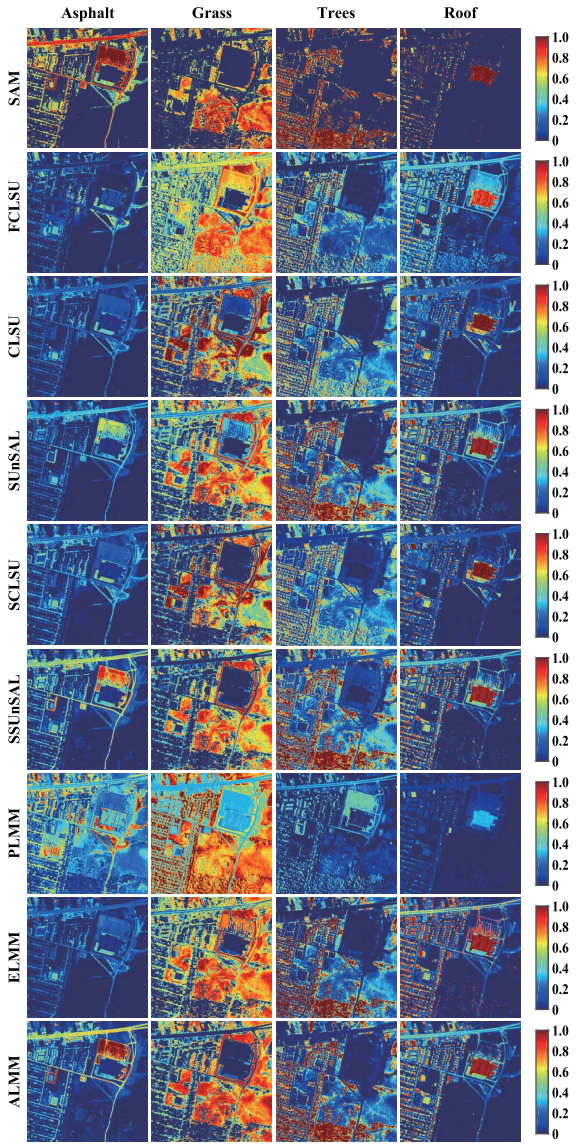
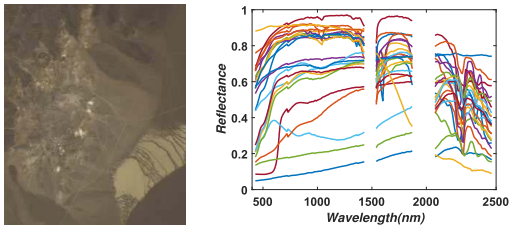


Fig. 12. The abundance maps comparison between the proposed method and the state-of-the-art methods.

maps. According to the quantitative and visual results, we analyze the performance of the different algorithms as follows. FCLSU performs rather poor estimation for the abundances, since spectral variability comes into play in the real data. Similarly, CLSU also fails to deal with spectral variability; however it outperforms FCLSU, as shown visually in the Fig. 12 and Table II, due to the relaxation of the ASC. When scaling factors are considered, there is better identification of the materials of asphalt, trees, and roof using SCLSU.



(a) (b)

Fig. 13. A false color image of the Cuprite data and the endmember dictionary constructed by spectral library. (a) A false color image. (b) Endmembers.

In particular, FCLSU and CLSU both fail to detect the material of asphalt, but SCLSU effectively does.

Although ELMM is able to detect some areas where only one scaling factor presents a difficulty for interpreting all endmembers and meanwhile obtains a relatively lower rRMSE and aSAM as listed in Table II, the non-convexity involved in the simultaneous estimation of the abundance maps and scaling factors prevents ELMM from achieving better performance (lower CMMS than that of CLSU and SCLSU). As shown in the third row of Fig. 12, ELMM obtains a purer identification for the materials of trees and roof, while there is still room for improvement in its abundance estimation of asphalt and grass. In Fig. 12, the performance of PLMM is relatively poor because it is not able to address the scaling factors, which is the main spectral variability in the study scene.

In this scene, there are many pure pixels, owing to high resolution; however, they are considered mixed pixels in the comparison of methods due to the existence of spectral variability. As shown in Fig. 12, the visual performance of the proposed ALMM method is superior to the other methods and consistent numerical evaluation is listed in Table II as well. More specifically, the asphalt is purely identified by ALMM, unlike the others; and a similar observation can be found in the grass as well. For the trees and the roof, the abundance maps estimated by ALMM show higher contrast than those estimated by other methods. This result implies that the proposed method successfully addresses spectral variability.

In order to further highlight the differences between the proposed method and CLSU, SCLSU, SUnSAL, and SSUnSAL, we emphatically focus on their abundance maps. Each material, by and large, becomes more purely identified and the corresponding abundance maps clearer by successively using CLSU, SCLSU, SUnSAL, SSUnSAL, and the proposed method. Observing each method's abundance maps separately, without considering the scaling factors, CLSU and SUnSAL encounter similar troubles, where the abundances generally exceed 1 as shown in the second and fourth rows of Fig. 12, which leads to difficulty distinguishing the abundance maps of CLSU and SUnSAL. Once the scaling factors are considered, SCLSU immediately shows a competitive result, although it still cannot match SSUnSAL, especially in the identification of asphalt and grass. Pure material identification and clear abundance maps are the more reasonable and desirable results given by our model.

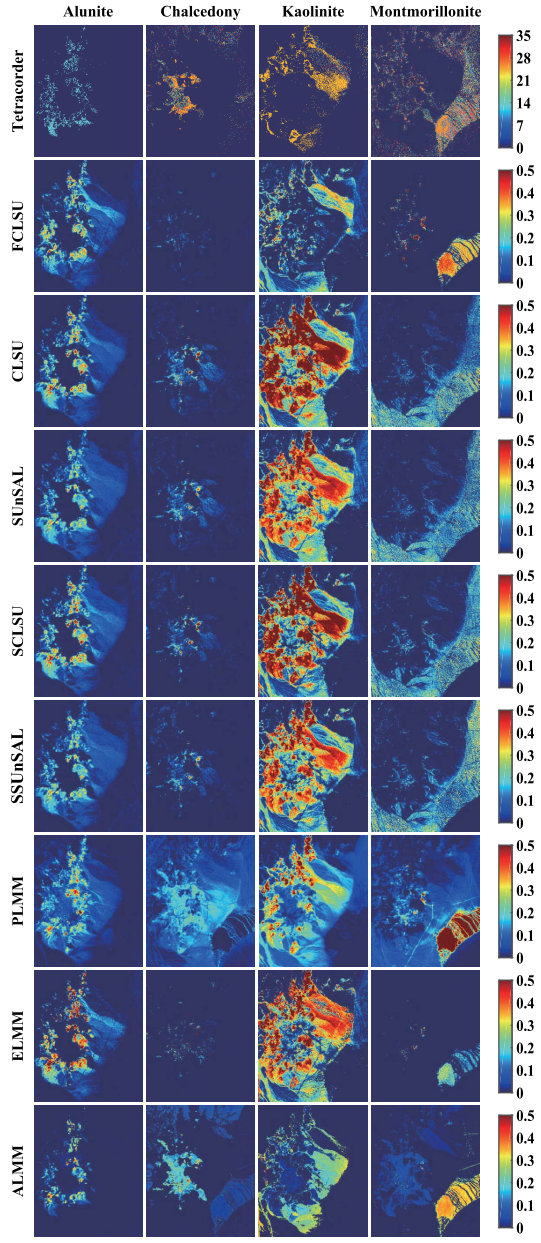


Fig. 14. The abundance maps estimated by different SU methods and the first row shows the so-called ground truth generated by Tetracorder.

C. Second Real Data (Cuprite)

1) Data Description: The second real dataset is the hyperspectral image acquired by the airborne visible-infrared imaging spectrometer (AVIRIS) over the Cuprite mining district in western Nevada, USA, which is composed of various minerals. We selected a sub-image composed of 304 × 257 pixels at a GSD of 20 m to evaluate the performance between the proposed method and the compared methods. The wavelength of 224 spectral bands ranges from 400 nm to 2500 nm with 10 nm spectral resolution. Before unmixing, bands eroded by water absorption, atmospheric effects, and noise (bands1-2, 104-113, 148-167, 221-224) were removed; 188 bands were used in the experiment. A false-color image of the Cuprite data is shown in Fig. 13(a).

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON THE CUPRITE DATA. THE BEST ONE IN SHOWN IN BOLD

Algorithm	FCLSU	CLSU	SUnSAL	SCLSU	SSUnSAL	PLMM	ELMM	ALMM
OA (%)	31.03	62.75	67.03	62.75	67.03	57.98	52.06	**74.17**
rRMSE	0.0470	0.0213	0.0206	0.0213	0.0206	0.367	0.0181	**0.0003**
aSAM	2.4588	1.9083	1.9278	1.9083	1.9278	2.1462	1.4138	**0.0052**

2) Experimental Setup: With the difficult challenges generated by the highly mixed property of the minerals and the low spatial resolution of the study image, this scene is commonly used for evaluating unmixing performance. Furthermore, data-driven endmember extraction is very challenging due to highly mixed effects. Therefore, we used the USGS spectral library to construct the endmember dictionary. The detailed procedures are as follows: First, VCA was applied for extracting 14 endmembers.[5] Then, material identification was performed using the USGS spectral library and spectral feature fitting [44]. Next, the endmember dictionary ($\mathbf{A}$) was constructed based on identified spectral signatures from the library, whose similarity scores are higher than a threshold.[6] Finally, 24 spectral signatures we selected from the spectral library as the endmember dictionary, as shown in Fig. 13(b). We used the same endmember matrix for all algorithms.

3) Results and Analysis: Similar to the first real data, we evaluate the performance for the Cuprite data both quantitatively and visually. The classification-based evaluation and the two reconstruction indicies are summarized in Table III. There is a difference in calculating OA. Owing to highly mixed effects of the minerals, it is quite difficult to exactly estimate the number of endmembers. Therefore, in order to effectively use OA for quantitatively assessing the performance of the different algorithms, we only considered four principal minerals, i.e., alunite, chalcedony, kaolinite, and montmorillonite.

The estimated abundance maps of the four minerals are shown in Fig. 14. The first row represents the reference classification maps generated by Tetracorder software [45]. Since FCLSU fails to take spectral variability into account and while strictly following the ANC and the ASC to the abundance maps, it yielded the unexpected result of the absence of certain material, as shown in the abundance map of the material *Chalcedony* of Fig. 14. Although the CLSU and SUnSAL algorithms can improve the visual effects by relaxing the ASC, particularly for the materials of kaolinite and montmorillonite, the range of abundances is obviously over 1, which makes no sense in reality. With the consideration of scaling factors, scaled versions of CLSU and SUnSAL effectively show the abundances to be in the understandable range. On the other hand, the reasonable assumption that the mixed spectral signature is sparsely represented by the endmember dictionary leads to a good visual result that approaches that of Tetracorder, as shown in the comparisons between CLSU and SUnSAL as well as their scaled versions.

PLMM and ELMM tend to specify the spectral variability. Considering the spectral variability as the perturbation information, PLMM is relatively hard to detect the pure area, since the main spectral variability (scaling factors) is ignored in this model. The estimated abundance maps of PLMM in Fig. 14 gives consistent results. While ELMM gives one scaling factor for each endmember, which yields much clearer results. The proposed method shows the best visual resemblance, compared with the results from the Tetracorder. The abundance maps generated by the proposed ALMM are more distinct and show greater contrast, and the distribution of each material is regional as well, which implies that various spectral variabilities could be learned effectively. Consistent with the analysis above, Table III gives a similar quantitative evaluation.

[5]The number of endmembers is estimated by Hysime on the subset of Cuprite.

[6]In our case, the threshold is experimentally set up as 0.93.

V. CONCLUSION

ELMM and PLMM have their respective drawbacks. ELMM ignores those spectral variabilities that cannot be explained only by scaling factors, and it is hard to obtain a good scaling estimation due to ELMM's non-convexity. With PLMM, the perturbation information is too general to model various spectral variabilities. To this end, we proposed a novel spectral mixture model, called ALMM, which considers not only the principal scaling factor but also other various spectral variabilities by introducing the spectral variability dictionary to expand the scalability of the endmember dictionary. To effectively promote spectral unmixing based on the proposed method, we modeled the spectral variability as low-coherent with the endmember dictionary and developed an algorithm for learning the spectral variability dictionary. By analyzing experimental results on a synthetic dataset and two real datasets, we found that the methods taking the spectral variability into consideration are generally superior to those that do not. More notably, the proposed method is able to obtain a more accurate abundance estimation compared to other state-of-the-art algorithms, since we separately model the spectral variability as scaling factors and other spectral variability according to their distinctive properties.

APPENDIX A
SOLUTION TO ALMM-BASED SPECTRAL UNMIXING

The object function in Eq. (20) is not convex with respect to all variables simultaneously, but it is a convex problem regarding the separate variable when other variables are fixed. As a result, we successively minimize $\mathscr{L}_U$ with respect to $\mathbf{x}_k, S_k, \mathbf{b}_k, \mathbf{g}_k, \mathbf{h}_k, \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k$ as follows:

Optimization with respect to $\mathbf{x}_k$ *and* S_k*:* Bundle $S_k\mathbf{A}$ as $\mathbf{D}$ and this subproblem can be written as

$$\underset{\mathbf{x}_k}{arg\,\min}\ \frac{1}{2}\|\mathbf{y}_k - \mathbf{D}\mathbf{x}_k - \mathbf{E}\mathbf{b}_k\|_2^2 + \boldsymbol{\lambda}_k^{\mathrm{T}}(\mathbf{g}_k - \mathbf{x}_k) + \boldsymbol{\nu}_k^{\mathrm{T}}(\mathbf{h}_k - \mathbf{x}_k) + \frac{\mu}{2}\|\mathbf{g}_k - \mathbf{x}_k\|_2^2 + \frac{\mu}{2}\|\mathbf{h}_k - \mathbf{x}_k\|_2^2, \quad (25)$$

which has a closed-form solution:

$$\mathbf{x}_k \leftarrow (\mathbf{D}^{\mathrm{T}}\mathbf{D} + 2\mu\mathbf{I})^{-1} \times (\mu\mathbf{g}_k + \boldsymbol{\lambda}_k + \mu\mathbf{h}_k + \boldsymbol{\nu}_k + \mathbf{D}^{\mathrm{T}}\mathbf{y}_k - \mathbf{D}^{\mathrm{T}}\mathbf{E}\mathbf{b}_k). \quad (26)$$

Inspired by SCLSU, we further update $\mathbf{x}_k$ in order to satisfy the sum-to-one constraint by

$$\mathbf{x}_k \leftarrow \mathbf{x}_k / \mathbf{1}^{\mathrm{T}}\mathbf{x}_k. \quad (27)$$

Hereinafter, bundle $\mathbf{A}\mathbf{x}_k$ as Z and S_k can be estimated by solving NNLS problem [36]:

$$\hat{S}_k = \underset{S_k \geq 0}{arg\,min} \frac{1}{2} \|(\mathbf{y}_k - \mathbf{E}\mathbf{b}_k) - S_k\mathbf{Z}\|_2^2. \quad (28)$$

Optimization with respect to $\mathbf{b}_k$*:* For $\mathbf{b}_k$, the optimization problem is

$$\underset{\mathbf{b}_k}{arg\,min} \frac{1}{2} \|\mathbf{y}_k - (S_k\mathbf{A})\mathbf{x}_k - \mathbf{E}\mathbf{b}_k\|_2^2 + \frac{\beta}{2} \|\mathbf{b}_k\|_2^2, \quad (29)$$

which is readily solved by

$$\mathbf{b}_k \leftarrow (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \beta\mathbf{I})^{-1}(\mathbf{E}^{\mathrm{T}}\mathbf{y}_k - S_k\mathbf{E}^{\mathrm{T}}\mathbf{A}\mathbf{x}_k). \quad (30)$$

Optimization with respect to $\mathbf{g}_k$*:* The subproblem of $\mathbf{g}_k$ can be written as

$$\underset{\mathbf{g}_k}{arg\,min} \frac{\alpha}{2} \|\mathbf{g}_k\|_1 + \boldsymbol{\lambda}_k^{\mathrm{T}}(\mathbf{g}_k - \mathbf{x}_k) + \frac{\mu}{2} \|\mathbf{g}_k - \mathbf{x}_k\|_2^2, \quad (31)$$

whose solution is the well-known *soft threshold* [46]:

$$\mathbf{g}_k \leftarrow \max\{\mathbf{0}, \|\mathbf{x}_k - \boldsymbol{\lambda}/\mu\|_1 - \alpha/\mu\} sign(\mathbf{x}_k - \boldsymbol{\lambda}/\mu), \quad (32)$$

where $sign(\bullet)$ is defined by

$$sign(\bullet) = \begin{cases} 1, & \bullet \succeq 0 \\ -1, & \bullet \prec 0. \end{cases} \quad (33)$$

Optimization with respect to $\mathbf{h}_k$*:* The optimization problem of $\mathbf{h}_k$ is

$$\underset{\mathbf{h}_k}{arg\,min}\, \boldsymbol{\nu}_k^{\mathrm{T}}(\mathbf{h}_k - \mathbf{x}_k) + \frac{\mu}{2} \|\mathbf{h}_k - \mathbf{x}_k\|_2^2 + l_R^+(\mathbf{h}_k). \quad (34)$$

Here the update rule for $\mathbf{h}_k$ is

$$\mathbf{h}_k \leftarrow \max\{\mathbf{0}, \mathbf{x}_k - \boldsymbol{\nu}/\mu\}. \quad (35)$$

Lagrange multipliers update $\boldsymbol{\lambda}_k$ *and* $\boldsymbol{\nu}_k$*:* Before stepping into the next iteration, Lagrange multipliers need to be updated by

$$\boldsymbol{\lambda}_k \leftarrow \boldsymbol{\lambda}_k + \mu(\mathbf{g}_k - \mathbf{x}_k), \quad \boldsymbol{\nu}_k \leftarrow \boldsymbol{\nu}_k + \mu(\mathbf{h}_k - \mathbf{x}_k). \quad (36)$$

APPENDIX B
SOLUTION TO ALMM-BASED SPECTRAL VARIABILITY DICTIONARY LEARNING

To solve Eq. (21), we have:

Optimization with respect to $\mathbf{M}$*:* The optimization problem can be formulated as follows:

$$\underset{\mathbf{M}}{arg\,min}\, \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{M} - \mathbf{E}\mathbf{B}\|_{\mathrm{F}}^2 + \boldsymbol{\Omega}^{\mathrm{T}}(\mathbf{M} - \mathbf{X}\mathbf{S}) + \frac{\xi}{2} \|\mathbf{M} - \mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2, \quad (37)$$

which can be quickly solved by

$$\mathbf{M} \leftarrow (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \xi\mathbf{I})^{-1}(\mathbf{A}^{\mathrm{T}}\mathbf{Y} - \mathbf{A}^{\mathrm{T}}\mathbf{E}\mathbf{B} + \xi\mathbf{X}\mathbf{S} - \boldsymbol{\Omega}). \quad (38)$$

Optimization with respect to $\mathbf{B}$*:* The analytical solution for $\mathbf{B}$ can be simply obtained by the matrix form of Eq. (30)

$$\mathbf{B} \leftarrow (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \beta\mathbf{I})^{-1}(\mathbf{E}^{\mathrm{T}}\mathbf{Y} - \mathbf{E}^{\mathrm{T}}\mathbf{A}\mathbf{M}). \quad (39)$$

Optimization with respect to $\mathbf{X}$*:* The optimization problem is expressed as follows:

$$\underset{\mathbf{X}}{arg\,min}\, \boldsymbol{\Lambda}^{\mathrm{T}}(\mathbf{G} - \mathbf{X}) + \boldsymbol{V}^{\mathrm{T}}(\mathbf{H} - \mathbf{X}) + \boldsymbol{\Omega}^{\mathrm{T}}(\mathbf{M} - \mathbf{X}\mathbf{S}) + \frac{\xi}{2} \|\mathbf{G} - \mathbf{X}\|_{\mathrm{F}}^2 + \frac{\xi}{2} \|\mathbf{H} - \mathbf{X}\|_{\mathrm{F}}^2 + \frac{\xi}{2} \|\mathbf{M} - \mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2. \quad (40)$$

The solution of Eq. (40) is given by

$$\mathbf{X} \leftarrow (\xi\mathbf{G} + \boldsymbol{\Lambda} + \xi\mathbf{H} + \boldsymbol{V} + \boldsymbol{\Omega}\mathbf{S}^{\mathrm{T}} + \xi\mathbf{M}\mathbf{S}^{\mathrm{T}}) \times (\xi\mathbf{S}\mathbf{S}^{\mathrm{T}} + 2\xi\mathbf{I})^{-1}. \quad (41)$$

In order to remove the scaling factors while satisfying the sum-to-one constraint, $\mathbf{X}$ is rewritten as a matrix form of Eq. (27).

$$\mathbf{X} \leftarrow \mathbf{X} \oslash (\mathbf{1}^{\mathrm{T}}\mathbf{X}), \quad (42)$$

where $\oslash$ is defined as a term-wise Hadamard division.

Optimization with respect to $\mathbf{S}$*:* Subsequently, the variable $\mathbf{S}$ can be collected by solving the following problem:

$$\underset{\mathbf{S}}{arg\,min}\, \boldsymbol{\Omega}^{\mathrm{T}}(\mathbf{M} - \mathbf{X}\mathbf{S}) + \boldsymbol{\Delta}^{\mathrm{T}}(\mathrm{T} - \mathbf{S}) + \frac{\xi}{2} \|\mathbf{M} - \mathbf{X}\mathbf{S}\|_{\mathrm{F}}^2 + \frac{\xi}{2} \|\mathbf{T} - \mathbf{S}\|_{\mathrm{F}}^2, \quad (43)$$

whose a closed-form solution can be obtained as

$$\mathbf{S} \leftarrow (\xi\mathbf{X}^{\mathrm{T}}\mathbf{X} + \xi\mathbf{I})^{-1}(\xi\mathbf{X}^{\mathrm{T}}\mathbf{M} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega} + \xi\mathrm{T} + \boldsymbol{\Delta}). \quad (44)$$

Optimization with respect to $\mathbf{E}$*:* The object function with respect to $\mathbf{E}$ is written as

$$\underset{\mathbf{E}}{arg\,min}\, \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{M} - \mathbf{E}\mathbf{B}\|_{\mathrm{F}}^2 + \boldsymbol{\Pi}^{\mathrm{T}}(\mathbf{Q} - \mathbf{E}) + \frac{\xi}{2} \|\mathbf{Q} - \mathbf{E}\|_{\mathrm{F}}^2, \quad (45)$$

which has the analytical solution of

$$\mathbf{E} \leftarrow (\mathbf{Y}\mathbf{B}^{\mathrm{T}} - \mathbf{A}\mathbf{M}\mathbf{B}^{\mathrm{T}} + \xi\mathbf{Q} + \boldsymbol{\Pi})(\mathbf{B}\mathbf{B}^{\mathrm{T}} + \xi\mathbf{I})^{-1}. \quad (46)$$

Optimization with respect to $\mathbf{Q}$*:* Inspired by [35], the optimization problem with the Gram matrix $\mathbf{E}^T\mathbf{E}$ can be effectively solved as follows: we define $\mathbf{Q}_p$ to be $\mathbf{Q}$ of the former step, so it can be regarded as a known matrix in the current step, and therefore the optimization problem for $\mathbf{Q}$ can be specified as

$$\underset{\mathbf{Q}}{arg\,min}\, \frac{\gamma}{2} \left\|\mathbf{A}^{\mathrm{T}}\mathbf{Q}\right\|_{\mathrm{F}}^2 + \frac{\eta}{2} \left\|\mathbf{Q}_p^{\mathrm{T}}\mathbf{Q} - \mathbf{I}\right\|_{\mathrm{F}}^2 + \boldsymbol{\Pi}^{\mathrm{T}}(\mathbf{Q} - \mathbf{E}) + \frac{\xi}{2} \|\mathbf{Q} - \mathbf{E}\|_{\mathrm{F}}^2, \quad (47)$$

which can be easily deduced as

$$\mathbf{Q} \leftarrow (\gamma\,\mathbf{A}\mathbf{A}^{\mathrm{T}} + \eta\mathbf{Q}_p\mathbf{Q}_p^{\mathrm{T}} + \xi\mathbf{I})^{-1}(\eta\mathbf{Q}_p + \xi\mathbf{E} - \boldsymbol{\Pi}). \quad (48)$$

Optimization with respect to $\mathbf{G}$ *and* $\mathbf{H}$*:* The two variables can be summarized using the matrix form of Eq. (32) and Eq. (35) as

$$\mathbf{G} \leftarrow \max\{\mathbf{0}, \|\mathbf{X} - \mathbf{\Lambda}/\xi\|_{1,1} - \alpha/\xi\} sign(\mathbf{X} - \mathbf{\Lambda}/\xi), \quad (49)$$

$$\mathbf{H} \leftarrow \max\{\mathbf{0}, \mathbf{X} - \boldsymbol{V}/\xi\}. \quad (50)$$

Optimization with respect to $\mathbf{T}$*:* The variable $\mathbf{T}$ can be updated by using the same rule as with $\mathbf{H}$:

$$\mathbf{T} \leftarrow \max\{\mathbf{0}, \mathbf{S} - \mathbf{\Delta}/\xi\}. \quad (51)$$

Lagrange multipliers update $\mathbf{\Lambda}$*,* $\boldsymbol{V}$*,* $\mathbf{\Omega}$*,* $\mathbf{\Pi}$ *and* $\mathbf{\Delta}$*:* Following the rule of Eq. (36), these Lagrange multipliers can be updated in each iteration:

$$\begin{aligned} &\mathbf{\Lambda} \leftarrow \mathbf{\Lambda} + \xi(\mathbf{G} - \mathbf{X}), \quad \boldsymbol{V} \leftarrow \boldsymbol{V} + \xi(\mathbf{H} - \mathbf{X}), \\ &\mathbf{\Delta} \leftarrow \mathbf{\Delta} + \xi(\mathbf{T} - \mathbf{S}), \quad \mathbf{\Pi} \leftarrow \mathbf{\Pi} + \xi(\mathbf{Q} - \mathbf{E}), \\ &\mathbf{\Omega} \leftarrow \mathbf{\Omega} + \xi(\mathbf{M} - \mathbf{XS}). \end{aligned} \quad (52)$$

ACKNOWLEDGMENTS

The authors would like to thank Pierre-Antoine Thouvenin from IRIT (Institut de Recherche en Informatique de Toulouse) for providing the PLMM code tested in our experiments, and the Hyperspectral Digital Imagery Collection Experiment (HYDICE) for sharing the urban dataset free of charge.

REFERENCES

[1] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "Learning a low-coherence dictionary to address spectral variability for hyperspectral unmixing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 235–239.

[2] L. Ma, X. Zhang, X. Yu, and D. Luo, "Spatial regularized local manifold learning for classification of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 9, no. 2, pp. 609–624, Feb. 2016.

[3] D. Hong, N. Yokoya, and X. Zhu, "Local manifold learning with robust neighbors selection for hyperspectral dimensionality reduction," in *Proc. IEEE IGARSS*, Jul. 2016, pp. 40–43.

[4] D. Hong, N. Yokoya, and X. X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Ses.*, vol. 10, no. 6, pp. 2960–2975, Jun. 2017.

[5] Z.-W. Pan, H.-L. Shen, C. Li, S. Chen, and J. H. Xin, "Fast multispectral imaging by spatial pixel-binning and spectral unmixing," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3612–3625, Aug. 2016.

[6] D. Hong and X. X. Zhu, "SULoRA: Subspace unmixing with low-rank attribute embedding for hyperspectral data analysis," *IEEE J. Sel. Topics Signal Process.*, to be published.

[7] S. Valero, P. Salembier, and J. Chanussot, "Hyperspectral image representation and processing with binary partition trees," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1430–1443, Apr. 2013.

[8] T. Matsuki, N. Yokoya, and A. Iwasaki, "Hyperspectral tree species classification of japanese complex mixed forest with the aid of lidar data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2177–2187, May 2015.

[9] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint and progressive learning from high-dimensional data for multi-label classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany. Springer, Sep. 2018, pp. 478–493.

[10] N. Yokoya and A. Iwasaki, "Object detection based on sparse representation and Hough voting for optical remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2053–2062, May 2015.

[11] D. Hong, W. Liu, J. Su, Z. Pan, and G. Wang, "A novel hierarchical approach for multispectral palmprint recognition," *Neurocomputing*, vol. 151, pp. 511–521, Mar. 2015.

[12] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[13] D. M. Rogge, B. Rivard, J. Zhang, and J. Feng, "Iterative spectral unmixing for optimizing per-pixel endmember sets," *IEEE Trans. Geosci. Remote Sens.*, vol.44 , no. 12, pp. 3725–3736, Dec. 2006.

[14] J. B Adams *et al.*, "Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the brazilian amazon," *Remote Sens. Environ.*, vol. 52, no. 2, pp. 137–154, May 1995.

[15] B. Somers, G. P. Asner, L. Tits, and P. Coppin, "Endmember variability in spectral mixture analysis: A review," *Remote Sens. Environ.*, vol. 115, no. 7, pp. 1603–1616, Jul. 2011.

[16] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 95–104, Jan. 2014.

[17] L. Drumetz, J. Chanussot, and C. Jutten, "Variability of the endmembers in spectral unmixing: Recent advances," in *Proc. WHISPERS*, Aug. 2016, pp. 1–5.

[18] A. Halimi, P. Honeine, and J. M. Bioucas-Dias, "Hyperspectral unmixing in presence of endmember variability, nonlinearity, or mismodeling effects," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4565–4579, Oct. 2016.

[19] T. Uezato, R. J. Murphy, A. Melkumyan, and A. Chlingaryan, "Incorporating spatial information and endmember variability into unmixing analyses to improve abundance estimates," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5563–5575, Dec. 2016.

[20] O. Eches, N. Dobigeon, C. Mailhes, and J-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. Application to hyperspectral imagery," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1403–1413, Jun. 2010.

[21] X. Du, A. Zare, P. Gader, and D. Dranishnikov, "Spatial and spectral unmixing using the beta compositional model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1994–2003, Jun. 2014.

[22] X. Fu, W.-K. Ma, J. M. Bioucas-Dias, and T.-H. Chan, "Semiblind hyperspectral unmixing in the presence of spectral library mismatches," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5171–5184, Sep. 2016.

[23] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 525–538, Jan. 2016.

[24] M. A. Veganzones *et al.* "A new extended linear mixing model to address spectral variability," in *Proc. WHISPERS*, Jun. 2014, pp. 1–4.

[25] L. Drumetz, M.-A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3890–3905, Aug. 2016.

[26] D. C. Heinz and C.-I Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, Mar. 2001.

[27] J. M. P. Nascimento and J. M. B. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.

[28] F. Zhu, Y. Wang, S. Xiang, B. Fan, and C. Pan, "Structured sparse method for hyperspectral unmixing," *ISPRS J. Photogramm. Remote Sens.*, vol. 88, no. 1, pp. 101–118, Feb. 2014.

[29] Y. Wang, C. Pan, S. Xiang, and F. Zhu, "Robust hyperspectral unmixing with correntropy-based metric," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4027–4040, Nov. 2015.

[30] X. Liu, W. Xia, B. Wang, and L. Zhang, "An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 757–772, Feb. 2011.

[31] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, Jul. 2008.

[32] E. C. Johnson and D. L. Jone, "Joint recovery of sparse signals and parameter perturbations with parameterized measurement models," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 5900–5904.

[33] M S. C. Almeida and M. A. T. Figueiredo, "Blind image deblurring with unknown boundaries using the alternating direction method of multipliers," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, VIC, Australia, Sep. 2013, pp. 586–590.

[34] S. Henrot, J. Chanussot, and C. Jutten, "Dynamical spectral unmixing of multitemporal hyperspectral images," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3219–3232, Jul. 2016.
[35] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 2055–2065, Apr. 2013.
[36] D. Kim, S. Sra, and I. Dhillon, "Tackling box-constrained optimization via a new projected quasi-Newton approach," *SIAM J. Sci. Comput.*, vol. 32, no. 6, pp. 3548–3563, Dec. 2010.
[37] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2014
[38] F. Wang, W. Cao, and Z. Xu. (May 2015). "Convergence of multi-block Bregman ADMM for nonconvex composite problems." [Online]. Available: https://arxiv.org/abs/1505.03063
[39] Q. Liu, X. Shen, and Y. Gu. (May 2017). "Linearized ADMM for non-convex non-smooth optimization with convergence analysis." [Online]. Available: https://arxiv.org/abs/1705.02502
[40] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Front. Math. China*, vol. 7, no. 2, pp. 365–384, Apr. 2012.
[41] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 676–683.
[42] P. Zhou, C. Zhang, and Z. Lin, "Bilevel model-based discriminative dictionary learning for recognition," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1173–1187, Mar. 2017.
[43] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.
[44] R. N. Clark and T. L. Roush, "Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications," *J. Geophys. Res.*, vol. 89, no. B7, pp. 6329–6340, Jul. 1984.
[45] R. N. Clark *et al.*, " Imaging spectroscopy: Earth and planetary remote sensing with the USGS tetracorder and expert systems," *J. Geophys. Res. Planets*, vol. 108, no. E12, pp. 5131–5146, Dec. 2003.
[46] J. M. B.-Dias and M. A. T. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. WHISPERS*, Jun. 2010, pp. 1–4.

Danfeng Hong (S'16) received the B.Sc. degree in computer science and technology from the Neusoft College of Information, Northeastern University, China, in 2012, and the M.Sc. degree in computer vision from Qingdao University, China, in 2015. He is currently pursuing the Ph.D. degree in signal processing in earth observation with the Technical University of Munich, Munich Germany, and the Remote Sensing Technology Institute, German Aerospace Center, Oberpfaffenhofen, Germany.

His research interests include signal/image processing and analysis, pattern recognition, machine/deep learning and their applications in Earth Vision.

Naoto Yokoya (S'10–M'13) received the M.Sc. and Ph.D. degrees in aerospace engineering from The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

From 2012 to 2013, he was a Research Fellow with the Japan Society for the Promotion of Science, Tokyo. From 2013 to 2017, he was an Assistant Professor with The University of Tokyo. From 2015 to 2017, he was also an Alexander von Humboldt Research Fellow with the German Aerospace Center, Oberpfaffenhofen, and the Technical University of Munich, Munich, Germany. Since 2018, he has been leading the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo. His research interests include image analysis and data fusion in remote sensing.

In 2017, he received the Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. Since 2017, he has been a Co-Chair of the IEEE Geoscience and Remote Sensing Image Analysis and Data Fusion Technical Committee. His model was the most accurate among over 800 submissions.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998. In 1999, he was with the Geography Imagery Perception Laboratory for the Delegation Generale de l'Armement (DGA–French National Defense Department). Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He was a Visiting Scholar with Stanford University, KTH (Sweden), and NUS (Singapore). Since 2013, he has been an Adjunct Professor with the University of Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles, Los Angeles. He is conducting his research at the GIPSA-Lab. His research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008). He was a member of the Institut Universitaire de France (2012–2017). He is the founding President of the IEEE Geoscience and Remote Sensing French Chapter (2007–2010) which received the 2010 IEEE GRSS Chapter Excellence Award. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He was the Co-Chair (2005–2008) and the Chair (2009–2011) of the GRS Data Fusion Technical Committee. He was an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2005–2007) and *Pattern Recognition* (2006–2008). He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015). Since 2007, he has been an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. In 2013, he was a Guest Editor for the PROCEEDINGS OF THE IEEE and in 2014 a Guest Editor of the *IEEE Signal Processing Magazine*. Since 2018, he has also been an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Xiao Xiang Zhu (S'10–M'12–SM'14) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Professor of signal processing in earth observation with TUM and the German Aerospace Center (DLR); the Head of the Department "EO Data Science" with the DLR's Earth Observation Center; and the Head of the Helmholtz Young Investigator Group "SiPEO" at DLR and TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council, Naples, Italy, in 2009, Fudan University, Shanghai, China, in 2014, The University of Tokyo, Tokyo, Japan, in 2015, and the University of California at Los Angeles, Los Angeles, USA, in 2016, respectively. Her main research interests are remote sensing and earth observation, signal processing, machine learning and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Appendices

D Hong D., Zhu X. X., 2019. SULoRA: Subspace Unmixing with Low-Rank Attribute Embedding for Hyperspectral Data Analysis. IEEE Journal of Selected Topics in Signal Processing (JSTSP), 12(6): 1351-1363.

https://ieeexplore.ieee.org/document/8502105

SULoRA: Subspace Unmixing With Low-Rank Attribute Embedding for Hyperspectral Data Analysis

Danfeng Hong, *Student Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*—To support high-level analysis of spaceborne imaging spectroscopy (hyperspectral) imagery, spectral unmixing has been gaining significance in recent years. However, from the inevitable spectral variability, caused by illumination and topography change, atmospheric effects and so on make it difficult to accurately estimate abundance maps in spectral unmixing. Classical unmixing methods, e.g., linear mixing model (LMM) and extended LMM, fail to robustly handle this issue, particularly facing complex spectral variability. To this end, we propose a subspace-based unmixing model using low-rank learning strategy, called subspace unmixing with low-rank attribute embedding (SULoRA), robustly against spectral variability in inverse problems of hyperspectral unmixing. Unlike those previous approaches that unmix the spectral signatures directly in original space, SULoRA is a general subspace unmixing framework that jointly estimates subspace projections and abundance maps in order to find a raw subspace that is more suitable for carrying out the unmixing procedure. More importantly, we model such raw subspace with low-rank attribute embedding. By projecting the original data into a low-rank subspace, SULoRA can effectively address various spectral variabilities in spectral unmixing. Furthermore, we adopt an alternating direction method of multipliers based algorithm to solve the resulting optimization problem. Extensive experiments on synthetic and real datasets are performed to demonstrate the superiority and effectiveness of the proposed method in comparison with the previous state-of-the-art methods.**

***Index Terms*—Alternating direction method of multipliers, hyperspectral data analysis, low-rank attribute embedding, remote sensing, subspace unmixing, spectral variability.**

I. Introduction

HYPERSPECTRAL imagery (HSI) is characterized by very rich spectral information, which enables us to detect targets of interest and identify unknown materials more easily. Motivated by this, considerable attentions have been paid to hyperspectral data processing and analysis, such as dimensionality reduction [1], [2], image segmentation [3], land-cover and land-use classification [4], and target detection [5] and so on. However, most of pixels in HSI suffer from the effect of spectral mixing due to a lower spatial-resolution than that of traditional RGB or multispectral imagery. These material mixtures inevitably degrade the spectrally discriminative ability, particularly in some high-level applications. To overcome this, spectral unmixing is defined as that decomposes the reference spectral signatures into a collection of spectral signatures of pure materials (or *endmembers*) and their abundance fractions (or *abundance maps*). In remote sensing community, spectral unmixing techniques have been widely and successfully applied to a variety of tasks, including mineral exploration and identification [6], forest monitoring [7].

Assuming the absent of any spectral, spatial, and temporal variabilities as well as microscopic interaction (e.g. multiple scatting, intimate mixing, etc.) between the materials are negligible, then the mixed spectrum of each pixel in the HSI scene is approximately measured by a *linear mixing model* (LMM) [8]. There is, however, a main factor-spectral variability, propagating unpredictable errors to LMM. This further yields an inaccurate unmixing process, since these errors are basically absorbed by *endmembers* and *abundance maps*. Nonlinearity, i.e. nonlinearly mixing spectral signatures, resulting from, e.g. multiple scattering and intimate mixing, is one of the main causes of spectral variability. In addition, varying acquisition conditions (e.g. illumination, topography, atmospheric effects) as well as physically and chemically intrinsic change of the material possibly speed up spectral degradation, which can be seen as another kind of spectral variability.

Recently, enormous efforts modeling errors either from statistics-based or regression-based point of view have been made to address the spectral variability [9]. Two mainstream statistical methods, namely the normal composition model [10] and the beta compositional model [11], assume the endmember spectra following a given probability distribution. On the other hand, inspired by LMM-the regression-based seminal work, and its variations have been successively proposed to deterministically model the spectral variability. A perturbed linear mixing model (PLMM) was proposed in [12] to fit the spectral variability using a Gaussian prior with each endmember. Similarly, Fu *et al.* designed a dictionary-adjusted nonconvex sparsity-encouraging regression (DANSER) by modeling the mismatch between the spectral library and the observed spectrum under a Gaussian distribution [13]. Although these approaches attempt to model the spectral variability in a general way, only a given explicit distribution, i.e. Gaussian, is still insufficient. In most

Manuscript received April 15, 2018; revised August 9, 2018; accepted September 4, 2018. Date of publication October 22, 2018; date of current version December 17, 2018. This work was supported in part by the European Research Council under the European Union's Horizon 2020 research and innovation programme under Grant ERC-2016-StG-714087 and in part by the Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de). The guest editor coordinating the review of this paper and approving it for publication was Prof. Thierry Bouwmans. *(Corresponding author: Xiao Xiang Zhu.)*

The authors are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling 82234, Germany, and also with the Signal Processing in Earth Observation, Technical University of Munich, Munich 80333, Germany (e-mail: danfeng.hong@dlr.de; xiao.zhu@dlr.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2018.2877497

hyperspectral scenes, the spectral signature is frequently scaled due to illumination or topological change, hence the scaling factor, as a principal variability, is quite coherent with the corresponding spectral signature. Such attributed spectral variability is hardly represented by a Gaussian-guided term. Drumetz *et al.* proposed an extended LMM (ELMM) [14] by modeling the different scaling factors on each endmember, but is a significant shortcoming in that other spectral variabilities are not be involved correspondingly.

While aforementioned unmixing algorithms have been successively proposed and successfully applied to some specific datasets, the ability of robustness and generalization in handling various spectral variabilities still remains limited. For this reason, we propose a robust subspace-based unmixing method by jointly performing subspace learning and unmixing in a closed-loop. With low-rank attribute embedding, the spectral variability can be effectively removed in the learnt low-rank subspace, achieving a robust spectral unmixing. More specifically, our contributions can be unfolded as follows:

- We propose a general subspace-based unmixing framework by jointly low-rank subspace learning and unmixing, called **s**ubspace **u**nmixing with **lo**w-**r**ank **a**ttribute embedding (SULoRA), to achieve a robust unmixing in a proper subspace rather than in the original space. Moreover, mostly linear unmixing models can be considered as special cases in this general framework.
- With the low-rank attribute embedding, the proposed SULoRA can broadly mitigate the effects of various spectral variabilities by projecting the original data into a more representative low-rank subspace.
- An alternating direction method of multipliers (ADMM) is adopted to solve the resulting optimization problem.

The remainder of this paper is organized as follows. Section II briefly summarizes the related work in spectral unmixing and analyzes their advantages and disadvantages. In Section III, we first clarify the motivation and then propose our methodology of the SULoRA model as well as corresponding ADMM-based optimization algorithm. Section IV presents the experimental results on two different datasets (a synthetic data and a real urban data) and gives the intuitive analysis and discussion both qualitatively and quantitatively. Finally, Section V concludes with a summary.

II. Related Work

In this section, we review state-of-arts unmixing algorithms, emphatically introducing LMM-based unmixing models and its variations including fully constrained least squares unmixing (FCLSU) [15], partial constrained least squares unmixing (PCLSU) [16], sparse unmixing by variable splitting and augmented Lagrangian (SUnSAL) [17], as well as their scaled versions (scaled partial constrained least squares unmixing (SPCLSU) [18] and scaled sparse unmixing by variable splitting and augmented Lagrangian (SSUnSAL) [19]), ELMM and PLMM.

A. LMM

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, \mathbf{y}_N] \in \mathbb{R}^{D\times N}$ be an unfolded HSI with D bands and N pixels, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P] \in \mathbb{R}^{D\times P}$ be the endmembers with the size of $D \times P$. $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N] \in \mathbb{R}^{P\times N}$ is denoted as abundance maps, whose each column vector stands for the fractional abundance at each pixel. $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_i, \ldots, \mathbf{r}_N] \in \mathbb{R}^{D\times N}$ is the residual (e.g. noise, modeling errors and others) in the form of matrix. Under an ideal condition without any external disturbance, the spectral measurement for a given pixel, denoted by $\mathbf{y}_i \in \mathbb{R}^{D\times 1}$, is well approximated by a set of linear combination of endmember spectra weighted by theri corresponding fractional abundances, resulting in the LMM:

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{r}_i, \tag{1}$$

where $\mathbf{a}_i$ and $\mathbf{x}_i$ should be non-negative in order to meet the physical conditions in reality. Moreover, the fractional abundance $\mathbf{x}_i$, as the name indicated, represents the proportions occupied by the different endmembers. This means $\mathbf{x}_i$ should be also subject to a sum-to-one constraint. Therefore, Eq. (1) with the necessary constraints is expressed as

$$\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{r}_i, \quad \text{s.t. } \mathbf{A} \succeq \mathbf{0},\ \mathbf{x}_i \succeq 0,\ \sum_{i=1}^{N} \mathbf{x}_i = 1. \tag{2}$$

Collecting all pixels, a compact matrix form of Eq. (2) can be written as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{R}, \quad \text{s.t. } \mathbf{A} \succeq \mathbf{0},\ \mathbf{X} \succeq \mathbf{0},\ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \tag{3}$$

In the following, we will detail several popular unmixing algorithms based on LMM:

1) FCLSU: In practice, the endmembers ($\mathbf{A}$) can be pre-extracted from the given scene using endmember extraction methods, i.e. pixel purity index (PPI), vertex component analysis (VCA) [20]. This renders us to more effectively and conveniently estimate the abundance maps ($\mathbf{X}$) by degrading the Eq. (3) to least-square regression problem, leading to FCLSU:

$$\min_{\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 \quad \text{s.t. } \mathbf{X} \succeq \mathbf{0},\ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \right\}. \tag{4}$$

Considering the presence of spectral variability, FCLSU yields a poor performance. It mainly derives from the strong sum-to-constraint, as explained in [8]. A common way to this issue is to relax the abundance fractions sum to less or larger than one or to consider a part of full constraints.

2) PCLSU: Following the above solution, the resulting PCLSU can be formulated by solving

$$\min_{\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 \quad \text{s.t. } \mathbf{X} \succeq \mathbf{0}. \right\}. \tag{5}$$

The estimated variable $\mathbf{X}$ in Eq. (5) might be any scales, owing to a badly-conditioned observed matrix $\mathbf{Y}$. To alleviate the effects of the ill-posed problem, meaningfully physical assumptions have to be added in the form of regularization.

3) SUnSAL: As observed, the abundances on each endmember are theoretically supposed to be sparse. Bioucas-Dias *et al.* embedded this property into LMM and achieved a powerful SUnSAL algorithm. The resulting optimization problem can be

written as follows

$$\min_{\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{X}\|_{1,1} \text{ s.t. } \mathbf{X} \succeq \mathbf{0},\ \mathbf{1}^{\mathrm{T}}\mathbf{X} = \mathbf{1}. \right\}, \tag{6}$$

where $\|\mathbf{X}\|_{1,1} \equiv \sum_{k=1}^{N} \|\mathbf{x}_k\|_1$ is denoted as an approximation of sparsity-promoting term.

In view of effectiveness of SUnSAL, SUnSAL's variations have been subsequently proposed in recent years, such as SUnSAL with total variation spatial regularization (SUnSAL-TV) [21], collaborative sparse regression (CLSUnSAL) [22], etc. We have to admit, however, that these advanced methods are still subject to the framework of LMM that is sensitive to spectral variabilities.

B. ELMM

ELMM aims to modeling the principle spectral variability (scaling factors) to allow a pixel-wise variation at each endmember:

$$\mathbf{y}_i = \mathbf{A}\mathbf{S}_i\mathbf{x}_i + \mathbf{r}_i, \tag{7}$$

where $\mathbf{S}_i \in \mathbb{R}^{P\times P}$ is a diagonal matrix with the nonnegative constraint ($\mathbf{S}_i \succeq 0$). A matrix form of Eq. (7) can be repented as

$$\mathbf{Y} = \mathbf{A}(\mathbf{S}\odot\mathbf{X}) + \mathbf{R}, \tag{8}$$

here $\mathbf{S} \in \mathbb{R}^{P\times N}$ is a full matrix collecting the scaling factors from all pixels whose i^{th} column is $\mathbf{S}_i$. The operator $\odot$ is denoted as the Schur-Hadamard (termwise) product.

1) Unmixing Under the ELMM: Intuitively, the optimization problems in (7) and (8) are hardly to be analytically solved. In [14], a trick is employed by splitting the coupled variables ($\mathbf{S}$ and $\mathbf{X}$), then we have

$$\min_{\mathbf{X},\mathbf{S}\succeq 0,\underline{\mathbf{A}}} \left\{ \sum_{k=1}^{N} \left(\|\mathbf{y}_k - \mathbf{A}_k\mathbf{x}_k\|_2^2 + \lambda_S\|\mathbf{A}_k - \mathbf{A}_0\mathbf{S}_k\|_{\mathrm{F}}^2 \right) \right\}, \tag{9}$$

where $\mathbf{A}_0$ is the reference endmember spectrum, $\underline{\mathbf{A}} = \{\mathbf{A}_i\}$ is a collection of pixel-dependent endmember matrices, and λ_S plays a balance role between the two separated terms. Eq. (9) can be alternatively optimized with respect to each variable by alternating minimization strategy [23].

2) SPCLSU: Prior to ELMM, scaling factors have been investigated in a simple way, that is SPCLSU [18] in which endmembers are reasonably assumed by sharing a same scale as the scaling factors are strongly associated with topography. SPCLSU actually conducts a PCLSU in the beginning, and then normalizes the abundance maps to meet sum-to-one. This is a simple but effective strategy, which is also involved in our proposed method.

C. PLMM

As the name suggested, PLMM attempts to describe the spectral variability as an additive perturbation information. Both the pixel-wise and the corresponding matrix form of PLMM can be expressed, respectively

$$\mathbf{y}_i = (\mathbf{A} + \mathbf{\Delta}_i)\mathbf{x}_i + \mathbf{r}_i, \tag{10}$$

and

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \underbrace{[\mathbf{\Delta}_1\mathbf{x}_1|\ldots|\mathbf{\Delta}_i\mathbf{x}_i|\ldots|\mathbf{\Delta}_N\mathbf{x}_N]}_{\mathbf{\Delta}} + \mathbf{R}, \tag{11}$$

where $\mathbf{\Delta}$ is $[\mathbf{\Delta}_1\mathbf{x}_1|\ldots|\mathbf{\Delta}_i\mathbf{x}_i|\ldots|\mathbf{\Delta}_N\mathbf{x}_N]$ denotes the perturbation information of the endmembers.

1) Unmixing Under the PLMM: The optimization problem corresponding to PLMM-based unmixing can be given as

$$\min_{\mathbf{A},\mathbf{\Delta},\mathbf{X}} \left\{ \begin{array}{l} \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{X}-\mathbf{\Delta}\|_{\mathrm{F}}^2 + \alpha\mathbf{\Phi}(\mathbf{X}) + \beta\mathbf{\Psi}(\mathbf{A}) \\ \quad + \gamma\mathbf{\Upsilon}(\mathbf{\Delta}) \end{array} \right\}, \tag{12}$$

where $\mathbf{\Phi}$, $\mathbf{\Psi}$, and $\mathbf{\Upsilon}$ parameterized by α, β, and γ, are penalties with respect to variables $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{\Delta}$, receptively. Notably, $\mathbf{\Upsilon}$ term is modeled by a Frobenius norm.

2) DANSER: Likewise being generalized to PLMM framework, DANSER adopts a sparsity-encouraging regression technique for a dictionary-based spectral unmixing, where a perturbation-like information is explored to measure the mismatch between spectral dictionary and observed endmembers. This model, the DANSER, is formulated by

$$\min_{\mathbf{A}',\mathbf{X}} \left\{ \begin{array}{l} \frac{1}{2}\|\mathbf{Y}-\mathbf{A}'\mathbf{X}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{A}'-\mathbf{A}\|_{\mathrm{F}}^2 + \beta\|\mathbf{X}\|_{2,p}^p \\ \quad \text{s.t. } \mathbf{X} \succeq \mathbf{0} \end{array} \right\}, \tag{13}$$

where $\mathbf{A}'$ is a corrupted endmember matrix obtained by perturbing $\mathbf{A}$.

Although the aforementioned methods have shown an advancement in treating the spectral variability, especially facing main spectral variabilities (e.g. scaling factors), they are still lack of robustness and generalization to others that we are unknown. Jump out of this circle, a new insight is provided into this problem that we propose to conduct the spectral unmixing in a robust subspace instead of directly unmixing in original spectral space. Please go to next section for more details.

III. Subspace Unmixing With Low-Rank Attribute Embedding

A. General Motivation

There is a trade-off between spectral information gain and the spectral variability. On one hand, spectrum are expected to be spectrally discriminative. Conversely, this means that more complex spectral variabilities might get involved in hyperspectral data. A feasible solution to this issue is spectral unmixing in a 'raw' subspace rather than in the original space. In the learnt subspace, the pixels belonging to the same class are expected to be strongly correlated by using a low-rank attribute embedding. Further, this process can be mathematically modeled as

$$\begin{aligned} &\mathbf{Y} = \mathbf{Y}' + \mathbf{R}',\ \text{s.t. } \mathbf{Y}' = \mathbf{\Theta}\mathbf{Y}, \\ &\mathbf{Y}' = \mathbf{\Theta}\mathbf{A}\mathbf{X} + \mathbf{R}'', \end{aligned} \tag{14}$$

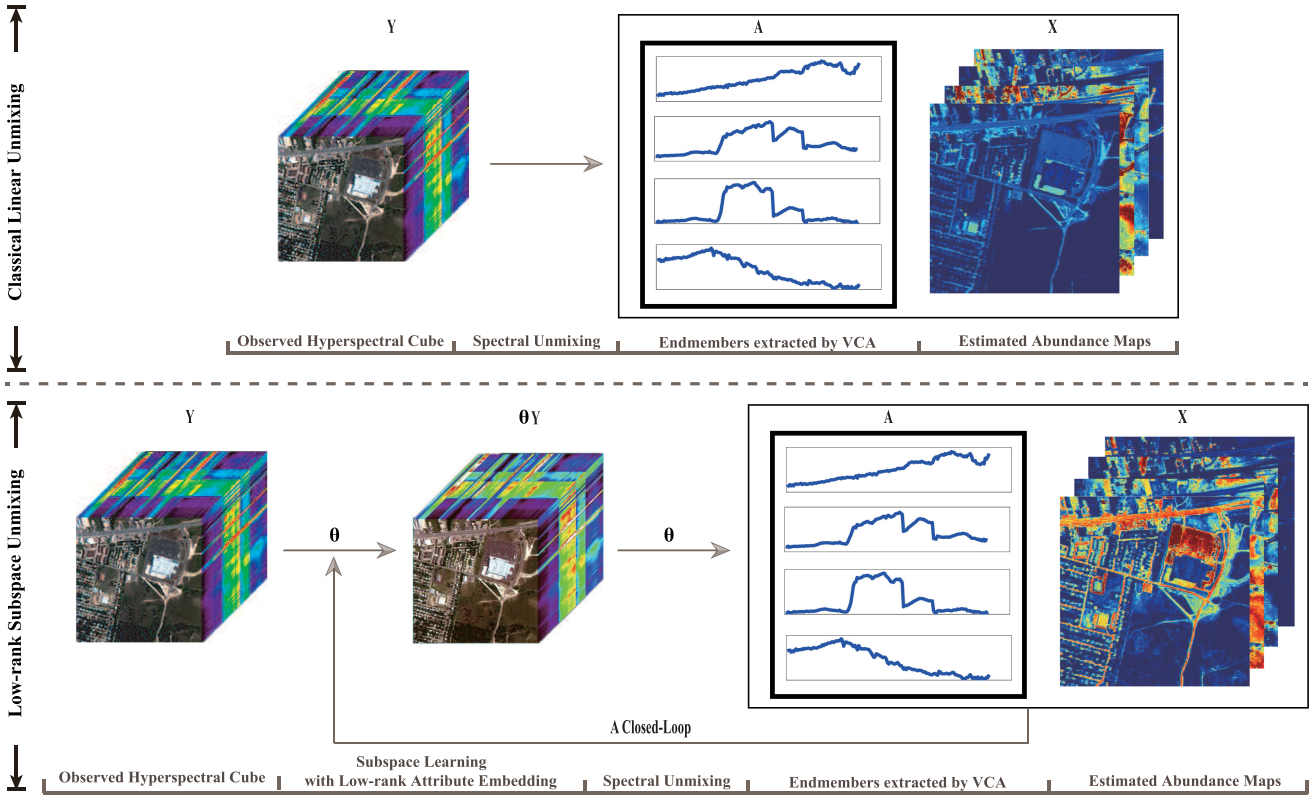


Fig. 1. A comparison of the holistic workflow between the original-space-based method and the proposed SULoRA.

where $\mathbf{\Theta}$ denotes the low-rank subspace projections, and $\mathbf{Y}^{'}$ is the spectrally subspace representation after embedding the low-rank attribute.

Fig. 1 shows a comparison in holistic workflow of spectral unmixing between using the original-space-based and the subspace-based (SULoRA) approaches.

B. Low-Rank Attribute Embedding

Inspired by [24] in which a novel strategy of low-rank attribute embedding is proposed with the application to person re-identification, we further improve this term by integrating our general motivation described above, making it more applicable to hyperspectral unmixing task.

Step by step, we will clarify the motivation of using the low-rank attribute embedding in great detail. It is well-known that hyperspectral imagery inevitably suffers from various spectral variabilities in the process of imaging. These spectral variabilities, which are generated due to illumination conditions, topography change, atmospheric effects, and material nonlinear mixing, are complex and even hardly represented using a common model. Instead of directly modeling such changeable property, we hypothetically treat the spectral variability as *an unknown complex noise*. Therefore, modeling the complex spectral variability could be converted to a special denoising problem. Noises in the data can be generally removed through a projection transformation. During this process, one is expected to be the projected or denoised data as close as possible with the original data, resulting in a mathematical expression ($\mathbf{Y} \doteq \mathbf{\Theta Y}$). Besides, we also expect to structurally maintain consistency between noisy data ($\mathbf{Y}$) and processed data ($\mathbf{\Theta Y}$), which might be achieved by correlative or collaborative filtering in order to emphasize the correlation and structural property between the samples. Low-rank representation has been widely and successfully applied for modeling the sample-based correlation [25]–[27], hence the estimated projection $\mathbf{\Theta}$ can be naturally endowed with a low-rank attribute (e.g., rank $(\mathbf{\Theta}) \preceq C$) in our case.

C. Problem Formulation

As introduced in Subsection III-A, our proposed SULoRA shown in Eq. (14) can be formulated as a following constrained optimization problem

$$\min_{\mathbf{X},\mathbf{\Theta}} \left\{ \begin{array}{l} \frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{AX})\|_{\mathrm{F}}^2 + \Phi(\mathbf{\Theta}) + \Upsilon(\mathbf{X}) \\ \quad \text{s.t. } \mathbf{X} \succeq \mathbf{0} \end{array} \right\}, \tag{15}$$

which aims at estimating the variables with respect to $\mathbf{X}$ and $\mathbf{\Theta}$. Since the problem (15) is undetermined, the variables $\mathbf{X}$ and $\mathbf{\Theta}$ should be regularized by reasonable prior knowledge. The two regularization terms $\Phi(\mathbf{\Theta})$ and $\Upsilon(\mathbf{X})$ are described below.

1) Subspace Regularization $\Phi(\mathbf{\Theta})$: According to the discussion and analysis in Section III-B, the subspace projections $\mathbf{\Theta}$ are characterized by a low-rank attribute in order to transfer the original hyperspectral data into a robust subspace, which can be approximately formulated by the form of $\|\mathbf{\Theta}\|_*$. Essentially, the main difference between those previously proposed low-rank representation learning and the proposed SULoRA lies in

the motive. More specifically, the former ones usually aim to robust clustering in subspace [25], [27] that needs to estimate the connectivity between samples, while our goal is to find or learn a low-rank subspace projection so that the learned projection can play a correlative filtering-like role robustly against various spectral variabilities, which is computationally efficient. Besides, we also hope to structurally maintain the spectral properties, making the learnt subspace as close as possible with the original space. This second prior can be formed by $\|\mathbf{Y}-\mathbf{\Theta Y}\|_F$. The final resulting expression of regularization with respect to $\mathbf{\Theta}$ is

$$\Phi(\mathbf{\Theta})=\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{\Theta Y}\|_{\mathrm{F}}^2+\beta\|\mathbf{\Theta}\|_*, \tag{16}$$

where α and β are the corresponding penalty parameters.

2) Abundance Regularization $\Upsilon(\mathbf{X})$*:* For a given hyperspectral scene, the spectral signature consists of limited kinds of materials, hence the abundances should be encouraged to be sparse. This term parameterized by γ can be expressed by

$$\Upsilon(\mathbf{X})=\gamma\|\mathbf{X}\|_{1,1}. \tag{17}$$

In our model, the non-negativity constraint ($\mathbf{X}\succeq\mathbf{0}$) has to be considered to satisfy the physical assumption. It should be noted, however, that the sum-to-one constraint is not directly considered in our optimization problem (Eq. (15)), since the hard constraint is too strong to yield a badly-estimated abundance maps. We adopt the same technique in SPCLSU [18] to force $\mathbf{X}$ to follow the sum-to-one constraint.

Different with matrix factorization-based unmixing approaches that simultaneously estimate the endmembers and the abundance maps, the proposed SULoRA first determines the number of endmembers via HySime [28], and then separately extracts the endmembers from the HSI scene with VCA and estimates the abundance maps. The benefits of the scheme in our model are two-fold. On one hand, the endmembers extracted from the data tend to preserve, to the greatest extent, spectrally physical significance, and thereby improve the stability of estimating the abundance maps. On the other hand, it effectively simplifies the model's complexity by optimizing fewer variables, finding a good solution easier.

D. Model Optimization Using ADMM-Based Algorithm

The optimization problem shown in Eq. (15) is convex, we adopt an ADMM-based optimization algorithm [29]–[31] for a fast and efficient solution. To facilitate the use of ADMM, we first convert Eq. (15) to an equivalent form introducing multiple auxiliary variables $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{J}$ to replace $\mathbf{\Theta}$, $\mathbf{X}$, and $\mathbf{X}$, respectively.

$$\min_{\mathbf{X},\mathbf{\Theta},\mathbf{G},\mathbf{H},\mathbf{J}}\left\{\begin{array}{l}\frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{AX})\|_{\mathrm{F}}^2+\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{\Theta Y}\|_{\mathrm{F}}^2\\ +\beta\|\mathbf{G}\|_*+\gamma\|\mathbf{H}\|_{1,1}+l_R^+(\mathbf{J})\\ \text{s.t. }\mathbf{\Theta}=\mathbf{G},\ \mathbf{X}=\mathbf{H},\ \mathbf{X}=\mathbf{J}\end{array}\right\}, \tag{18}$$

where $()^+$ denotes an operator that intercepts the positive part of each component of the matrix, and $l_R^+(\mathbf{J})$ is defined as $\mathbf{J}\succeq\mathbf{0}$. This problem can be equivalently solved by minimizing the following augmented Lagrangian function:

$$\begin{aligned}\mathscr{L}_U(\mathbf{X},\mathbf{\Theta},\mathbf{G},\mathbf{H},\mathbf{J},\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\mathbf{\Lambda}_3)&=\frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{AX})\|_{\mathrm{F}}^2\\&+\frac{\alpha}{2}\|\mathbf{Y}-\mathbf{\Theta Y}\|_{\mathrm{F}}^2+\beta\|\mathbf{G}\|_*+\gamma\|\mathbf{H}\|_{1,1}+l_R^+(\mathbf{J})\\&+\mathbf{\Lambda}_1^{\mathrm{T}}(\mathbf{G}-\mathbf{\Theta})+\mathbf{\Lambda}_2^{\mathrm{T}}(\mathbf{H}-\mathbf{X})+\mathbf{\Lambda}_3^{\mathrm{T}}(\mathbf{J}-\mathbf{X})\\&+\frac{\mu}{2}\|\mathbf{G}-\mathbf{\Theta}\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{H}-\mathbf{X}\|_{\mathrm{F}}^2+\frac{\mu}{2}\|\mathbf{J}-\mathbf{X}\|_{\mathrm{F}}^2,\end{aligned} \tag{19}$$

where $\{\mathbf{\Lambda_i}\}_{i=1}^3$ are Lagrange multipliers and μ is the penalty parameter. The specific optimization flow for solving the problem (19) is summarized in Algorithm 1, and the solution to each subproblem is detailed in the following.

Algorithm 1: Subspace Unmixing With Low-Rank Attribute Embedding (SULoRA).

Input: $\mathbf{Y}$, $\mathbf{A}$, $\mathbf{X}_0$, α, β, γ, $maxIter$.
Output: $\mathbf{X}$, $\mathbf{\Theta}$.
Initialization: $\mathbf{G}=\mathbf{0}$, $\mathbf{H}=\mathbf{0}$, $\mathbf{J}=\mathbf{0}$, $\mathbf{\Lambda}_1=\mathbf{0}$, $\mathbf{\Lambda}_2=\mathbf{0}$, $\mathbf{\Lambda}_3=\mathbf{0}$, $\mu=10^{-3}$, $\mu_m=10^6$, $\rho=1.5$, $\varepsilon=10^{-6}$, $t=1$.
while *not converged* or $t>maxIter$ **do**
  Fix other variables to update $\mathbf{\Theta}$ by
  $\mathbf{\Theta}=(\alpha\mathbf{YY}^{\mathrm{T}}+\mu\mathbf{G}+\mathbf{\Lambda}_1)\times(\alpha\mathbf{YY}^{\mathrm{T}}+(\mathbf{Y}-\mathbf{AX})(\mathbf{Y}-\mathbf{AX})^{\mathrm{T}}+\mu\mathbf{I})^{-1}$.
  Fix other variables to update $\mathbf{X}$ by
  $\mathbf{X}=((\mathbf{\Theta A})^{\mathrm{T}}(\mathbf{\Theta A})+2\mu\mathbf{I})^{-1}\times((\mathbf{\Theta A})^{\mathrm{T}}\mathbf{\Theta Y}+\mu\mathbf{H}+\mathbf{\Lambda}_2+\mu\mathbf{J}+\mathbf{\Lambda}_3)$.
  Fix other variables to update $\mathbf{G}$ by
  $[\mathbf{U},\mathbf{S},\mathbf{V}]=\mathrm{svd}(\mathbf{\Theta}-\mathbf{\Lambda}_1/\mu),\quad \mathbf{S}=\mathrm{diag}(\{s_k\}_{k=1}^r)$
  $\mathbf{G}=\mathbf{US}_\tau\mathbf{V},\quad \mathbf{S}_\tau=\mathrm{diag}(\max\{0,s_k-\beta/\mu\})$.
  Fix other variables to update $\mathbf{H}$ by
  $\mathbf{H}=\max\{\mathbf{0},|\mathbf{X}-\mathbf{\Lambda}_2/\mu|-\gamma/\mu\}\odot sign(\mathbf{X}-\mathbf{\Lambda}_2/\mu)$.
  Fix other variables to update $\mathbf{J}$ by
  $\mathbf{J}=\max\{\mathbf{0},\mathbf{X}-\mathbf{\Lambda}_3/\mu\}$.
  Update Lagrange multipliers by
  $\mathbf{\Lambda}_1\leftarrow\mathbf{\Lambda}_1+\mu(\mathbf{G}-\mathbf{\Theta}),\ \mathbf{\Lambda}_2\leftarrow\mathbf{\Lambda}_2+\mu(\mathbf{H}-\mathbf{X})$
  $\mathbf{\Lambda}_3\leftarrow\mathbf{\Lambda}_3+\mu(\mathbf{J}-\mathbf{X})$.
  Update penalty parameter by
  $\mu=\min(\rho\mu,\mu_m)$.
  Check the convergence conditions: **if** $\|\mathbf{G}-\mathbf{\Theta}\|_{\mathrm{F}}<\varepsilon$ *and* $\|\mathbf{G}-\mathbf{X}\|_{\mathrm{F}}<\varepsilon$ *and* $\|\mathbf{J}-\mathbf{X}\|_{\mathrm{F}}<\varepsilon$ **then**
    Stop iteration;
  else
    $t\leftarrow t+1$;
  end
end

We successively minimize $\mathscr{L}_U$ with respect to the variables $\mathbf{\Theta}$, $\mathbf{X}$, $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{J}$ as well as Lagrange multipliers $\{\mathbf{\Lambda_i}\}_{i=1}^3$ as follows:

Optimization with respect to $\mathbf{\Theta}$*:* The optimization problem for $\mathbf{\Theta}$ is

$$\min_{\mathbf{\Theta}} \left\{ \begin{array}{l} \frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{A}\mathbf{X})\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{Y}-\mathbf{\Theta}\mathbf{Y}\|_{\mathrm{F}}^2 \\ \quad + \mathbf{\Lambda}_1^{\mathrm{T}}(\mathbf{G}-\mathbf{\Theta}) + \frac{\mu}{2}\|\mathbf{G}-\mathbf{\Theta}\|_{\mathrm{F}}^2 \end{array} \right\}, \tag{20}$$

which has an analytical solution of

$$\begin{aligned} \mathbf{\Theta} \leftarrow\ & (\alpha\mathbf{Y}\mathbf{Y}^{\mathrm{T}} + \mu\mathbf{G} + \mathbf{\Lambda}_1) \\ & \times (\alpha\mathbf{Y}\mathbf{Y}^{\mathrm{T}} + (\mathbf{Y}-\mathbf{A}\mathbf{X})(\mathbf{Y}-\mathbf{A}\mathbf{X})^{\mathrm{T}} + \mu\mathbf{I})^{-1}. \end{aligned} \tag{21}$$

Optimization with respect to $\mathbf{X}$*:* For $\mathbf{X}$, the optimization problem can be expressed as

$$\min_{\mathbf{X}} \left\{ \begin{array}{l} \frac{1}{2}\|\mathbf{\Theta}(\mathbf{Y}-\mathbf{A}\mathbf{X})\|_{\mathrm{F}}^2 + \mathbf{\Lambda}_2^{\mathrm{T}}(\mathbf{H}-\mathbf{X}) + \mathbf{\Lambda}_3^{\mathrm{T}}(\mathbf{J}-\mathbf{X}) \\ \quad + \frac{\mu}{2}\|\mathbf{H}-\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\mu}{2}\|\mathbf{J}-\mathbf{X}\|_{\mathrm{F}}^2 \end{array} \right\}, \tag{22}$$

whose a closed-form solution is

$$\begin{aligned} \mathbf{X} \leftarrow\ & ((\mathbf{\Theta}\mathbf{A})^{\mathrm{T}}(\mathbf{\Theta}\mathbf{A}) + 2\mu\mathbf{I})^{-1} \\ & \times ((\mathbf{\Theta}\mathbf{A})^{\mathrm{T}}\mathbf{\Theta}\mathbf{Y} + \mu\mathbf{H} + \mathbf{\Lambda}_2 + \mu\mathbf{J} + \mathbf{\Lambda}_3). \end{aligned} \tag{23}$$

Optimization with respect to $\mathbf{G}$*:* The objective function for $\mathbf{G}$ is written as

$$\min_{\mathbf{G}} \left\{ \beta\|\mathbf{G}\|_* + \mathbf{\Lambda}_1^{\mathrm{T}}(\mathbf{G}-\mathbf{\Theta}) + \frac{\mu}{2}\|\mathbf{G}-\mathbf{\Theta}\|_{\mathrm{F}}^2 \right\}, \tag{24}$$

which is solved via the Singular Value Thresholding (SVT) operator [32]:

- *Step 1:* Input a matrix $\mathbf{M}$ of rank r and consider the singular value decomposition (SVD):

$$\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}, \quad \mathbf{S} = \mathrm{diag}(\{s_k\}_{1\le k\le r}). \tag{25}$$

- *Step 2:* For each $\tau \ge 0$, we define the soft-thresholding operator $\mathcal{D}_\tau$ as follows

$$\mathcal{D}(\mathbf{M}) := \mathbf{U}\mathcal{D}_\tau(\mathbf{S})\mathbf{V}, \quad \mathcal{D}_\tau(\mathbf{S}) = \mathrm{diag}(\{s_k-\tau\}^+). \tag{26}$$

Using Eq. 26, $\|\mathbf{M}\|_*$ can be computed by $\|\mathcal{D}_\tau(\mathbf{S})\|_{1,1}$.

Optimization with respect to $\mathbf{H}$*:* The optimization problem of $\mathbf{H}$ is

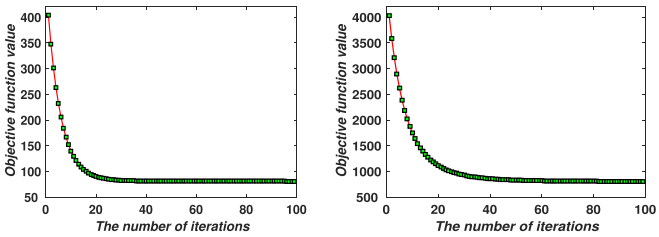
$$\min_{\mathbf{H}} \left\{ \gamma\|\mathbf{H}\|_{1,1} + \mathbf{\Lambda}_2^{\mathrm{T}}(\mathbf{H}-\mathbf{X}) + \frac{\mu}{2}\|\mathbf{H}-\mathbf{X}\|_{\mathrm{F}}^2 \right\}, \tag{27}$$

its solution is nothing but a well-known *soft threshold* [17]:

$$\mathbf{H} \leftarrow \max\{\mathbf{0}, |\mathbf{X}-\mathbf{\Lambda}_2/\mu| - \gamma/\mu\} \odot \mathrm{sign}(\mathbf{X}-\mathbf{\Lambda}_2/\mu). \tag{28}$$

Optimization with respect to $\mathbf{J}$*:* The subproblem of $\mathbf{J}$ can be given by

$$\min_{\mathbf{J}} \left\{ \mathbf{\Lambda}_3^{\mathrm{T}}(\mathbf{J}-\mathbf{X}) + \frac{\mu}{2}\|\mathbf{J}-\mathbf{X}\|_{\mathrm{F}}^2 + l_R^+(\mathbf{J}) \right\}, \tag{29}$$



(a) Synthetic Data (b) Urban Data

Fig. 2. Convergence analysis of SULoRA are experimentally performed on a synthetic data and a real urban data.

$\mathbf{J}$ can be updated using the following rule

$$\mathbf{J} \leftarrow \max\{\mathbf{0}, \mathbf{X}-\mathbf{\Lambda}_3/\mu\}. \tag{30}$$

Lagrange multipliers update $\{\mathbf{\Lambda_i}\}_{i=1}^3$*:* In each iteration, Lagrange multipliers need to be updated by

$$\begin{aligned} & \mathbf{\Lambda}_1 \leftarrow \mathbf{\Lambda}_1 + \mu(\mathbf{G}-\mathbf{\Theta}),\ \mathbf{\Lambda}_2 \leftarrow \mathbf{\Lambda}_2 + \mu(\mathbf{H}-\mathbf{X}) \\ & \mathbf{\Lambda}_3 \leftarrow \mathbf{\Lambda}_3 + \mu(\mathbf{J}-\mathbf{X}). \end{aligned} \tag{31}$$

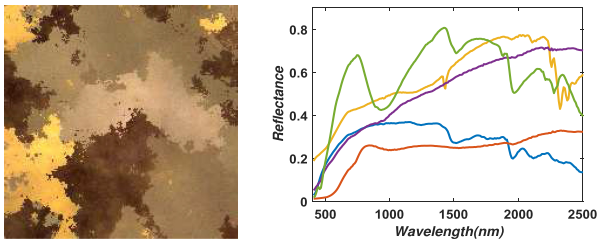
E. Convergence Analysis and Computational Cost

ADMM used in our optimization problem can be actually generalized to *inexact* Augmented Lagrange Multiplier (ALM) [33], whose convergence has been well studied when the number of block is less than three [29]. There is still not a *generally and strictly* theoretical proof in multi-blocks case. Fortunately for our case, its convergence is similarly guaranteed and supported in [32], [34]–[37]. Moreover, we experimentally record the objective function values in each iteration to draw the convergence curves of SULoRA on two used hyperspectral scenes (see Fig. 2).

As observed from Section III-D, the computational cost in the SULoRA algorithm is dominated by matrix products, and then the computational complexity of each subproblem in Eq. (18) with respect to the variables $\mathbf{X}$, $\mathbf{\Theta}$, $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{J}$ are, in each iteration, $\mathcal{O}(D^2N)$, $\mathcal{O}(D^2N)$, $\mathcal{O}(D^3)$, $\mathcal{O}(PN)$, and $\mathcal{O}(PN)$, respectively, where the most costly step is solving $\mathbf{\Theta}$, hence yielding an overall $\mathcal{O}(D^2N)$ computational cost for Eq. (18).

IV. Experiments

In this section, we quantitatively and visually evaluate the unmixing performance of the proposed SULoRA on a synthetic dataset presented in [14] and two real datasets over the areas of Urban and MUFFLE Gulfport Campus, in comparison with eight classical and state-of-the-art methods, including FCLSU, PCLSU, SPCLSU, SUnSAL, SSUnSAL (scaled SUnSAL), SLRU (sparse and low-rank unmixing) [38], PLMM and ELMM. We experimentally and empirically choose the regularization parameters to maximize performance of above methods. To make fair visual comparisons, we fix a display range of the abundance maps from 0 to 1 in Figs. 4 and 9. Because there are some algorithms ignoring the effects of scaling factors, resulting in the abundances that show the maximum of the display range but actually exceed it.



(a) A false color image (b) Endmembers

Fig. 3. A false color image of the synthetic data and corresponding five endmembers used for data simulation.

A. Synthetic Data

1) Data Description: Spectral simulation in the synthetic data is performed using five reference endmembers randomly selected from the spectral library of United States Geological Survey (USGS) with the size of 200 × 200 abundance maps generated using Gaussian fields, which strictly satisfies the abundance non-negative constraint (ANC) and the abundance sum-to-one constraint (ASC). The image consists of 200 × 200 pixels with 224 spectral bands in the wavelength from 400 nm to 2500 nm with spectral resolution. Fig. 3 shows a false color image of the synthetic data and five endmembers used for data simulation. The details of data simulation process can be unfolded as follows: Firstly, given five reference endmembers from USGS library, we multiply randomly-generated scaling factors ranging in [0.75, 1.25] by the spectral signatures, then a 25 dB white Gaussian noise was added to these scaled reference endmembers. Secondly, we linearly mix them with the generated abundance maps. Finally, an additive 25 dB white Gaussian was again added to the mixed spectrum. Using this simulation process, the spectral signature of each pixel in this dataset should be able to have a complex spectral variability consisting of endmember-dependent scaling factors and complex noise. Therefore, this simulated data with such spectral variability will give us a proper scenario to validate the proposed approach. More details for generating the simulated data can be found in [14].

2) Experimental Setup: Assuming the presence of pure endmembers in HSI scene, VCA, which is one of the most popular endmember extraction methods, is adopted in this paper to construct the endmember dictionary, while Hysime is used to estimate the number of endmembers. Next, these extracted endmembers can be effectively identified using the spectral angle compared to five reference endmembers.

To fairly assess the unmixing performance, we set the optimal parameters for the different algorithms. Both SUnSAL and SSUnSAL are parameterized by $2e-3$ on the sparsity-promoting term, while three regularization parameters [12] for abundances, endmembers, and perturbation in the PLMM are set to be $1e-2$, $1e-2$, and 1, respectively. The regularization parameter λ_S [14] in the ELMM is set to be 0.5. We also set the parameters of SLRU's sparse and low-rank terms to $2e-3$ and $1e-2$. α, β, and γ in Eqs. (16) and (17) can be set to 0.1, 0.01, and $8e-3$, respectively to maximize the performance of SULoRA.

Considering a fact that our method is an alternating minimizing optimization problem for multi-variables, a proper initialization would lead to a fast and reasonable solution. The abundance maps ($\mathbf{X}_0$) is initialized using the output of SPCLSU. Please refer to Algorithm 1 for more parameter settings.

We draw on three criteria of [14] to quantify the unmixing results, that is abundance overall root mean square error (aRMSE), reconstruction overall root mean square error (rRMSE), and average spectral angle mapper (aSAM). When the groundtruth of abundance maps ($\mathbf{X}^g = [\mathbf{x}_1^g, \ldots, \mathbf{x}_i^g, \ldots, \mathbf{x}_N^g] \in \mathbb{R}^{P\times N}$) is given, and then the estimated abundance maps ($\mathbf{X}^e = [\mathbf{x}_1^e, \ldots, \mathbf{x}_i^e, \ldots, \mathbf{x}_N^e] \in \mathbb{R}^{P\times N}$) can be measured by aRMSE defined as

$$aRMSE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{P}\sum\nolimits_{p=1}^{P}(\mathbf{x}_{pi}^e - \mathbf{x}_{pi}^g)^2}. \tag{32}$$

If without the reference of abundance maps, the other two rules (rRMSE and aSAM) are used by computing reconstruction errors between the observed hyperspectral data $\mathbf{Y}^o = [\mathbf{y}_1^o, \ldots, \mathbf{y}_i^o, \ldots, \mathbf{y}_N^o] \in \mathbb{R}^{D\times N}$ and its reconstruction $\mathbf{Y}^r = [\mathbf{y}_1^r, \ldots, \mathbf{y}_i^r, \ldots, \mathbf{y}_N^r] \in \mathbb{R}^{D\times N}$. The former is defined by

$$rRMSE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{D}\sum\nolimits_{l=1}^{D}(\mathbf{y}_{di}^r - \mathbf{y}_{di}^o)^2}, \tag{33}$$

while the latter is aSAM, expressed as

$$aSAM = \frac{1}{N}\sum_{k=1}^{N} arccos\left(\frac{\mathbf{y}_i^{r\mathrm{T}}\mathbf{y}_i^o}{\|\mathbf{y}_i^r\|\|\mathbf{y}_i^o\|}\right). \tag{34}$$

For a fair and reasonable comparison, we average the results of the three criteria out of 10 runs for the different algorithms, because VCA cannot always guarantee the same estimations in each round.

3) Results and Discussion: Fig. 4 shows the estimated abundance maps of the different algorithms, while Table I correspondingly lists the quantitative assessment for three different indices (aRMSE, rRMSE, and aSAM) and computational cost for each algorithm. Since the visual difference of Fig. 4 is not salient, we highlight the differences by the abundance difference maps displayed in Fig. 5.

Visually, FCLSU and PLMM yield a poor performance due to the presence of the spectral variability in the simulated scene. More precisely, the abundance maps estimated by FCLSU fully absorb the spectral variabilities, attributing to the sum-to-one constraint. Taking the rest of algorithms by and large, those of modeling scaling factors outperform those without considering ones. A similar quantitative trend also can be found in Table I. In details, the performance of PCLSU is better than that of FCLSU, since the PCLSU's abundances can be reasonably estimated in a cone not in a simplex by dropping the ASC. Actually the spectral variability is not eliminated by PCLSU, but still partially absorbed by the abundances. Fig. 4 provides a convincing evidence that the abundances for some pixels are higher than 1, and this violates the ASC. By trickily alleviating the effects of scaling factors, the abundances estimated by SPCLSU are more accurate than PCLSU's. Putting the sparse prior on the

TABLE I
THE QUANTITATIVE COMPARISON OF UNMIXING PERFORMANCE FOR THE DIFFERENT ALGORITHMS ON THE SYNTHETIC DATA. THE BEST ONE IS MARKED IN BOLD

Algorithm	FCLSU	PCLSU	SUnSAL	SPCLSU	SSUnSAL	SLRU	PLMM	ELMM	SURoLA
aRMSE	0.0630	0.0421	0.0399	0.0263	0.0243	0.0239	0.0621	0.0323	**0.0220**
rRMSE	0.0150	0.0123	0.0123	0.0123	0.0123	0.0123	0.0129	0.0058	**0.0011**
aSAM	1.9836	1.7717	1.7726	1.7717	1.7726	1.7712	1.8427	0.8392	**0.1789**
$\mathcal{O}$	DPN	DPN	DPN	DPN	DPN	DPN	DP^2N	DP^2N	D^2N

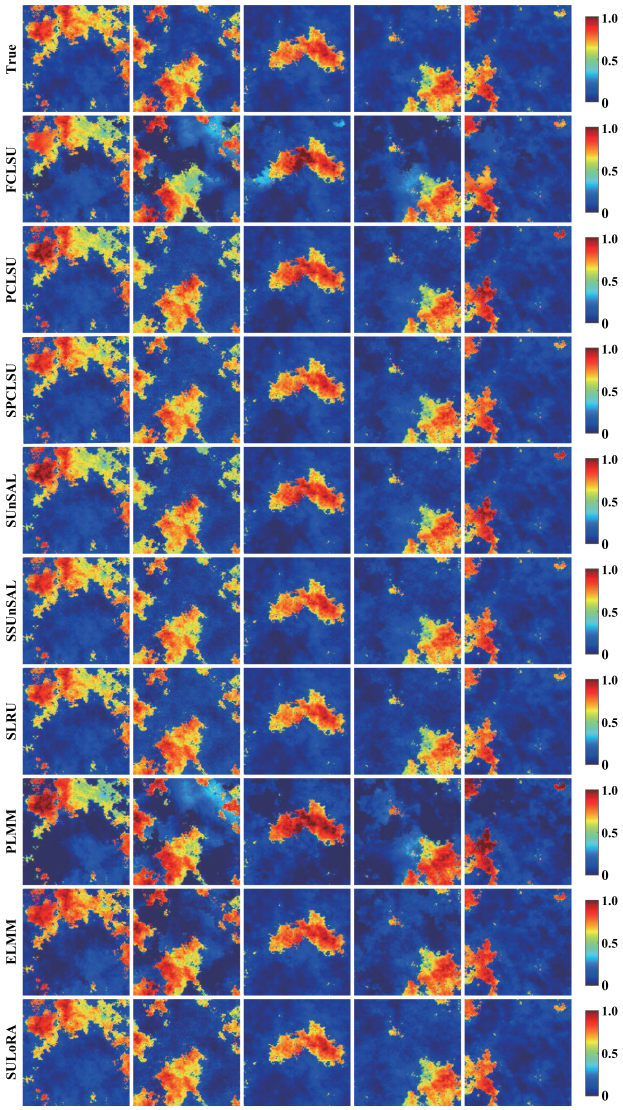


Fig. 4. Abundances estimated by different SU methods (each column corresponds to one endmember extracted by VCA) and the first row shows the ground truth.

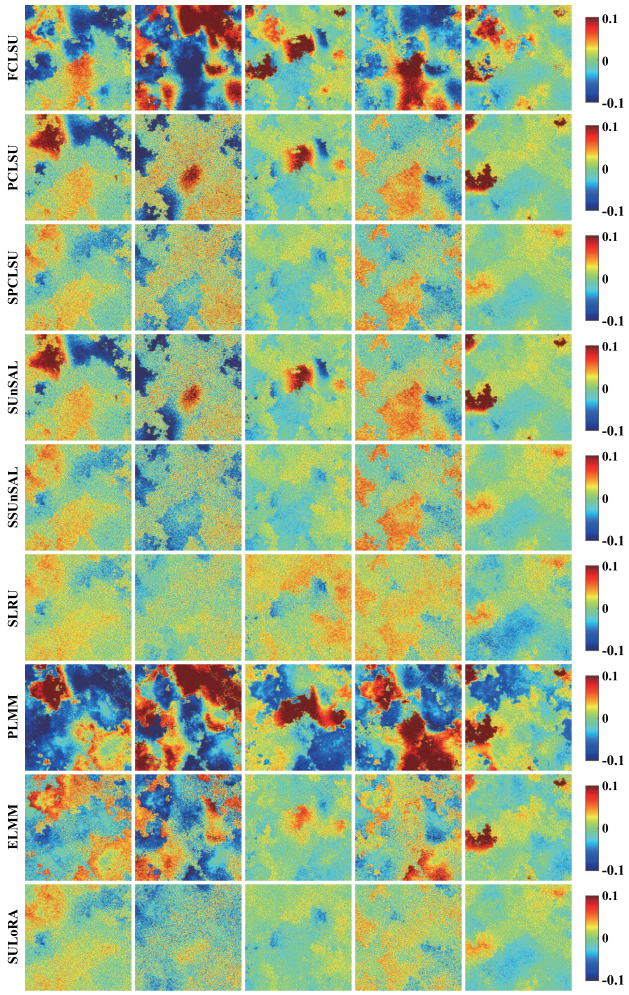


Fig. 5. Difference abundance maps using different spectral unmixing methods corresponding to Fig. 4.

abundance maps, SUnSAL and its scaled version (SSUnSAL) can further improve the performance compared to those without the sparsity-promoting term. This indirectly demonstrates that each pixel in HSI is composed of a few materials. In SLRU, the abundance maps are simultaneously constrained to be sparse and low-rank, leading to a slight improvement compared to only sparsity-promoting SUnSAL algorithm.

The ability in handling the other spectral variability that scaling factors can not be explained limits the ELMM. Furthermore, ELMM needs to simultaneously estimate a coupled set of variables (the scaling factors and abundance maps), this leads to a non-convex optimization problem, which easily drops to a local minimum. In a local region of HSI, the scaling factors for the different endmembers are highly correlated, because the endmember variability is dominated by the topography structure. This is possibly another factor that hinders the performance of the ELMM improving. For the PLMM, it attempts to model the spectral variabilities in a general way, but only a perturbed

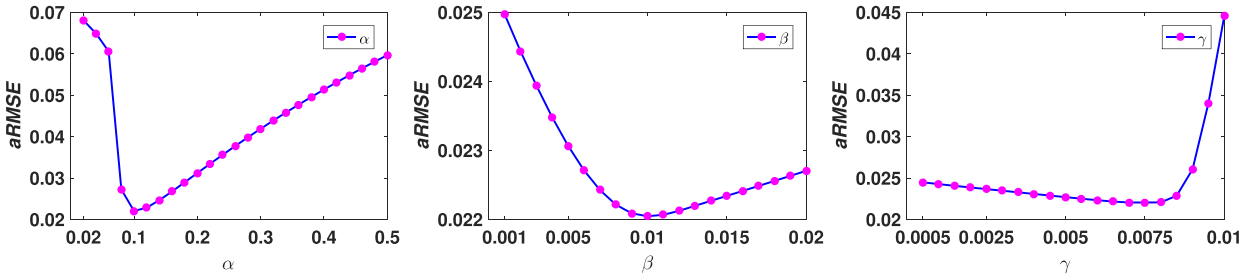


Fig. 6. Sensitivity analysis of three regularization parameters (e.g., α, β, and γ) in SULoRA (Eq. 18).

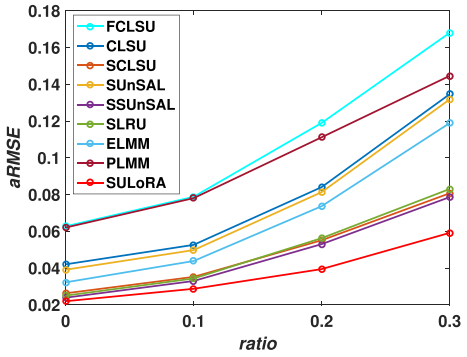
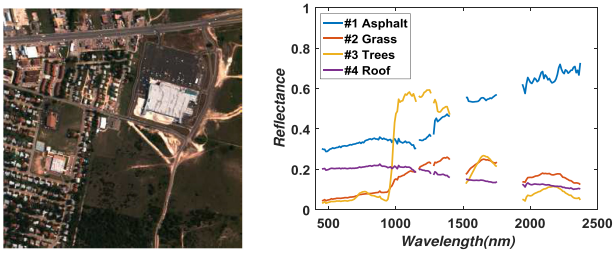


Fig. 7. Robustness evaluation of these compared algorithms using a RMSE at the different sparse noise ratio.



(a) A false color image (b) Endmembers

Fig. 8. A false color image of the Urban data and four extracted endmembers used in spectral unmixing.

information assumed by a Gaussian prior fails to represent the spectral variability (e.g. scaling factors).

As expected, the performance of the subspace-based spectral unmixing (the proposed SULoRA) is superior to that of other algorithms unmixing in the original hyperspectral space, indicating its superiority and effectiveness in dealing with the spectral variability. Fig. 5 highlights a more significant comparison using abundance difference maps between the groundtruth and the estimated abundance maps, where there are lower difference values in SULoRA than in others.

4) Parameters Sensitivity Analysis: The performance of the proposed SULoRA algorithm in Eq. (18) is, to some extent, sensitive to the setting of three regularization parameters (α, β, and γ), it is, as a result, indispensable to search a set of optimal parameter combination. For this reason, the corresponding experiments are conducted to investigate the parameters effects on the performance of estimating abundance maps (measured by aRMSE), as specifically shown in Fig. 6 where the optimal parameter combination in SULoRA is $\alpha = 0.1$, $\beta = 0.01$, and $\gamma = 8e-3$, respectively.

5) Robustness Analysis to Sparse Noise: We further investigate the robustness of the SULoRA against *sparse noise*. For this purpose, the simulated data is corrupted by sparse noise with different corrupted levels, namely $ratio = 0, 0.1, 0.2, 0.3$, where $ratio = 0$ denotes no additional sparse noise is added to the simulated data while $ratio = 0.1$, for instance, means that the 10% of total pixels are corrupted by additional sparse noise. Please refer to [39]–[41] for more experimental setting. As can be seen from Fig. 7, with the increase of *sparse noise ratio*, the performance of most compared approaches dramatically degrades, yet SULoRA still holds a stable and robust performance.

B. Real Data Over Urban Area

1) Data Description: This dataset was acquired by the Hyperspectral Digital Imagery Collection Experiment (HYDICE) over an urban area of Copperas Cove, Texas, USA. The entire image consists of 307×307 pixels at a ground sampling distance (GSD) of 2 m, and 58 noisy bands are removed, so that a total of 162 bands covering the spectral rank from 400 nm to 2500 nm with spectral resolution of 10 nm is selected by removing 58 noisy bands corrupted by water absorption and atmospheric effects in our experiments. This dataset used in hyperspectral unmixing has been widely reported in [42]–[44]. Additionally, we use a latest data version issued by Geospatial Research Laboratory (USA) and Engineer Research and Development Center (USA) in 2015.[1] Fig. 8(a) shows a false color image of the study scene and the endmembers are extracted by VCA.

2) Experimental Setup: There are four main endmembers in the scene: asphalt (road and parking lot), grass, trees, and roof. Please see the references [42] and [44] for more details. Similarly to the first data, HySime and VCA are adopted to determine the number of endmembers and extract the endmembers, respectively. Fig. 8(b) shows the endmembers used in spectral unmixing. The endmembers can be simply identified by comparing with the reference endmembers.[2]

According to two indices of aRMSE and aSAM, we select the optimal parameters for these compared algorithms. The parameters for the sparse and low-rank regularization terms in SLRU are set to $1e-2$ and $1e-2$. The sparsity-promoting term in SUnSAL and SSUnSAL is penalized by $6e-3$, while for PLMM, three regularization parameters for abundances, endmembers,

[1] http://www.tec.army.mil/Hypercube

[2] The reference endmembers can be introduced in [44] and [43].

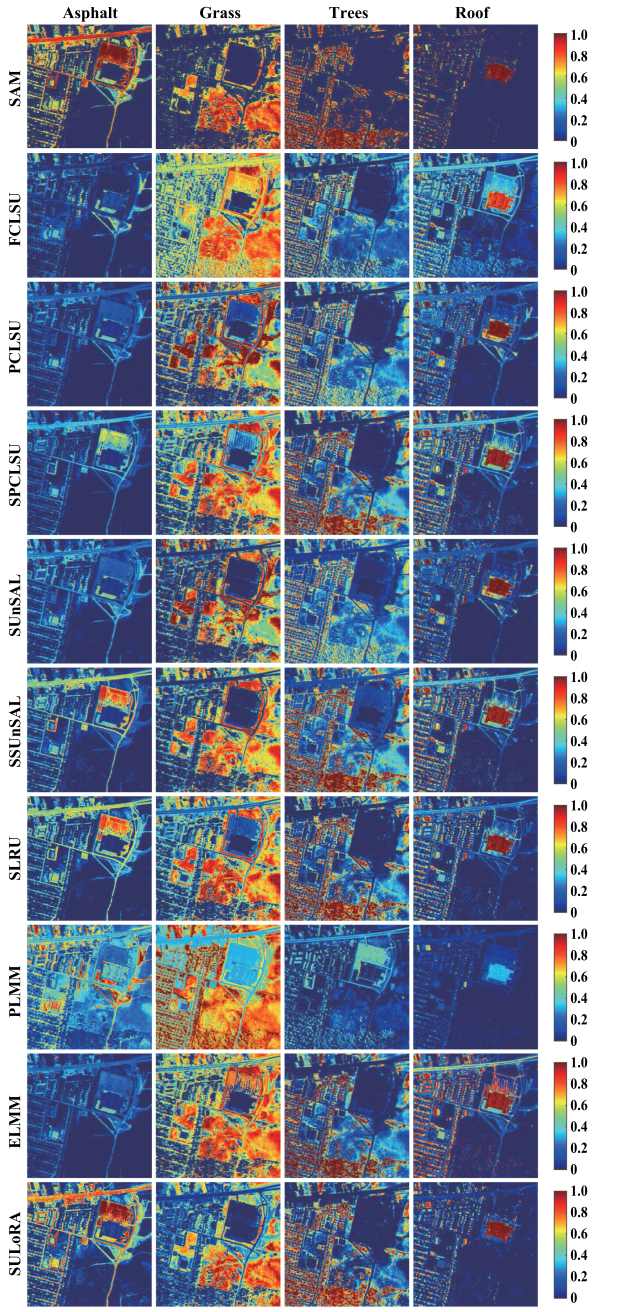


Fig. 9. Abundance maps comparison between the proposed method and the state-of-art methods.

and perturbation are selected to be $1e-2$, $1e-3$, and 1, respectively. The balance parameter λ_S in the ELMM is still 0.5. We finely tune α, β, and γ in SULoRA to 0.1, 0.01, and $5e-3$, respectively.

3) Results and Analysis: As there are no references of the abundance maps for the urban dataset, we propose to apply classification maps, i.e. overall accuracy (OA), to approximately assess the abundance maps. By comparing with the reference endmembers, the spectral angle mapper (SAM) is used to roughly generate classification results, as shown in the first row of Fig. 9 where the positive samples are marked in cosine similarity, while negative samples are masked out with 0. More specifically, we classify each pixel into an endmember with a maximum abundance response. As a result, OA can be regarded as a new index for evaluating the different methods, as listed in Table II. FCLSU performs a worse estimation in the abundances compared to other algorithms, since a more complex spectral variability comes into play in the real data. PCLSU sill fails to well deal with such spectral variability, despite a better performance than FCLSU. As visually shown in the Fig. 9, SPCLSU can effectively identify the materials of asphalt, trees, and roof, while considering scaling factors. As a comparison, neither FCLSU nor PCLSU detects the material of the asphalt, but SPCLSU successfully does. The regular pattern is also applicable to SUnSAL and SSUnSAL. By additionally considering a low-rank prior in the process of estimating abundance maps, SLRU performs better than SUnSAL, but it still fails to address the complex spectral variability.

Although ELMM is able to detect some areas, e.g. trees and roof, the complex spectral variability in the real scenario can not be fully interpreted only by scaling factors. This results in a relatively lower rRMSE and aSAM, as listed in Table II. On the other hand, the hard optimization problem in ELMM is another drawback, limiting ELMM up to a better performance. The main factor for the poor performance of PLMM is lack of a powerful fitting ability in the spectral variability by analyzing the visual and quantitative results from both Fig. 9 and Table II.

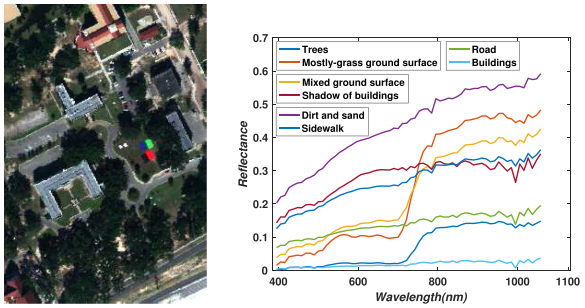
Thanks to the high-resolution of the urban HSI, we can find many pure pixels, but they are mistaken as mixed pixels with the existence of spectral variability. This easily makes many pixels misclassified using the aforementioned methods. Different with them, SULoRA can estimate the abundance maps in a robust subspace, so that its visual effect is superior to others', as shown in Fig. 9, and a consistent numerical evaluation is also listed in Table II. For instance, the asphalt and grass can be purely identified by SULoRA, unlike the others. The abundance maps of the tree and roof estimated by SULoRA show higher contrast as well. These phenomena can objectively explain the robustness and effectiveness of the proposed method.

C. Real Data (MUUFL Gulfport Campus)

1) Data Description: As introduced in [45], [46], the labeled hyperspectral image can be used for ultimately assessing the unmixing performance, hence the MUUFL Gulfport dataset is chosen as the second real data in our case, collected over the campus area in University of Southern Mississippi-Gulfpark Campus, Long Beach, Mississippi, USA [47]. It consists of 325×220 pixels at a GDS of 1 m. There are 11 classes in this study scene, but we just consider 8 main classes as they have enough number of pixels and clear spatial structure for a easier visualization, that is *#1 trees*, *#2 mostly-grass ground surface*, *#3 mixed ground surface*, *#4 dirt and sand*, *#5 road*, *#6 buildings*, *#7 shadow of buildings*, and *#8 sidewalk*. The 8 noisy bands were removed, resulting in a total of 64 bands left in the spectral range from 375 nm to 1050 nm. Fig. 10 shows a RGB image and the endmembers extracted by VCA of the used scene.

TABLE II
THE QUANTITATIVE COMPARISON OF UNMIXING PERFORMANCE FOR THE DIFFERENT ALGORITHMS ON THE REAL URBAN DATA. THE BEST ONE IS MARKED IN BOLD

Algorithm	FCLSU	PCLSU	SUnSAL	SPCLSU	SSUnSAL	SLRU	PLMM	ELMM	SULoRA
OA (%)	54.66	68.08	71.26	68.08	71.26	73.59	58.55	62.41	**86.71**
rRMSE	0.0397	0.0086	0.0086	0.0086	0.0086	0.0084	0.0111	0.0072	**0.0019**
aSAM	8.6734	2.9569	2.9582	2.9569	2.9582	2.9472	3.6240	2.0378	**0.6491**



(a) A false color image (b) Endmembers

Fig. 10. A RGB image of the MUFFLE dataset and eight extracted endmembers used in spectral unmixing.

2) Experimental Setup: Likewise, the number of endmembers can be estimated by HySime and the endmembers can be extracted by VCA. The extracted endmembers are handily identified using SAM, as massive labeled samples for each class are available.

The optimal parameters for all compared methods and the proposed SULoRA are detailed in the following. The l_1-norm term in SUnSAL and SSUnSAL is parameterized by $3e-4$, while the parameters for SLRU are $2e-4$ and 0.1, respectively. Three regularization parameters in PLMM are set to be $1e-3$, $1e-2$, and 1, respectively, while the parameter λ_S in ELMM plays a role in balancing the two fidelity terms, which is assigned to 0.5 in our case. For SULoRA, α, β, and γ are experimentally assigned to 0.8, 0.1, and $6e-4$, respectively.

3) Results and Analysis: Given these labeled classification maps of each class as shown in the first row of Fig. 11, classification (e.g., OA) can be explored as a potential way to evaluate the quality of estimated abundance maps. Correspondingly, Table III quantitatively lists the performance assessment (three indices: OA, rRMSE, and aSAM) for all algorithms.

FCLSU shows a poor estimation in abundance maps, since it fails to model the complex spectral variabilities. For those algorithms that provide different priors in estimating the abundance maps, e.g., scaling (SPCLSU, SSUnSAL), sparse (SUnSAL, SSUnSAL), low-rank (SLRU), etc., there is a moderate performance improvement compared to those without considering prior knowledge. One thing to be noted is that PLMM obtains desirable results of rRMSE and aSAM in comparison with previous methods (expect our proposed SULoRA), but interestingly it yields a poorest OA. The reason for this mainly lies in that only perturbation information hardly represents the complex spectral variability, and meanwhile such modeling strategy could also corrupt some important spectral attributes misdeemed as certain

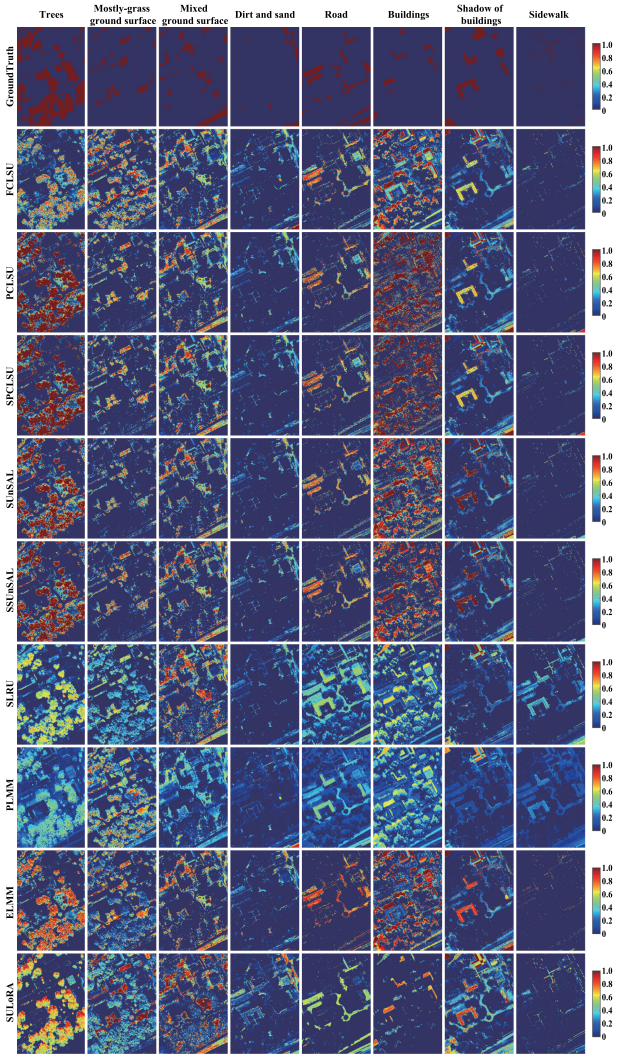


Fig. 11. Abundance maps comparison between the proposed method and the state-of-art methods.

spectral variability. As can be seen from Fig. 11, ELMM obtains a good abundance estimation, since it is good at handling the scaling factors (principle spectral variability). But unfortunately, ELMM's performance is limited by the presence of other spectral variabilities. In a word, these previously proposed methods basically pay more attentions on somewhat special spectral variability, lacking of generalization ability. Considering the complexity of the spectral variability in real-world, the proposed SULoRA accounts for spectral variability in a generalized fashion by embedding the low-rank attribute, resulting in more robust

TABLE III
THE QUANTITATIVE COMPARISON OF UNMIXING PERFORMANCE FOR THE DIFFERENT ALGORITHMS ON THE MUFFLE GULFPORT CAMPUS DATA. THE BEST ONE IS MARKED IN BOLD

Algorithm	FCLSU	PCLSU	SUnSAL	SPCLSU	SSUnSAL	SLRU	PLMM	ELMM	SULoRA
OA (%)	55.52	62.21	66.61	62.21	66.61	66.58	51.65	63.16	**85.86**
rRMSE	0.0135	0.0110	0.0109	0.0110	0.0109	0.0162	0.0099	0.0127	**0.0058**
aSAM	5.5854	4.7699	4.7830	4.7699	4.7830	6.6452	4.0436	5.1839	**2.6003**

and effective unmixing results visually and quantitatively (see Fig. 11).

V. CONCLUSION

This paper is motivated by the fact that the spectral signature in the original hyperspectral space inevitably suffers from largely and diversely spectral variabilities. To address this issue, we propose to unmix the HSI in a subspace instead of in the original space. This results in a general subspace unmixing framework that jointly learns a subspace projection and abundance maps. With the low-rank attribute embedding, we further develop a low-rank subspace unmixing approach, called **s**pectral **u**nmixing with **lo**w-**r**ank **a**ttribute embedding (SULoRA). Experimental results demonstrate that SULoRA is able to obtain a higher unmixing performance both visually and quantitatively, than other state-of-the-art algorithms. In the future, we would like to cast the subspace-based framework to advanced unmxing methods designed in the original spectral space, aiming at a more robust spectral unmixing.

ACKNOWLEDGMENT

The authors would like to thank P.-A. Thouvenin from Institut de Recherche en Informatique de Toulouse and Prof. J. M. Bioucas-Dias for providing MATLAB codes with respect to PLMM and SUnSAL tested in their experiments, and the Hyperspectral Digital Imagery Collection Experiment for sharing the urban dataset free of charge. The authors would like to express their appreciation to Prof. J. Chanussot for providing the simulated dataset used in their first experiment as well as MATLAB codes for ELMM.

REFERENCES

[1] D. Hong, N. Yokoya, and X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, Jul. 2017.

[2] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 469–484.

[3] M. Veganzones, G. Tochon, M. Dalla-Mura, A. Plaza, and J. Chanussot, "Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3574–3589, Aug. 2014.

[4] T. Matsuki, N. Yokoya, and A. Iwasaki, "Hyperspectral tree species classification of Japanese complex mixed forest with the aid of lidar data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2177–2187, Apr. 2015.

[5] Z. Wang, R. Zhu, K. Fukui, and J. Xue, "Matched shrunken cone detector (MSCD): Bayesian derivations and case studies for hyperspectral target detection," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5447–5461, Aug. 2017.

[6] D. A. Roberts, M. Gardner, R. Church, S. Ustin, G. Scheer, and R. O. Green, "Mapping chaparral in the Santa Monica mountains using multiple endmember spectral mixture models," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 267–279, 1998.

[7] C. Yang, J. H. Everitt, Q. Du, B. Luo, and J. Chanussot, "Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture," *Proc. IEEE*, vol. 101, no. 3, pp. 582–592, Jul. 2013.

[8] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Feb. 2012.

[9] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 95–104, Dec. 2014.

[10] O. Eches, N. Dobigeon, C. Mailhes, and J.Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1403–1413, Jun. 2010.

[11] X. Du, A. Zare, P. Gader, and D. Dranishnikov, "Spatial and spectral unmixing using the beta compositional model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1944–2003, Jun. 2014.

[12] P. A. Thouvenin, N. Dobigeon, and J. Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 525–538, Jun. 2016.

[13] X. Fu, W. K. Ma, J. M. Bioucas-Dias, and T. H. Chan, "Semiblind hyperspectral unmixing in the presence of spectral library mismatches," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5171–5184, May 2016.

[14] L. Drumetz, M. A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3890–3905, Aug. 2016.

[15] D. C. Heinz and C. I. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 3, pp. 529–545, Mar. 2001.

[16] R. Heylen, D. Burazerovic, and P. Scheunders, "Fully constrained least squares spectral unmixing by simplex projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4112–4122, Nov. 2011.

[17] J. M. Bioucas-Dias and M. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Proc. IEEE Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2010, pp. 1–4.

[18] M. A. Veganzones *et al.*, "A new extended linear mixing model to address spectral variability," in *Proc. IEEE Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2014, pp. 1–5.

[19] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "Learning low-coherence dictionary to address spectral variability for hyperspectral unmixing," in *Proc. IEEE Int. Conf. Image Process.*, Beijing, China, Sep. 2017, pp. 235–239.

[20] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 898–910, Apr. 2005.

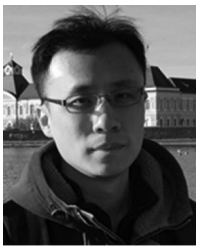
[21] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4484–4502, Nov. 2012.

[22] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 341–354, Jan. 2014.

[23] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM. J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, Jul. 2008.

[24] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3739–3747.

[25] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[26] N. Shahid, N. Perraudin, V. Kalofolias, G. Puy, and V. P. Andergheynst, "Fast robust PCA on graphs," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 740–756, Jun. 2016.
[27] L. Zhang, W. Wei, Y. Zhang, C. Shen, A. van den Hengel, and Q. Shi, "Cluster sparsity field: An internal hyperspectral imagery prior for reconstruction," *Int. J. Comput. Vis.*, vol. 126, pp. 797–821, 2018.
[28] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Jul. 2008.
[29] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2010, arXiv:1009.5055.
[30] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.
[31] L. Yang *et al.*, "Image reconstruction via manifold constrained convolutional sparse coding for image sets," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1072–1081, Oct. 2017.
[32] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Apr. 2013.
[33] L. Chen, X. Li, D. Sun, and K. Toh, "On the equivalence of inexact proximal ALM and ADMM for a class of convex composite programming," 2018, arXiv:1803.10803.
[34] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers Math. China*, vol. 7, no. 2, pp. 365–384, Apr. 2012.
[35] Y. Zhang, Z. Jiang, and L. Davi, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 676–683.
[36] D. Fortun, P. Guichard, N. Chu, and M. Unser, "Reconstruction from multiple poses in fluorescence imaging: Proof of concept," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 1, pp. 61–70, Feb. 2016.
[37] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 608–622, Jun. 2016.
[38] P. V. Giampouras, K. E. Themelisand A. A. Rontogiannis, and K. D. Koutroumbas, "Simultaneously sparse and low-rank abundance matrix estimation for hyperspectral image unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4775–4789, Aug. 2016.
[39] W. He, H. Zhang, and L. Zhang, "Sparsity-regularized robust non-negative matrix factorization for hyperspectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4267–4279, Sep. 2016.
[40] Y. Chen, Y. Guo, Y. Wang, D. Wang, C. Peng, and G. He, "Denoising of hyperspectral images using nonconvex low rank matrix approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5366–5380, Sep. 2017.
[41] W. He, H. Zhang, H. Shen, and L. Zhang, "Hyperspectral image denoising using local low-rank matrix recovery and global spatial–spectral total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 713–729, Mar. 2018.
[42] F. Zhu, Y. Wang, S. Xiang, B. Fan, and C. Pan, "Structured sparse NMF for hyperspectral unmixing," *ISPRS J. Photogramm. Remote Sens.*, vol. 88, no. 1, pp. 101–118, Feb. 2014.
[43] Y. Wang, C. Pan, S. Xiang, and F. Zhu, "Robust hyperspectral unmixing with correntropy based metric," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4027–4040, Jul. 2015.
[44] X. Liu, W. Xia, B. Wang, and L. Zhang, "An approach based on constrained nonnegative matrix factorization to unmix hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 757–772, Feb. 2011.
[45] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.
[46] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 4959–4962.
[47] X. Du and A. Zare, "Technical report: Scene label ground truth map for MUUFL Gulfport data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. 20170417, Apr. 2017.



Danfeng Hong (S'16) received the B.Sc. degree in computer science and technology from Northeastern University, Shenyang, China, in 2012, and the M.Sc. degree in computer vision from Qingdao University, Qingdao, China, in 2015. He is currently working toward the Ph.D. degree with the Signal Processing in Earth Observation, Technical University of Munich, Munich, Germany, and the Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling , Germany. His research interests include image processing, pattern recognition, and machine learning and their applications to hyperspectral data analysis.

Xiao Xiang Zhu (S'10–M'12–SM'14) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from Technical University of Munich, Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently a Professor with the Signal Processing in Earth Observation (www.sipeo.bgu.tum.de) at Technical University of Munich, Munich, Germany, and German Aerospace Center (DLR), Wessling, Germany, the head of the department "EO Data Science" at DLR's Earth Observation Center, and the head of the Helmholtz Young Investigator Group "SiPEO" at DLR and TUM. He was a Guest Scientist or visiting professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests are remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Appendices

E Hong D., Yokoya N., Chanussot J., Zhu X. X., 2019. CoSpace: Common Subspace Learning from Hyperspectral-Multispectral Correspondences. IEEE Transactions on Geoscience and Remote Sensing (TGRS), 57(7): 4349-4359.

https://ieeexplore.ieee.org/document/8672122

CoSpace: Common Subspace Learning From Hyperspectral-Multispectral Correspondences

Danfeng Hong, *Student Member, IEEE*, Naoto Yokoya, *Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— With a large amount of open satellite multispectral (MS) imagery (e.g., Sentinel-2 and Landsat-8), considerable attention has been paid to global MS land cover classification. However, its limited spectral information hinders further improving the classification performance. Hyperspectral imaging enables discrimination between spectrally similar classes but its swath width from space is narrow compared to MS ones. To achieve accurate land cover classification over a large coverage, we propose a cross-modality feature learning framework, called common subspace learning (CoSpace), by jointly considering subspace learning and supervised classification. By locally aligning the manifold structure of the two modalities, CoSpace linearly learns a shared latent subspace from hyperspectral-MS (HS-MS) correspondences. The MS out-of-samples can be then projected into the subspace, which are expected to take advantages of rich spectral information of the corresponding hyperspectral data used for learning, and thus leads to a better classification. Extensive experiments on two simulated HS-MS data sets (University of Houston and Chikusei), where HS-MS data sets have tradeoffs between coverage and spectral resolution, are performed to demonstrate the superiority and effectiveness of the proposed method in comparison with previous state-of-the-art methods.**

***Index Terms*— Common subspace learning (CoSpace), cross-modality learning, hyperspectral, landcover classification, multispectral (MS), remote sensing.**

I. Introduction

RECENTLY, the launch of operational optical broadband [multispectral (MS)] satellites has successfully boosted the usage of MS data for various tasks such as urban monitoring, management of natural resources, ecosystem, and disasters prediction. There has been a growing interest in large-scale land cover mapping of urban [1], agriculture monitoring [2], [3], and mineral exploration [4], since high-quality MS satellite imagery is openly available on a global scale (e.g., Sentinel-2 and Landsat-8). However, MS data fail to discriminate spectrally similar classes due to its broad spectral bandwidth. Hyperspectral imaging can acquire richer spectral information that enables high discrimination ability but its coverage from space is much narrower than the one of MS imaging due to the limitations of imaging devices and satellite techniques. This tradeoff naturally motivates us to ponder a question: *can HS imagery covering only a limited part of the MS imagery be explored to improve the classification of the entire area covered by the MS imagery?* This is as a typical cross-modal feature learning problem.

Researchers have proposed a variety of multimodal feature learning algorithms by introducing additional information, which can be roughly categorized into two parts: fusion-based joint feature learning (FJFL) [5], [6] and alignment-based shared feature learning (ASFL) [7]. The main difference between FJFL and ASFL is illustrated in Fig. 1.

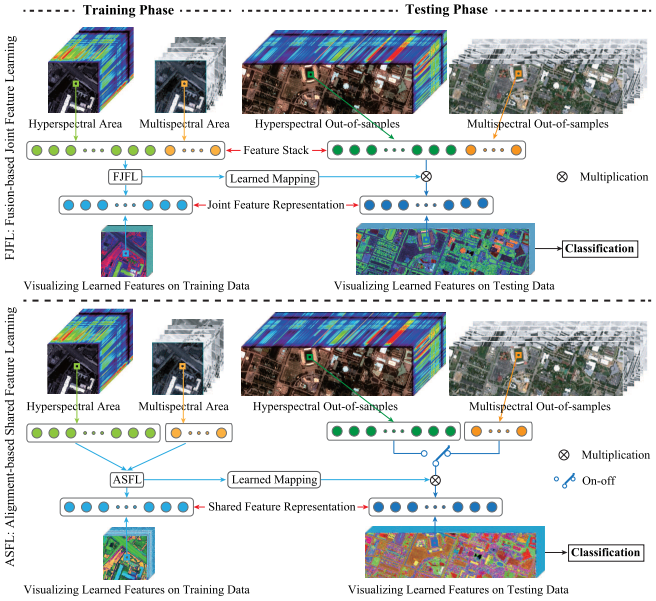


Fig. 1. Illustration of the two kinds of multimodal feature learning frameworks, where the switch (on–off) means that only one modality is involved as the testing samples to meet the hypothesis of cross-modal learning.

Manuscript received April 18, 2018; revised October 11, 2018; accepted December 29, 2018. Date of publication March 20, 2019; date of current version June 24, 2019. This work was supported in part by the European Research Council through the European Union's Horizon 2020 Research and Innovation Programme under Grant ERC-2016-StG-714087, in part by Helmholtz Association through the Framework of the Young Investigators Group "Signal Processing in Earth Observation (SiPEO)" under Grant VH-NG-1018, and in part by the German Research Foundation (DFG) under Grant ZH 498/7-2. The work of N. Yokoya was supported in part by the Japan Society for the Promotion of Science under Grant Grants-in-Aid for Scientific Research (KAKENHI) 15K20955 and in part by Alexander von Humboldt Fellowship for Post-Doctoral Researchers. *(Corresponding author: Xiao Xiang Zhu.)*

D. Hong and X. X. Zhu are with the Remote Sensing Technology Institute (IMF), German Aerospace Center, 82234 Weßling, Germany, and also with Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: danfeng.hong@dlr.de; xiaoxiang.zhu@dlr.de).

N. Yokoya is with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: naoto.yokoya@riken.jp).

J. Chanussot is with the GIPSA-Lab, CNRS, Grenoble INP, Université Grenoble Alpes, F-38000 Grenoble, France, and also with the Faculty of Electrical and Computer Engineering, University of Iceland, 101 Reykjavik, Iceland (e-mail: jocelyn@hi.is).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2018.2890705

FJFL aims to learn discriminative features by absorbing the different properties from multimodal data. FJFL fuses the different sources at the data level to diversify the information and then to further learn the higher level feature representation. One intuitive way for FJFL is to directly learn a joint data representation at the feature level. At present, this is the mainstream approach for multimodal data analysis [8]. For example, by embedding the height information from light detection and ranging (LiDAR) into MS (HS) data, Ghamisi *et al.* [9] learned multifold features from HS and LiDAR correspondences for a multimodal classification task. Iyer *et al.* [10] provided a graph-based new perspective for feature extraction and segmentation of multimodal images and achieved a desirable result. The resulting discriminative features are beneficial for improving the performance of some high-level applications, especially classification [11], [12], object detection [13], image/video analysis [14], and spectral unmixing [15]. Image fusion can also be regarded as a part of FJFL when feature learning is applied subsequently. For instance, hyperspectral and MS (HS-MS) data fusion enhances the spectral resolution of MS data by fusing it with the low-spatial-resolution HS data [5]. The fused HS-MS product can be then seen as a new input for further discriminative feature learning.

Behind the advancement of FJFL, the *complete data correspondence* is the prerequisite. This limitation undoubtedly results in a poor fit for cross-modal data analysis, in particular, for cross-modal feature learning [16].[1] In our MS-HS case, the cross-modal learning refers to a problem that given a large-scale MS image and a limited HS area partially overlapping with the MS data (see Fig. 2, for example), we learn the low-dimensional embedding representation from the limited amount of MS-HS correspondences and transfer the learned features to the rest of MS data for improving the performance of large-scale land-cover and land-use mapping. During the process, we expect to transfer the discrimination capability learned from the rich spectral information into MS data through the learned common subspace in order to more effectively identify some challenging classes that are hardly recognized by MS data due to its poor spectral information. Please note that we just start a preliminary investigation of cross-modal learning (MS-HS) in this paper, that is, the MS and HS images share the same land-cover classes.

Unlike FJFL, ASFL is more apt for cross-modal feature learning, since ASFL can adaptively shuttle back and forth between the different modalities or domains by means of the learned common subspace. Matasci *et al.* [17] linearly projected the hyperspectral data of the source and target domains into a common feature space where the gap between domains in hyperspectral image classification is expected to be reduced. Kulis *et al.* [18] addressed the issue of visual domain adaption by learning a nonlinear transformation in kernel space, with the application to general object recognition. In [19], a probabilistic framework was proposed to align the class distributions of two domains for robust hyperspectral image classification. Manifold alignment (MA) [20] is also a powerful tool for modeling this kind of issue. Inspired by MA, Tuia *et al.* [7] aligned multiview remote sensing images on manifolds by fully allowing for the spectral variabilities between the different angle imageries, yielding a significant improvement of classification performance.

[1]In contrast to multimodal learning (bimodality, for example), cross-modal learning trains on single modality and tests on bimodality, or *vice versa* (train on bimodality and test on single modality).

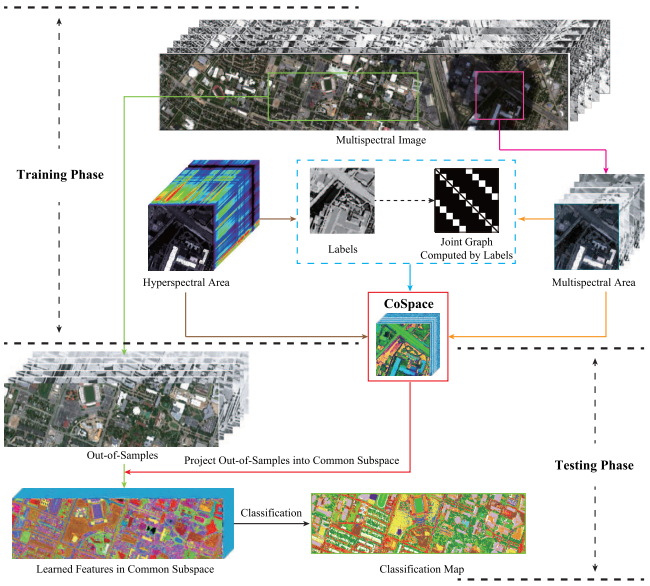


Fig. 2. Holistic workflow of the proposed CoSpace.

It should be noted that these methods mentioned earlier only consider the differences of a unimodality between the source and target domains at the level of original features, but they fail to investigate the transferability of multimodality since the different modalities usually hold the different feature dimensions. Although these approaches can build connections between features or instances, a poorly connected relationship between the learned common subspace and label information is still hindering the low-dimensional feature representation from being more discriminative.

We propose a cross-modality feature learning framework, called common subspace learning (CoSpace), that learns the shared feature representation (common subspace) from partial HS-MS correspondences. Extensive experiments are conducted on simulated MS and partially overlapped real HS data based on two airborne HS data sets: the University of Houston and Chikusei data sets. MS data are generated from HS data by using the spectral response functions (SRFs) of Sentinel-2. We relabel the training and testing classes on the data sets to meet the problem setting of cross-modal feature learning and further to make them more challenging (see Section III for details). Our contributions can be specifically unfolded as follows.

1) We propose a novel CoSpace approach by jointly considering the subspace learning and classification in order to effectively bridge the learned features and label information, aiming at addressing the HS-MS cross-modal feature learning issue.

2) By locally aligning HS-MS data on the low-dimensional manifolds where the features of HS and MS share the same dimension, CoSpace linearly learns a latent shared subspace from HS-MS correspondences, where samples are expected to be classified better. Because of the subspace learned in a linear way, the out-of-samples data can be simply and smoothly embedded.
3) An optimization algorithm based on the alternating direction method of multipliers (ADMMs) is designed to solve the proposed model.

The remainder of this paper is organized as follows. In Section II, we first clarify our motivation and then propose the methodology of the CoSpace model, finally, elaborate on the corresponding ADMM-based optimization algorithm. Section III presents the experimental results and analysis on two different HS-MS data sets both qualitatively and quantitatively. Finally, some conclusions are drawn in Section IV.

II. CoSpace: Common Subspace Learning

To take the benefit of HS imagery covering only a limited part of the MS imagery and, subsequently, improve the classification results of the entire area covered by the MS imagery, our idea is to learn an HS-MS common subspace, in which the data from one domain can be adaptively transferred to another domain.

Our solution to the problem is to learn an HS-MS common subspace, in which the data from one domain can be adaptively transferred to another domain.

Fig. 2 shows the holistic diagram of the proposed CoSpace method.

A. Problem Formulation

Let $\mathbf{X}_M \in \mathbb{R}^{d_M \times N}$ and $\mathbf{X}_H \in \mathbb{R}^{d_H \times N}$ be the observed MS image with d_M bands by N pixels and the HS image with d_H bands by N pixels, respectively. $\mathbf{Y} \in \mathbb{R}^{L \times N}$ is the label matrix represented by one-hot encoding. $\Theta_M \in \mathbb{R}^{d \times d_M}$ ($\Theta_H \in \mathbb{R}^{d \times d_H}$) is denoted as the projection matrix for connecting the MS (HS) data and the latent subspace. The variable $\mathbf{P} \in \mathbb{R}^{L \times d}$ is the weighted matrix specified by bridging the latent subspace and label information. Accordingly, $\widetilde{\mathbf{Y}} = [\mathbf{Y}, \mathbf{Y}] \in \mathbb{R}^{L \times 2N}$ can be modeled as follows.

The CoSpace can be modeled as follows:

$$\widetilde{\mathbf{Y}} = \mathbf{P}\Theta\widetilde{\mathbf{X}} + \mathbf{E} \tag{1}$$

where $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_H \end{bmatrix} \in \mathbb{R}^{(d_M+d_H)\times 2N}$, and $\Theta = [\Theta_M, \Theta_H] \in \mathbb{R}^{d\times(d_M+d_H)}$. $\mathbf{E} \in \mathbb{R}^{L\times 2N}$ is the corresponding residual matrix containing the additive noise and other errors.

Since (1) is a typically ill-posed problem because of more degrees of flexibility involved (e.g., latent subspace estimation), several assumptions (or prior knowledge) should be introduced into CoSpace using regularization technique. Followed by a popular joint learning framework proposed in [21], we formulate the CoSpace as the following constrained optimization problem:

$$\min_{\mathbf{P},\Theta} \left\{ \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \Phi(\mathbf{P}) + \Psi(\Theta) \right\} \quad \text{s.t.} \quad \Theta\Theta^{\mathrm{T}} = \mathbf{I}. \tag{2}$$

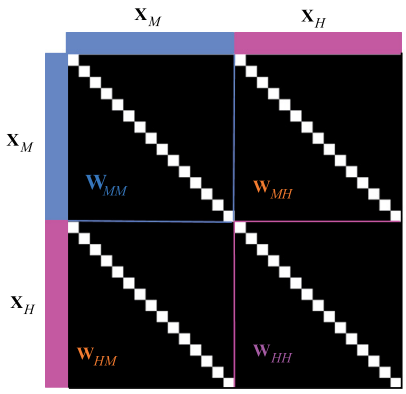


Fig. 3. Example to clarify the joint adjacency matrix.

The two regularization terms in (2) are detailed in the following.

To achieve a reliable generalization of our model, the variable $\mathbf{P}$ parameterized by α can be regularized by a Frobenius norm

$$\Phi(\mathbf{P}) = \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \tag{3}$$

and the prior knowledge with respect to Θ, resulting in a multimodal MA regularization, can be expressed with a joint graph structure as

$$\Psi(\Theta) = \frac{\beta}{2}\,\mathrm{tr}(\Theta\widetilde{\mathbf{X}}\mathbf{L}(\Theta\widetilde{\mathbf{X}})^{\mathrm{T}}) \tag{4}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W} \in \mathbb{R}^{2N\times 2N}$ stands for a joint Laplacian matrix, $\mathbf{W}$ that is a corresponding adjacency matrix can be directly inferred from label information in the form of the linear discriminant analysis (LDA)-like graph [22]

$$\mathbf{W}_{i,j} = \begin{cases} 1/N_k, & \text{if } \mathbf{X}_i \text{ and } \mathbf{X}_j \text{ belong to the } k\text{th class} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

and then $\mathbf{D}$ is computed by $\mathbf{D}_{ii} = \sum_{i\neq j} \mathbf{W}_{i,j}$. Fig. 3 illustrates the joint graph structure.

B. Model Optimization

Considering the nonconvexity of problem (2), an iterative alternating optimization strategy is adopted to solve the convex subproblems of each variable $\mathbf{P}$ and Θ. An implementation of CoSpace is given in Algorithm 1.

Optimization with respect to $\mathbf{P}$: This is a typical least-squares problem with the Tikhonov regularization that can be formulated as

$$\min_{\mathbf{P}} \left\{ \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \right\} \tag{6}$$

which has a closed-form solution

$$\mathbf{P} = (\widetilde{\mathbf{Y}}\mathbf{Q}^{\mathrm{T}})(\mathbf{Q}\mathbf{Q}^{\mathrm{T}} + \alpha\mathbf{I})^{-1} \tag{7}$$

where $\mathbf{Q} = \Theta\widetilde{\mathbf{X}}$.

Algorithm 1: CoSpace

Input: $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{X}}$, $\mathbf{L}$, and parameters α, β, *maxIter*.
Output: $\mathbf{P}$, Θ
$t = 1$, $\zeta = 1e-4$;
Initializating $\mathbf{P}$ and Θ
while *not converged* or $t >$ *maxIter* **do**
  Fix other variables to update $\mathbf{P}$ by (7)
  Fix other variables to update Θ by **Algorithm 2**
  Compute the objective function value E^{t+1} and check the convergence condition: **if** $|\frac{E^{t+1}-E^t}{E^t}| < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t+1$;
  end
end

Algorithm 2: Solving the Subproblem for Θ

Input: $\widetilde{\mathbf{Y}}$, $\mathbf{P}$, $\mathbf{J}$, $\widetilde{\mathbf{X}}$, $\mathbf{L}$, β, *maxIter*.
Output: Θ.
Initialization: $\Theta = \mathbf{0}$, $\mathbf{G} = \mathbf{0}$, $\Lambda_1 = \mathbf{0}$, $\Lambda_2 = \mathbf{0}$, $\mu = 10^{-3}$, $\mu_{\max} = 10^6$, $\rho = 1.5$, $\varepsilon = 10^{-6}$, $t = 1$.
while *not converged* or $t >$ *maxIter* **do**
  Fix other variables to update $\mathbf{J}$ by

$$\mathbf{J} = (\mathbf{P}^{\mathrm{T}}\mathbf{P} + \mu\mathbf{I})^{-1}(\mathbf{P}^{\mathrm{T}}\widetilde{\mathbf{Y}} + \mu\Theta\widetilde{\mathbf{X}} - \Lambda_1).$$

  Fix other variables to update Θ by

$$\Theta = (\mu\mathbf{J}\widetilde{\mathbf{X}}^{\mathrm{T}} + \Lambda_1\widetilde{\mathbf{X}}^{\mathrm{T}} + \mu\mathbf{G} + \Lambda_2) \times(\mu\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}} + \mu\mathbf{I} + \beta\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^{\mathrm{T}})^{-1}.$$

  Fix other variables to update $\mathbf{G}$ by

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{svd}(\Theta - \Lambda_2/\mu), \quad \mathbf{G} = \mathbf{U}\mathbf{I}_{n\times m}\mathbf{V}.$$

  Update Lagrange multipliers by

$$\Lambda_1 \leftarrow \Lambda_1 + \mu(\mathbf{J} - \Theta\widetilde{\mathbf{X}}), \quad \Lambda_2 \leftarrow \Lambda_2 + \mu(\mathbf{G} - \Theta).$$

  Update penalty parameter by

$$\mu = \min(\rho\mu, \mu_{\max}).$$

  Check the convergence conditions: **if** $\|\mathbf{J} - \Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{G} - \Theta\|_{\mathrm{F}} < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t+1$;
  end
end

Optimization with respect to Θ: The optimization problem for Θ can be formulated as

$$\min_{\Theta}\left\{\begin{array}{c}\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\Theta\widetilde{\mathbf{X}}\mathbf{L}(\Theta\widetilde{\mathbf{X}})^{\mathrm{T}}) \\ \text{s.t.} \quad \Theta\Theta^{\mathrm{T}} = \mathbf{I}\end{array}\right\}. \tag{8}$$

In order to solve (8) effectively with ADMM, we consider an equivalent form by introducing auxiliary variables $\mathbf{J}$ and $\mathbf{G}$ to replace $\Theta\widetilde{\mathbf{X}}$ and Θ, respectively,

$$\min_{\Theta,\mathbf{J},\mathbf{G}}\left\{\begin{array}{l}\frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{J}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\Theta\widetilde{\mathbf{X}}\mathbf{L}(\Theta\widetilde{\mathbf{X}})^{\mathrm{T}}) \\ \text{s.t.} \quad \mathbf{J} = \Theta\widetilde{\mathbf{X}}, \ \mathbf{G} = \Theta, \ \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I}\end{array}\right\}. \tag{9}$$

The augmented Lagrangian version of (9) is

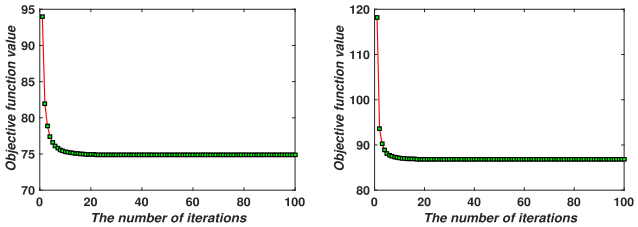
$$\begin{aligned}&\mathscr{L}_C(\Theta, \mathbf{J}, \mathbf{G}, \Lambda_1, \Lambda_2)\\ &\quad= \frac{1}{2}\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{J}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\Theta\widetilde{\mathbf{X}}\mathbf{L}(\Theta\widetilde{\mathbf{X}})^{\mathrm{T}}) + \Lambda_1^{\mathrm{T}}(\mathbf{J} - \Theta\widetilde{\mathbf{X}})\\ &\qquad+ \Lambda_2^{\mathrm{T}}(\mathbf{G} - \Theta) + \frac{\mu}{2}\|\mathbf{J} - \Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \frac{\mu}{2}\|\mathbf{G} - \Theta\|_{\mathrm{F}}^2\\ &\qquad\text{s.t.} \quad \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I}\end{aligned} \tag{10}$$

where Λ_1 and Λ_2 are the Lagrange multipliers and μ is the penalty parameter. Algorithm 2 summarizes the specific procedures for solving the problem (9), and the solution to each subproblem is detailed in Appendix A.

Finally, we repeat these optimization procedures until a stopping criterion is satisfied.

C. Convergence Analysis

The iterative alternating strategy used in Algorithms 1 and 2 is a block coordinate descent, whose convergence is theoretically guaranteed as long as each subproblem of (2) is strictly convex, which can be exactly minimized [23]. Moreover, we experimentally display an illustration to clarify the convergence of CoSpace on both HS-MS data sets, where the objective function value is recorded in each iteration (see Fig. 4).



(a) (b)

Fig. 4. Convergence analysis of CoSpace is experimentally performed on the two HS-MS data sets. (a) University of houston HS-MS data sets. (b) Chikusei HS-MS data sets.

III. Experiments

In this section, we quantitatively and qualitatively evaluate the performance of the proposed method on two HS-MS data sets taken over the University of Houston and Chikusei. To validate the transferability of learned features by our CoSpace method, classification is explored as a potential application. Therefore, three different classifiers, namely, the nearest neighbor (NN) based on the Euclidean distance, linear support vector machines (LSVMs), and canonical correlation forest (CCF) [24], are selected for this task. As a variant of random forest [25], CCF has shown its effectiveness in various tasks [26]–[28] because of supervised feature extraction via canonical correlation analysis when constructing each decision tree. Furthermore, we compare the proposed method (CoSpace)

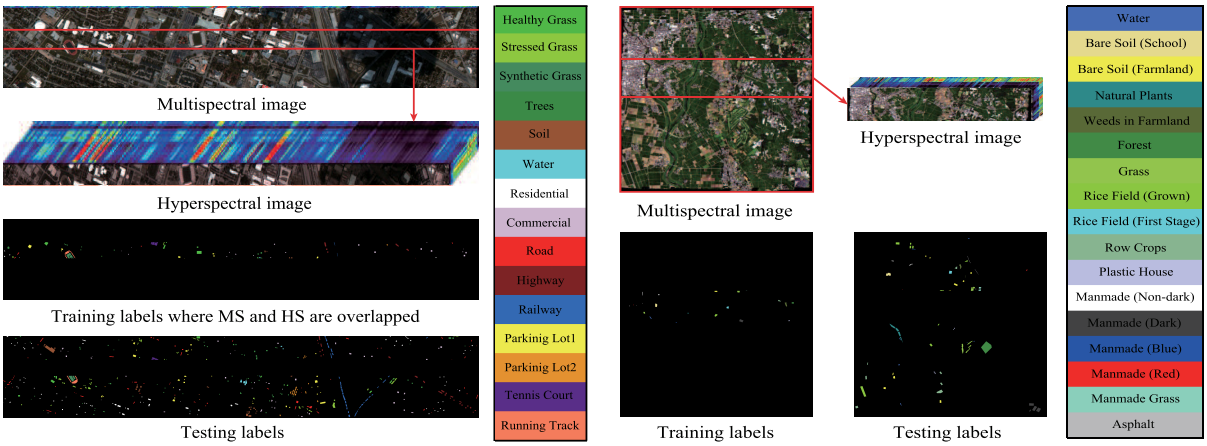


Fig. 5. MS image and its corresponding hyperspectral image that partially covers the same area, as well as training and testing labels, for (a) University of Houston HS-MS data set and (b) Chikusei HS-MS data set, respectively.

TABLE I

NUMBER OF TRAINING AND TESTING SAMPLES FOR THE UNIVERSITY OF HOUSTON MS-HS DATA SET

Class No.	Class Name	Training	Testing
1	Healthy Grass	537	699
2	Stressed Grass	61	1154
3	Synthetic Grass	340	357
4	Tree	209	1035
5	Soil	74	1168
6	Water	22	303
7	Residential	52	1203
8	Commercial	320	924
9	Road	76	1149
10	Highway	279	948
11	Railway	33	1185
12	Parking Lot1	329	904
13	Parking Lot2	20	449
14	Tennis Court	266	162
15	Running Track	279	381
	Total	2897	12021

TABLE II

NUMBER OF TRAINING AND TESTING SAMPLES FOR THE CHIKUSEI MS-HS DATA SET

Class No.	Class Name	Training	Testing
1	Water	301	858
2	Bare Soil (School)	992	1867
3	Bare Soil (Farmland)	455	4397
4	Natural Plants	150	4272
5	Weeds in Farmland	928	1108
6	Forest	486	11904
7	Grass	989	5526
8	Rice Field (Grown)	813	8816
9	Rice Field (First Stage)	667	1268
10	Row Crops	377	5961
11	Plastic House	165	475
12	Manmade (Non-dark)	170	568
13	Manmade (Dark)	1291	6373
14	Manmade (Blue)	111	431
15	Manmade (Red)	35	187
16	Manmade Grass	21	1019
17	Asphalt	384	417
	Total	8335	55447

with several classical approaches, which are suitable for the cross-modal feature learning task, including principle component analysis based on joint dimensionality reduction (P-JDR for short) [29], locality preserving projection (LPP) based on unsupervised MA (L-USMA for short) [30], and LPP-based supervised MA (L-SMA) [31] as well as the original MS (baseline). Tables I and II list the number of training and test samples on two used data sets.

A. University of Houston HS-MS Data Sets

1) Data Description: The HS data were acquired by the ITRES CASI-1500 sensor over an urban area around the campus of the University of Houston, Houston, TX, USA, which was provided in the 2013 IEEE GRSS data fusion contest [32]. The image consists of 349 × 1905 pixels with 144 spectral bands in the wavelength from 364 to 1046 nm with spectral resolution of 10 nm at a ground sampling distance of 2.5 m. Spectral simulation is performed to generate the MS image by degrading the full HS image in the spectral domain using the MS (SRFs of Sentinel-2 as filters (for more details refer to [5]). Following this, the MS data with dimensions of $349\times1905\times10$ are generated. The MS image and the corresponding partial HS image over the University of Houston scene are shown in Fig. 5(a).

2) Experimental Setup: Initially, we redistribute the training and testing samples, as shown in Fig. 5(a) and, more specifically, listed in Table I, to meet our problem setting that there is a large amount of the MS data (complete low-quality data) together with a limited amount of the HS data (incomplete high-quality data).

For the performance assessment of the algorithms, we adopt three criteria to quantify experiential results as follows.

1) *Overall Accuracy* (OA): This index is defined by the ratio between the number of MS samples that are correctly classified and the number of corresponding test samples.

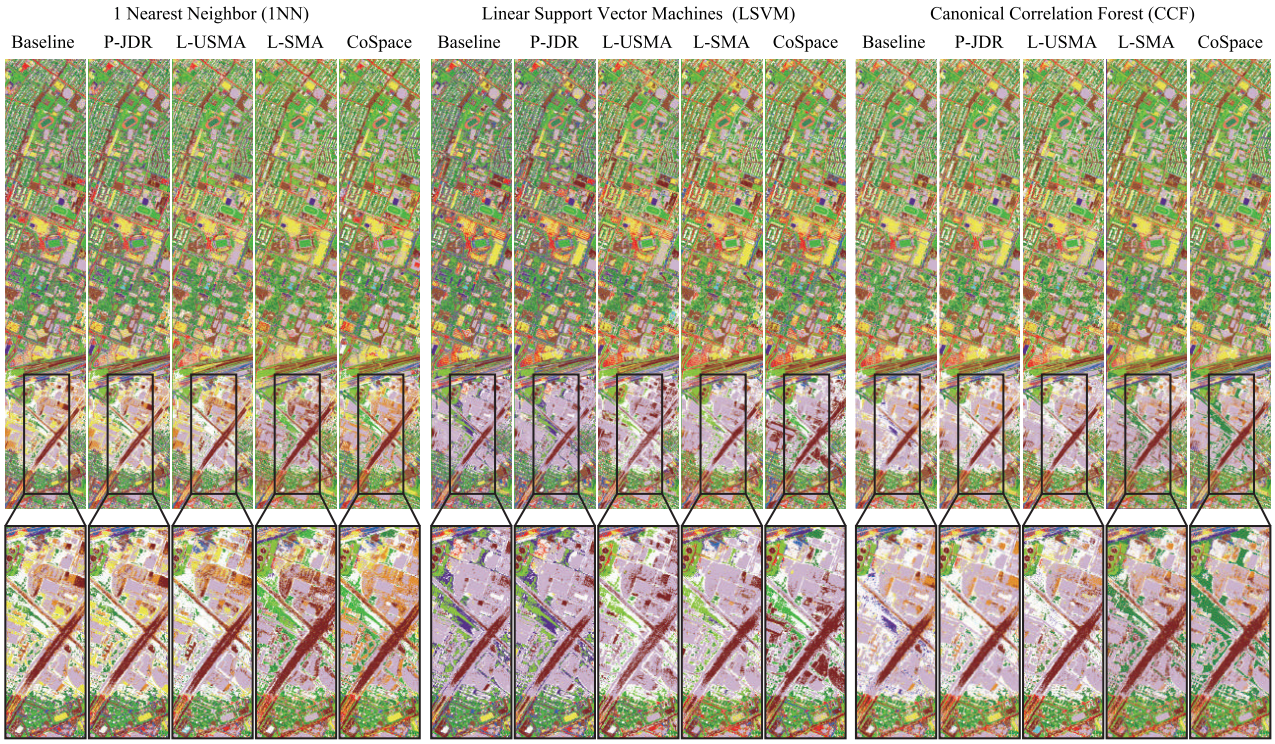


Fig. 6. Classification maps and corresponding highlighted subareas of the different algorithms obtained using three kinds of classifiers on the University of Houston data set.

2) *Average Accuracy* (AA): We collect the classification accuracy of each class and average them to achieve an AA-based evaluation.
3) *Kappa Coefficient* (κ): It statistically measures the agreement between the final classification map and the ground-truth map. Generally speaking, κ is more robust and convincing than a simple percent-based agreement calculation (e.g., OA and AA), since the agreement occurring by chance is fully considered.

Furthermore, we experimentally maximize the performance of the different algorithms by tuning their parameters, such as dimension (d), regularization parameters (α, β, γ), and so on, using ten-fold cross-validation on training data. For the dimension (d) which is a common parameter for all algorithms, they can be selected ranging from 10 to 50 at an interval of 10. For the number of NNs (k) and the standard deviation of Gaussian kernel function (σ) in L-USMA, we select them in the range of $\{10, 20, \ldots, 50\}$ and $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, respectively, and two regularization parameters (α, β) in CoSpace are both chosen from $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$.

3) Results and Analysis: Fig. 6 shows the classification maps of compared algorithms using three different classifiers, while Table III details the quantitative assessment results under the optimal parameters determined by cross-validation.

Overall, after absorbing partial HS information, those ASFL approaches are prone to obtain a better classification result, compared to the baseline (only MS data). P-JDR steadily outperforms the baseline, especially using 1NN and LSVM classifiers, although its classification accuracy using CCF is slightly lower than that of baseline. By embedding local topological structure of data, L-USMA performs better than baseline, and even P-JDR, showing stable results for three kinds of classifiers. With a more discriminative supervised information, L-SMA obtains more competitive results by locally constructing LDA-like graph, whose performance is basically superior to that of the baseline, P-JDR, and L-USMA. Unlike L-SMA that only aligns different modalities on a common subspace, the proposed CoSpace learns a latent subspace by aligning different modalities and also bridges the learned subspace with label information, achieving the best classification accuracy. Compared to baseline, P-JDR, L-USMA, and L-SMA, CoSpace increases the OAs of 7.12%, 2.45%, 2.49%, and 3.78%, respectively, with 1NN classifier, and 7.26%, 5.07%, 3.84%, and 1.37%, respectively, with LSVM classifier, as well as 3.96%, 5.04%, 3.75%, and 2.58%, respectively, with CCF classifier. Likewise, there are similar trends for the other indices of AA and κ, which indicate that CoSpace tends to learn semantically meaningful features.

We can also observe from Fig. 5(a) and Table I that the training samples collected in a very limited area badly results in the data unbalance between different classes. For instance, the number of training samples in *Health Grass* is dozens of times as much as that in *Water*, *Railway*, *Residential*, *Commercial*, and *Parking Lot2*. This might make the classifier impossible to be trained effectively, since more attentions are paid on those classes with large-size samples, and, contrariwise, the small-scale classes play relatively less and even nothing.

TABLE III

QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON THE UNIVERSITY OF HOUSTON DATA. THE BEST ONE IS SHOWN IN BOLD

Algorithm	Baseline (%)			P-JDR (%)			L-USMA (%)			L-SMA (%)			CoSpace (%)		
Parameter	/			$d = 30$			$(k, \sigma, d) = (10, 1, 20)$			$d = 30$			$(\alpha, \beta, d) = (0.01, 0.01, 30)$		
Classifier	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF
OA	61.70	62.12	68.21	66.37	64.31	67.13	66.33	65.54	68.41	65.04	68.01	69.59	68.82	69.38	**72.17**
AA	65.57	65.97	70.47	69.23	67.50	69.64	69.37	68.81	70.84	68.15	70.50	71.02	71.29	71.69	**73.56**
κ	0.5842	0.5889	0.6543	0.6345	0.6118	0.6430	0.6342	0.6251	0.6565	0.6202	0.6520	0.6695	0.6613	0.6672	**0.6975**
Class1	69.10	76.39	67.95	73.39	71.82	70.24	80.40	78.68	67.67	81.69	75.25	68.53	**88.56**	75.54	69.96
Class2	79.20	80.59	78.08	81.20	81.72	72.53	78.94	79.90	75.04	90.99	**97.57**	77.90	73.48	73.74	77.99
Class3	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Class4	85.22	85.51	92.27	87.63	91.21	93.24	94.30	96.23	92.85	88.60	94.78	**98.74**	93.62	**98.74**	98.26
Class5	98.89	99.06	99.40	98.72	98.72	99.32	**99.49**	99.40	98.89	99.40	98.97	99.14	98.97	99.40	99.40
Class6	86.14	86.14	86.14	86.47	86.14	76.57	86.14	**86.47**	**86.47**	**86.47**	**86.47**	70.96	85.15	85.48	85.15
Class7	36.49	50.62	63.76	59.6	53.78	51.04	49.79	50.21	63.18	58.19	72.32	77.14	69.99	73.98	**80.05**
Class8	50.76	56.49	56.06	59.42	59.42	59.09	54.76	**66.23**	56.82	55.74	62.01	62.23	58.77	63.53	62.01
Class9	60.92	56.22	70.58	64.14	63.01	**72.32**	63.45	65.19	69.10	47.17	49.96	61.27	64.40	59.79	64.93
Class10	40.93	45.36	45.25	44.41	49.16	45.36	44.09	53.90	47.47	49.47	58.12	52.32	45.15	**64.14**	57.70
Class11	39.16	27.43	43.88	36.03	35.53	39.41	45.91	29.79	43.71	35.36	28.86	36.46	44.14	36.54	**47.26**
Class12	41.37	31.64	56.08	51.55	25.66	**65.82**	45.8	29.76	62.06	65.04	35.84	62.5	50.22	46.79	62.72
Class13	0.45	0.00	0.67	0.67	0.00	2.45	0.22	0.00	1.11	**4.68**	0.00	0.00	0.89	0.00	0.45
Class14	97.53	97.53	98.77	98.15	98.77	99.38	99.38	98.77	**100.00**	97.53	**100.00**	**100.00**	98.15	**100.00**	99.38
Class15	97.38	96.59	**98.16**	97.11	97.64	97.9	97.9	97.64	**98.16**	97.9	97.38	**98.16**	97.9	97.64	**98.16**

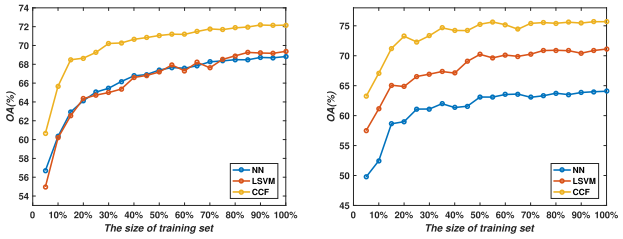


Fig. 7. Sensitivity analysis to the training set size using three different classifiers on the two MS-HS data sets. (a) Houston HS-MS data sets. (b) Chikusei HS-MS data sets.

A feasible solution to the problem is to enhance data representative ability by jointly feature learning from multimodalities. For the performance evaluation of classifying those small-scale classes, e.g., *Residential*, *Commercial*, and *Railway*, a direct evidence has been shown in Table III that those ASFL-based approaches (e.g., P-JDR, L-USMA, and L-SMA as well as CoSpace) obviously perform better on these small-scale samples than directly using original MS data (baseline). As expected, CoSpace dramatically outperforms the others, particularly on *Residential* and *Railway*. There is no denying, however, that CoSpace is superior to other algorithms to a larger extent, although it fails to effectively identify *Parking Lot2* as same with others.

To visually highlight the classification differences for the different methods, we enlarge the classification maps of a subarea overshadowed by the cloud, as shown in Fig. 6 where we can see that the methods with considering the hyperspectral information are able to generate the more discriminative features than the baseline, while the proposed CoSpace yields a better performance in identifying the materials in the shadow area, particularly for vegetation (e.g., *Grass*), *Residential*, and *Commercial* that are easily misclassified by the traditional methods.

4) Sensitivity Analysis to the Training Set Size: As the performance of the CoSpace largely depends on the number of training samples, it is, therefore, indispensable to investigate the sensitivity of the training set size. In detail, we conduct the classification using the CoSpace by fixing the test set and setting a series of new training sets randomly selected from the original training set with the different percentages ranging from 5% to 100% at a 5% interval. As can be seen in Fig. 7(a), there is a similar trend in OAs using different classifiers, that is, the classification accuracy improves with the training set size, faster in the early, and later basically stabilized.

B. Chikusei HS-MS Data Sets

1) Data Description: The Headwall's hyperspectral visible and near-infrared series C (VNIR-C) imaging sensor acquired the airborne HS data set over the agricultural and urban areas of Chikusei, Ibaraki, Japan, in 2014. This VNIR-C sensor collected 128 bands covering the wavelength range from 363 to 1018 nm with spectral resolution of 10 nm, and the scene consists of 2517×2335 pixels at ground sample distance of 2.5 m. The data set was made available to the scientific community recently, and more details regarding the data acquisition and processing can be found in [33]. Similarly, the MS image with the size of $2517 \times 2335 \times 10$ was simulated by spectrally down-sampling the full HS image using the known SRFs of Sentinel-2. The generated MS image and the partial HS image over the Chikusei scene are shown in Fig. 5(b).

2) Experimental Setup: Fig. 5(b) shows the latest training and testing labeling of the Chikusei data set, which is quantified in Table II. Three indices: OA, AA, and κ introduced earlier are calculated to quantitatively assess the classification performance. Similar to the case of the University of Houston data set, the parameters for those given algorithms are determined by the tenfold cross-validation on the training samples and the same range setting with those used for the University of Houston data set is also conducted to the Chikusei data set.

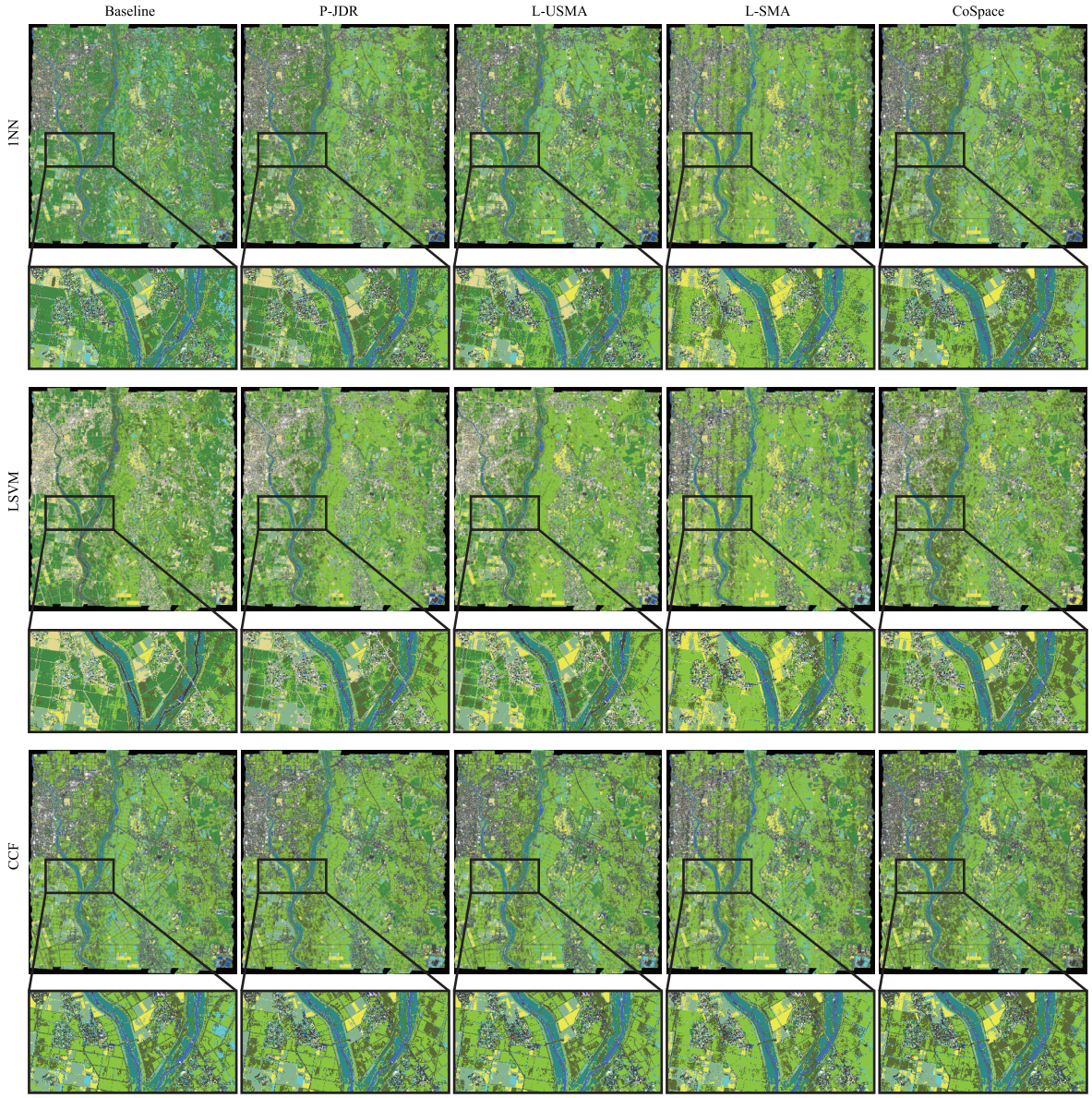


Fig. 8. Classification maps and corresponding highlighted subareas of the different algorithms obtained using three kinds of classifiers on the Chikusei data set.

3) Results and Analysis: Similar to the University of Houston scene, we evaluate the performance for the Chikusei data both quantitatively and visually. Three classification indices with optimal parameters for different algorithms are summarized in Table IV. For visual comparison, we give the corresponding classification maps in the full scene with those comparative algorithms under the different classifiers, as shown in Fig. 8.

As the classes in the Chikusei scene are more challenging classes and the distribution of training samples is inhomogeneous, directly using original MS data as input fails to identify certain materials, such as *Forest*, *Man-made (Dark)*, and *Man-made (Grass)*, yielding a poor performance in OA, AA, and κ. Especially while using 1NN classifier, P-JDR and L-USMA, which belong to the unsupervised feature learning method, observably exceed baseline in classification accuracy by 4.46% and 5.49%, respectively. For LSVM and CCF classifiers, a similar trend is also demonstrated in Table IV. Because of the limited training samples and their distribution unbalance, the subspace projection learned by L-SMA easily traps into over-fitting, despite only having a weak performance improvement compared to these previously compared algorithms. By jointly performing subspace learning and classification, CoSpace not only aligns the different modalities in a latent common subspace but also connects the subspace with label information formulated by training data. As a result, CoSpace obtains a higher classification accuracy than other algorithms, as listed in Table IV. This might attribute to

TABLE IV

QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON THE CHIKUSEI DATA. THE BEST ONE IS SHOWN IN BOLD

Algorithm	Baseline (%)			P-JDR (%)			L-USMA (%)			L-SMA (%)			CoSpace (%)		
Parameter	l			$d = 30$			$(k, \sigma, d) = (10, 1, 30)$			$d = 20$			$(\alpha, \beta, d) = (0.1, 0.01, 30)$		
Classifier	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF	1NN	LSVM	CCF
OA	55.99	60.20	71.11	60.45	68.19	71.87	61.48	67.31	72.33	62.44	67.90	71.53	64.07	71.12	**75.69**
AA	62.39	69.42	70.40	64.95	71.71	71.02	64.95	70.06	71.49	64.52	70.79	66.47	68.05	7**3.96**	71.46
κ	0.5084	0.5523	0.6761	0.5614	0.6414	0.6834	0.5715	0.6318	0.6893	0.5812	0.6391	0.6802	0.5995	0.6746	**0.7260**
Class1	79.49	78.21	80.54	97.2	98.25	81.93	98.48	**99.18**	81.00	80.89	98.72	82.52	80.19	92.54	79.25
Class2	**95.02**	94.43	82.7	93.2	93.95	94.86	93.25	93.68	93.95	92.98	93.20	92.50	**95.02**	93.47	94.91
Class3	11.37	23.54	50.06	14.40	16.03	24.13	31.07	39.37	56.74	61.81	62.57	55.31	80.10	**80.40**	77.71
Class4	91.64	92.13	92.56	88.65	90.87	93.38	89.49	92.21	94.90	86.54	90.57	91.53	93.47	90.59	**96.23**
Class5	87.00	**97.65**	94.68	67.78	70.94	76.26	63.90	34.57	79.87	25.36	28.43	16.06	74.28	83.94	66.52
Class6	65.36	62.01	81.48	55.18	75.11	**87.77**	50.90	67.15	79.42	48.29	62.52	78.91	53.49	63.61	79.02
Class7	99.26	99.67	99.93	91.95	98.43	**99.98**	91.02	97.00	99.71	95.46	96.87	97.79	88.74	97.74	99.75
Class8	43.42	57.11	93.40	85.32	95.58	97.95	91.64	96.32	**99.29**	93.93	95.59	93.49	65.61	95.05	92.72
Class9	99.92	**100.00**	**100.00**	98.34	98.66	99.53	98.82	99.45	99.92	99.21	99.53	99.13	**100.00**	98.66	99.76
Class10	16.96	**24.81**	19.56	21.36	23.75	22.24	19.78	20.13	17.16	21.15	21.39	15.48	22.68	22.35	18.00
Class11	2.11	0.00	2.11	0.63	1.05	**6.53**	0.00	0.00	3.37	0.00	0.00	0.00	0.00	0.00	0.00
Class12	87.85	90.32	88.91	86.44	88.20	89.96	81.51	88.38	87.32	84.43	90.14	85.92	86.62	**90.32**	80.46
Class13	33.08	33.11	33.09	33.08	33.11	33.11	32.36	33.01	33.11	25.47	32.61	56.25	33.11	33.11	**67.90**
Class14	67.75	94.20	85.38	59.40	58.24	**99.77**	58.70	59.40	89.56	71.93	72.85	59.40	59.40	59.40	52.44
Class15	97.33	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	96.26	96.79	**100.00**	94.12	93.58	**100.00**	**100.00**	93.58	97.86
Class16	66.24	74.88	88.62	95.39	95.58	99.71	99.80	98.43	**100.00**	99.61	99.71	99.51	93.13	97.84	**100.00**
Class17	16.79	58.03	3.84	15.83	**81.29**	0.24	7.19	76.26	0.00	15.59	65.23	7.91	30.94	64.75	0.00

the learned common subspace, since the features projected in the subspace can absorb various properties from different modalities.

Similarly, we also make a visual comparison by giving a salient region in which the CoSpace's superiority in classifying complex and similar land-cover classes is further shown as detailed in Fig. 8. Compared to other alignment-based methods, CoSpace is capable of better transferring HS information into MS data by means of joint subspace learning and classification, yielding a more discriminative low-dimensional embedding. The learned features can recognize those classes of holding very similar features in MS data, such as *Bare Soil (Farmland)* and *Row Crops*, *Weeds in Farmland*, and *Rice Field (Grown)*, more effectively. As shown in Fig. 8, CoSpace performs more reasonable and competitive classification results, that is, on the one hand, the *Weeds in Farmland* and *Rice Field (Grown)* are most likely to be coexisted in a scene; on the other hand, the *Bare Soil (Farmland)* and *Row Crops* are separated more correctly. This can be explained by a powerful transferability of HS information in the proposed CoSpace.

4) Sensitivity Analysis to the Training Set Size: Similar to the MS-HS Houston data sets, we apply the same investigating strategy and observe the trend of classification performance using CoSpace with different sizes of training sets on the MS-HS Chikusei data sets in Fig. 7(b). There is a very substantial change in classification accuracy with the increase of the training set size ranging from 5% to 40% of total training samples, while the performance tends to be stable after the training set size is over 50%.

IV. CONCLUSION

The tradeoff between MS and HS imaging in terms of observation ranges and spectral resolution motivates us to ponder whether HS data partially overlapping MS data can contribute to improving the classification performance of the whole MS imagery. For this purpose, we proposed CoSpace to achieve the property transferring in the different domains by learning a latent common subspace. Moreover, an effective joint strategy that simultaneously considers subspace learning and classification is embedded into the proposed method to tightly bridge the gap between the learned subspace and label information, leading to a more discriminative feature representation. The superior classification performance using CoSpace is demonstrated on two different data sets, compared to using other state-of-art methods.

We performed transfer learning on homogeneous data sets in the considered MS-HS case in the sense that both data sources are optical images covering similar spectral ranges and thus the HS information can be transferred into the MS one linearly. The CoSpace's ability in handling heterogeneous data sources remains limited due to its linearized modeling. In the future work, we will develop a more general system by integrating some powerful and emerging nonlinear tools (e.g., deep learning) into our framework.

In addition, we just assumed to share the same land-cover classes across MS and HS images in this paper. In reality, the number of land-cover classes in the large-scale MS scene might be usually more than the one in the overlapped area of MS and HS images. This naturally motivates us to generalize our model in the future work.

APPENDIX
SOLUTION TO PROBLEM (8) WITH RESPECT TO Θ

The solution to problem (8) can be transferred to equivalently solve the problem (10) with ADMM. Considering the fact that the object function in (10) is not convex with respect to all variables simultaneously, but it is a convex problem regarding the separate variable when other variables are fixed, therefore, we successively minimize $\mathscr{L}_C$ with respect to $\Theta, \mathbf{J}, \mathbf{G}, \Lambda_1, \Lambda_2$ as follows.

Optimization with respect to Θ*:* The optimization problem for Θ can be written as

$$\min_{\Theta}\left\{\begin{array}{c}\dfrac{\beta}{2}\mathrm{tr}(\Theta\widetilde{\mathbf{X}}\mathbf{L}(\Theta\widetilde{\mathbf{X}})^{\mathrm{T}})+\Lambda_1^{\mathrm{T}}(\mathbf{J}-\Theta\widetilde{\mathbf{X}})\\ +\Lambda_2^{\mathrm{T}}(\mathbf{G}-\Theta)+\dfrac{\mu}{2}\|\mathbf{J}-\Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2+\dfrac{\mu}{2}\|\mathbf{G}-\Theta\|_{\mathrm{F}}^2\}\end{array}\right\} \tag{11}$$

which has a closed-form solution

$$\Theta\leftarrow(\mu\mathbf{J}\widetilde{\mathbf{X}}^{\mathrm{T}}+\Lambda_1\widetilde{\mathbf{X}}^{\mathrm{T}}+\mu\mathbf{G}+\Lambda_2)\times(\mu\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}+\mu\mathbf{I}+\beta\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^{\mathrm{T}})^{-1}. \tag{12}$$

Optimization with respect to $\mathbf{J}$*:* The variable $\mathbf{J}$ can be estimated by solving the following problem:

$$\min_{\mathbf{J}}\left\{\frac{1}{2}\|\widetilde{\mathbf{Y}}-\mathbf{P}\mathbf{J}\|_{\mathrm{F}}^2+\Lambda_1^{\mathrm{T}}(\mathbf{J}-\Theta\widetilde{\mathbf{X}})+\frac{\mu}{2}\|\mathbf{J}-\Theta\widetilde{\mathbf{X}}\|_{\mathrm{F}}^2\right\} \tag{13}$$

its analytical solution is given by

$$\mathbf{J}\leftarrow(\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mu\mathbf{I})^{-1}(\mathbf{P}^{\mathrm{T}}\widetilde{\mathbf{Y}}+\mu\Theta\widetilde{\mathbf{X}}-\Lambda_1). \tag{14}$$

Optimization with respect to $\mathbf{G}$*:* For $\mathbf{G}$, the optimization problem with orthogonal constraint can be formulated as

$$\min_{\mathbf{G}}\left\{\Lambda_2^{\mathrm{T}}(\mathbf{G}-\Theta)+\frac{\mu}{2}\|\mathbf{G}-\Theta\|_{\mathrm{F}}^2\},\ \text{s.t. }\mathbf{G}\mathbf{G}^{\mathrm{T}}=\mathbf{I}\right\} \tag{15}$$

which can be effectively solved using the strategy of splitting orthogonality constraints [34] in two steps.

The first step is to perform the *singular value decomposition (SVD) factorization*

$$[\mathbf{U},\mathbf{S},\mathbf{V}]=\mathrm{svd}(\Theta-\Lambda_2/\mu). \tag{16}$$

The second step is to update $\mathbf{G}$ with satisfying orthogonal constraint

$$\mathbf{G}\leftarrow\mathbf{U}\mathbf{I}_{n\times m}\mathbf{V}. \tag{17}$$

Lagrange multipliers (Λ_1, Λ_2) *and penalty parameter* (μ) *update:* Before stepping into the next iteration, Lagrange multipliers need to be updated by

$$\Lambda_1\leftarrow\Lambda_1+\mu(\mathbf{J}-\Theta\widetilde{\mathbf{X}}),\quad \Lambda_2\leftarrow\Lambda_2+\mu(\mathbf{G}-\Theta) \tag{18}$$

and penalty parameter be updated by

$$\mu\leftarrow\min(\rho\mu,\mu_{\max}). \tag{19}$$

ACKNOWLEDGMENT

The authors would like to thank the Hyperspectral Image Analysis Group and the NSF Funded Center for Airborne Laser Mapping at the University of Houston for providing the CASI University of Houston data set. They would also like to thank Prof. D. Cai and Dr. C. Wang for providing MATLAB codes for LPP and MA algorithms.

REFERENCES

[1] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of difference images for change detection over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1076–1086, Aug. 2012.

[2] C. Yang, J. H. Everitt, Q. Du, B. Luo, and J. Chanussot, "Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture," *Proc. IEEE*, vol. 101, no. 3, pp. 582–592, Mar. 2013.

[3] H. Xie *et al.*, "Dynamic monitoring of agricultural fires in China from 2010 to 2014 using MODIS and GlobeLand30 data," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 10, p. 172, 2016.

[4] F. D. Van der Meer, H. M. A. Van der Werff, and F. J. A. Van Ruitenbeek, "Potential of ESA'S Sentinel-2 for geological applications," *Remote Sens. Environ.*, vol. 148, pp. 124–133, Apr. 2014.

[5] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, Jun. 2017.

[6] Y. Chen, C. Li, P. Ghamisi, X. Jia, and Y. Gu, "Deep fusion of remote sensing data for accurate classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1253–1257, Aug. 2017.

[7] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7708–7720, Dec. 2014.

[8] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[9] P. Ghamisi, B. Höfle, and X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Dec. 2017.

[10] G. Iyer, J. Chanussot, and A. L. Bertozzi, "A graph-based approach for feature extraction and segmentation of multimodal images," in *Proc. ICIP*, Sep. 2017, pp. 3320–3324.

[11] D. Hong, N. Yokoya, and X. Zhu, "Local manifold learning with robust neighbors selection for hyperspectral dimensionality reduction," in *Proc. IGARSS*, 2016, pp. 40–43.

[12] D. Hong, N. Yokoya, and X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, Jul. 2017.

[13] S. Liu, Q. Du, X. Tong, A. Samat, L. Bruzzone, and F. Bovolo, "Multi-scale morphological compressed change vector analysis for unsupervised multiple change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4124–4137, 2017.

[14] G. Tochon, J. Chanussot, M. D. Mura, and A. L. Bertozzi, "Object tracking by hierarchical decomposition of hyperspectral video sequences: Application to chemical gas plume tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4567–4585, May 2017.

[15] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "Learning a low-coherence dictionary to address spectral variability for hyperspectral unmixing," in *Proc. ICIP*, Sep. 2017, pp. 1–5.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. IMLS*, Jul. 2011, pp. 689–696.

[17] G. Matasci, M. Volpi, D. Tuia, and M. Kanevski, "Transfer component analysis for domain adaptation in image classification," *Proc. SPIE*, vol. 8180, p. 81800F, Oct. 2011.

[18] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. CVPR*, 2011, pp. 1785–1792.

[19] X. Zhou and S. Prasad, "Domain adaptation for robust classification of disparate hyperspectral images," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 822–836, Dec. 2017.

[20] C. Wang, P. Krafft, and S. Mahadevan, *Chapter of Manifold Learning: Theory and Applications-Manifold Alignment*. New York, NY, USA: CSC Press, 2011.

[21] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proc. IJCAI*, Jul. 2009, pp. 1077–1082.

[22] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2011, pp. 1294–1299.

[23] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.

[24] T. Rainforth and F. Wood. (2015). "Canonical correlation forests." [Online]. Available: https://arxiv.org/abs/1507.05444

[25] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[26] J. Xia, N. Yokoya, and A. Iwasaki, "Hyperspectral image classification with canonical correlation forests," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 421–431, Jan. 2017.
[27] N. Yokoya *et al.*, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.
[28] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & Progressive learning from high-dimensional data for multi-label classification," in *Proc. ECCV*, 2018, pp. 469–484.
[29] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[30] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS*, Dec. 2004, pp. 153–160.
[31] C. Wang and S. Mahadevan, "A general framework for manifold alignment.," in *Proc. AAAI*. Nov. 2009, pp. 53–58.
[32] C. Debes *et al.*, "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.
[33] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, 2016.
[34] R. Lai and S. Osher, "A splitting method for orthogonality constrained problems," *J. Sci. Comput.*, vol. 58, no. 2, pp. 431–449, 2014.

Danfeng Hong (S'16) received the B.Sc. degree in computer science and technology from the Neusoft College of Information, Northeastern University, Shenyang, China, in 2012, and the M.Sc. degree in computer vision from Qingdao University, Qingdao, China, in 2015. He is currently pursuing the Ph.D. degree in signal processing in earth observation from the Technical University of Munich, Munich, Germany, and the Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany.

His research interests include signal/image processing and analysis, pattern recognition, machine/deep learning and their applications in Earth Vision.

Naoto Yokoya (S'10–M'13) received the M.Sc. and Ph.D. degrees in aerospace engineering from The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

From 2012 to 2013, he was a Research Fellow with the Japan Society for the Promotion of Science, Tokyo. From 2013 to 2017, he was an Assistant Professor with The University of Tokyo. From 2015 to 2017, he was also an Alexander von Humboldt Research Fellow with the German Aerospace Center (DLR), Weßling, Germany, and Technical University of Munich, Munich, Germany. Since 2018, he has been leading the Geoinformatics Unit with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo. His research interests include image analysis and data fusion in remote sensing.

Dr. Yokoya is a Co-Chair of IEEE Geoscience and Remote Sensing Image Analysis and Data Fusion Technical Committee since 2017. In 2017, he was a recipient of the Data Fusion Contest 2017 organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. His model was the most accurate among over 800 submissions.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Universit de Savoie, Annecy, France, in 1998. In 1999, he was with the Geography Imagery Perception Laboratory for the Delegation Generale de l'Armement (French National Defense Department). Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; KTH, Stockholm, Sweden; and NUS, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles, Los Angeles, CA, USA. He is conducting his research at GIPSA-Lab. His research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing.

Dr. Chanussot was a member of the IEEE Geoscience and Remote Sensing Society AdCom from 2009 to 2010, in charge of membership development and Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008. He is a member of the Institut Universitaire de France from 2012 to 2017. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing. He was the Chair from 2009 to 2011 and the Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is the Founding President of IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010 which received the 2010 IEEE GRSS Chapter Excellence Award. He was a co-recipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 and 2015 Symposium Best Paper Award, the IEEE GRSS 2012 Transactions Prize Paper Award, and the IEEE GRSS 2013 Highest Impact Paper Award. He was an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2005 to 2007 and *Pattern Recognition* from 2006 to 2008. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. Since 2007, he has been an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and since 2018, he has also been an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Guest Editor for the PROCEEDINGS OF THE IEEE in 2013 and *IEEE Signal Processing Magazine* in 2014.

Xiao Xiang Zhu (S'10–M'12–SM'14) received the M.Sc., Dr.-Ing., and the Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Professor of Signal Processing in Earth Observation, TUM, and German Aerospace Center (DLR), Weßling, Germany; the Head of the Department EO Data Science with the DLR's Earth Observation Center; and the Head of the Helmholtz Young Investigator Group SiPEO with DLR and TUM. She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests include remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of theIEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Appendices

F Hong D., Yokoya N., Ge N., Chanussot J., Zhu X. X., 2019. Learnable Manifold Alignment (LeMA): A Semi-supervised Cross-modality Learning Framework for Land Cover and Land Use Classification. ISPRS Journal of Photogrammetry and Remote Sensing, 147: 193-205

https://www.sciencedirect.com/science/article/pii/S0924271618302843

ISPRS Journal of Photogrammetry and Remote Sensing 147 (2019) 193–205

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification

Danfeng Hong[a,b], Naoto Yokoya[c], Nan Ge[a], Jocelyn Chanussot[d], Xiao Xiang Zhu[a,b,*]

[a] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[b] Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany
[c] Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan
[d] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France

ARTICLE INFO

Keywords:
Cross-modality
Graph learning
Hyperspectral
Manifold alignment
Multispectral
Remote sensing
Semi-supervised learning

ABSTRACT

In this paper, we aim at tackling a general but interesting cross-modality feature learning question in remote sensing community—*can a limited amount of highly-discriminative (e.g., hyperspectral) training data improve the performance of a classification task using a large amount of poorly-discriminative (e.g., multispectral) data?* Traditional semi-supervised manifold alignment methods do not perform sufficiently well for such problems, since the hyperspectral data is very expensive to be largely collected in a trade-off between time and efficiency, compared to the multispectral data. To this end, we propose a novel semi-supervised cross-modality learning framework, called learnable manifold alignment (LeMA). LeMA learns a joint graph structure directly from the data instead of using a given fixed graph defined by a Gaussian kernel function. With the learned graph, we can further capture the data distribution by graph-based label propagation, which enables finding a more accurate decision boundary. Additionally, an optimization strategy based on the alternating direction method of multipliers (ADMM) is designed to solve the proposed model. Extensive experiments on two hyperspectral-multispectral datasets demonstrate the superiority and effectiveness of the proposed method in comparison with several state-of-the-art methods.

1. Introduction

Multispectral (MS) imagery has been receiving an increasing interest in the urban area (e.g. a large-scale land-cover mapping (Huang et al., 2014; Hong et al., 2016), building localization (Kang et al., 2018)), agriculture (Yang et al., 2013), and mineral products (Van der Meer et al., 2014), as operational optical broadband (multispectral) satellites (e.g. Sentinel-2 and Landsat-8 (Yokoya et al., 2017)) enable the multispectral imagery openly available on a global scale. In general, a reliable classifier needs to be trained on a large amount of labeled, discriminative, and high-quality samples. Unfortunately, labeling data, in particular large-scale data, is very gruelling and time-consuming. A natural alternative way to this issue is to consider tons of unlabeled data, yielding a semi-supervised learning. On the other hand, MS data fails to spectrally discriminate similar classes due to its broad spectral bandwidth. A simple way is to improve the data quality by fusing high-discriminative hyperspectral (HS) data (Yokoya et al., 2017). Although such data is expensive to collect, we may be able to expect a small amount of such data available. The aforementioned two points motivate us to raise a question related to transfer learning and cross-modality learning: *Can a limited amount of HS training data partially overlapping MS data improve the performance of a classification task using a large coverage of MS testing data?*

Over the past decades, land-cover and land-use classification tasks of optical remote sensing imagery has received increasing attention in the unsupervised (Hong et al., 2017; Li et al., 2014; Tarabalka et al., 2009), supervised (Zhang et al., 2012; Hong et al., 2018), and semi-supervised ways (Xia et al., 2014; Tuia et al., 2014). To our best knowledge, the classifying ability in unsupervised learning (or dimensionality reduction) still remains limited, due to missing label information. By fully considering the variability of intra-class and inter-class from labels, supervised learning is able to perform the classification task better. In reality, a limited number of labeled samples usually hinders the trained classier towards a high classification performance, further leading to a possible failure in some challenging classification or transferring tasks owing to the lack of generalization and

* Corresponding author at: Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany and Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany.
E-mail address: xiaoxiang.zhu@dlr.de (X.X. Zhu).

https://doi.org/10.1016/j.isprsjprs.2018.10.006
Received 11 May 2018; Received in revised form 8 August 2018; Accepted 16 October 2018
Available online 30 November 2018

representability. Alternatively, semi-supervised learning draws into plenty of unlabeled data in learning process. This is capable of better capturing the distribution of different categories in order to find an accurate decision boundary.

On the other hand, considerable work related to transfer learning (TL) or domain adaptation (DA) has been successfully developed and applied in the remote sensing community (Bruzzone and Marconcini, 2010; Banerjee et al., 2015; Matasci et al., 2015; Tuia et al., 2016; Samat et al., 2016, 2017). According to the different transferred objects, the TL or DA approaches can be roughly categorized into three groups, including parameter adaptation, instance-based transfer, and feature-based alignment or representation.

The seminal work dealing with parameter adaptation was presented in Khosla et al. (2012) and Woodcock et al. (2001), aiming at transferring an existing classifier (or parameters) trained or learned from the source domain to the target domain. Differently, the instance-based transferring technique transfers the knowledge by reweighting (Jiang and Zhai, 2007) or resampling (Sugiyama et al., 2008) the samples of the source domain to those of the target domain. A similar idea based on active learning (Samat et al., 2016) has also been proposed to address this issue, by selecting the most informative samples in the target domain to replace with those samples of the source domain that do not match the data distribution of the target domain (Persello and Bruzzone, 2012).

For the final group of feature-based alignment or representation, manifold alignment (MA) is one of the most popular semi-supervised learning framework (Wang et al., 2011) that facilitates transfer learning. MA has been successfully applied to various tasks in remote sensing community, e.g. classification (Tuia et al., 2016), data visualization (Liao et al., 2016), multi-modality data analysis (Tuia et al., 2014), etc. The key idea of MA can be generalized as learning a common (or shared) subspace where different data can be aligned to learn a joint feature representation. Generally, existing MA methods can be approximately categorized into unsupervised, supervised, and semi-supervised approaches. The unsupervised approach usually fails to align multimodal data sufficiently well, as their corresponding low-dimensional embeddings may be quite diverse (Wang and Mahadevan, 2009). In the supervised case, only aligning the limited number of training samples to learn a common subspace leads to weak transferability. While preserving a joint manifold structure created by both labeled and unlabeled data, semi-supervised alignment allows different data sources to be better transformed into the common subspace (Wang and Mahadevan, 2011).

Although the joint manifold structure used in conventional semi-supervised MA approaches can relate features or instances, poor connections between the common subspace and label information still hinder the low-dimensional feature representation from being more discriminative. More importantly, in most graph-based semi-supervised learning algorithms (e.g. graph-based label propagation (GLP) (Zhu et al., 2003), semi-supervised manifold alignment (S-SMA (Tuia et al., 2014)) (Wang and Mahadevan, 2011)), the topology of unlabeled samples is merely given by a fixed Gaussian kernel function, which is computed in the original space rather than in the common space. This makes it difficult to adaptively transfer unlabeled samples into the learned common subspace, particularly when applied to multimodal data due to different numbers of dimensions. To address these issues, we propose a learnable manifold alignment (LeMA) by a data-driven graph learning directly from a common subspace so as to make the multimodal data comparable as well as improve the explainability of the learned common subspace, which further results in a better transferability. More specifically, our contributions can be summarized as follows:

- We propose a novel semi-supervised cross-modality learning framework called learnable manifold alignment (LeMA) for a large-scale land-cover classification task. One spectrally-poor MS and one spectrally rich HS data are considered as two different modalities and applied for this task, where the spatial extent of the former is a true superset of that of the latter.
- Unlike jointly feature learning in which the model is both trained and tested from completed HS-MS correspondences, LeMA learns an aligned feature subspace from the labeled HS-MS correspondences and partially unlabeled MS data, and allows to identify out-of-samples using either MS data or HS data; Such the learnt subspace is a good fit for our case of cross-modality learning.[1]
- Instead of directly computing graph structure with a Gaussian kernel function, a data-driven graph learning method is exploited behind LeMA in order to strengthen the abilities of transferring and generalization;
- An optimization framework based on the alternating direction method of multipliers (ADMM) is designed to fast and effectively solve the proposed model.

The remainder of this paper is organized as follows. Section 2 elaborates on our motivation and proposes the methodology for the LeMA and the corresponding optimization algorithm. In Section 3, we present the experimental results on two HS-MS datasets over the areas of the University of Houston and Chikusei, respectively, and meanwhile discuss the qualitative and quantitative analysis. Section 4 concludes with a summary.

2. Learnable Manifold Alignment (LeMA)

In this section, a cross-modality learning problem is firstly casted and the motivation is stated in the following. Accordingly, we formulate the methodology of our proposed and then elucidate an ADMM-based optimization algorithm to solve it.

2.1. Problem statement and motivation

For many high-level data analysis tasks in remote sensing community, such as land-cover classification, data collection plays an important role, since information-rich training samples enable us to easily find an optimal decision boundary.

There is, however, a typical bottleneck in collecting a large amount of labeled and discriminative data. Despite the MS data available at a global scale from the satellites of Sentinel-2 and Landsat-8, the identification and discrimination of materials are unattainable at an accuracy level by MS data, resulting from its poorly spectral information. On the contrary, HS data is characterized by rich spectral information, but only can be acquired in very small areas, due to the limitations of imaging sensors. This issue naturally guides us to jointly utilize the HS and MS bi-modal data, specifically leading to the following interesting and challenging question *can a limited number of HS training data contribute to the classification task of a large-scale MS data?*

A feasible solution to the issue can be unfolded to two parts: (1) *cross-modality learning*: learning a common subspace where the features are expected to absorb the different properties from the HS-MS modalities and meanwhile the HS and MS data can be transferred each other; (2) *semi-supervised learning*: Embedding massive unlabeled MS samples which are relatively in large quantities and easy to be collected, so as to learn a more discriminative feature representation. Fig. 1 illustrates the workflow of LeMA.

2.2. Problem formulation

To effectively model the aforementioned issue, we intend to develop

[1] In contrast to multi-modal learning (bi-modality for example), cross-modal learning trains on single modality and tests on bi-modality, or *vice versa* (train on bi-modality and test on single modality).

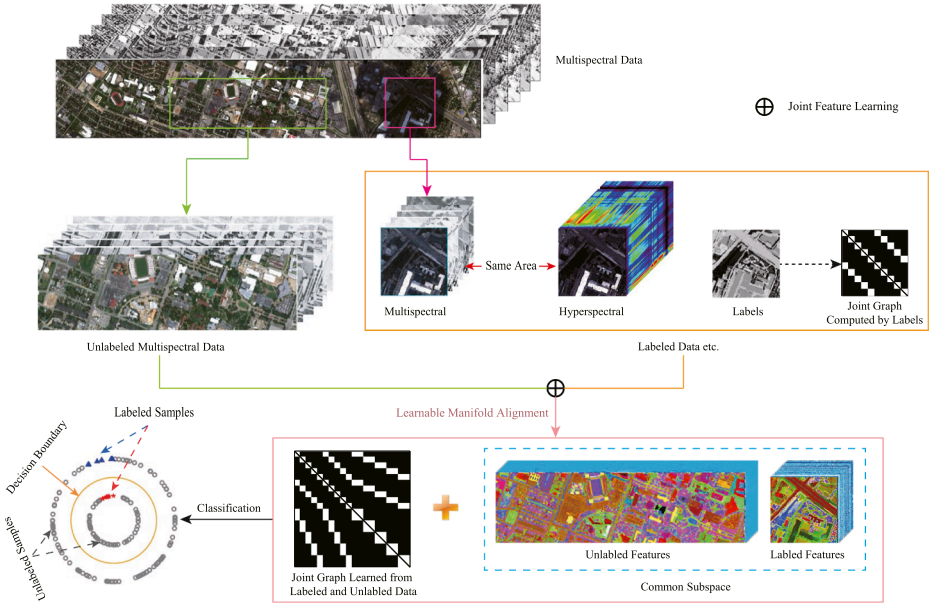


Fig. 1. An illustration of the proposed LeMA method.

a joint learning framework which better learns a discriminative common subspace from high-quality HS data and low-quality MS data. Intuitively, such a common subspace can be shaped by selectively absorbing the benefits of both high-quality data with more details and low-quality data with more structural information. Therefore, following a popular joint learning framework (Ji and Ye, 2009), we formulate the common subspace learning problem as

$$\min_{\mathbf{P},\mathbf{\Theta}} \frac{1}{2}\left\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\mathbf{P}\right\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}(\mathbf{E}\mathbf{L}\mathbf{E}^{\mathrm{T}}) \text{ s.t. } \mathbf{E} = \mathbf{\Theta}\widetilde{\mathbf{X}}, \quad \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}, \tag{1}$$

where $\widetilde{\mathbf{Y}} = [\mathbf{Y}, \mathbf{Y}] \in \mathbb{R}^{d\times 2N}$ and $\mathbf{Y} \in \mathbb{R}^{d\times N}$ is the label matrix represented by one-hot encoding, $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_M \end{bmatrix} \in \mathbb{R}^{(d_H+d_M)\times 2N}$ and $\mathbf{X}_H$ and $\mathbf{X}_M$ stand respectively for the data from hyperspectral and multispectral domains, $\mathbf{\Theta} = [\mathbf{\Theta}_H, \mathbf{\Theta}_M]$ and $\mathbf{P}$ are respectively the common subspace projection and the linear projection to bridge the common subspace and label information. $\mathbf{L} = \mathbf{D} - \mathbf{W} \in \mathbb{R}^{2N\times 2N}$ stands for a joint Laplacian matrix, $\mathbf{W}$ is an adjacency matrix and $\mathbf{D}_{ii} = \sum_{i\neq j} \mathbf{W}_{i,j}$. $\mathbf{W}$ is generally used to measure the similarity between samples. With the orthogonal constraint ($\mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}$), the global optimal solutions with respect to the variables $\mathbf{\Theta}$ and $\mathbf{P}$ can be theoretically guaranteed (Ji and Ye, 2009).

Algorithm 1. Learnable Manifold Alignment (LeMA)

Input: $\widetilde{\mathbf{Y}}$, $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{X}}'$, $\widetilde{\mathbf{L}}$, α, β, *maxIter*.
Output: $\mathbf{P}$, $\mathbf{\Theta}$, $\widetilde{\mathbf{L}}$

$t = 1, \zeta = 1e-4$;
Initializating $\mathbf{P}$ and $\mathbf{\Theta}$
while *not converged* or $t >$ *maxIter* **do**
  Fix other variables to update $\mathbf{P}$ by Eq. (6)
  Fix other variables to update $\mathbf{\Theta}$ by **Algorithm 2**
  Fix other variables to update $\widetilde{\mathbf{L}}$ by equivalently optimizing $\widetilde{\mathbf{W}}$ in a distributed fashion:
    1. update $\widetilde{\mathbf{W}}_{HU}$ by **Algorithm 3**;
    2. update $\widetilde{\mathbf{W}}_{MU}$ by **Algorithm 3**;
    3. align $\widetilde{\mathbf{W}}_{HU}$ and $\widetilde{\mathbf{W}}_{MU}$ by $\max(\widetilde{\mathbf{W}}_{HU}, \widetilde{\mathbf{W}}_{MU})$;
    4. update $\widetilde{\mathbf{W}}_{UU}$ by **Algorithm 4**
    5. compute $\widetilde{\mathbf{L}} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}$, $\widetilde{\mathbf{D}}_{ii} = \sum_{i\neq j} \widetilde{\mathbf{W}}_{ij}$
  Compute the objective function value E^{t+1} and check the convergence condition: **if** $|\frac{E^{t+1}-E^t}{E^t}| < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t+1$;
  end
end

The first term of Eq. (1) is a fidelity term, and the regularization term $\frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2$ parameterized by α aims to achieve a reliable generalization of the proposed model. The third term acts as supervised manifold alignment (SMA) (Wang et al., 2011). We refer to the proposed framework for joint common subspace learning as CoSpace.

To further exploit the information of unlabeled samples, we extend the CoSpace in Eq. (1) to LeMA by learning a joint Laplacian matrix, which can be formulated as follows with extra constraints related to necessary conditions of $\widetilde{\mathbf{L}}$:

$$\min_{\mathbf{P},\boldsymbol{\Theta},\widetilde{\mathbf{L}}} \frac{1}{2}\left\|\widetilde{\mathbf{Y}}-\mathbf{P}\boldsymbol{\Theta}\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2+\frac{\alpha}{2}\left\|\mathbf{P}\right\|_{\mathrm{F}}^2+\frac{\beta}{2}\mathrm{tr}\left(\mathbf{H}\widetilde{\mathbf{L}}\mathbf{H}^{\mathrm{T}}\right)$$
$$\text{s.t. } \mathbf{H}=\boldsymbol{\Theta}\widetilde{\mathbf{X}}',\ \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\mathbf{I},\ \widetilde{\mathbf{L}}=\widetilde{\mathbf{L}}^{\mathrm{T}},\ \widetilde{\mathbf{L}}_{i,j,i\neq j}\leq 0,\ \widetilde{\mathbf{L}}_{i,j,i=j}\geq 0,\ \mathrm{tr}(\widetilde{\mathbf{L}})=s, \tag{2}$$

where $\widetilde{\mathbf{X}}'=\begin{bmatrix}\mathbf{X}_H & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_M & \mathbf{X}_U\end{bmatrix}\in\mathbb{R}^{(d_H+d_M)\times(2N+N_U)}$, $\widetilde{\mathbf{L}}\in\mathbb{R}^{(2N+N_U)\times(2N+N_U)}$, and $\mathbf{X}_U\in\mathbb{R}^{d_M\times N_U}$ represents the unlabeled MS samples and $s>0$ controls the scale. Note that a feasible and effective approach to choose the unlabeled data with respect to the variable $\widetilde{\mathbf{X}}'$ is to group total samples besides the training samples into some landmarks (cluster centers). These landmarks are used as the unlabeled data, which can fully take into account the available information and meanwhile effectively reduce the computational cost. Due to the use of clustering technique in unlabeled data, we experimentally and empirically set the ratio of labeled and unlabeled data to approximately be 1:1.

The model in Eq. (2) can be simplified by optimizing the adjacency matrix ($\widetilde{\mathbf{W}}$) instead of directly solving a hard optimization problem of $\widetilde{\mathbf{L}}$, then we have

$$\mathrm{tr}\left(\mathbf{H}\widetilde{\mathbf{L}}\mathbf{H}^{\mathrm{T}}\right)=\frac{1}{2}\mathrm{tr}\left(\widetilde{\mathbf{W}}\mathbf{Z}\right)=\frac{1}{2}\left\|\widetilde{\mathbf{W}}\odot\mathbf{Z}\right\|_{1,1}, \tag{3}$$

where $\widetilde{\mathbf{W}}\in\mathbb{R}^{(2N+N_U)\times(2N+N_U)}$, $\mathbf{Z}\in\mathbb{R}^{(2N+N_U)\times(2N+N_U)}$ is defined as *a pairwise Euclidean distance matrix*: $\mathbf{Z}_{i,j}=\|\mathbf{H}_i-\mathbf{H}_j\|^2$. $\odot$ denotes the Schur-Hadamard (termwise) product.

Algorithm 2. Solving the subproblem for $\boldsymbol{\Theta}$

Input: $\widetilde{\mathbf{Y}}$, $\mathbf{P}$, $\mathbf{J}$, $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{X}}'$, $\widetilde{\mathbf{L}}$, β, *maxIter*.
Output: $\boldsymbol{\Theta}$.
Initialization: $\boldsymbol{\Theta}=\mathbf{0}$, $\mathbf{G}=\mathbf{0}$, $\boldsymbol{\Lambda}_1=\mathbf{0}$, $\boldsymbol{\Lambda}_2=\mathbf{0}$, $\mu=10^{-3}$, $\mu_{\max}=10^6$, $\rho=1.5$, $\varepsilon=10^{-6}$, $t=1$.
while *not converged* or $t>$ *maxIter* **do**
  Fix other variables to update $\mathbf{J}$ by $\mathbf{J}=(\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mu\mathbf{I})^{-1}(\mathbf{P}^{\mathrm{T}}\widetilde{\mathbf{Y}}+\mu\boldsymbol{\Theta}\widetilde{\mathbf{X}}-\boldsymbol{\Lambda}_1)$.
  Fix other variables to update $\boldsymbol{\Theta}$ by
$$\boldsymbol{\Theta}=(\mu\mathbf{J}\widetilde{\mathbf{X}}^{\mathrm{T}}+\boldsymbol{\Lambda}_1\widetilde{\mathbf{X}}^{\mathrm{T}}+\mu\mathbf{G}+\boldsymbol{\Lambda}_2)\times(\mu\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}+\mu\mathbf{I}+\beta\widetilde{\mathbf{X}}'\widetilde{\mathbf{L}}\widetilde{\mathbf{X}}'^{\mathrm{T}})^{-1}.$$
  Fix other variables to update $\mathbf{G}$ by
$$[\mathbf{U},\mathbf{S},\mathbf{V}]=\mathrm{svd}(\boldsymbol{\Theta}-\boldsymbol{\Lambda}_2/\mu),\quad \mathbf{G}=\mathbf{U}\mathbf{I}_{n\times m}\mathbf{V}.$$
  Update Lagrange multipliers by
$$\boldsymbol{\Lambda}_1\leftarrow\boldsymbol{\Lambda}_1+\mu(\mathbf{J}-\boldsymbol{\Theta}\widetilde{\mathbf{X}}),\quad \boldsymbol{\Lambda}_2\leftarrow\boldsymbol{\Lambda}_2+\mu(\mathbf{G}-\boldsymbol{\Theta}).$$
  Update penalty parameter by $\mu=\min(\rho\mu,\mu_{\max})$.
  Check the convergence conditions: **if** $\|\mathbf{J}-\boldsymbol{\Theta}\widetilde{\mathbf{X}}\|_{\mathrm{F}}<\varepsilon$ *and* $\|\mathbf{G}-\boldsymbol{\Theta}\|_{\mathrm{F}}<\varepsilon$ **then**
    Stop iteration;
  else
    $t\leftarrow t+1$;
  end
end

Using Eq. (3), we can equivalently convert the optimization problem of smooth manifold in (2) to that of graph sparsity

$$\min_{\mathbf{P},\boldsymbol{\Theta},\widetilde{\mathbf{W}}} \frac{1}{2}\left\|\widetilde{\mathbf{Y}}-\mathbf{P}\boldsymbol{\Theta}\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2+\frac{\alpha}{2}\left\|\mathbf{P}\right\|_{\mathrm{F}}^2+\frac{\beta}{4}\left\|\widetilde{\mathbf{W}}\odot\mathbf{Z}\right\|_{1,1}$$
$$\text{s.t. } \mathbf{H}=\boldsymbol{\Theta}\widetilde{\mathbf{X}}',\ \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{T}}=\mathbf{I},\ \widetilde{\mathbf{W}}=\widetilde{\mathbf{W}}^{\mathrm{T}},\ \widetilde{\mathbf{W}}_{ij}\geq 0,\ \|\widetilde{\mathbf{W}}\|_{1,1}=s, \tag{4}$$

where $\|\widetilde{\mathbf{W}}\odot\mathbf{Z}\|_{1,1}$ can be interpreted as a *weighted ℓ_1-norm of $\widetilde{\mathbf{W}}$ which enforces weighted sparsity*.

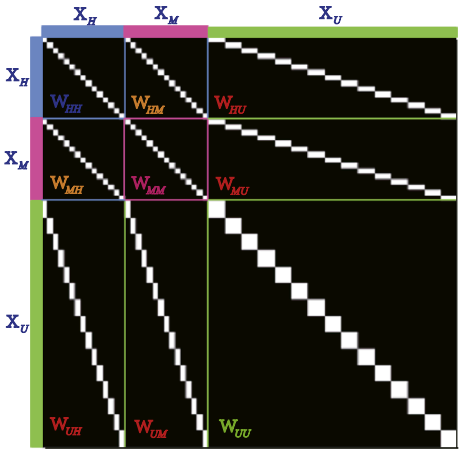
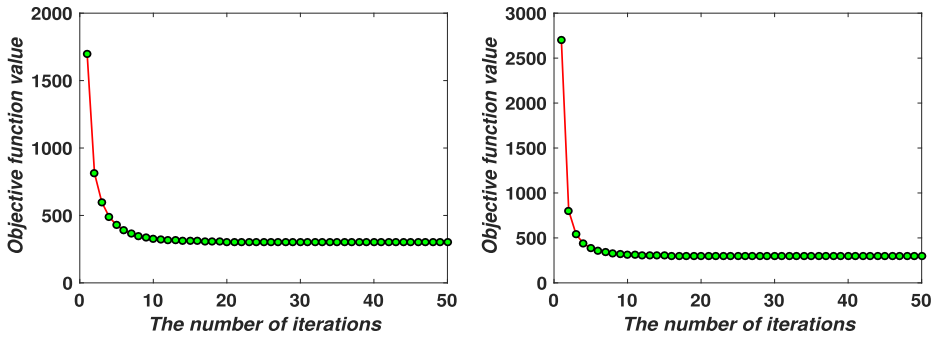


Fig. 2. An example for the joint adjacency matrix $\widetilde{\mathbf{W}}$.

We further elaborate the relationship between the proposed LeMA model and our motivation in an easy-understanding way. In general, we aim at finding a common subspace by learning a pair of projections ($\boldsymbol{\Theta}_M$ and $\boldsymbol{\Theta}_H$) corresponding to two kinds of different modalities (e.g., MS and HS), respectively. In order to effectively improve the discriminative ability of the learned subspace, we make a connection between the subspace and label information by jointly estimating the regression coefficient $\mathbf{P}$ and common projections $\boldsymbol{\Theta}$, as formulated in Eq. (1). What's more, the alignment behavior of different modalities can be represented by $\mathbf{W}$'s connectivity, that is, if the i^{th} sample $\mathbf{X}_i$ and the j^{th} sample $\mathbf{X}_j$ are connected ($\mathbf{W}_{i,j}=1$), and then the two samples belong to the same class; *vice versa*. Besides, we construct an extra adjacency matrix based on those unlabeled samples in order to globally capture the data distribution. The matrix is usually obtained by a Gaussian kernel function (semi-supervised CoSpace) and also can be learned from the data (LeMA as formulated in Eq. (2)).

Algorithm 3. Solving the subproblem for $\widetilde{\mathbf{W}}_{HU(MU)}$

Input: $\mathbf{Z}_{H(M)}$, $\mathbf{Z}_U$, $\widetilde{\mathbf{W}}$, β, *maxIter*.
Output: $\widetilde{\mathbf{W}}$.
Initialization: $\mathbf{M} = \widetilde{\mathbf{W}}$, $\mathbf{S} = \mathbf{U} = \mathbf{K} = \mathbf{0}$, $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{\Lambda}_3 = \mathbf{\Lambda}_4 = \mathbf{0}$, $\mu = 10^{-2}$, $\mu_{\max} = 10^6$, $\rho = 2$, $\varepsilon = 10^{-6}$, $t = 1$.
Compute Z: $\mathbf{Z}_{i,j} = \|\mathbf{Z}^i_{H(M)} - \mathbf{Z}^j_U\|^2_{\mathrm{F}}$.
while *not converged* or $t > maxIter$ **do**
  Fix other variables to update $\widetilde{\mathbf{W}}$ by
$$\widetilde{\mathbf{W}} = (\mathbf{M} + \mathbf{S} + \mathbf{U} + \mathbf{K} + \mathbf{\Lambda}_1 + \mathbf{\Lambda}_2 + +\mathbf{\Lambda}_3 + \mathbf{\Lambda}_4)/(4\mu).$$
  Fix other variables to update $\mathbf{U}$ by $\mathbf{U} = \max(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_1/\mu, 0)$.
  Fix other variables to update $\mathbf{M}$ by
$$\mathbf{M} = \max(\|\widetilde{\mathbf{W}} - \mathbf{\Lambda}_2/\mu\|_{1,1} - (\beta\mathbf{Z}/4\mu), 0) \odot \mathrm{sign}(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_2/\mu).$$
  Fix other variables to update $\mathbf{S}$ by $\mathbf{S} = \mathrm{prox}(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_3/\mu)$.
  Fix other variables to update $\mathbf{K}$ by $\mathbf{K} = \min(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_4/\mu, 1/N_k)$.
  Update Lagrange multipliers by
$$\mathbf{\Lambda}_1 = \mathbf{\Lambda}_1 + \mu(\mathbf{U} - \widetilde{\mathbf{W}}), \quad \mathbf{\Lambda}_2 = \mathbf{\Lambda}_2 + \mu(\mathbf{M} - \widetilde{\mathbf{W}}),$$
$$\mathbf{\Lambda}_3 = \mathbf{\Lambda}_3 + \mu(\mathbf{S} - \widetilde{\mathbf{W}}), \quad \mathbf{\Lambda}_4 = \mathbf{\Lambda}_4 + \mu(\mathbf{K} - \widetilde{\mathbf{W}}).$$
  Update penalty parameter by $\mu = \min(\rho\mu, \mu_{\max})$. Check the convergence conditions: **if** $\|\mathbf{U} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{M} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{S} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{K} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\widetilde{\mathbf{W}}^{t+1} - \widetilde{\mathbf{W}}^{t}\|_{F} < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
  end
end



(a) The University of Houston MS-HS Datasets (b) The Chikusei MS-HS Datasets

Fig. 3. Convergence analysis of LeMA are experimentally performed on the two MS-HS datasets.

2.3. Model optimization

Considering the complexity of the non-convex problem (4), an iterative alternating optimization strategy is adopted to solve the convex subproblems of each variable $\mathbf{P}$, $\mathbf{\Theta}$, and $\mathbf{W}$. An implementation of LeMA is given in Algorithm 1.

Optimization with respect to $\mathbf{P}$: This is a typical least-squares problem with Tikhonov regularization, which can be formulated as

$$\min_{\mathbf{P}} \frac{1}{2}\left\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\mathbf{P}\right\|_{\mathrm{F}}^2, \tag{5}$$

which has a closed-form solution

$$\mathbf{P} = (\widetilde{\mathbf{Y}}\mathbf{E}^{\mathrm{T}})(\mathbf{E}\mathbf{E}^{\mathrm{T}} + \alpha\mathbf{I})^{-1}, \tag{6}$$

where $\mathbf{E} = \mathbf{\Theta}\widetilde{\mathbf{X}}$.

Optimization with respect to $\mathbf{\Theta}$: the optimization problem for $\mathbf{\Theta}$ can be formulated as

$$\min_{\mathbf{\Theta}} \frac{1}{2}\left\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}\widetilde{\mathbf{X}}\right\|_{\mathrm{F}}^2 + \frac{\beta}{2}\mathrm{tr}\left(\mathbf{H}\widetilde{\mathbf{L}}\mathbf{H}^{\mathrm{T}}\right) \text{ s.t. } \mathbf{H} = \mathbf{\Theta}\widetilde{\mathbf{X}}', \ \mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}} = \mathbf{I}. \tag{7}$$

In order to solve (7) effectively with ADMM, we consider an equivalent form by introducing auxiliary variables $\mathbf{J}$ and $\mathbf{G}$ to replace $\mathbf{\Theta}\widetilde{\mathbf{X}}$ and $\mathbf{\Theta}$, respectively.

$$\min_{\mathbf{\Theta},\mathbf{J},\mathbf{G}} \tfrac{1}{2}\left\|\widetilde{\mathbf{Y}} - \mathbf{P}\mathbf{J}\right\|_{\mathrm{F}}^2 + \tfrac{\beta}{2}\mathrm{tr}\left(\mathbf{\Theta}\widetilde{\mathbf{X}}'\widetilde{\mathbf{L}}(\mathbf{\Theta}\widetilde{\mathbf{X}}')^{\mathrm{T}}\right)$$
$$\text{s.t. } \mathbf{J} = \mathbf{\Theta}\widetilde{\mathbf{X}}, \ \mathbf{G} = \mathbf{\Theta}, \ \mathbf{G}\mathbf{G}^{\mathrm{T}} = \mathbf{I}. \tag{8}$$

Algorithm 2 lists the more detailed procedures for solving the problem (8).

Algorithm 4. Solving the subproblem for $\widetilde{\mathbf{W}}_{UU}$

Input: $\mathbf{Z}_U$, $\widetilde{\mathbf{W}}$, γ, *maxIter*.
Output: $\widetilde{\mathbf{W}}$.
Initialization: $\mathbf{M} = \widetilde{\mathbf{W}}$, $\mathbf{U} = \mathbf{V} = \mathbf{S} = \mathbf{K} = \mathbf{T} = \mathbf{0}$, $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{\Lambda}_3 = \mathbf{\Lambda}_4 = \mathbf{\Lambda}_5 = \mathbf{\Lambda}_6 = \mathbf{\Lambda}_7 = \mathbf{0}$, $\mu = 10^{-2}$, $\mu_{\max} = 10^6$, $\rho = 2$, $\varepsilon = 10^{-6}$, $t = 1$.
Compute Z: $\mathbf{Z}_{i,j} = \|\mathbf{Z}_U^i - \mathbf{Z}_U^j\|_{\mathrm{F}}^2$.
while *not converged* or $t > maxIter$ **do**
  Fix other variables to update $\widetilde{\mathbf{W}}$ by
    $\widetilde{\mathbf{W}} = (\mathbf{V} + \mathbf{U}^{\mathrm{T}} + \mathbf{M} + \mathbf{S} + \mathbf{K} + \mathbf{T} + \mathbf{\Lambda}_1 + \mathbf{\Lambda}_2^{\mathrm{T}} + \mathbf{\Lambda}_3 + \mathbf{\Lambda}_4 + \mathbf{\Lambda}_5 + \mathbf{\Lambda}_7)/(6\mu)$.
  Fix other variables to update $\mathbf{U}$ by $\mathbf{U} = \big(\widetilde{\mathbf{W}}^{\mathrm{T}} + \mathbf{V} - (\mathbf{\Lambda}_1 + \mathbf{\Lambda}_6)\big)/(2\mu)$.
  Fix other variables to update $\mathbf{V}$ by $\mathbf{V} = \big(\widetilde{\mathbf{W}} + \mathbf{U} - (\mathbf{\Lambda}_2 + \mathbf{\Lambda}_6)\big)/(2\mu)$.
  Fix other variables to update $\mathbf{M}$ by
    $\mathbf{M} = \max(\|\widetilde{\mathbf{W}} - \mathbf{\Lambda}_3/\mu\|_{1,1} - \gamma\mathbf{Z}/(4\mu), 0) \odot \mathrm{sign}(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_3/\mu)$.
  Fix other variables to update $\mathbf{S}$ by $\mathbf{S} = \mathrm{prox}(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_4/\mu)$.
  Fix other variables to update $\mathbf{K}$ by $\mathbf{K} = \max(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_5/\mu, 0)$.
  Fix other variables to update $\mathbf{T}$ by $\mathbf{T} = \min(\widetilde{\mathbf{W}} - \mathbf{\Lambda}_7/\mu, 1/N_k)$.
  Update Lagrange multipliers by
    $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_1 + \mu(\mathbf{U} - \widetilde{\mathbf{W}}^{\mathrm{T}})$, $\mathbf{\Lambda}_2 = \mathbf{\Lambda}_2 + \mu(\mathbf{V} - \widetilde{\mathbf{W}})$,
    $\mathbf{\Lambda}_3 = \mathbf{\Lambda}_3 + \mu(\mathbf{M} - \widetilde{\mathbf{W}})$, $\mathbf{\Lambda}_4 = \mathbf{\Lambda}_4 + \mu(\mathbf{S} - \widetilde{\mathbf{W}})$,
    $\mathbf{\Lambda}_5 = \mathbf{\Lambda}_5 + \mu(\mathbf{K} - \widetilde{\mathbf{W}})$, $\mathbf{\Lambda}_6 = \mathbf{\Lambda}_6 + \mu(\mathbf{U} - \mathbf{V})$,
    $\mathbf{\Lambda}_7 = \mathbf{\Lambda}_7 + \mu(\mathbf{T} - \widetilde{\mathbf{W}})$.
  Update penalty parameter by $\mu = \min(\rho\mu, \mu_{\max})$.
  Check the convergence conditions: **if** $\|\mathbf{U} - \widetilde{\mathbf{W}}^{\mathrm{T}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{V} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{M} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{S} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{K} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{U} - \mathbf{V}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\mathbf{T} - \widetilde{\mathbf{W}}\|_{\mathrm{F}} < \varepsilon$ *and* $\|\widetilde{\mathbf{W}}^{t+1} - \widetilde{\mathbf{W}}^{t}\|_{\mathrm{F}} < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t + 1$;
  end
end

Optimization with respect to $\widetilde{\mathbf{W}}$: $\widetilde{\mathbf{W}}$ is a joint adjacency matrix and consists mainly of nine parts as shown in Fig. 2. Among the nine parts, $\widetilde{\mathbf{W}}_{HH}$, $\widetilde{\mathbf{W}}_{HM}$, $\widetilde{\mathbf{W}}_{MH}$ and $\widetilde{\mathbf{W}}_{MM}$ can be directly inferred from label information in the form of the LDA-like graph (Gu et al., 2011):

$$\widetilde{\mathbf{W}}_{i,j} = \begin{cases} 1/N_k, & \text{if } \mathbf{X}_i \text{ and } \mathbf{X}_j \text{ belong to the } k\text{-th class;} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Given the symmetry of $\widetilde{\mathbf{W}}$, (i.e., $\widetilde{\mathbf{W}}_{HM} = \widetilde{\mathbf{W}}_{MH}$, $\widetilde{\mathbf{W}}_{MU} = \widetilde{\mathbf{W}}_{UM}$, and $\widetilde{\mathbf{W}}_{MU} = \widetilde{\mathbf{W}}_{UM}$), we only need to update three of out nine parts, namely $\widetilde{\mathbf{W}}_{HU}$, $\widetilde{\mathbf{W}}_{MU}$, and $\widetilde{\mathbf{W}}_{UU}$. The optimization problems of $\widetilde{\mathbf{W}}_{HU}$ and $\widetilde{\mathbf{W}}_{MU}$ can be formulated by

$$\min_{\widetilde{\mathbf{W}}_{HU(MU)}} \frac{\beta}{4}\left\|\widetilde{\mathbf{W}} \odot \mathbf{Z}\right\|_{1,1} \text{ s.t. } 1/N_k \geq \widetilde{\mathbf{W}}_{i,j} \geq 0, \ \left\|\widetilde{\mathbf{W}}\right\|_{1,1} = s, \tag{10}$$

which can be solved by ADMM. More details can be found in Algorithm 3, where $\mathbf{Z}_{H(M)}$ and $\mathbf{Z}_U$ represent respectively the subspace features of $\mathbf{X}_{H(M)}$ and $\mathbf{X}_U$, prox stands for the proximal operator for $\|\widetilde{\mathbf{W}}\|_{1,1} = s$ (Heide et al., 2015). We technically add the constraint $\widetilde{\mathbf{W}}_{i,j} \leq 1/N_k$ in order to share the same unit level with LDA-like graph.

For $\widetilde{\mathbf{W}}_{UU}$, the objective function can be written as

$$\min_{\widetilde{\mathbf{W}}_{UU}} \frac{\beta}{4}\left\|\widetilde{\mathbf{W}} \odot \mathbf{Z}\right\|_{1,1} \text{ s.t. } \widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}^{\mathrm{T}}, \ 1/N_k \geq \widetilde{\mathbf{W}}_{i,j} \geq 0, \ \left\|\widetilde{\mathbf{W}}\right\|_{1,1} = s, \tag{11}$$

which can be effectively solved using Algorithm 4.

Finally, we repeat these optimization procedures until a stopping criterion is satisfied.

2.4. Convergence analysis

The alternative alternating strategy used in Algorithm 1 is nothing but a block coordinate descent (BCD), which has been theoretically supported to converge to a stationary point as long as each subproblem in Eq. (4) is exactly minimized (Bertsekas, 1999). As observed, these subproblems with respect to the variables $\mathbf{P}$, $\mathbf{\Theta}$ and $\widetilde{\mathbf{W}}$ are strongly convex, and hence each independent task can ideally find a unique minimum when the Lagrangian parameter is updated within finitely iterative steps (Boyd et al., 2011). Besides, ADMM used in each subproblem optimization is actually generalized to *inexact* Augmented Lagrange Multiplier (ALM) (Chen et al., 2018), whose convergence has been well studied when the number of block is less than three (Lin et al., 2010) (e.g. Algorithm 2). Although there is still not a *generally and strictly* theoretical proof in multi-blocks case, yet the convergence analysis for some common cases such as our Algorithms 3 and 4 has been well conducted in Hong et al. (2017), Liu et al. (2013), Zhong et al. (2016) and Zhou et al. (2017). We also experimentally record the objective function values in each iteration to draw the convergence curves of LeMA on two used HS-MS datasets (see Fig. 3).

3. Experiments

In this section, we quantitatively and qualitatively evaluate the performance of the proposed method on two simulated HS-MS datasets (University of Houston and Chikusei) and a real multispectral-lidar and hyperspectral dataset provided by 2018 IEEE GRSS data fusion contest (DFC2018), by the form of classification using two commonly used and high-performance classifiers, namely linear support vector machines (LSVM), and canonical correlation forest (CCF) (Tom and Frank, 2015). Three indices: overall accuracy (OA), average accuracy (AA), kappa coefficient (κ), are calculated to quantitatively assess the classification performance. Moreover, we compare the performance of the proposed LeMA and several other state-of-art algorithms, i.e. GLP (Zhu et al., 2003), SMA, S-SMA (Wang and Mahadevan, 2009), CoSpace and Semi-supervised

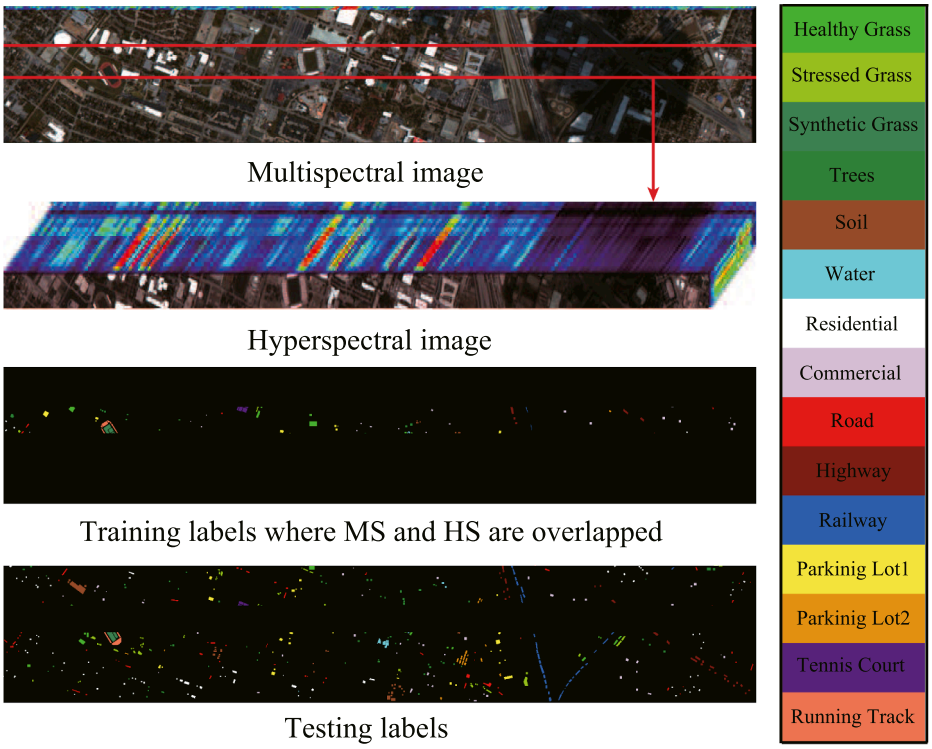


Fig. 4. The multispectral image and its corresponding hyperspectral image that partially covers the same area, as well as training and testing labels, for University of Houston dataset.

CoSpace (S-CoSpace). The original MS data is used as a baseline. SMA constructs an LDA-like joint graph using label information. Besides label information, S-SMA method also uses unlabeled samples to generate the joint graph by computing the similarity based on Euclidean distance. The same strategy of graph construction is adopted for CoSpace and S-CoSpace.

3.1. The simulated MS-HS datasets over the University of Houston

3.1.1. Data description

The HS data in the simulated *Houston MS-HS datasets* was acquired by the ITRES-CASI-1500 sensor with the size of 349 × 1905 at a ground sampling distance (GSD) of 2.5 m over the University of Houston campus and its neighboring urban areas. This data was provided for the 2013 IEEE GRSS data fusion contest, with 144 bands covering the wavelength range from 364 nm to 1046 nm. Spectral simulation is performed to generate the MS image by degrading the HS image in the spectral domain using the MS spectral response functions (SRFs) of Sentinel-2 as filters (for more details refer to Yokoya et al., 2017). The MS data we used is generated with dimensions of 349 × 1905 × 10.

3.1.2. Experimental setup

To meet our problem setting, a HS image partially overlapping MS image and a whole MS image are used in our experiments, and meanwhile the corresponding training and test samples can be re-assigned, as shown in Fig. 4. In detail, since the total labels are available, we seek

Table 1
The number of training and testing samples for the two used MS-HS datasets.

Class No.	Houston MS-HS dataset			Chikusei MS-HS dataset		
	Class Name	Training	Testing	Class Name	Training	Testing
1	Healthy Grass	537	699	Water	301	858
2	Stressed Grass	61	1154	Bare Soil (School)	992	1867
3	Synthetic Grass	340	357	Bare Soil (Farmland)	455	4397
4	Tree	209	1035	Natural Plants	150	4272
5	Soil	74	1168	Weeds in Farmland	928	1108
6	Water	22	303	Forest	486	11904
7	Residential	52	1203	Grass	989	5526
8	Commercial	320	924	Rice Field (Grown)	813	8816
9	Road	76	1149	Rice Field (First Stage)	667	1268
10	Highway	279	948	Row Crops	377	5961
11	Railway	33	1185	Plastic House	165	475
12	Parking Lot1	329	904	Manmade (Non-dark)	170	568
13	Parking Lot2	20	449	Manmade (Dark)	1291	6373
14	Tennis Court	266	162	Manmade (Blue)	111	431
15	Running Track	279	381	Manmade (Red)	35	187
16	/	/	/	Manmade Grass	21	1019
17	/	/	/	Asphalt	384	417
	Total	2897	12021	Total	8335	55447

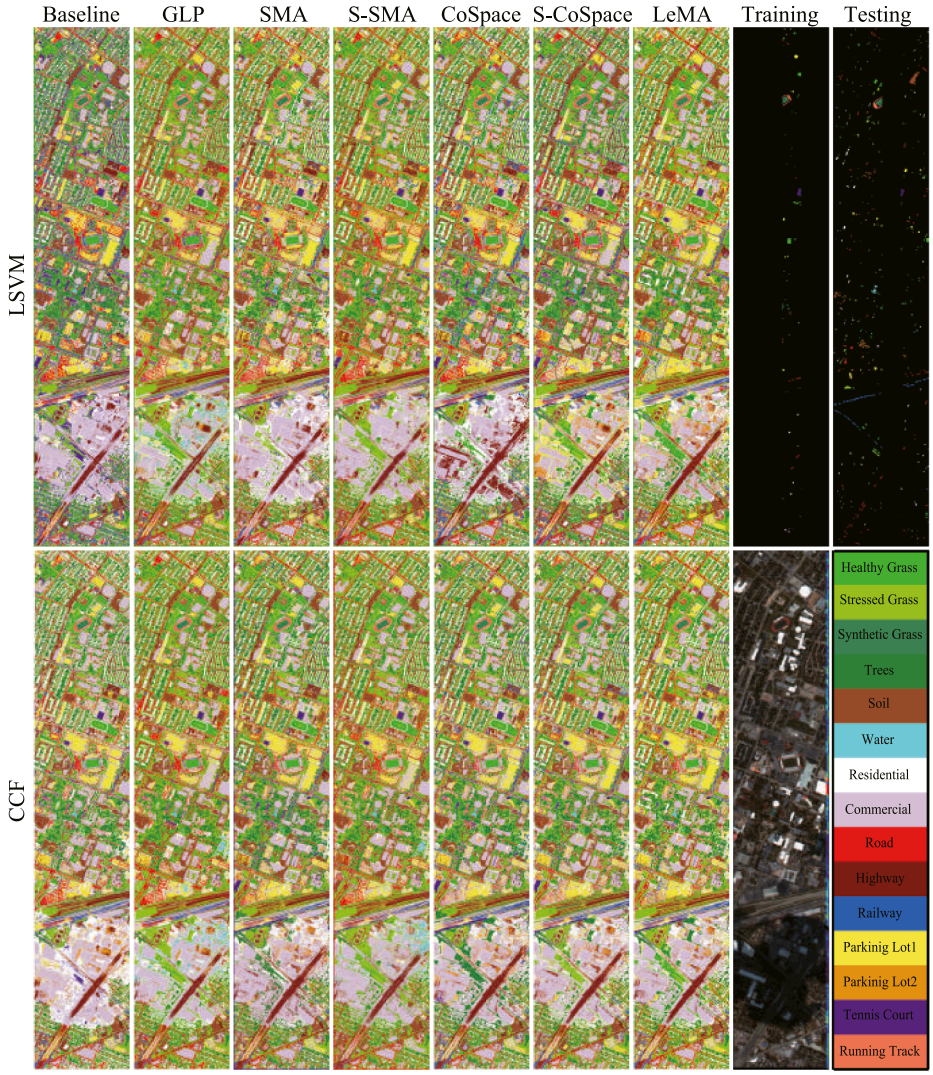


Fig. 5. Classification maps of the different algorithms obtained using two kinds of classifiers on the University of Houston dataset.

Table 2
Quantitative performance comparison with the different algorithms on the University of Houston data. The best one is shown in bold.

Methods Parameter	Baseline (%) d 10		GLP (%) (k, σ, d) (10, 1, 10)		SMA (%) d 30		S-SMA (%) (k, σ, d) (10, 0.1, 30)		CoSpace (%) (α, β, d) (0.01, 0.01, 30)		S-CoSpace (%) (α, β, d) (0.1, 0.01, 30)		LeMA (%) (α, β, d) (0.01, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	62.12	68.21	64.71	70.01	68.01	69.59	69.29	70.10	69.38	72.17	70.41	73.75	73.42	**76.35**
AA	65.97	70.47	68.18	72.18	70.50	71.02	72.00	72.88	71.69	73.56	73.12	75.61	74.76	**77.18**
κ	0.5889	0.6543	0.6164	0.6728	0.6520	0.6695	0.6659	0.6754	0.6672	0.6975	0.6784	0.7146	0.7110	**0.7428**
Class1	76.39	67.95	77.83	77.97	75.25	68.53	74.25	73.53	75.54	69.96	**91.85**	87.98	89.56	85.84
Class2	80.59	78.08	93.85	**98.01**	97.57	77.9	97.57	93.67	73.74	77.99	90.12	91.59	93.67	93.85
Class3	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Class4	85.51	92.27	89.66	96.62	94.78	98.74	95.85	98.55	98.74	98.26	92.75	97.29	97.49	**99.61**
Class5	99.06	99.4	99.49	**99.66**	98.97	99.14	99.32	99.4	99.4	99.4	99.4	**99.66**	99.49	99.57
Class6	86.14	86.14	96.37	99.01	86.47	70.96	**99.67**	99.67	85.48	85.15	**99.67**	96.70	86.47	86.47
Class7	50.62	63.76	48.63	64.01	72.32	77.14	72.15	69.66	73.98	80.05	75.06	80.96	83.21	**88.03**
Class8	56.49	56.06	56.60	59.85	62.01	62.23	**64.61**	63.85	63.53	62.01	55.84	60.39	62.77	62.01
Class9	56.22	70.58	69.63	69.02	49.96	61.27	50.57	45.00	59.79	64.93	65.8	**71.54**	64.49	61.88
Class10	45.36	45.25	45.46	49.89	58.12	52.32	58.33	63.61	**64.14**	57.70	58.97	51.79	60.97	53.59
Class11	27.43	43.88	22.45	38.65	28.86	36.46	36.46	34.77	36.54	47.26	35.78	38.65	41.27	**49.96**
Class12	31.64	56.08	31.75	37.83	35.84	62.50	34.18	55.2	46.79	62.72	34.29	58.52	45.02	**76.88**
Class13	0.00	0.67	0.00	1.11	0.00	0.00	0.00	0.45	0.00	0.45	0.00	0.89	0.00	**1.78**
Class14	97.53	98.77	94.44	92.59	**100.00**	**100.00**	99.38	98.15	**100.00**	99.38	99.38	**100.00**	99.38	**100.00**
Class15	96.59	98.16	96.59	**98.43**	97.38	98.16	97.64	97.64	97.64	98.16	97.90	98.16	97.64	98.16

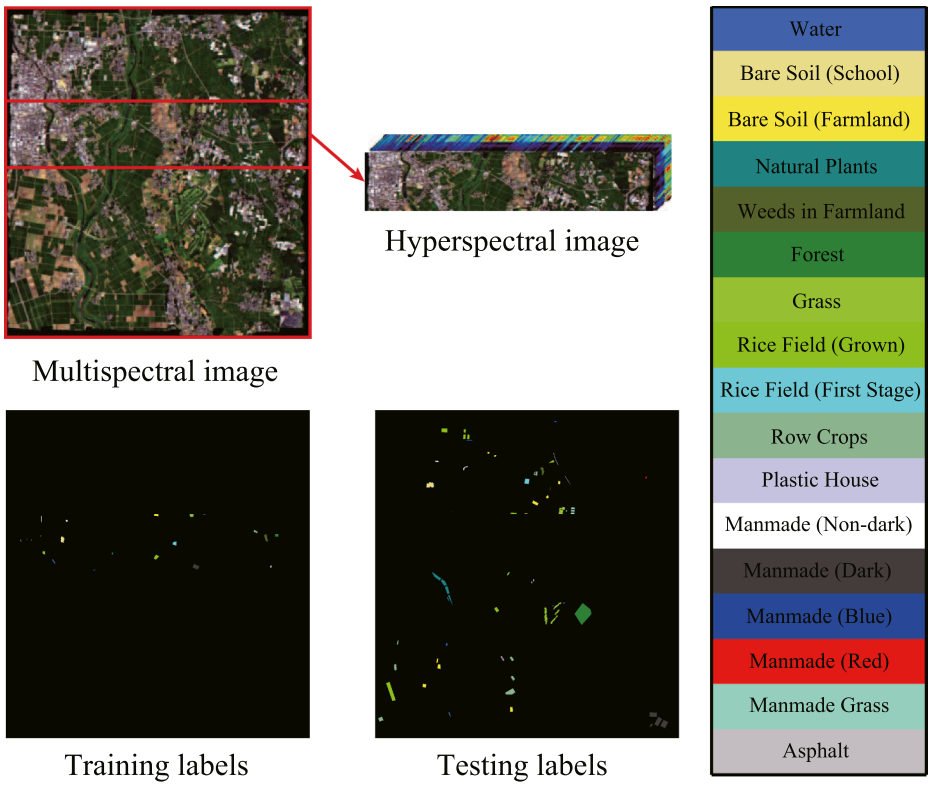


Fig. 6. The multispectral image and its corresponding hyperspectral image that partially covers the same area, as well as training and testing labels, for Chikusei Dataset.

out a region where all kinds of classes are involved. The labels in the region are selected as the training set and the rest are seen as the test set, as shown in Fig. 4 and specifically quantified in Table 1.

The parameters of the different methods are determined by a 10-fold cross-validation on the training data. More specifically, we tune the parameters of the different algorithms to maximize their performances, e.g. dimension (d), penalty parameters (α, β), etc. The dimension (d) is a common parameter for all compared algorithms, and it can be determined covering the range from 10 to 50 at an interval of 10. For the number of nearest neighbors (k) and the standard deviation of Gaussian kernel function (σ) in artificially computing the adjacency matrix ($\mathbf{W}$) of GLP, SMA, and S-SMA, we select them in the range of $\{10, 20, ..., 50\}$ and $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, respectively, Similarly to CoSpace, S-CoSpace and LeMA, we set the two regularization parameters (α, β) ranging from $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$.

3.1.3. Results and analysis

Fig. 5 shows the classification maps of compared algorithms using LSVM and CCF classifiers, while Table 2 lists the specific quantitative assessment results with optimal parameters obtained by 10-fold cross-validation.

Overall, the methods based on manifold alignment outperform baseline and GLP using the different classifiers. This means that the limited amount of HS data can guide the corresponding MS data towards better discriminative feature representations. More specifically when compared with S-SMA, SMA yields a relatively poor performance since it only considers the correspondences of MS-HS labeled data. This indicates that reasonably embedding unlabeled samples into the manifold alignment framework can effectively help us capture the real data distribution, and thereby obtain more accurate decision boundaries. Unfortunately, these approaches only attempt to align different data in a common subspace, but they hardly take the connections between the common subspace and label information into account,[2] which leads to a lack of discriminative ability. With regards to this, our proposed joint learning framework "CoSpace" and its semi-supervised version "S-CoSpace" achieve the desired results on the given MS-HS datasets.

By fully considering the connectivity of the common subspace, label information, and unlabeled information encoded by the learned graph structure, the performance of LeMA is much more superior to that of any other methods as can be observed in Table 2. This demonstrates that LeMA is likely to learn a more discriminative feature representation and to find a better decision boundary.

As observed from Fig. 4 and Table 2, the training samples are relatively a few and meanwhile the distribution between different classes is extremely unbalanced. While training the classifier, more attentions are paid on those classes with large-size samples, and some small-scale classes possibly play less and even nothing. For this reason, we propose to consider those large-scale unlabeled data, achieving a semi-supervised learning. Using this strategy, the semi-supervised methods, i.e. GLP, S-SMA, S-CoSpace, obviously perform better than baseline and their supervised ones (SMA and CoSpace). Moreover, we can see from Table 2 that there is a significant improvement of classification performance in some classes (e.g.*Stressed Grass*, *Water*) after accounting for unlabeled samples, particularly between SMA and S-SMA as well as CoSpace and S-CoSpace. However, these aforementioned semi-supervised methods carry out the label propagation on a given graph manually computed by gaussian kernel function, limiting the adaptiveness and discriminability of the algorithms. LeMA can adaptively learn a data-driven graph structure where the labels tend to spread

[2] The connectivity in manifold alignment is not strictly equivalent to the similarity of the two samples.

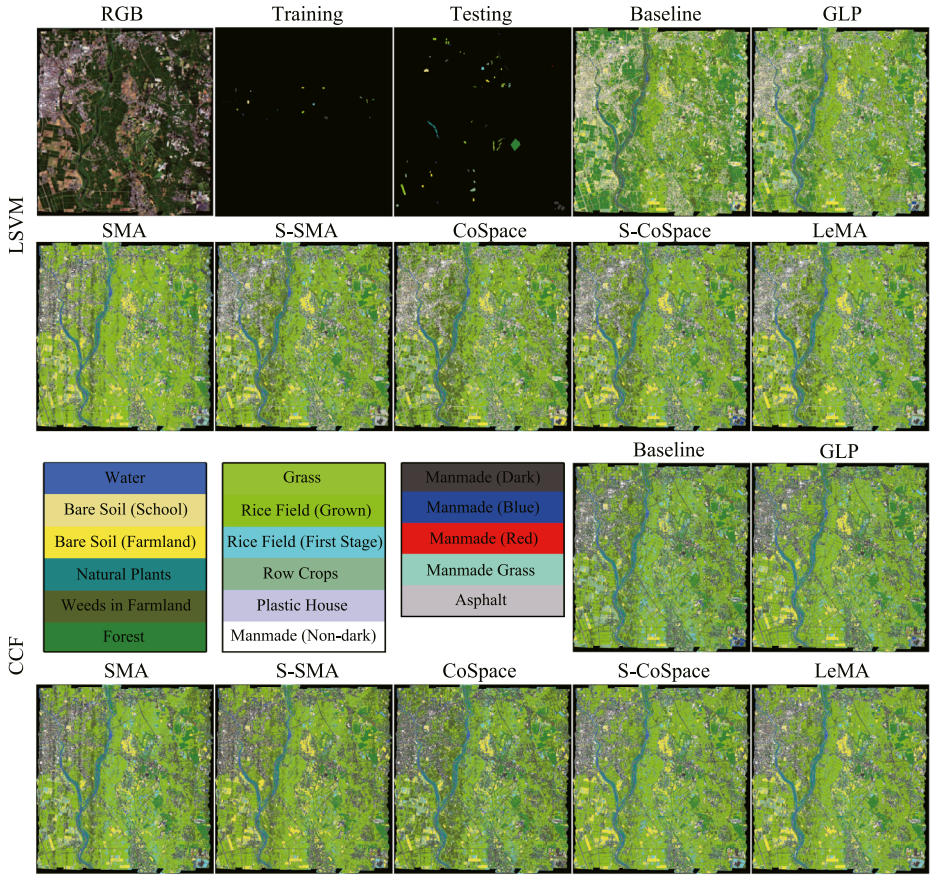


Fig. 7. Classification maps of the different algorithms obtained using two kinds of classifiers on the Chikusei dataset.

more smoothly, which can result in a more effective material identification for those challenging classes (few training samples), such as *Trees*, *Residential*, *Railway*, *Parking Lot1*. In addition, we can also observe an easily overlooked phenomenon that the LeMA's ability in identifying certain classes still remains limited, such as *Parking Lot2* (only 1.78%) and *Railway* (49.96%). *Parking Lot2* is basically classified to *Commercial* and *Parking Lot1*, while *Railway* is largely identified as *Road* and *Commercial*. This might be explained by the limited number of training samples as well as fairly similar spectral properties between several classes.

Table 3
Quantitative performance comparison with the different algorithms on the Chikusei data. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	d		(k, σ, d)		d		(k, σ, d)		(α, β, d)		(α, β, d)		(α, β, d)	
	10		(10, 1, 10)		20		(10, 0.1, 20)		(0.1, 0.01, 30)		(0.1, 0.01, 30)		(0.1, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	60.20	71.11	62.30	72.26	67.90	71.53	69.68	73.27	71.12	75.69	72.60	77.11	75.11	**81.71**
AA	69.42	70.40	69.80	70.71	70.79	66.47	72.27	70.01	73.96	71.46	71.64	71.33	75.29	**75.73**
κ	0.5523	0.6761	0.5784	0.6894	0.6391	0.6802	0.6602	0.6818	0.6746	0.7260	0.6911	0.7420	0.7194	**0.7933**
Class1	78.21	80.54	78.09	80.42	98.72	82.52	**99.53**	97.90	92.54	79.25	98.83	98.37	98.25	98.83
Class2	94.43	82.70	94.11	93.84	93.20	92.50	93.20	93.09	93.47	**94.91**	87.04	93.63	93.20	93.79
Class3	23.54	50.06	37.75	76.87	62.57	55.31	68.41	76.55	80.40	77.71	80.65	77.23	89.29	**89.90**
Class4	92.13	92.56	92.23	95.72	90.57	91.53	92.51	88.76	90.59	96.23	94.64	92.49	95.11	**96.96**
Class5	**97.65**	94.68	96.84	88.45	28.43	16.06	24.01	32.85	83.94	66.52	51.81	43.32	60.74	67.78
Class6	62.01	81.48	57.47	69.67	62.52	78.91	68.27	79.67	63.61	79.02	72.34	**88.48**	76.34	87.27
Class7	99.67	99.93	99.66	**100.00**	96.87	97.79	95.40	99.37	97.74	99.75	98.41	99.87	97.63	99.80
Class8	57.11	93.40	69.06	98.93	95.59	93.49	96.88	96.53	95.05	92.72	**99.48**	98.45	99.27	99.18
Class9	**100.00**	**100.00**	**100.00**	99.92	99.53	99.13	99.45	99.21	98.66	99.76	99.21	98.34	99.76	**100.00**
Class10	24.81	19.56	**26.64**	19.06	21.39	15.48	20.94	13.09	22.35	18.00	22.75	14.83	26.47	26.46
Class11	0.00	2.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	5.47	0.63	**5.68**
Class12	**90.32**	88.91	**90.32**	89.61	90.14	85.92	90.14	89.44	90.32	80.46	89.96	89.44	88.38	90.14
Class13	33.11	33.09	33.11	36.50	32.61	56.25	31.32	30.88	33.11	67.90	33.11	54.93	33.11	**68.73**
Class14	**94.20**	85.38	79.12	59.40	72.85	59.40	**94.20**	86.31	59.40	52.44	14.39	49.19	45.01	53.60
Class15	**100.00**	**100.00**	**100.00**	**100.00**	93.58	**100.00**	**100.00**	**100.00**	93.58	97.86	**100.00**	**100.00**	**100.00**	**100.00**
Class16	74.88	88.62	74.19	93.52	99.71	99.51	99.80	98.82	97.84	**100.00**	97.35	97.25	98.04	95.78
Class17	58.03	3.84	58.03	0.24	65.23	7.91	62.11	7.67	64.75	0.00	77.70	11.27	**78.66**	13.43

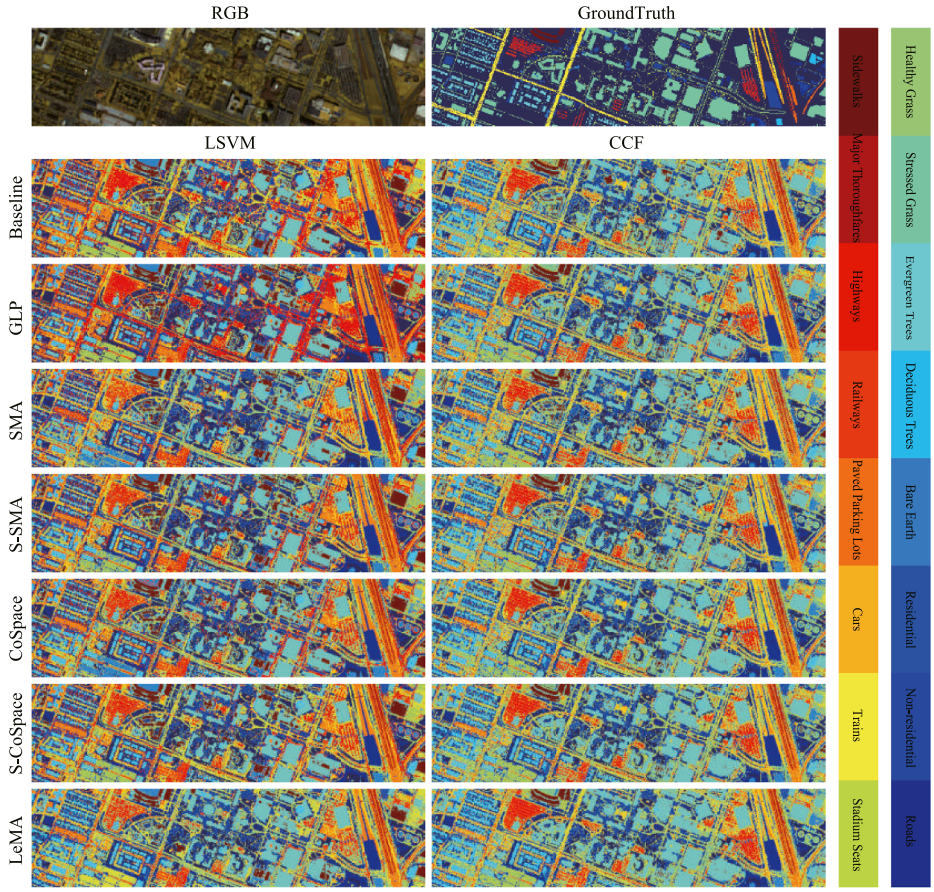


Fig. 8. Classification maps of the different algorithms obtained using two kinds of classifiers on the real dataset of DFC2018 (Multispectral-Lidar and Hyperspectral data).

3.2. The simulated MS-HS datasets over Chikusei

3.2.1. Data description

Similarly to Houston data, the MS data with dimensions of $2517 \times 2335 \times 10$ at a GSD of 2.5 m was simulated by the HS data acquired by the Headwall's Hyperspec-VNIR-C sensor over Chikusei area, Ibaraki, Japan. It consists of 128 bands in the spectral range from 363 nm to 1018 nm with the 10 nm spectral resolution. The dataset has been made available to the scientific research (Yokoya and Iwasaki, 2016).

3.2.2. Experimental setup

Fig. 6 shows the corresponding MS and partial HS images as well as selected training labels and test labels. Again, the overlapped region between MS and HS, which should include all the classes listed in Table 1, is chosen based on the given ground truth (Yokoya and Iwasaki, 2016). Additionally, the parameters configuration for all algorithms can be adaptively completed by a 10-fold cross-validation on the training set, which is more generalized to different datasets. Regarding how to run the cross-validation for parameters setting, please refer to Section 3.1.2 for more details.

3.2.3. Results and analysis

We assess the classification performance of the different algorithms for the Chikusei MS-HS data both quantitatively and visually, as shown in Fig. 7 and Table 3.

Similarly to the University of Houston MS-HS data, there is a basically consistent trend for the different algorithms in the Chikusei MS-HS data. On the whole, the original MS data (baseline) fails to identify some specific materials such as *Plastic House*, *Manmade (Dark)*, *Rice Field (Grown)*, *Bare Soil (Farmland)*, and *Forest*, due to its poor spectral information and a limited number of training samples. GLP utilizes the unlabeled samples to augment the training samples in a semi-supervised way, yet it is still limited by the low-discriminative spectral signatures. By aligning the MS and HS data, these alignment-based approaches (e.g. SMA, S-SMA, CoSpace, S-CoSpace, and LeMA) are able to find a common subspace in which the learnt features are expected to absorb the different properties from two modalities, resulting in a better performance. Compared to the supervised methods (SMA and CoSpace), their corresponding semi-supervised versions (S-SMA and S-CoSpace) obtain higher classification accuracies on both classifiers, which is detailed in Table 3. As expected, the performance of the LeMA is significantly superior to that of others, thanks to the great contributions of a common subspace learning from MS-HS data, a data-driven graph learning and the semi-supervised learning strategy. Despite so, the LeMA still fails to recognize some challenging classes, such as *Weeds in Farmland*, *Row Crops*, *Plastic House*, and *Asphalt*. The reasons could be twofold. On one hand, the performance of LeMA is limited, to some extent, by the unbalanced data sets. On the other hand, LeMA' transferring ability would sharply degrade when a great spectral variability between training and test samples exists.

3.3. The real multispectral-lidar and hyperspectral datasets in DFC2018

Although we follow strict simulation procedures, yet the two MS-HS datasets used above (Houston and Chikusei) essentially originate from a similar data source (homogeneous), which means there is a strong correlation in their spectral features. This makes the information of the different modalities transferred more effectively, but could limit the generalization ability in practice. To this end, we apply a real bi-modal dataset – multispectral-lidar and hyperspectral (heterogeneous) provided by the latest IEEE GRSS data fusion contest 2018 (DFC2018).

Table 4
Quantitative performance comparison with the different algorithms on the DFC2018 data. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	d		(k, σ, d)		d		(k, σ, d)		(α, β, d)		(α, β, d)		(α, β, d)	
	7		(10, 1, 7)		30		(10, 1, 30)		(0.1, 0.1, 30)		(0.1, 0.01, 30)		(0.1, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	51.35	72.84	52.28	73.15	52.73	70.37	54.69	72.13	55.56	74.04	58.65	76.59	61.69	**79.98**
AA	59.46	78.64	60.57	81.64	58.06	77.78	65.34	78.72	66.16	80.46	67.72	83.67	65.54	**88.82**
κ	0.4194	0.6534	0.4289	0.6587	0.4366	0.6256	0.4598	0.6441	0.4670	0.6682	0.4987	0.6990	0.5284	**0.7414**
Class1	91.70	84.62	96.15	93.12	84.01	85.43	94.13	90.89	95.14	89.07	94.74	95.14	92.31	**100.00**
Class2	33.90	80.17	35.62	80.74	73.00	82.40	69.57	80.17	61.32	80.37	69.73	81.52	78.09	**87.90**
Class3	94.92	96.16	96.02	96.57	95.06	95.06	96.30	96.30	93.83	97.26	94.79	96.30	96.57	**99.45**
Class4	83.00	92.50	85.50	97.50	85.50	90.00	84.50	94.00	83.00	91.00	85.50	98.00	79.00	**100.00**
Class5	43.71	90.42	30.54	87.43	53.29	87.43	52.10	85.03	61.08	92.22	45.51	92.22	30.54	**100.00**
Class6	80.44	90.60	81.32	91.82	78.79	87.77	82.80	87.98	83.94	90.35	85.24	91.27	89.71	**96.50**
Class7	59.26	82.01	61.11	81.52	57.62	78.21	58.66	82.45	59.89	82.37	63.95	85.14	69.56	**87.47**
Class8	14.07	31.98	10.75	36.00	21.71	28.00	20.83	35.16	26.64	38.71	11.77	39.51	31.43	**49.96**
Class9	48.54	54.14	50.77	58.40	44.87	56.96	52.60	53.49	47.94	63.30	53.69	**68.55**	40.47	62.26
Class10	10.16	42.07	8.00	31.70	6.77	37.82	5.55	29.21	11.02	36.67	24.21	**38.40**	12.93	38.04
Class11	23.54	72.03	25.96	79.07	79.07	74.45	45.88	75.45	34.21	76.26	54.12	81.49	62.58	**100.00**
Class12	93.85	85.85	92.92	94.46	92.00	87.08	85.85	90.15	85.54	86.15	74.15	95.38	66.46	**100.00**
Class13	60.50	74.96	57.31	87.56	59.33	73.45	60.17	77.98	63.03	79.33	64.71	87.06	70.59	**99.83**
Class14	39.93	87.15	55.21	90.63	17.71	86.11	47.22	85.76	66.32	89.58	75.69	90.63	55.21	**99.65**
Class15	95.39	96.77	97.70	**100.00**	93.55	98.16	99.54	97.70	99.54	98.62	99.54	**100.00**	95.85	**100.00**
Class16	78.39	96.77	84.19	99.68	77.74	96.13	89.68	97.74	86.13	96.13	86.13	98.06	77.42	**100.00**

3.3.1. Data description

Multi-source optical remote sensing data, such as multispectral-lidar data, hyperspectral data, and very high-resolution RGB data, is provided in the contest. More specifically, the multispectral-lidar imagery consists of 1202×4768 pixels with 7 bands (3 intensity bands and 4 DSMs-related bands (Saux et al., 2018)) collected from 1550 nm, 1064 nm, and 532 nm at a 0.5 m GSD, while the hyperspectral data comprises 48 bands covering a spectral range from 380 nm to 1050 nm at 1 m GSD, and its size is 601×2384. In our case, our LeMA model is trained on partial multispectral-lidar and hyperspectral correspondences and tested only using multispectral-lidar data, in order to meet the requirement of our cross-modality learning task. The first row of Fig. 8 shows the RGB image of this scene and the labeled ground truth image.

3.3.2. Experimental setup

Our aim is, once again, to investigate whether the limited amount of hyperspectral data can improve the performance of another modality, e.g., multispectral data (homogeneous) or multispectral-lidar data (heterogeneous). Therefore, we randomly assign 10% of total labeled samples as training set and the rest of it as test set in the experiment. Moreover, 16 main classes are selected out of 20 (see Fig. 8), by removing several small classes with too few samples, e.g. *Artificial Turf*, *Water*, *Crosswalks*, and *Unpaved Parking Lots*. Likewise, we automatically configure the parameters of the proposed LeMA and the compared algorithms by a 10-fold cross-validation on the training set, which is detailed in Section 3.1.2.

3.3.3. Results and analysis

We show the averaged results of the different algorithms out of 10 runs to obtain a relatively stable and meaningful performance comparison, because the training and test sets are randomly generated from total samples in each round, as listed in Table 4. Correspondingly, Fig. 8 visually highlights the differences of classification maps for the different methods.

Generally speaking, hyperspectral information embedding can effectively improve the classification performance of the multispectral-lidar data, which implies that the models based common subspace learning (e.g., SMA, S-SMA, CoSpace, S-CoSpace, and LeMA) can transfer the knowledge from one modality to another modality to some extent. We also observe from Table 4 that the semi-supervised methods which consider the unlabeled samples (e.g., GLP, S-SMA, S-CoSpace, and LeMA) always perform better than those purely supervised ones. Not unexpectedly, LeMA integrating rich spectral information and unlabeled samples achieves a superior performance, which demonstrates that the learning-based graph structure is more applicable to capturing the data distribution and further find a potential optimal decision boundary.

One thing to be noted, however, is that compared to the performance of the different algorithms in the simulated MS-HS datasets from similar sources (homogeneous), the knowledge transferring ability of these algorithms in handling the real multispectral-lidar and hyperspectral datasets from different sources (heterogeneous) remains limited, since all listed methods including our LeMA are modeled in a linearized way. Unfortunately, a single linear transformation fails to fit the gap between heterogeneous modalities well, despite a limited performance improvement.

4. Conclusions

In real-world problems, a large amount of low-quality data (e.g. MS data) can often be easily collected. On the contrary, high-quality data (e.g. HS data) are usually expensive and difficult to obtain. This motivates us to investigate whether a limited amount of high-quality data can contribute to relevant tasks with a large amount of low-quality data. For this purpose, we propose a novel semi-supervised learning framework called LeMA, which effectively connects the common subspace and label information, and automatically embeds the unlabeled information into the proposed framework by adaptively learning a Laplacian matrix from the data. Extensive experiments are conducted using the LeMA on two homologous MS-HS simulated datasets and a heterogenous multispectral-lidar and hyperspectral real dataset in comparison with the other state-of-arts algorithms, demonstrating the superiority and effectiveness of the LeMA in the knowledge transferring ability. We have to admit, however, that despite a significant performance improvement in LeMA, yet its representative ability is still

limited by linearly modeling way, especially facing highly-nonlinear heterogenous data. Towards this issue, we will continue to improve our model to a nonlinear version and simultaneously consider the spatial information (e.g., morphological profiles) to further strengthen the feature representation ability.

Acknowledgements

The authors would like to thank the Hyperspectral Image Analysis group and the NSF Funded Center for Airborne Laser Mapping (NCALM) at the University of Houston for providing the CASI University of Houston dataset. The authors would like to express their appreciation to Prof. D. Cai and Dr. C. Wang for providing MATLAB codes for LPP and manifold alignment algorithms.

This work was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No [ERC-2016-StG-714087]) and from Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de). The work of N. Yokoya was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI 15K20955 and Alexander von Humboldt Fellowship for postdoctoral researchers.

References

Banerjee, B., Bovolo, F., Bhattacharya, A., Bruzzone, L., Chaudhuri, S., Buddhiraju, K.M., 2015. A novel graph-matching-based approach for domain adaptation in classification of remote sensing image pair. IEEE Trans. Geosci. Remote Sens. 53 (7), 4045–4062.

Bertsekas, D.P., 1999. Nonlinear Programming. Athena Scientific Belmont.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. 3 (1), 1–122.

Bruzzone, L., Marconcini, M., 2010. Domain adaptation problems: a dasvm classification technique and a circular validation strategy. IEEE Trans. Pattern Anal. Mach. Intell. 32 (5), 770–787.

Chen, L., Li, X., Sun, D., Toh, K., 2018. On the equivalence of inexact proximal alm and admm for a class of convex composite programming. arXiv preprint arXiv:1803.10803.

Gu, Q., Li, Z., Han, J., 2011. Joint feature selection and subspace learning. In: Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1294–1299.

Heide, F., Heidrich, W., Wetzstein, G., 2015. Fast and flexible convolutional sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5135–5143.

Hong, D., Yokoya, N., Zhu, X., 2016. The k-lle algorithm for nonlinear dimensionality reduction of large-scale hyperspectral data. In: Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2016 8th Workshop on. IEEE, pp. 1–5.

Hong, D., Yokoya, N., Chanussot, J., Zhu, X., 2017. Learning low-coherence dictionary to address spectral variability for hyperspectral unmixing. In: Proceedings of IEEE International Conference on Image Processing (ICIP), pp. 1–5.

Hong, D., Yokoya, N., Zhu, X., 2017. Learning a robust local manifold representation for hyperspectral dimensionality reduction. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 10 (6), 2960–2975.

Hong, D., Yokoya, N., Xu, J., Zhu, X., 2018. Joint and progressive learning from high-dimensional data for multi-label classification. In: European Conference on Computer Vision (ECCV). Springer, pp. 478–493.

Huang, X., Lu, Q., Zhang, L., 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. ISPRS J. Photogrammetry Remote Sens. 90, 36–48.

Ji, S., Ye, J., 2009. Linear dimensionality reduction for multi-label classification. In: Proceedings of the 21th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1077–1082.

Jiang, J., Zhai, X., 2007. Instance weighting for domain adaptation in nlp. In: Proceedings of ACL, pp. 264–271.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X., 2018. Building instance classification using street view images. ISPRS J. Photogrammetry Remote Sens.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A., Torralba, A., 2012. Undoing the damage of dataset bias. In: European Conference on Computer Vision (ECCV). Springer, pp. 158–171.

Liao, D., Qian, Y., Zhou, J., Tang, Y., 2016. A manifold alignment approach for hyperspectral image visualization with natural color. IEEE Trans. Geosci. Remote Sens. 54 (6), 3151–3162.

Li, J., Zhang, H., Zhang, L., 2014. Column-generation kernel nonlocal joint collaborative representation for hyperspectral image classification. ISPRS J. Photogrammetry Remote Sens. 94, 25–36.

Lin, Z., Chen, M., Ma, Y., 2010. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y., 2013. Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 171–184.

Matasci, G., Volpi, M., Kanevski, M., Bruzzone, L., Tuia, D., 2015. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 53 (7), 3550–3564.

Persello, C.C., Bruzzone, L., 2012. Active learning for domain adaptation in the supervised classification of remote sensing images. IEEE Trans. Geosci. Remote Sens. 50 (11), 4468–4483.

Samat, A., Gamba, P., Abuduwaili, J., Liu, S., Miao, Z., 2016. Geodesic flow kernel support vector machine for hyperspectral image classification by unsupervised subspace feature transfer. Remote Sens. 8 (3), 234.

Samat, A., Gamba, P., Liu, S., Du, P., Abuduwaili, J., 2016. Jointly informative and manifold structure representative sampling based active learning for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 54 (11), 6803–6817.

Samat, A., Persello, C., Gamba, P., Liu, S., Abuduwaili, J., Li, E., 2017. Supervised and semi-supervised multi-view canonical correlation analysis ensemble for heterogeneous domain adaptation in remote sensing image classification. Remote Sens. 9 (4), 337.

Saux, B.L., Yokoya, N., Hansch, R., Prasad, S., 2018. 2018 IEEE GRSS data fusion contest: multimodal land use classification [technical committees]. IEEE Geosci. Remote Sens. Mag. 6 (1), 52–54.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., Kawanabe, M., 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In: Advances in Neural Information Processing Systems (NIPS), pp. 1433–1440.

Tarabalka, Y., Benediktsson, J., Chanussot, J., 2009. Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques. IEEE Trans. Geosci. Remote Sens. 47 (8), 2973–2987.

Tom, R., Frank, W., 2015. Canonical correlation forests. arXiv preprint arXiv:1507.05444.

Tuia, D., Volpi, M., Trolliet, M., Camps-Valls, G., 2014. Semisupervised manifold alignment of multimodal remote sensing images. IEEE Trans. Geosci. Remote Sens. 52 (12), 7708–7720.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: an overview of recent advances. IEEE Geosci. Remote Sens. Mag. 4 (2), 41–57.

Tuia, D., Marcos, D., Camps-Valls, G., 2016. Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization. ISPRS J. Photogrammetry Remote Sens. 120, 1–12.

Van der Meer, F.D., Van der Werff, H.M.A., Van Ruitenbeek, F.J.A., 2014. Potential of esa's sentinel-2 for geological applications. Remote Sens. Environ. 148, 124–133.

Wang, C., Mahadevan, S., 2009. A general framework for manifold alignment. In: AAAI Fall Symposium on Manifold Learning and its Applications (AAAI).

Wang, C., Mahadevan, S., 2011. Heterogeneous domain adaptation using manifold alignment. In: Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1541–1546.

Wang, C., Krafft, P., Mahadevan, S., 2011. Chapter of Manifold Learning: Theory and Applications-Manifold alignment. CSC Press.

Woodcock, C., Macomber, S.A., Pax-Lenney, M., Cohen, W.B., 2001. Monitoring large areas for forest change using landsat: generalization across space, time and landsat sensors. Remote Sens. Environ. 78 (1–2), 194–203.

Xia, J., Chanussot, J., Du, P., He, X., 2014. Semi-supervised probabilistic principal component analysis for hyperspectral remote sensing image classification. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 7 (6), 2224–2236.

Yang, C., Everitt, J.H., Du, Q., Luo, B., Chanussot, J., 2013. Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture. Proc. IEEE 101 (3), 582–592.

Yokoya, N., Iwasaki, A., 2016. Airborne hyperspectral data over chikusei. Tech. Rep. SAL-2016-05-27.

Yokoya, N., Grohnfeldt, C., Chanussot, J., 2017. Hyperspectral and multispectral data fusion: a comparative review. IEEE Geosci. Remote Sens. Mag. 5 (2), 29–56.

Zhang, L., Zhang, L., Tao, D., Huang, X., 2012. On combining multiple features for hyperspectral remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 50 (3), 879–893.

Zhong, Y., Wang, X., Zhao, L., Feng, R., Zhang, L., Xu, Y., 2016. Blind spectral unmixing based on sparse component analysis for hyperspectral remote sensing imagery. ISPRS J. Photogrammetry Remote Sens. 119, 49–63.

Zhou, P., Zhang, C., Lin, Z., 2017. Bilevel model based discriminative dictionary learning for recognition. IEEE Trans. Image Process. 26 (3), 1173–1187.

Zhu, X., Ghahramani, Z., Lafferty, J.D., 2003. Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning (ICML), pp. 912–919.