THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

Orthogonal Constrained Minimization with Tensor Group Sparsity Regularization for Hyperspectral Image Restoration

Shijie Yu

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Applied Mathematics

Orthogonal Constrained Minimization with Tensor Group Sparsity Regularization for Hyperspectral Image Restoration

Shijie Yu

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

December 2024

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

YU Shijie (Name of student)

Dedicated to my family.

Abstract

Hyperspectral images (HSIs) captured by hyperspectral sensors often suffer from noise, blurring, and other degradations, which can significantly reduce their visual quality and the accuracy of the subsequent tasks. Traditional HSI restoration methods typically process the spatial information of HSIs on a band-by-band basis, which neglects the spectral information inherent. Also, those methods often use elementwise sparsity measures to characterize sparse components, which fail to recognize the linear structures within these components. This thesis aims to develop new approaches based on (nonlocal) low-rank tensor regularization and tensor group sparsity $\ell_{2,p}$ norm $(0 < p < 1)$, along with some spatial and spectral priors, to provide a more comprehensive method that preserves the structure of HSIs. It includes two optimization models and two algorithms for solving two important problems in HSI processing.

Firstly, we present a class of orthogonal constrained minimization problems to tackle HSI restoration problems, such as removing mixed noise like Gaussian noise, stripes, and dead lines. The proposed class of models employs two types of regularization terms. One is a tensor group sparsity regularization term for removing structured noise. We use the tensor $\ell_{2,p}$ norm, extended from the matrix $\ell_{2,p}$ norm, and provide a solution for the proximal operator of the tensor $\ell_{2,p}$ norm. The other term is a new sparsity-enhanced nonlocal low-rank tensor regularization for removing Gaussian noise. This regularization term exploits the spatial nonlocal self-similarity

and spectral correlation in HSIs to enhance restoration, ensuring that similar patterns in distant regions are jointly considered for improved denoising. Specifically, we propose a weighted tensor $\ell_{2,p}$ norm to enhance sparsity in the core tensor, promoting low-rankness in nonlocal similar matching blocks.

Secondly, we adopt a proximal block coordinate descent (P-BCD) algorithm to solve the proposed nonconvex nonsmooth minimization with orthogonal constraints. The solution to each subproblem in the P-BCD algorithm can be efficiently computed. The first order optimality condition of the problem is defined by substationarity, symmetry, and feasibility. We prove that any accumulation point of the generated sequence by the P-BCD algorithm is a first order stationary point.

Thirdly, we apply the proposed approach to HSI denoising and destriping, and conduct numerical experiments to validate the superiority of our proposed approach. We test it on simulated noisy HSIs generated from several datasets under various mixed noise conditions, as well as on a real dataset. The results demonstrate that our method outperforms others in metrics such as mean peak signal-to-noise ratio. In terms of visual quality, our method effectively restores HSIs by preserving important image details and removing noise, particularly highly structured noise like stripes and dead lines.

Lastly, we combine the proposed model with a deep neural network to incorporate an implicit proximal denoiser prior. Specifically, for detecting anomaly objections in noisy HSIs, the tensor $\ell_{2,p}$ norm in the original model is utilized to characterize the anomalies, while the implicit proximal denoiser prior is employed to remove Gaussian noise. The P-BCD method remains effective for solving the newly proposed model, with certain steps updated using a proximal denoiser within a plug-and-play (PnP) framework. We evaluate this PnP version of the P-BCD method (PnP-PBCD) on anomaly detection in HSI contaminated with or without Gaussian noise. The results demonstrate that the proposed method can effectively detect anomalous

objects, whereas competing methods may mistakenly identify noise as anomalies or incorrectly match the anomalous objects due to noise interference.

In summary, the orthogonal constrained minimization models with tensor group sparsity regularization are well-suited for various image restoration problems. Additionally, the P-BCD method and its PnP version are reliable with convergence guarantees.

This thesis contains some research results of the following paper during the period of my PhD study.

- F. He, R. K. M. Chun, Z. Qiu, S. Yu, Y. Shi, C. H. To, and X. Chen, *Choroid segmentation of retinal OCT images based on CNN classifier and l2-lq fitter*, Computational and Mathematical Methods in Medicine, (2021), p. 8882801.

- X. Liu, S. Yu, J. Lu, and X. Chen, *Orthogonal constrained minimization with tensor $\ell_{2,p}$ regularization for HSI denoising and destriping*, submitted to SIAM Journal on Imaging Sciences (under review), arXiv preprint arXiv:2407.03605, (2024).

- X. Liu and S. Yu, *Provably convergent plug-and-play proximal block coordinate descent method for hyperspectral anomaly detection*, submitted to International Journal of Numerical Analysis and Modeling (under review), arXiv preprint arXiv:2412.14824, (2024).

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Xiaojun Chen, for her invaluable guidance, support, and encouragement throughout my research journey. Her profound expertise, insightful advice, and unwavering patience have been instrumental in shaping this thesis. I am also sincerely grateful to my co-supervisor, Professor Jian Lu, for his constructive feedback, support, and inspiring discussions, which have significantly contributed to the progress of this work.

I extend my heartfelt appreciation to Dr. Xiaoxia Liu, whose mentorship and expertise have played a pivotal role in my research. Her advice and encouragement have been a constant source of motivation, especially during challenging moments. I am equally thankful to Dr. Chao Li for her support and insightful discussions, which have greatly enriched this work.

I am deeply indebted to my colleagues and friends in the Department of Applied Mathematics (AMA) at The Hong Kong Polytechnic University and our collaborators at Shenzhen University for their support and collaboration. Their contributions have made this research a fulfilling and enjoyable experience.

Lastly, I would like to thank my family and friends for their unwavering support, patience, and encouragement throughout my academic journey. Their belief in me has been a cornerstone of my motivation and perseverance. This work is dedicated to all those who have guided and supported me along the way.

Contents

List of Figures

List of Tables

List of Notation

$\mathbb{R}$	The set of real numbers
$\mathbb{R}^{I_1}$	The set of I_1-dimensional real vectors
$\mathbb{R}^{I_1 \times I_2}$	The set of $I_1 \times I_2$ real matrices
$\mathbb{R}^{I_1 \times I_2 \times I_3}$	The set of $I_1 \times I_2 \times I_3$ real tensors
$\mathbb{R}_+$	The set of nonnegative real numbers
$\mathbb{R}_{++}$	The set of positive real numbers
$\mathbb{N}_+$	The set of positive integers
$\mathbb{S}_{m,n}$	The Stiefel manifold, i.e., the set of $m \times n$ matrices with orthonormal columns
$\odot$	The component-wise multiplication
$\langle \cdot, \cdot \rangle$	The inner product
$\times_n$	The mode-n product of a tensor with a matrix
$\| \cdot \|$	The ℓ_2 norm for vectors
$\| \cdot \|_F$	The Frobenius norm for matrices or tensors
$\| \cdot \|_{2,p}$	The tensor $\ell_{2,p}$ (quasi-)norm with $p \in (0, 1)$ of a third order tensor
$\| \cdot \|_{1,w}$	The weighted tensor (component-wise) ℓ_1 norm
$a^\top, A^\top$	The transpose of vector a or matrix A
$\mathrm{Diag}(\cdot)$	The diagonal matrix of a given vector

$[\,\cdot\,]$	The stack of independent matrices or independent third order tensors
$[\,\cdot\,]^{(j)}$	The j-th independent matrix or tensor in a stack $[\,\cdot\,]$
$\mathbf{I}_{n_i}$	The identity matrix of size $n_i \times n_i$
Id	The identity map on $\mathbb{R}^{I_1 \times I_2 \times I_3}$
$\mathbf{R}(\cdot)$	The similar blocks extraction operator
$\mathcal{A}_{(i)}$	The mode-i unfolding of a tensor $\mathcal{A}$

Chapter 1

Introduction

Hyperspectral images (HSIs) are collected by hyperspectral sensors across the electromagnetic spectrum. For a three-dimensional (3-D) HSI, the first two dimensions represent spatial information, and the third dimension represents the spectral information of a scene. An illustration of an HSI is shown in Figure 1.1. HSIs are widely used for various applications [64, 71, 65, 108] such as object detection [98], material identification [10, 27], etc.

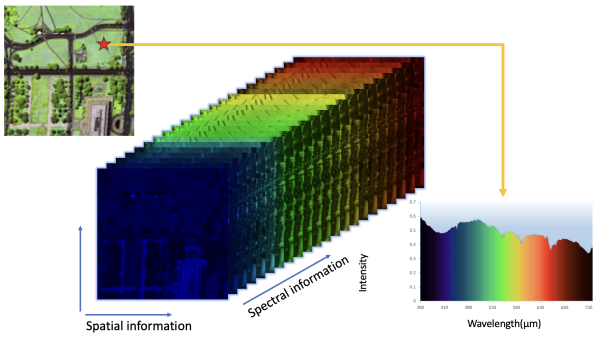


Figure 1.1: An illustrative figure of an HSI.

HSIs are often contaminated by various types of noise during acquisition, which can severely degrade the quality and reliability of the extracted spectral information. Common noise types include Gaussian noise, impulse noise, and stripe noise, which may result from atmospheric interference, sensor limitations, or operational conditions. Such mixed noise presents a challenge in retaining the spatial and spec-

tral integrity of HSI data, necessitating advanced denoising techniques to effectively recover the underlying true signal.

In addition to denoising, anomaly detection is another critical task in HSIs, particularly for identifying objects or materials that deviate from the typical background. Anomalies might represent rare minerals in geological surveys, contaminants in agricultural products, or specific targets in defense applications. Detecting these anomalies is complicated by the high-dimensional nature of HSI data and the potential for similar anomalies to be masked by noise. Effective anomaly detection thus requires methods capable of isolating distinctive spectral signatures without being misled by noise or common background features.

Both denoising and anomaly detection share overlapping challenges, as they aim to preserve essential spectral-spatial structures while suppressing noise and distinguishing subtle anomalies. This duality necessitates robust algorithms that leverage spatial and spectral redundancies in HSI data. The development of models that can simultaneously address both noise and anomaly detection issues is key to advancing the field of hyperspectral imaging and maximizing its applicability in real-world scenarios.

1.1 HSI Denoising and Destriping

HSI denoising and destriping have become an essential step for tasks in HSI processing. Many methods have been proposed to remove Gaussian noise in HSIs. Conventional 2-D methods [15, 21], processing HSIs band by band, do not fully utilize the strong correlation between adjacent bands. 3-D methods such as block-matching 4-D filtering (BM4D) [55], spectral-spatial adaptive hyperspectral total variation (SSAHTV) model [99], and sparse representation methods [107, 92] incorporate both spatial and spectral information and outperform conventional methods.

However, those methods may fail to remove non-Gaussian noise.

In real-world scenarios, HSIs are often contaminated by more than one type of noise due to atmospheric effects and instrument noise. Various methods have been proposed to remove the mixed noise, including low-rank matrix based methods, low-rank tensor based methods, and deep learning based methods. Low-rank matrix based methods [29, 102] reshape an HSI into a matrix and impose low-rankness on the reshaped HSI. Zhang et al. [103] formulated the HSI denoising problem as a low-rank matrix factorization problem and solved it by the "Go Decomposition" algorithm; Zhang et al. [102] proposed a double low-rank matrix decomposition which utilizes the ℓ_1 norm for the impulse noise and the matrix nuclear norm for stripes, and adopted augmented Lagrangian method (ALM) to solve the model; Yang et al. [97] also used double low-rankness, but added spatial-spectral total variation (SSTV) to the model and performed the decomposition on full band blocks (FBBs) rather than the entire HSI.

Low-rank tensor based methods [47, 18, 61, 26, 108] view HSIs as tensors and perform tensor low-rank decompositions while preserving the spatial-spectral correlations. Wang et al. [83] used Tucker tensor decomposition and an anisotropic SSTV regularization to characterize the piece-wise smooth structures of the HSI; Chen et al. [14] proposed a low-rank tensor decomposition (LRTD) method, which utilized the higher-order singular value decomposition (HOSVD) for low-rankness and the $\ell_{2,1}$ norm for characterizing the stripes, and adopted ALM for solving the optimization model; Cao et al. [9] proposed a subspace-based nonlocal low-rank and sparse factorization (SNLRSF) method for removing mixed noise in HSI, which conducted nonlocal low-rank factorization via successive singular value decomposition (SVD); Xiong et al. [93] proposed the LRTFL0 method using low-rank block term decomposition and spectral-spatial ℓ_0 gradient regularization to achieve gradient smoothness.

Recently, some deep neural networks [7, 100, 86] have been proposed to denoise

HSIs. Quasi-recurrent neural networks (QRNN) [7] combine recurrent neural networks (RNN) with convolutional neural networks (CNN); the 3-D version of QRNN (QRNN3D) [86] can effectively embed the correlation of HSIs; CNN can also be used as a denoiser in a plug-and-play (PnP) fashion for HSI denoising [74].

1.2 Hyperspectral Anomaly Detection

Hyperspectral anomaly detection aims to identify pixels or regions in HSIs that significantly differ from the surrounding background without prior knowledge of the target spectral information. These pixels, often referred to as anomalies, could represent objects or materials such as aircraft, ships, vehicles, or other structures that deviate from the natural background. Detecting such anomalies is crucial due to their significance in various applications. For example, in environmental monitoring, anomalies may indicate areas affected by pollution or disease in vegetation [63]; in the food industry, anomalies may be detected for quality control by identifying physical defects and inconsistencies in products [88]. By leveraging the rich spectral information provided by HSIs, the accuracy and reliability of anomaly detection can be enhanced, thereby improving decision-making processes in fields such as security, agriculture, and resource management.

In hyperspectral anomaly detection, the Reed-Xiaoli (RX) method, introduced by Reed and Xiaoli in 1990 [67], is a foundational method known for its simplicity and widespread adoption. The RX method assumes that background spectral features follow a multivariate Gaussian distribution and identifies anomalies by calculating the Mahalanobis distance from the background. Over time, RX has inspired several variants to address its limitations in real-world applications. For example, the local RX method [58] enhances localized anomaly detection using sliding windows for background estimation; the kernel RX method [45] maps data into

high-dimensional feature spaces to better adapt to nonlinear distributions; and the weighted RX method [30] introduces pixel-level weighting for improving robustness against noise. While RX and its variants are computationally efficient and serve as benchmarks in the field, they often rely on Gaussian assumptions and are sensitive to noise and outliers, limiting their performance in complex scenes.

In contrast to statistical approaches like RX, representation-based methods focus on explicitly modeling the structure of HSIs without assuming a predefined distribution. Li et al. [48] proposed the background joint sparse representation detection (BJSRD) method, which reconstructs each background pixel using a sparse set of coefficients from a dictionary. Xu et al. [94] introduced the low-rank and sparse representation (LRASR) method, which models the background as a low-rank component while representing anomalies as sparse components. Feng et al. [24] developed the local spatial constraint and total variation (LSC-TV) method, which combines low-rank modeling with superpixel segmentation and total variation (TV) regularization to effectively separate anomalies in complex scenes. To preserve the intrinsic 3D structure of HSIs, the low-rank component is characterized using tenor low-rank representation. For example, the tensor low-rank and sparse representation (TLRSR) method [81] utilizes the tensor singular value decomposition (t-SVD), while the method proposed in [23] employs the tensor ring decomposition.

Deep learning methods have significantly improved hyperspectral anomaly detection by extracting hierarchical features from high-dimensional data using deep neural networks. Among these, the Auto-AD method [82], a fully convolutional autoencoder, autonomously reconstructs the background and highlights anomalies through reconstruction errors, eliminating the need for manual parameter tuning or preprocessing. Other neural network models, such as stacked denoising autoencoders (SDAs) [106] and spectral-constrained adversarial autoencoders (SC-AAE) [91], use manifold learning and adversarial strategies to enhance anomaly detection capabil-

ities. These approaches are highly effective in nonlinear and complex environments but often require large datasets and significant computational resources, which can pose challenges for real-time applications.

1.3 Preliminaries

1.3.1 Tucker Decomposition

We first present some necessary notations for tensors and preliminaries on tensor operations, then we introduce the Tucker decomposition.

For a third order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, we let $x_{i_1 i_2 i_3}$ denote its (i_1, i_3, i_3)-th entry, let $x_{i_1 \ldots i_{k-1} : i_{k+1} \ldots i_3}$ denote its $(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_3)$-th mode-$k$ fiber and let $X_{::i_3}$ denote its i_3-th frontal slice. The mode-k unfolding of a third order tensor $\mathcal{X}$ is denoted as $\mathcal{X}_{(k)} = \text{unfold}_{(k)}(\mathcal{X})$, which is the process to linearize all indexes except index k. The dimensions of $\mathcal{X}_{(k)}$ are $I_k \times \prod_{j=1, j \neq k}^{3} I_j$. An element $x_{i_1 i_2 i_3}$ of $\mathcal{X}$ corresponds to the position of (i_k, j) in matrix $\mathcal{X}_{(k)}$, where $j = 1 + \sum_{l=1, l \neq k}^{3} (i_l - 1) \prod_{m=1, m \neq k}^{l-1} I_m$. Note that the mode-$k$ unfolding is also called the mode-k matricization. And the inverse process of the mode-k unfolding of a tensor $\mathcal{X}$ is denoted by $\mathcal{X} = \text{fold}_{(k)}(\mathcal{X}_{(k)})$. The mode-$k$ product of $\mathcal{X}$ and a matrix $Y \in \mathbb{R}^{J \times I_k}$, denoted by $\mathcal{X} \times_k Y$, is a new tensor $\mathcal{Z} \in \mathbb{R}^{I_1 \times \cdots \times I_{k-1} \times J \times I_{k+1} \cdots \times I_3}$ where

$$z_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_3} = \sum_{i_k=1}^{I_k} x_{i_1 \ldots i_{k-1} i_k i_{k+1} \ldots i_3} y_{j i_k}.$$

Tucker decomposition was introduced by Ledyard R. Tucker [77] in 1966. It generalizes SVD to tensors by decomposing a tensor into a smaller core tensor and a set of factor matrices. The Tucker decomposition of a third order tensor $\mathcal{X}$ can be written as

$$\mathcal{X} = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3,$$

where $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is the core tensor, $U_k \in \mathbb{R}^{I_k \times R_k}$ are the factor matrices, and $\times_k$ denotes the mode-k product. As Tucker decomposition is not unique, several variants have been proposed. For example, nonnegative Tucker decomposition [43] imposes the nonnegativity on the core tensor and the factor matrices and HOSVD [17] imposes the orthogonality on the factor matrices.

1.3.2 Block Coordinate Descent Algorithms

We first provide some preliminaries on optimization including the definitions for (limiting) subdifferentials and proximal operators, and then review some existing works on block coordinate descent (BCD) algorithms.

First, let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function with a finite lower bound. The (limiting) subdifferential of f at $x \in \operatorname{dom} f := \{x \in \mathbb{R}^d : f(x) < \infty\}$, denoted by $\partial f(x)$, is defined as

$$\partial f(x) := \{u \in \mathbb{R}^d : \exists x^k \to x, f(x^k) \to f(x) \text{ and } u^k \to u \text{ with } u^k \in \hat{\partial} f(x^k) \text{ as } k \to \infty\},$$

where $\hat{\partial} f(x)$ denotes the Fréchet subdifferential of f at $x \in \operatorname{dom} f$, which is the set of all $u \in \mathbb{R}^d$ satisfying

$$\liminf_{y \neq x, y \to x} \frac{f(y) - f(x) - \langle u, y - x \rangle}{\|y - x\|} \geq 0. \tag{1.1}$$

One can also observe that $\{u \in \mathbb{R}^d : \exists x^k \to x, f(x^k) \to f(x) \text{ and } u^k \to u \text{ with } u^k \in \partial f(x^k) \text{ as } k \to \infty\} \subseteq \partial f(x)$. The function f is μ-strongly convex if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex with $\mu \geq 0$; f is ρ-weakly convex if $f + \frac{\rho}{2}\|\cdot\|^2$ is convex with $\rho \geq 0$. The proximal operator of f with parameter $\lambda > 0$ evaluated at $x \in \mathbb{R}^d$, denoted as $\operatorname{prox}_{\lambda f}(x)$, is defined as

$$\operatorname{prox}_{\lambda f}(x) := \operatorname*{argmin}_{u \in \mathbb{R}^d} \left[f(u) + \frac{1}{2\lambda}\|u - x\|^2 \right].$$

Note that $\text{prox}_{\lambda f}$ is a set-valued map, when the minimizer is not unique, for example, for some nonconvex functions. And $\text{prox}_{\lambda f}$ reduces to a single-valued map, when the minimizer is unique, for example, for convex functions.

Second, we review some existing BCD algorithms for solving the following minimization problem with an objective function $f(x_1, x_2, \ldots, x_m)$ that can be expressed as $g(x_1, x_2, \ldots, x_m)$ plus a sum of m block functions:

$$\min_{x_1, x_2, \ldots, x_m} f(x_1, x_2, \ldots, x_m) := g(x_1, x_2, \ldots, x_m) + \sum_{i=1}^{m} h_i(x_i),$$

where $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_m} \to (-\infty, +\infty]$ and $h_i : \mathbb{R}^{d_i} \to (-\infty, +\infty]$, $i = 1, 2, \ldots, m$.

The origins of BCD methods date back to the 1950s, primarily in the context of solving decomposition problems and dynamic programming. Early works focused on convex functions, assuming strong convexity or pseudoconvexity to ensure global convergence. Hildreth [32] and Warga [85] formalized the idea of cyclic optimization over block variables, laying the foundation for BCD methods. Around 2000, the work by Tseng [76] marked a significant advancement by extending BCD to nonconvex and nonsmooth optimization problems under conditions such as pseudoconvexity or unique minimizers in block updates. For example, the BCD algorithm operates by iteratively optimizing the objective function with respect to one block of variables while fixing the others. At each iteration, a variable block i is selected, typically using cyclic or randomized strategies. The update rule of the BCD algorithm with a cyclic strategy is as follows

$$x_i^{k+1} \in \underset{x_i}{\text{argmin}} f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^{k}, \ldots, x_m^{k}).$$

Recently, a variant of the BCD algorithm called the block prox-linear (BPL) method was proposed in [95]. Then function g is assumed to be continuously differ-

entiable and its gradient with respect to the block x_i is represented as $\nabla_{x_i} g$, and the update rule is given via the proximal operator as follows

$$x_i^{k+1} \in \mathrm{prox}_{\alpha h_i}\big(x_i^k - \alpha \nabla_{x_i} g(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_m^k)\big).$$

The convergence of the BPL method is guaranteed under certain conditions. For example, the subsequence convergence can be achieved by assuming the block-wise Lipschitz continuity of the gradient of g, well-defined proximal operators for h_i, and boundedness of the level sets of f. By further imposing the Kurdyka-Łojasiewicz (KL) property on the objective function, the whole sequence convergence can also be obtained.

1.3.3 Optimization over the Stiefel Manifold

We first provide some preliminaries on the Stiefel Manifold, and then review some existing algorithms for solving optimization problems over the Stiefel Manifold.

First, the Stiefel manifold is defined as $\mathbb{S}_{m,n} := \{X \in \mathbb{R}^{m\times n} : X^\top X = \mathbf{I}_n\}$ with $m \geq n$, and its tangent space at a point $X \in \mathbb{R}^{m\times n}$ is given by $\mathcal{T}_X\mathbb{S}_{m,n} := \{Y \in \mathbb{R}^{m\times n} : Y^\top X + X^\top Y = 0\}$. A Riemannian metric on the Stiefel manifold can be defined using the metric induced from the Euclidean inner product.

Consider the following optimization problem constrained on the Stiefel manifold

$$\min_{X\in\mathbb{R}^{m\times n}} f(X) \quad \text{s.t. } X^\top X = \mathbf{I}_n,$$

where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ with $m \geq n$. The Riemannian gradient of a smooth function f at X is given by $\mathrm{grad}\, f(X) := \mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(\nabla f(X))$, where $\mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(Y) := (\mathbf{I}_m - XX^\top)Y + \frac{1}{2}X(X^\top Y - Y^\top X)$ projects a matrix Y onto the tangent space $\mathcal{T}_X\mathbb{S}_{m,n}$. The optimality conditions for the problem can be summarized as

$$\left\{\begin{array}{rcl} (\mathbf{I}_m - XX^\top)\nabla f(X) &=& 0, \\ X^\top \nabla f(X) &=& \nabla f(X)^\top X, \\ X^\top X &=& \mathbf{I}_n. \end{array}\right.$$

Second, we review some algorithms for optimization over the Stiefel manifold. Optimization problems over the Stiefel manifold $\mathbb{S}_{m,n}$, the set of matrices with orthonormal columns, can be viewed as eigenvalue problems and matrix factorization problems. For example, computational techniques, such as QR decomposition, played a critical role in ensuring orthogonality [68], forming the foundation of constrained optimization algorithms. However, the nonconvexity of the feasible region poses significant challenges in finding global minimizers. Ensuring orthogonality during iterations, particularly for large-scale problems, further introduces computational overhead. Traditional methods like explicit reorthogonalization via QR decomposition are often computationally expensive and struggle to scale efficiently.

To overcome these challenges, innovative approaches have been developed, including retraction-based methods [68] and projection-based algorithms [60]. Retraction-based methods map tangent vectors back to the manifold using approximations such as Cayley transformations or QR decompositions, while projection-based methods solve subproblems in the tangent space before projecting back to the manifold. Recent advancements, such as exact penalty function methods, simplify the problem by reformulating it to preserve the global minimizers of the original problem.

Algorithmic frameworks like gradient projection and columnwise block coordinate descent have demonstrated improved computational efficiency while maintaining feasibility. Parallelization techniques, such as penalty-based methods like PLAM [75] and PenCF [90], mitigate the need for costly reorthogonalization by leveraging distributed computing, making them particularly effective for large-scale problems, such as those encountered in electronic structure calculations.

Various optimization methods designed for matrix manifold problems have also been applied to this problem, including gradient-based methods [56, 59, 2], conjugate gradient methods [20, 1], trust region methods [96], and (Quasi-)Newton methods [20, 37]. These methods aim to find a feasible point with a lower function value than the

current iterate, addressing challenges in global convergence [46, 87].

1.4 Summary of Contributions of the Thesis

In this thesis, we develop new approaches based on (nonlocal) low-rank tensor regularization and tensor group sparsity $\ell_{2,p}$ norm $(0 < p < 1)$, along with some spatial and spectral priors, to provide a more comprehensive method that preserves the structure of HSIs. It includes two optimization models and two iterative algorithms with convergence guarantees for solving two important problems in hyperspectral image processing. Our main contributions are summarized as follows.

- We propose a class of orthogonal constrained minimization problems to tackle HSI restoration problems, such as removing mixed noise like Gaussian noise, stripes, and dead lines. The proposed class of models employs two types of regularization terms, which are a new sparsity-enhanced low-rank regularization and a generalized tensor group sparsity measure. The sparsity-enhanced low-rank regularization exploits the spatial nonlocal self-similarity and spectral correlation in HSIs to enhance restoration, ensuring that similar patterns in distant regions are jointly considered for improved denoising. The generalized tensor group sparsity measure, with a specific example being the tensor $\ell_{2,p}$ norm with $p \in (0, 1)$, measures the group sparsity in HSIs for characterizing the linear patterns in HSIs.

- We propose a proximal block coordinate descent (P-BCD) algorithm for solving the models. Each subproblem of the P-BCD algorithm has an exact solution, which either has a closed-form solution or is easy to compute. To show the convergence of the P-BCD algorithm, we define the stationary point of the proposed model using three equalities of substationarity, symmetry, and feasibility for orthogonal constraints. We prove that any accumulation point of the

sequence generated by the proposed algorithm is a stationary point.

- We show the proposed nonlocal low-rank tensor regularized $\ell_{2,p}$ (NLTL2p) approach for HSI denoising and destriping can outperform other state-of-the-art methods even a deep learning based method on the numerical experiments tested on both simulated and real HSI datasets, in metrics such as mean peak signal-to-noise ratio. In terms of visual quality, our method effectively restores HSIs by preserving important image details and removing noise, particularly highly structured noise like stripes and dead lines.

- We combine the proposed model with a deep neural network to incorporate an implicit proximal denoiser prior for hyperspectral anomaly detection and propose a PnP version of the P-BCD method (PnP-PBCD), in which certain steps are updated using a proximal denoiser. The tensor $\ell_{2,p}$ norm in the model is utilized to characterize the anomalies, while the implicit proximal denoiser prior is employed to remove Gaussian noise. The results, tested on HSIs contaminated with or without Gaussian noise, demonstrate that the proposed method can effectively detect anomalous objects, whereas competing methods may mistakenly identify noise as anomalies or incorrectly match the anomalous objects due to noise interference.

1.5 Organization of the Thesis

This thesis is organized as follows. Chapter 1 introduces the background of HSI restoration and anomaly detection, and presents some preliminaries used in this thesis. Chapter 2 develops a class of orthogonal constrained minimization models with tensor group sparsity regularization. Chapter 3 presents the P-BCD algorithm for solving the resulting nonconvex and nonsmooth optimization problem, along with a detailed convergence analysis. In Chapter 4, the proposed model is applied to HSI de-

noising and destriping, demonstrating its effectiveness through extensive experiments on simulated and real datasets. Chapter 5 extends the methodology to anomaly detection in noisy HSIs, integrating the low-rank tensor regularization with a proximal denoiser within a PnP framework to enhance anomaly detection performance under various noise conditions. Finally, Chapter 6 concludes the thesis by summarizing the key contributions and discussing potential future research directions.

Chapter 2

Orthogonal Constrained Minimization with Tensor Group Sparsity Regularization

2.1 Problem Statement for HSI Restoration

In HSI restoration, an observed HSI $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ can be decomposed into three components: a low-rank tensor $\mathcal{L}$, a sparse tensor $\mathcal{S}$, and a noise tensor $\mathcal{N}$, represented as follows [9]

$$\mathcal{D} = \mathcal{L} + \mathcal{S} + \mathcal{N},$$

in which $I_1 \times I_2$ are the spatial dimensions (height and width), and I_3 is the number of spectral bands.

In the context of HSI denoising and destriping, $\mathcal{L}$ represents the clean, low-rank HSI, encapsulating the primary spectral-spatial structure of the scene. The sparse component $\mathcal{S}$ corresponds to structured noise, such as stripe noise or dead lines, which is typically sparse and spatially localized. Meanwhile, $\mathcal{N}$ represents Gaussian noise affecting all spectral bands. The objective is to recover an accurate $\mathcal{L}$, enhancing the clarity and fidelity of the spectral information. To evaluate the effectiveness of denoising algorithms, metrics such as the peak signal-to-noise ratio (PSNR) [34] are employed to measure the quality of $\mathcal{L}$.

For anomaly detection, $\mathcal{L}$ denotes the low-rank background, representing consistent spectral features across the scene. The sparse component $\mathcal{S}$ represents anomalies with distinctive spectral features that deviate from the background, such as rare materials or targets, which are spatially sparse. And $\mathcal{N}$ accounts for Gaussian noise that affects the quality of detection. The primary focus of anomaly detection is the accuracy of $\mathcal{S}$, with evaluation metrics such as the area under the curve (AUC) [36] used to assess detection performance, emphasizing the identification of anomalies over background quality.

By exploiting the low-rank nature of hyperspectral backgrounds and the sparsity of noise or anomalies, a unified model can be developed for addressing fundamental tasks in HSI processing such as mixed noise removal and anomaly detection.

2.2 Model Formulation

To tackle the problems in HSI restoration such as mixed noise removal, we propose an optimization model utilizing low-rank tensor regularization and a group sparsity measure. The proposed model is formulated as follows

$$\begin{aligned} \min_{\substack{\mathcal{S},[X_1],[X_2],\\ [X_3],[\mathcal{G}],\mathcal{L}}} \quad & \frac{\delta}{2}\|\mathbf{R}(\mathcal{L}+\mathcal{S}-\mathcal{D})\|_F^2+\gamma\|\sqrt{\mathcal{W}_{\mathbf{R}}}\odot\mathcal{S}\|_{2,\psi}+\|[\mathcal{G}]\|_{1,w} \\ & +\frac{1}{2}\|\mathbf{R}(\mathcal{L})-[\mathcal{G}]\times_1[X_1]\times_2[X_2]\times_3[X_3]\|_F^2 \\ \text{s.t.} \quad & [X_i]^\top[X_i]=[\mathbf{I}_{n_i}], i=1,2,3, \end{aligned} \tag{2.1}$$

where

- $\mathcal{S},\mathcal{L},\mathcal{D}\in\mathbb{R}^{I_1\times I_2\times I_3}$, $[X_i]\in\mathbb{R}^{m_i\times n_i\times N}$, $i=1,2,3$, and $[\mathcal{G}]\in\mathbb{R}^{n_1\times n_2\times n_3\times N}$;
- $\mathbf{R}:\mathbb{R}^{I_1\times I_2\times I_3}\to\mathbb{R}^{m_1\times m_2\times m_3\times N}$ denotes the similar blocks extraction operator, and $\mathbf{R}^\top:\mathbb{R}^{m_1\times m_2\times m_3\times N}\to\mathbb{R}^{I_1\times I_2\times I_3}$ denotes the transpose of $\mathbf{R}$ satisfying $\langle\mathbf{R}(\mathcal{L}),[\mathcal{Y}]\rangle=\langle\mathcal{L},\mathbf{R}^\top([\mathcal{Y}])\rangle$ for any $\mathcal{L}\in\mathbb{R}^{I_1\times I_2\times I_3}$, $[\mathcal{Y}]\in\mathbb{R}^{m_1\times m_2\times m_3\times N}$;

- $\mathcal{W}_{\mathbf{R}} \in \mathbb{R}_{++}^{I_1 \times I_2 \times I_3}$ is a weight tenor such that the component-wise multiplication with $\mathcal{W}_{\mathbf{R}}$ being an equivalent operation of $\mathbf{R}^\top \mathbf{R}$, that is, $\mathcal{W}_{\mathbf{R}} \odot \mathrm{Id} = \mathbf{R}^\top \mathbf{R}$, $\odot$ denotes the component-wise multiplication, Id denotes the identity mapping on $\mathbb{R}^{I_1 \times I_2 \times I_3}$, and $\sqrt{\mathcal{W}_{\mathbf{R}}}$ denotes the component-wise square root of $\mathcal{W}_{\mathbf{R}}$, that is, the (i_1, i_2, i_3)-th entry of $\sqrt{\mathcal{W}_{\mathbf{R}}}$ is equal to $\sqrt{(\mathcal{W}_{\mathbf{R}})_{i_1 i_2 i_3}}$;

- $\|\mathcal{S}\|_{2,\psi}$ denotes the generalized tensor group sparsity measure of a third order tensor $\mathcal{S}$ defined by

$$\|\mathcal{S}\|_{2,\psi} = \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \psi\left(\|s_{:i_2 i_3}\|_2\right) \tag{2.2}$$

  with $\psi : \mathbb{R} \to [0, +\infty)$ being a continuous sparsity-promoting function [69] and $\|s_{:i_2 i_3}\|_2 = \left(\sum_{i_1=1}^{I_1} s_{i_1 i_2 i_3}^2\right)^{\frac{1}{2}}$;

- $[\mathcal{G}] \times_1 [X_1] \times_2 [X_2] \times_3 [X_3]$ denotes an independent 3-D HOSVD with $[\mathcal{G}] \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times N}$, and $[X_i] \in \mathbb{R}^{m_i \times n_i \times N}$, $i = 1, 2, 3$, such that $[X_i]$ is independently orthogonal, i.e., $[X_i]^\top [X_i] = [\mathbf{I}_{n_i}]$, with $\mathbf{I}_{n_i}$ representing the identity matrix of size $n_i \times n_i$;

- $\|[\mathcal{G}]\|_{1,w}$ denotes the weighted tensor (component-wise) ℓ_1 norm for a fourth order tensor $[\mathcal{G}]$ defined by

$$\|[\mathcal{G}]\|_{1,w} = \sum_{j=1}^{N} w_j \|[\mathcal{G}]^{(j)}\|_1 \tag{2.3}$$

  with a weight vector $w \in \mathbb{N}_+$.

The resulting model is a nonconvex nonsmooth minimization problem with orthogonal constraints. In particular, the first term of the model (2.1) is a data fidelity term to remove Gaussian noise, the second term is a group sparsity measure to remove

sparse noise with linear structures, and the last two terms are the sparsity-enhanced nonlocal low-rank tensor regularization terms.

In the following subsections, we will provide more details on the generalized tensor group sparsity regularization term and the sparsity-enhanced nonlocal low-rank tensor regularization used in model (2.1).

2.3 Generalized Tensor Group Sparsity Regularization

As shown in (2.2), the generalized group sparsity measure groups the sparse component $\mathcal{S}$ along a specific direction, for example, mode-1, measures the magnitude of $s_{:i_2i_3}$ using the ℓ_2 norm, and then characterizes the group sparsity using a sparsity-promoting function [69] for ψ. We give some examples of continuous sparsity-promoting functions for ψ:

(i) ℓ_1 norm: $\psi(t) = |t|$;

(ii) ℓ_p norm: $\psi(t) = |t|^p$, $p \in (0, 1)$;

(iii) Relaxed ℓ_p norm: $\psi(t) = (|t| + \varepsilon)^p - \varepsilon^p$, $p \in (0, 1)$, $\varepsilon > 0$;

(iv) Minimax concave penalty (MCP) [101]: for $\theta > \lambda$,

$$\psi_{\lambda,\theta}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\theta}, & |t| \le \theta\lambda, \\ \frac{\theta\lambda^2}{2}, & \text{otherwise;} \end{cases}$$

(v) Smoothly clipped absolute deviation (SCAD) [22]: for $\lambda > 0$ and $\theta > 2$,

$$\psi_{\lambda,\theta}(t) = \begin{cases} \lambda|t|, & |t| \le \lambda, \\ \frac{-t^2+2\theta\lambda|t|-\lambda^2}{2(\theta-1)}, & \lambda < |t| \le \theta\lambda, \\ \frac{(\theta+1)\lambda^2}{2}, & \text{otherwise.} \end{cases}$$

Note that (i) is convex, (iii) is $p\varepsilon^{p-1}$-weakly convex, (iv) is $\frac{1}{\theta}$-weakly convex, and (v) is $\frac{1}{\theta-1}$-weakly convex, according to [5].

Since $\|\cdot\|_{2,\psi}$ is separable, the proximal operator of $\mu\|\cdot\|_{2,\psi}$ at $\widetilde{\mathcal{S}}$ as follows

$$\mathcal{S} \in \text{prox}_{\mu\|\cdot\|_{2,\psi}}(\widetilde{\mathcal{S}})$$

can be computed via the proximal operator of $\mu\psi \circ \|\cdot\|_2$ at the (i_2, i_3)-th mode-1 fiber $\tilde{s}_{:i_2i_3}$, i.e.,

$$s_{:i_2i_3} \in \text{prox}_{\mu\psi\circ\|\cdot\|_2}(\tilde{s}_{:i_2i_3}),$$

where $s_{:i_2i_3}$ is the (i_2, i_3)-th mode-1 fiber of $\mathcal{S}$. According to Theorem 4.1 in [98], we have

$$\text{prox}_{\mu\psi\circ\|\cdot\|_2}(\tilde{s}) = \begin{cases} \text{prox}_{\mu\psi}(\|\tilde{s}\|_2)\frac{\tilde{s}}{\|\tilde{s}\|_2}, & \|\tilde{s}\|_2 \neq 0, \\ 0, & \|\tilde{s}\|_2 = 0. \end{cases}$$

Depending on the choice of ψ, the proximal operator of $\text{prox}_{\mu\psi\circ\|\cdot\|}$ is computed differently.

2.3.1 Tenosr $\ell_{2,p}$ Norm

To measure the linear structural sparsity of the sparse noise tensor $\mathcal{S}$, we extend the matrix $\ell_{2,p}$ norm for group sparsity to its tensor form. As the stripes and dead lines often align the first dimension, we define the tensor $\ell_{2,p}$ $(0 < p < 1)$ norm of a third order tensor $\mathcal{S}$ as follows

$$\|\mathcal{S}\|_{2,p} = \left(\sum_{i_2=1}^{I_2}\sum_{i_3=1}^{I_3}\|s_{:i_2i_3}\|_2^p\right)^{\frac{1}{p}} = \left(\sum_{i_3=1}^{I_3}\sum_{i_2=1}^{I_2}\left(\sum_{i_1=1}^{I_1}s_{i_1i_2i_3}^2\right)^{\frac{p}{2}}\right)^{\frac{1}{p}}. \tag{2.4}$$

The tensor $\ell_{2,p}$ norm is exactly equal to the matrix $\ell_{2,p}$ norm of the unfolding matrix of $\mathcal{S}$ along the first dimension, that is, $\|\mathcal{S}\|_{2,p} = \|S_{(1)}\|_{2,p}$. The matrix $\ell_{2,p}$ norm is a nonconvex and nonsmooth function. And it has been applied to image

processing [50], machine learning [54, 19], feature selection [49, 79], multi-view classification [84], etc.

The generalized tensor group sparsity measure given in (2.2) reduces to $\|\mathcal{S}\|_{2,p}^p$, the tensor $\ell_{2,p}$ norm with a power of p, if we choose ψ as the ℓ_p norm with $p \in (0,1)$, i.e., $\psi(t) = |t|^p$, $p \in (0,1)$. In the following, we summarize some results for minimizing the tensor group sparsity measure $\|\mathcal{S}\|_{2,p}^p$ as follows

$$\min_{\mathcal{S}} \mu\|\mathcal{S}\|_{2,p}^p + \frac{1}{2}\|\mathcal{S} - \widetilde{\mathcal{S}}\|_F^2, \tag{2.5}$$

where $\widetilde{\mathcal{S}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is a given tensor and parameter $\mu > 0$. Since $\|\cdot\|_{2,p}^p$ and $\|\cdot\|_F^2$ are both group-separable, solving problem (2.5) for $\mathcal{S}$ is equivalent to solving the following subproblem for each (i_2, i_3)-th vector of $\mathcal{S}$ along the first dimension

$$\min_{s} \mu\|s\|_2^p + \frac{1}{2}\|s - \tilde{s}\|_2^2, \tag{2.6}$$

where $s \in \mathbb{R}^{I_1}$ and $\tilde{s} \in \mathbb{R}^{I_1}$, for simplicity, represent $s_{:i_2i_3}$ and $\tilde{s}_{:i_2i_3}$, respectively, and $\|s\|_2^p = (s_1^2 + s_2^2 + \cdots + s_{I_1}^2)^{\frac{p}{2}}$. It follows from the triangle inequality that the objective function of (2.6) satisfies the following inequality for any $s \in \mathbb{R}^{I_1}$

$$\mu\|s\|_2^p + \frac{1}{2}\|s - \tilde{s}\|_2^2 \geq \mu\|s\|_2^p + \frac{1}{2}(\|s\|_2 - \|\tilde{s}\|_2)^2. \tag{2.7}$$

And the equality holds if and only if $s = t\tilde{s}$ for some $t \geq 0$ or $\tilde{s} = 0$. Observe that the right-hand side of the inequality is only related to $\|s\|_2$ and $\|\tilde{s}\|_2$. If $\tilde{s} = 0$, the solution of problem (2.6) is $s = 0$. If $\tilde{s} \neq 0$, we can view s as $s = t\|\tilde{s}\|_2 v$ with $t \geq 0$ being a scalar and $v \in \mathbb{R}^{I_1}$ being a unit vector. When we restrict the minimization problem (2.6) by $\|s\|_2 = t\|\tilde{s}\|_2$ with a fixed t, according to (2.7), the solution of the restricted problem of (2.6) is obtained only when $v = \frac{\tilde{s}}{\|\tilde{s}\|_2}$. Hence, if $\tilde{s} \neq 0$, the solution of (2.6) is $s = t\tilde{s}$, where t is a minimizer of the following problem

$$\min_{t \in [0,\infty)} \nu t^p + \frac{1}{2}(t-1)^2, \tag{2.8}$$

with $\nu = \mu\|\tilde{s}\|_2^{p-2}$. That is, it only requires solving a one-dimensional problem (2.8) for computing the solutions of problem (2.6).

Next, we show a lemma and a solver for computing solutions of problem (2.6). Let $g(t) = \nu|t|^p + \frac{1}{2}(t-1)^2$. Note that $g(t) > g(|t|)$ for $\forall t < 0$ and $g(t) > g(1)$ for $\forall t > 1$. Then problem (2.8) can be relaxed to an unconstrained problem with g being the objective function, which can be solved using Theorem 1 in [57]. Also, problem (2.8) can be reduced to a box constrained problem with constraint $t \in [0, 1]$, which can be solved using Lemma 4.1 in [51]. We summarize the results for (2.8) in the following lemma.

Lemma 2.1. *Let $p \in (0, 1)$ and $\nu > 0$. Let*

$$\nu_0 := \frac{(2(1-p))^{1-p}}{(2-p)^{2-p}} \quad \text{and} \quad \tau(\nu) := (2\nu(1-p))^{\frac{1}{2-p}}.$$

Then the set of optimal solutions of problem (2.8), denoted as $\Omega^*(\nu)$, is given by

$$\Omega^*(\nu) = \begin{cases} \{0\}, & \text{if } \nu > \nu_0, \\ \{0, \tau(\nu_0)\}, & \text{if } \nu = \nu_0, \\ \{t^*\}, & \text{if } 0 < \nu < \nu_0, \end{cases}$$

where $t^* \in (\tau(\nu), 1)$ is the unique solution of the equation

$$\nu p t^{p-1} + t - 1 = 0 \tag{2.9}$$

with $t \in (\tau(\nu), \infty)$.

Proof. Let $g : [0, \infty) \to \mathbb{R}$ be defined as the objective function of problem (2.8). When $t = 0$, $g(0) = \frac{1}{2}$. When $t > 0$, we define $u : (0, \infty) \to \mathbb{R}$ as $u(t) = \nu t^{p-1} + \frac{t}{2} - 1$ and then we have

$$g(t) = tu(t) + \frac{1}{2}. \tag{2.10}$$

We compute the first and second order derivatives of $u(t)$ as follows

$$u'(t) = \nu(p-1)t^{p-2} + \frac{1}{2}$$

$$u''(t) = \nu(p-1)(p-2)t^{p-3}.$$

For $0 < p < 1$, we have $u'(\tau(\nu)) = 0$ and $u''(t) > 0$. Then by the second derivative test, $u(t)$ obtains a global minimum at $t = \tau(\nu)$ and the minimum value is

$$\begin{aligned} u(\tau(\nu)) &= \nu(2\nu(1-p))^{\frac{p-1}{2-p}} + \frac{1}{2}(2\nu(1-p))^{\frac{1}{2-p}} - 1 \\ &= \nu^{\frac{1}{2-p}}(2(1-p))^{\frac{p-1}{2-p}}(2-p) - 1 \\ &= \left(\frac{\nu}{\nu_0}\right)^{\frac{1}{2-p}} - 1. \end{aligned}$$

(i) $0 \in \Omega^*(\nu)$ if and only if $g(t) - g(0) = tu(t) \geq 0$, $\forall\ t \in (0, \infty)$, if and only if $u(t) \geq 0$, $\forall\ t \in (0, \infty)$, if and only if $\nu \geq \nu_0$. In fact, $g(t) - g(0) > 0$, for $\forall t \in (0, \infty)$, if and only if $\nu > \nu_0$. Hence, $\Omega^*(\nu) = \{0\}$ if $\nu > \nu_0$.

(ii) $t^* \in \Omega^*(\nu) \neq \varnothing$ with $t^* \in (0, \infty)$ if and only if t^* is a solution of (2.9), according to Fermat's rule, i.e., $g'(t^*) = 0$. We compute the first and second order derivatives of $g(t)$ on $(0, \infty)$ as follows

$$g'(t) = \nu p t^{p-1} + t - 1$$

$$g''(t) = \nu p(p-1)t^{p-2} + 1.$$

Let $t_0 = (\nu p(1-p))^{\frac{1}{2-p}}$. It can be verified that $g''(t_0) = 0$, $g''(t) < 0$ on $(0, t_0)$ and $g''(t) > 0$ on (t_0, ∞). Then $g'(t)$ is strictly decreasing on $(0, t_0)$ and strictly increasing on (t_0, ∞). Since $\tau(\nu) > t_0$, we have $g'(t)$ is strictly increasing on $(\tau(\nu), \infty)$. Also, it follows from (2.10) that $g'(\tau(\nu)) = u(\tau(\nu)) + tu'(\tau(\nu)) = (\frac{\nu}{\nu_0})^{\frac{1}{2-p}} - 1$. And we have $g'(1) = \nu p > 0$. Hence, $t^* \in \Omega^*(\nu)$, where $t^* \in (\tau(\nu), 1)$ is the unique solution of (2.9), if and only if $g'(\tau(\nu)) \leq 0$, i.e., $\nu \leq \nu_0$. In particular, if $\nu < \nu_0$, we have

$g'(\tau(\nu)) < 0$ and then $\Omega^*(\nu) = \{t^*\}$ where $t^* \in (\tau(\nu), 1)$ is the unique solution of (2.9).

(iii) If $\nu = \nu_0$, by (i), we have $0 \in \Omega^*(\nu_0)$ and, by (ii), we have $t^* = \tau(\nu_0) \in \Omega^*(\nu_0)$, where t^* is the unique solution of (2.9). Hence, $\Omega^*(\nu_0) = \{0, \tau(\nu_0)\}$. □

According to Lemma 2.1, when $\nu = \nu_0$, there are two minimizers for problem (2.8). For simplicity, we will choose 0 in this case. When $\nu \in (0, \nu_0)$, the minimizer of problem (2.8) is unique and can obtained by solving (2.9). If p is chosen as, for example, $p = 1/2$, (2.9) has a closed-form root. Otherwise, we estimate the unique root t^* by Newton's method with an initial value of $t_0 = (\tau(\nu) + 1)/2$. Altogether, we summarize a proximal operator of the tensor $\ell_{2,p}$ norm in the following theorem.

Theorem 2.1. *Let $p \in (0, 1)$ and $\mu > 0$. Define the operator $\Gamma_\mu : \mathbb{R} \to \mathbb{R}$ by*

$$\Gamma_\mu(\beta) := \begin{cases} 0, & \text{if } \beta \leq \frac{\beta_0(2-p)}{2(1-p)}, \\ t^*, & \text{otherwise,} \end{cases}$$

where $\beta_0 = (2\mu(1-p))^{\frac{1}{2-p}}$, and $t^* \in [\beta_0\beta, 1)$ is the unique solution of

$$\mu\beta^{p-2}pt^{p-1} + t - 1 = 0, \quad t \in [\beta_0\beta, \infty)\,.$$

Then a solution of the proximal operator of the tensor $\ell_{2,p}$ norm at $\widetilde{\mathcal{S}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ can be computed by

$$\Gamma_\mu(\|\tilde{s}_{:i_2i_3}\|_2)\tilde{s}_{:i_2i_3} \in \operatorname{prox}_{\mu\|\cdot\|_2^p}(\tilde{s}_{:i_2i_3}) = \left(\operatorname{prox}_{\mu\|\cdot\|_{2,p}^p}(\widetilde{\mathcal{S}})\right)_{:i_2i_3},$$

for $i_2 = 1, 2, \ldots, I_2$, $i_3 = 1, 2, \ldots, I_3$.

2.4 Nonlocal Low-rank Tensor Regularization

Nonlocal low-rank tensor regularization is a technique used in image processing to enhance image restoration by exploiting the inherent structures and redundancies

present in the images. For HSIs, it follows from the spectral correlation and the spatial nonlocal self-similarity that, for an image block, we can find enough nonlocal similar blocks across the image or within a local window [8]. Stacking and reforming similar tensors into a higher-order tensor via the similar blocks extraction operator $\mathbf{R}$, a clean HSI can be approximated by nonlocal low-rank tensors [55, 52]. In particular, we propose a nonlocal low-rank tensor regularization using the independent 3-D HOSVD with sparsity enhancements on the independent core tensors.

2.4.1 Independent 3-D HOSVD

Independent 3-D HOSVD is a tensor decomposition method that extends the concept of matrix SVD to higher dimensions.

We first introduce the definition of a 3-D HOSVD and then define an independent 3-D HOSVD using the notation of $[\,\cdot\,]$. For a third order tensor $\mathcal{Y} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$, the (truncated) 3-D HOSVD of $\mathcal{Y}$ is to approximate $\mathcal{Y}$ in the following form

$$\mathcal{Y} \approx \mathcal{G} \times_1 X_1 \times_2 X_2 \times_3 X_3, \tag{2.11}$$

where $\mathcal{G} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is the core tensor, and $X_i \in \mathbb{R}^{m_i \times n_i}$ is the i-th factor matrix such that $X_i^\top X_i = \mathbf{I}_{n_i}$. Note that $m_i \geq n_i$ and X_i belongs to a Stiefel manifold, that is, $X_i \in \mathbb{S}_{m_i,n_i}$. By imposing orthogonality on the factor matrices, the decomposition in (2.11) can inherit many nice properties from the matrix SVD. For example, the core can have the all-orthogonality and the ordering property [11].

When a fourth order tensor has little correlation across the last mode, we view the fourth order tensor as a stack of independent third order tensors. Using the notation of $[\,\cdot\,]$, we denote such a fourth order tensor as $[\mathcal{Y}] \in \mathbb{R}^{m_1 \times m_2 \times m_3 \times N}$ and its j-th third order tensor as $[\mathcal{Y}]^{(j)} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$, $j = 1, 2, \ldots, N$. Also, a stack of independent matrices is denoted as $[X] \in \mathbb{R}^{m \times n \times N}$. And we call $[X]$ is independently orthogonal if $[X]^\top [X] = [\mathbf{I}_n]$, meaning $([X]^{(j)})^\top [X]^{(j)} = \mathbf{I}_n$, or equivalently $[X] \in [\mathbb{S}_{m,n}]$, meaning

$[X]^{(j)} \in \mathbb{S}_{m,n}$. Then we define an independent 3-D HOSVD of $[\mathcal{Y}]$ as

$$[\mathcal{Y}] \approx [\mathcal{G}] \times_1 [X_1] \times_2 [X_2] \times_3 [X_3],$$

where $[\mathcal{G}] \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times N}$, $[X_i] \in \mathbb{R}^{m_i \times n_i \times N}$, and $[\mathcal{Y}]^{(j)} \approx [\mathcal{G}]^{(j)} \times_1 [X_1]^{(j)} \times_2 [X_2]^{(j)} \times_3 [X_3]^{(j)}$ with $([X_i]^{(j)})^\top [X_i]^{(j)} = \mathbf{I}_{n_i}$, $i = 1, 2, 3$, $j = 1, 2, \ldots, N$. Similarly, we extend the notation of $[\,\cdot\,]$ to other operations acting on independent tensors. That is, performing an operation on an independent tensor means performing the operation on each lower order tensor independently. For example, performing $[X]_{[(n)]}$ means independently performing $[X]_{(n)}^{(j)}$ for each j.

2.4.2 Nonlocal Low-rank Tensor Regularization with Sparsity Enhancements

The nonlocal low-rank tensor regularization consists of the extraction of nonlocal similar tensors that may have similar features and the characterization of the low-rankness of the tensor.

First, we apply block matching to find similar blocks and then stack them into a fourth order nonlocal similar tensor. Given an HSI $\mathcal{L}$, we divide it into a total number of N overlapping blocks. For the j-th block, we search within a local window for a total of m_2 blocks that are similar to the reference block based on Euclidean distance. Then the j-th nonlocal similar sub-tensor of $\mathcal{L}$, denoted as $\mathbf{R}_j(\mathcal{L})$, can be formed by unfolding all the nonlocal similar blocks in the j-th group and then stacking them together. As the nonlocal similar block sub-tensors are independent of each other, we can further stack them together into a fourth order nonlocal similar group tensor, denoted as $\mathbf{R}(\mathcal{L})$, and $[\mathbf{R}(\mathcal{L})]^{(j)} = \mathbf{R}_j(\mathcal{L})$.

Second, we impose the tensor low-rankness on the nonlocal similar tensor. In the nonlocal sub-tensor $\mathbf{R}_j(\mathcal{L})$ that we construct, the first dimension indicates the spatial information, the second dimension reveals the nonlocal self-similarity, and

the third dimension reflects the spectral correlation. We adopt the independent 3-D HOSVD to obtain a low-rank approximation of $\mathbf{R}(\mathcal{L})$, that is,

$$\mathbf{R}(\mathcal{L}) \approx [\mathcal{G}] \times_1 [X_1] \times_2 [X_2] \times_3 [X_3],$$

where $[\mathcal{G}] \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times N}$ denotes independent core tensors, and $[X_i] \in \mathbb{R}^{m_i \times n_i \times N}$ denotes the i-th factor matrices such that $[X_i]^\top [X_i] = [\mathbf{I}_{n_i}]$.

To further boost the low-rankness of $\mathbf{R}(\mathcal{L})$, we propose a sparsity-enhanced non-local low-rank tensor regularization term as follows

$$\frac{1}{2}\|[\mathcal{G}] \times_1 [X_1] \times_2 [X_2] \times_3 [X_3] - \mathbf{R}(\mathcal{L})\|_F^2 + \|[\mathcal{G}]\|_{1,w}, \tag{2.12}$$

where $\|[\mathcal{G}]\|_{1,w}$ is given in (2.3) with $w \in \mathbb{N}_+$. In particular, the first term of (2.12) measures the closeness between $\mathbf{R}(\mathcal{L})$ and the approximated low-rank tensor, and the second term measures the sparsity of the independent core tensors $[\mathcal{G}]$.

Chapter 3

The Proximal Block Coordinate Descent Algorithm and its Convergence Analysis

In the following, we present the P-BCD algorithm for solving the proposed nonlocal low-rank tensor model (2.1), which is a nonconvex and nonsmooth optimization problem over Stiefel manifolds. We will give the details of each update in the P-BCD algorithm and conduct a convergence analysis for the proposed P-BCD algorithm in this chapter.

Let F denote the objective function of model (2.1), that is,

$$F(\mathcal{S},[X_1],[X_2],[X_3],[\mathcal{G}],\mathcal{L}) := \frac{\delta}{2}\|\mathbf{R}(\mathcal{L}+\mathcal{S}-\mathcal{D})\|_F^2 + \gamma\left\|\sqrt{\mathcal{W}_{\mathbf{R}}}\odot\mathcal{S}\right\|_{2,\psi} + \Phi([\mathcal{G}]) + H([X_1],[X_2],[X_3],[\mathcal{G}],\mathcal{L}),$$

where $\Phi([\mathcal{G}]) = \|[\mathcal{G}]\|_{1,w}$ and

$$H([X_1],[X_2],[X_3],[\mathcal{G}],\mathcal{L}) := \frac{1}{2}\|\mathbf{R}(\mathcal{L}) - [\mathcal{G}]\times_1[X_1]\times_2[X_2]\times_3[X_3]\|_F^2. \tag{3.1}$$

3.1 The Proximal Block Coordinate Descent Algorithm

For solving model (2.1), the P-BCD algorithm is summarized as follows

$$\mathcal{S}^{k+1} \in \underset{\mathcal{S}}{\operatorname{argmin}}\ F(\mathcal{S}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_{\mathcal{S}}}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S} - \mathcal{S}^k)\|_F^2,$$

$$[X_1^{k+1}] \in \underset{[X_1]\in[\mathbb{S}_{m_1,n_1}]}{\operatorname{argmin}}\ F(\mathcal{S}^{k+1}, [X_1], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_1] - [X_1^k]\|_F^2,$$

$$[X_2^{k+1}] \in \underset{[X_2]\in[\mathbb{S}_{m_2,n_2}]}{\operatorname{argmin}}\ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_2] - [X_2^k]\|_F^2,$$

$$[X_3^{k+1}] \in \underset{[X_3]\in[\mathbb{S}_{m_3,n_3}]}{\operatorname{argmin}}\ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_3] - [X_3^k]\|_F^2,$$

$$[\mathcal{G}^{k+1}] = \underset{[\mathcal{G}]}{\operatorname{argmin}}\ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}], \mathcal{L}^k) + \frac{\alpha_{\mathcal{G}}}{2}\|[\mathcal{G}] - [\mathcal{G}^k]\|_F^2,$$

$$\mathcal{L}^{k+1} = \underset{\mathcal{L}}{\operatorname{argmin}}\ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^{k+1}], \mathcal{L}),$$

where $\alpha_{\mathcal{S}}, \alpha_X, \alpha_{\mathcal{G}} > 0$.

In the following, we present the details for computing each update.

3.1.1 The Update of $\mathcal{S}$

Recall that

$$\|\mathbf{R}(\mathcal{L})\|_F^2 = \langle \mathcal{L}, \mathbf{R}^\top \mathbf{R}(\mathcal{L})\rangle_F = \langle \mathcal{L}, \mathcal{W}_{\mathbf{R}} \odot \mathcal{L}\rangle_F = \|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{L}\|_F^2.$$

Then we have

$$\begin{aligned}
&F(\mathcal{S}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_{\mathcal{S}}}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S} - \mathcal{S}^k)\|_F^2 \\
=&\frac{\delta}{2}\|\mathbf{R}(\mathcal{L}^k + \mathcal{S} - \mathcal{D})\|_F^2 + \gamma \left\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{S}\right\|_{2,\psi} + \frac{\alpha_{\mathcal{S}}}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S} - \mathcal{S}^k)\|_F^2 \\
=&\frac{\delta}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S} + \mathcal{L}^k - \mathcal{D})\|_F^2 + \gamma \left\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{S}\right\|_{2,\psi} + \frac{\alpha_{\mathcal{S}}}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S} - \mathcal{S}^k)\|_F^2.
\end{aligned}$$

Hence, $\mathcal{S}^{k+1}$ is computed by

$$\mathcal{S}^{k+1} \in \underset{\mathcal{S}}{\operatorname{argmin}}\ \tilde{\gamma} \left\| \sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{S} \right\|_{2,\psi} + \frac{1}{2} \left\| \sqrt{\mathcal{W}_{\mathbf{R}}} \odot \left(\mathcal{S} - \left(\mathcal{S}^k - \tilde{\alpha}_{\mathcal{S}} \left(\mathcal{S}^k + \mathcal{L}^k - \mathcal{D} \right) \right) \right) \right\|_F^2,$$

where $\tilde{\gamma} = \frac{\gamma}{\delta + \alpha_{\mathcal{S}}}$ and $\tilde{\alpha}_{\mathcal{S}} = \frac{\delta}{\delta + \alpha_{\mathcal{S}}}$. Rescaling $\mathcal{S}$ using $\sqrt{\mathcal{W}_{\mathbf{R}}}\odot$, $\mathcal{S}^{k+1}$ can be written in terms of the proximal operator of $\|\cdot\|_{2,\psi}$ as follows

$$\mathcal{S}^{k+1} \in \left(\sqrt{\mathcal{W}_{\mathbf{R}}} \right)^{-1} \odot \operatorname{prox}_{\tilde{\gamma}\|\cdot\|_{2,\psi}} \left(\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \left(\mathcal{S}^k - \tilde{\alpha}_{\mathcal{S}} \left(\mathcal{S}^k + \mathcal{L}^k - \mathcal{D} \right) \right) \right), \tag{3.2}$$

where the (i_1, i_2, i_3)-th entry of $\left(\sqrt{\mathcal{W}_{\mathbf{R}}} \right)^{-1}$ is equal to $1/\sqrt{(\mathcal{W}_{\mathbf{R}})_{i_1 i_2 i_3}}$.

3.1.2 The Update of $[X_i]$

Before we solve the optimization subproblem in terms of $[X_i]$ over independent Stiefel manifolds, we can rewrite its objective function using the following useful fact for unfolding of tensors

$$\mathcal{Y} = \mathcal{G} \times_i X \quad \text{ if and only if } \quad Y_{(i)} = X G_{(i)}.$$

Then by applying $X \in \mathbb{S}_{m,n}$, the Frobenious norm of tensors can be rewritten into the Frobenious norm of matrices

$$\begin{aligned} \|\mathcal{G} \times_i X - \mathcal{L}\|_F^2 =& \|X G_{(i)} - L_{(i)}\|_F^2 \\ =& \|G_{(i)}\|_F^2 - 2\langle X G_{(i)}, L_{(i)} \rangle + \|L_{(i)}\|_F^2, \end{aligned} \tag{3.3}$$

where $G_{(i)}$ and $L_{(i)}$ denote the mode-i unfolding of $\mathcal{G}$ and $\mathcal{L}$, respectively. Since $\langle X G_{(i)}, L_{(i)} \rangle = \langle X, L_{(i)} G_{(i)}^\top \rangle$ and $\|X\|_F^2 = n$, minimizing $\|\mathcal{G} \times_i X - \mathcal{L}\|_F^2$ over X on the Stiefel manifold is equivalent to minimizing $\|X - L_{(i)} G_{(i)}^\top\|_F^2$ over the Stiefel manifold.

Taking $[X_1]$ as an example, we have

$$\begin{aligned}
&F(\mathcal{S}^{k+1}, [X_1], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_1] - [X_1^k]\|_F^2 \\
=&\frac{1}{2}\|\mathbf{R}(\mathcal{L}^k) - [\mathcal{G}^k] \times_1 [X_1] \times_2 [X_2^k] \times_3 [X_3^k]\|_F^2 + \frac{\alpha_X}{2}\|[X_1] - [X_1^k]\|_F^2 \\
=&\frac{1}{2}\|(\mathbf{R}(\mathcal{L}^k))_{[(1)]} - [X_1]([\mathcal{G}^k] \times_2 [X_2^k] \times_3 [X_3^k])_{[(1)]}\|_F^2 + \frac{\alpha_X}{2}\|[X_1] - [X_1^k]\|_F^2.
\end{aligned}$$

Then $[X_1^{k+1}]$ can be computed via the projection of unfolding matrices onto the Stiefel manifolds independently as follows

$$\begin{aligned}
[X_1^{k+1}] &\in \underset{[X_1]\in[\mathbb{S}_{m_1,n_1}]}{\operatorname{argmin}} \frac{1}{2}\|[X_1] - [P_1^k][Q_1^k]^\top\|_F^2 + \frac{\alpha_X}{2}\|[X_1] - [X_1^k]\|_F^2 \\
&= \underset{[X_1]\in[\mathbb{S}_{m_1,n_1}]}{\operatorname{argmin}} \frac{1}{2}\|[X_1] - ([X_1^k] - \tilde{\alpha}_X([X_1^k] - [P_1^k][Q_1^k]^\top))\|_F^2 \\
&= \operatorname{Proj}_{[\mathbb{S}_{m_1,n_1}]}\left([X_1^k] - \tilde{\alpha}_X\left([X_1^k] - [P_1^k][Q_1^k]^\top\right)\right),
\end{aligned}$$

where $[P_1^k] = (\mathbf{R}(\mathcal{L}^k))_{[(1)]}$, $[Q_1^k] = \left([\mathcal{G}^k] \times_2 [X_2^k] \times_3 [X_3^k]\right)_{[(1)]}$, and parameter $\tilde{\alpha}_X = \frac{1}{1+\alpha_X}$.

Similarly, we have

$$\begin{aligned}
[X_2^{k+1}] &\in \underset{[X_2]\in[\mathbb{S}_{m_2,n_2}]}{\operatorname{argmin}} F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_2] - [X_2^k]\|_F^2 \\
&= \operatorname{Proj}_{[\mathbb{S}_{m_2,n_2}]}\left([X_2^k] - \tilde{\alpha}_X\left([X_2^k] - [P_2^k][Q_2^k]^\top\right)\right),
\end{aligned}$$

and

$$\begin{aligned}
[X_3^{k+1}] &\in \underset{[X_3]\in[\mathbb{S}_{m_3,n_3}]}{\operatorname{argmin}} F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3], [\mathcal{G}^k], \mathcal{L}^k) + \frac{\alpha_X}{2}\|[X_3] - [X_3^k]\|_F^2 \\
&= \operatorname{Proj}_{[\mathbb{S}_{m_3,n_3}]}\left([X_3^k] - \tilde{\alpha}_X\left([X_3^k] - [P_3^k][Q_3^k]^\top\right)\right),
\end{aligned}$$

where $[P_2^k] = (\mathbf{R}(\mathcal{L}^k))_{[(2)]}$, $[P_3^k] = (\mathbf{R}(\mathcal{L}^k))_{[(3)]}$, $[Q_2^k] = \left([\mathcal{G}^k] \times_1 [X_1^{k+1}] \times_3 [X_3^k]\right)_{[(2)]}$, $[Q_3^k] = \left([\mathcal{G}^k] \times_1 [X_1^{k+1}] \times_2 [X_2^{k+1}]\right)_{[(3)]}$, and parameter $\tilde{\alpha}_X = \frac{1}{1+\alpha_X}$.

In summary, $[X_i^{k+1}]$ can be computed by

$$[X_i^{k+1}] \in \mathrm{Proj}_{[\mathbb{S}_{m_i,n_i}]}\left([X_i^k] - \tilde{\alpha}_X\left([X_i^k] - [P_i^k][Q_i^k]^\top\right)\right), \tag{3.4}$$

where $[P_i^k] = (\mathbf{R}(\mathcal{L}^k))_{[(i)]}$, $[Q_i^k] = \left([\mathcal{G}^k] \times_1 [X_1^{k+1}] \cdots \times_{i-1} [X_{i-1}^{k+1}] \times_{i+1} [X_{i+1}^k] \cdots \times_3 [X_3^k]\right)_{[(i)]}$, and parameter $\tilde{\alpha}_X = \frac{1}{1+\alpha_X}$.

In the following, we present a lemma for finding the projection onto a Stiefel manifold, which is given in Theorem 4.1 in [31] and proved in [3].

Lemma 3.1. *[31, 3] Given $A \in \mathbb{R}^{m\times n}$, $m \geq n$, consider the following Stiefel manifold projection problem*

$$\begin{aligned} &\min_{X\in\mathbb{R}^{m\times n}} \|X - A\|_F^2 \\ &\text{s.t.} \quad X^\top X = \mathbf{I}_n. \end{aligned} \tag{3.5}$$

Then the set of optimal solutions of problem (3.5)*, denoted as $\Omega^*(A)$, is given by*

$$\begin{aligned} \Omega^*(A) = \{UV^\top | A = U\Sigma V^\top, U \in \mathbb{R}^{m\times n}, \Sigma \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{n\times n} \\ \text{such that } U^\top U = V^\top V = \mathbf{I}_n \text{ and } \Sigma = \mathrm{Diag}(\sigma(A))\}, \end{aligned}$$

where $U\Sigma V^\top$ is a reduced SVD of A and $\sigma(A) \in \mathbb{R}^n$ is a vector of all the singular values of A. In particular, if A is of full column rank n, then $\Omega^*(A)$ is a singleton.

According to Lemma 3.1, problem (3.5) has a closed form solution, even though it may have multiple solutions when the given matrix does not have full column rank. Hence, if $[U][\Sigma][V]^\top$ is an independent reduced SVD of $[A] \in \mathbb{R}^{m\times n\times N}$, then $[U][V]^\top \in \mathrm{Proj}_{[\mathbb{S}_{m,n}]}([A])$.

3.1.3 The Update of $[\mathcal{G}]$

The subproblem for updating $[\mathcal{G}]$ can be reformulated by using the following property for any $X \in \mathbb{S}_{m,n}$, $m \geq n$,

$$\begin{aligned}
&\|\mathcal{G} \times_i X - \mathcal{L}\|_F^2 \\
=&\|G_{(i)} - X^\top L_{(i)}\|_F^2 - \|X^\top L_{(i)}\|_F^2 + \|L_{(i)}\|_F^2 \\
=&\|\mathcal{G} - \mathcal{L} \times_i X^\top\|_F^2 - \|\mathcal{L} \times_i X^\top\|_F^2 + \|\mathcal{L}\|_F^2,
\end{aligned}$$

which is derived from (3.3) and the constraint that $X^\top X = \mathbf{I}_n$.

Since we have

$$\begin{aligned}
&F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}], \mathcal{L}^k) + \frac{\alpha_\mathcal{G}}{2}\|[\mathcal{G}] - [\mathcal{G}^k]\|_F^2 \\
=&\frac{1}{2}\|\mathbf{R}(\mathcal{L}^k) - [\mathcal{G}] \times_1 [X_1^{k+1}] \times_2 [X_2^{k+1}] \times_3 [X_3^{k+1}]\|_F^2 + \Phi([\mathcal{G}]) + \frac{\alpha_\mathcal{G}}{2}\|[\mathcal{G}] - [\mathcal{G}^k]\|_F^2,
\end{aligned}$$

$[\mathcal{G}^{k+1}]$ can be computed by

$$\begin{aligned}
[\mathcal{G}^{k+1}] &= \operatorname*{argmin}_{[\mathcal{G}]} \frac{1}{2}\|[\mathcal{G}] - \mathbf{R}(\mathcal{L}^k) \times_1 [X_1^{k+1}]^\top \times_2 [X_2^{k+1}]^\top \times_3 [X_3^{k+1}]^\top\|_F^2 + \Phi([\mathcal{G}]) \\
&\qquad + \frac{\alpha_\mathcal{G}}{2}\|[\mathcal{G}] - [\mathcal{G}^k]\|_F^2 \\
&= \operatorname*{argmin}_{[\mathcal{G}]} \frac{1}{2}\|[\mathcal{G}] - ([\mathcal{G}^k] - \tilde{\alpha}_\mathcal{G}([\mathcal{G}^k] - [\mathcal{O}^k]))\|_F^2 + \tilde{\alpha}_\mathcal{G}\Phi([\mathcal{G}]) \\
&= \operatorname{prox}_{\tilde{\alpha}_\mathcal{G}\Phi}\left([\mathcal{G}^k] - \tilde{\alpha}_\mathcal{G}([\mathcal{G}^k] - [\mathcal{O}^k])\right), \qquad (3.6)
\end{aligned}$$

where $[\mathcal{O}^k] = \mathbf{R}(\mathcal{L}^k) \times_1 [X_1^{k+1}]^\top \times_2 [X_2^{k+1}]^\top \times_3 [X_3^{k+1}]^\top$, and $\tilde{\alpha}_\mathcal{G} = \frac{1}{1+\alpha_\mathcal{G}}$.

3.1.4 The Update of $\mathcal{L}$

After computing $[X_i^{k+1}]$ and $[\mathcal{G}^{k+1}]$, we can obtain the approximated low-rank group tensor, denoted as $[\mathcal{Y}^{k+1}]$, as follows

$$[\mathcal{Y}^{k+1}] = [\mathcal{G}^{k+1}] \times_1 [X_1^{k+1}] \times_2 [X_2^{k+1}] \times_3 [X_3^{k+1}]. \qquad (3.7)$$

Since we have

$$\begin{aligned}
&F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^{k+1}], \mathcal{L}) \\
=&\frac{\delta}{2}\|\mathbf{R}(\mathcal{L}+\mathcal{S}^{k+1}-\mathcal{D})\|_F^2 + \frac{1}{2}H([X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^{k+1}], \mathcal{L}) \\
=&\frac{\delta}{2}\|\mathbf{R}(\mathcal{L}+\mathcal{S}^{k+1}-\mathcal{D})\|_F^2 + \frac{1}{2}\|\mathbf{R}(\mathcal{L}) - [\mathcal{Y}^{k+1}]\|_F^2,
\end{aligned}$$

$\mathcal{L}^{k+1}$ can be computed in a unique closed form as follows

$$\begin{aligned}
\mathcal{L}^{k+1} = \underset{\mathcal{L}}{\operatorname{argmin}}\ &\frac{\delta}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{L}+\mathcal{S}^{k+1}-\mathcal{D})\|_F^2 \\
&+ \frac{1}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{L} - (\sqrt{\mathcal{W}_{\mathbf{R}}})^{-1} \odot \mathbf{R}^\top([\mathcal{Y}^{k+1}])\|_F^2 \\
= \tilde{\delta}\mathcal{W}_{\mathbf{R}}^{-1}&\mathbf{R}^\top([\mathcal{Y}^{k+1}]) + (1-\tilde{\delta})(\mathcal{D}-\mathcal{S}^{k+1}),
\end{aligned} \tag{3.8}$$

where $\tilde{\delta} = \frac{1}{1+\delta}$ and $\mathcal{W}_{\mathbf{R}}^{-1}$ denotes the component-wise inverse of $\mathcal{W}_{\mathbf{R}}$, that is, the (i_1, i_2, i_3)-th entry of $\mathcal{W}_{\mathbf{R}}^{-1}$ is equal to the reciprocal of $(\mathcal{W}_{\mathbf{R}})_{i_1 i_2 i_3}$.

3.1.5 Summary of the P-BCD Algorithm

The proposed P-BCD algorithm for model (2.1) is summarized in Algorithm 1.

Algorithm 1 Proximal BCD (P-BCD) algorithm for model (2.1)

1: Initialize $(\mathcal{S}^0, [X_1^0], [X_2^0], [X_3^0], [\mathcal{G}^0], \mathcal{L}^0)$ with $[X_i^0] \in [\mathbb{S}_{m_i,n_i}]$.
2: Set the tensor extraction operator $\mathbf{R}$.
3: Set parameters $\alpha_{\mathcal{S}}, \alpha_X, \alpha_{\mathcal{G}} > 0$.
4: Set $k = 0$.
5: **repeat**
6: Compute $\mathcal{S}^{k+1}$ by (3.2).
7: Compute $[X_i^{k+1}]$ by (3.4), $i = 1, 2, 3$.
8: Compute $[\mathcal{G}^{k+1}]$ by (3.6).
9: Compute $\mathcal{L}^{k+1}$ by (3.8).
10: $k \leftarrow k + 1$.
11: **until** the stopping criterion is met.
Output: $(\mathcal{S}^k, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k)$.

The proposed P-BCD algorithm aims to solve a particular optimization problem of the form as in (2.1) and an optimal solution for each subproblem is obtained in our algorithm as shown in the previous subsections. The P-BCD algorithm can be viewed as a special variant of the proximal alternating linearization minimization (PALM) [6] extended for multiple blocks or the block coordinate update with prox-linear approximation [95], called the block prox-linear method. In the next section, we present the convergence results of the P-BCD algorithm.

3.2 Convergence Analysis of the P-BCD Algorithm

In this section, we first define the first order optimality condition of problem (2.1), then prove that any accumulation point is a first order stationary point.

3.2.1 The First Order Optimality Condition

Let $\mathcal{Z} := (\mathcal{S}, [X_1], [X_2], [X_3], [\mathcal{G}], \mathcal{L})$. We define the first order optimality condition of the orthogonal constrained optimization problem (2.1). The point $\bar{\mathcal{Z}} := (\bar{\mathcal{S}}, [\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}})$ is a first order stationary point of problem (2.1) if $0 \in \partial F(\bar{\mathcal{Z}})$, that is,

$$
\begin{aligned}
&0 \in \delta(\bar{\mathcal{L}} + \bar{\mathcal{S}} - \mathcal{D}) + \gamma \partial \|\cdot\|_{2,\psi}(\bar{\mathcal{S}}),\\
&0 = \operatorname{grad}_{[X_i]} H([\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}}), \quad [\bar{X}_i]^\top [\bar{X}_i] = [\mathbf{I}_{n_i}], i = 1, 2, 3, \qquad (3.9)\\
&0 \in \nabla_{[\mathcal{G}]} H([\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}}) + \partial \Phi([\bar{\mathcal{G}}]),\\
&0 = \delta(\bar{\mathcal{L}} + \bar{\mathcal{S}} - \mathcal{D}) + \nabla_{\mathcal{L}} H([\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}}),
\end{aligned}
$$

where $\operatorname{grad}_{[X_i]} H([\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}})$ denotes the Riemannian gradient of H with respect to $[X_i]$ evaluated at $([\bar{X}_1], [\bar{X}_2], [\bar{X}_3], [\bar{\mathcal{G}}], \bar{\mathcal{L}})$, $i = 1, 2, 3$, and $\partial \|\cdot\|_{2,\psi}$ and $\partial \Phi$ denote the subdifferentials of $\|\cdot\|_{2,\psi}$ and Φ, respectively. Then we compute the gradients of H explicitly and replace the optimality condition for orthogonal constraints as in (3.9) using an equivalent condition introduced in [25]. Hence, we

call $\bar{\mathcal{Z}}$ is a first order stationary point of problem (2.1) if

$$0 \in \delta(\bar{\mathcal{L}} + \bar{\mathcal{S}} - \mathcal{D}) + \gamma \left(\sqrt{\mathcal{W}_{\mathbf{R}}}\right)^{-1} \odot \partial \| \cdot \|_{2,\psi}(\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \bar{\mathcal{S}}), \tag{3.10a}$$

$$0 = ([\mathbf{I}_{m_i}] - [\bar{X}_i][\bar{X}_i]^\top)[\bar{H}_i], \tag{3.10b}$$

$$0 = [\bar{H}_i]^\top[\bar{X}_i] - [\bar{X}_i]^\top[\bar{H}_i], \tag{3.10c}$$

$$[\bar{X}_i]^\top[\bar{X}_i] = [\mathbf{I}_{n_i}], \tag{3.10d}$$

$$0 \in [\bar{\mathcal{G}}] - [\bar{\mathcal{O}}] + \partial\Phi([\bar{\mathcal{G}}]), \tag{3.10e}$$

$$0 = \delta(\bar{\mathcal{L}} + \bar{\mathcal{S}} - \mathcal{D}) + \bar{\mathcal{L}} - \mathcal{W}_{\mathbf{R}}^{-1} \odot \mathbf{R}^\top([\bar{\mathcal{Y}}]), \tag{3.10f}$$

where $i = 1, 2, 3$, and

$$\begin{aligned}
[\bar{P}_i] =& (\mathbf{R}(\bar{\mathcal{L}}))_{[(i)]}, \\
[\bar{Q}_1] =& ([\bar{\mathcal{G}}] \times_2 [\bar{X}_2] \times_3 [\bar{X}_3])_{[(1)]}, \\
[\bar{Q}_2] =& ([\bar{\mathcal{G}}] \times_1 [\bar{X}_1] \times_3 [\bar{X}_3])_{[(2)]}, \\
[\bar{Q}_3] =& ([\bar{\mathcal{G}}] \times_1 [\bar{X}_1] \times_2 [\bar{X}_2])_{[(3)]}, \\
[\bar{H}_i] =& ([\bar{X}_i][\bar{Q}_i] - [\bar{P}_i])[\bar{Q}_i]^\top, \\
[\bar{\mathcal{O}}] =& \mathbf{R}(\bar{\mathcal{L}}) \times_1 [\bar{X}_1]^\top \times_2 [\bar{X}_2]^\top \times_3 [\bar{X}_3]^\top, \\
[\bar{\mathcal{Y}}] =& [\bar{\mathcal{G}}] \times_1 [\bar{X}_1] \times_2 [\bar{X}_2] \times_3 [\bar{X}_3].
\end{aligned}$$

3.2.2 Non-increasing Monotonicity

Next, we prove the non-increasing monotonicity of the objective sequence $\{F(\mathcal{Z}^k)\}$ and the boundedness of the sequence $\{\mathcal{Z}^k\}$ generated by Algorithm 1.

Theorem 3.1. *Let $\{\mathcal{Z}^k\}$ be the sequence generated by Algorithm 1. Then the following statements hold.*

(i) The sequence $\{F(\mathcal{Z}^k)\}$ of function values at the iteration points decreases

monotonically, and

$$\begin{aligned} &F(\mathcal{Z}^k) - F(\mathcal{Z}^{k+1}) \\ \geq &\frac{\alpha}{2}\|\mathcal{S}^{k+1} - \mathcal{S}^k\|_F^2 + \frac{\alpha}{2}\sum_{i=1}^{3}\|[X_i^{k+1}] - [X_i^k]\|_F^2 + \frac{\alpha}{2}\|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F^2. \end{aligned} \tag{3.11}$$

(ii) The sequence $\{\mathcal{Z}^k\}$ is bounded.

(iii) $\lim_{k\to\infty}\|\mathcal{S}^{k+1}-\mathcal{S}^k\|_F = 0$, $\lim_{k\to\infty}\|[X_i^{k+1}]-[X_i^k]\|_F = 0$, *and* $\lim_{k\to\infty}\|[\mathcal{G}^{k+1}]-[\mathcal{G}^k]\|_F = 0$, *for any* $i = 1, 2, 3$.

Proof. (i) According to the update of $\mathcal{S}$, we have

$$\begin{aligned} &F(\mathcal{Z}^k) - F(\mathcal{S}^{k+1}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) \\ =&F(\mathcal{S}^k, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) - F(\mathcal{S}^{k+1}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) \\ \geq &\frac{\alpha_{\mathcal{S}}}{2}\|\sqrt{\mathcal{W}_{\mathbf{R}}} \odot (\mathcal{S}^{k+1} - \mathcal{S}^k)\|_F^2 \\ \geq &\frac{c_1^2\alpha_{\mathcal{S}}}{2}\|\mathcal{S}^{k+1} - \mathcal{S}^k\|_F^2, \end{aligned}$$

where $c_1 = \min(\sqrt{\mathcal{W}_{\mathbf{R}}})_{i_1 i_2 i_3}$.

Next, it follows from the update of $[X_i]$ and Lemma 3.1 that

$$\begin{aligned} &F(\mathcal{S}^{k+1}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) - F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^k], \mathcal{L}^k) \\ =&F(\mathcal{S}^{k+1}, [X_1^k], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) - F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) \\ &+ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^k], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) - F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) \\ &+ F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^k], [\mathcal{G}^k], \mathcal{L}^k) - F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^k], \mathcal{L}^k) \\ \geq &\frac{\alpha_X}{2}\sum_{i=1}^{3}\|[X_i^{k+1}] - [X_i^k]\|_F^2. \end{aligned}$$

Then by the updates of $[\mathcal{G}]$ and $\mathcal{L}$, we have

$$\begin{aligned} &F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^k], \mathcal{L}^k) \\ &- F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^{k+1}], \mathcal{L}^k) \geq \frac{\alpha_{\mathcal{G}}}{2}\|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F^2 \end{aligned}$$

and

$$F(\mathcal{S}^{k+1}, [X_1^{k+1}], [X_2^{k+1}], [X_3^{k+1}], [\mathcal{G}^{k+1}], \mathcal{L}^k) - F(\mathcal{Z}^{k+1}) \geq 0.$$

Combining the inequalities above, we obtain (3.11) with $\alpha = \min\{c_1^2\alpha_{\mathcal{S}}, \alpha_{\mathcal{G}}, \alpha_X\}$.

(ii) Since $[X_i^k]^\top[X_i^k] = [\mathbf{I}_{n_i}]$ for each $i = 1, 2, 3$, we have the sequence $\{[X_i^k]\}$ is bounded. By (i), we have $F(\mathcal{Z}^k) \leq F(\mathcal{Z}^0)$. Also, we observe that $F(\mathcal{Z}^k) \geq \gamma c_1^p \|\mathcal{S}^k\|_{2,\psi} + \Phi([\mathcal{G}^k]) \geq 0$. Since

$$\lim_{\|\mathcal{S}\|_F \to \infty} \|\mathcal{S}\|_{2,\psi} = \infty \quad \text{and} \quad \lim_{\|[\mathcal{G}]\|_F \to \infty} \Phi([\mathcal{G}]) = \infty,$$

we must have the sequences $\{\mathcal{S}^k\}$ and $\{[\mathcal{G}^k]\}$ are bounded. As shown in (3.8) that $\mathcal{L}^k$ is uniquely determined by $\mathcal{S}^k, [X_1^k], [X_2^k], [X_3^k]$ and $[\mathcal{G}^k]$, the sequence $\{\mathcal{L}^k\}$ is also bounded.

(iii) Let K be an arbitrary integer. Summing (3.11) from $k = 1$ to $K - 1$, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \|\mathcal{S}^{k+1} - \mathcal{S}^k\|_F^2 + \sum_{k=0}^{K-1} \sum_{i=1}^{3} \|[X_i^{k+1}] - [X_i^k]\|_F^2 + \sum_{k=0}^{K-1} \|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F^2 \\
\leq &\frac{2}{\alpha} \left(F(\mathcal{Z}^0) - F(\mathcal{Z}^K)\right) \\
\leq &\frac{2}{\alpha} F(\mathcal{Z}^0).
\end{aligned}$$

Taking the limits of both sides of the inequality as $K \to \infty$, we have $\sum_{k=0}^{\infty} \|\mathcal{S}^{k+1} - \mathcal{S}^k\|_F^2 < \infty$, $\sum_{k=0}^{\infty} \|[X_i^{k+1}] - [X_i^k]\|_F^2 < \infty$ and $\sum_{k=0}^{\infty} \|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F^2 < \infty$. Then assertion (iii) immediately holds. □

In addition to the assertions presented in Theorem 3.1, more assertions can be derived in the following corollary.

Corollary 3.1. *Let $\{\mathcal{Z}^k\}$ be the sequence generated by Algorithm 1. Then $\lim_{k\to\infty} \|[\mathcal{Y}^{k+1}] - [\mathcal{Y}^k]\|_F = 0$ and $\lim_{k\to\infty} \|\mathcal{L}^{k+1} - \mathcal{L}^k\|_F = 0$.*

Proof. Since $\|[\mathcal{G}^{k+1}] \times_i [X_i] - [\mathcal{G}^k] \times_i [X_i]\|_F = \|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F$ for any $[X_i] \in [\mathbb{S}_{m_i,n_i}]$ and $\|[\mathcal{G}] \times_i [X_i^{k+1}] - [\mathcal{G}] \times_i [X_i^k]\|_F \leq \|[\mathcal{G}]\|_F \|[X_i^{k+1}] - [X_i^k]\|_F$, we have

$$\begin{aligned}\|[\mathcal{Y}^{k+1}] - [\mathcal{Y}^k]\|_F \leq & \|[\mathcal{G}^{k+1}] - [\mathcal{G}^k]\|_F + c_2\|[X_1^{k+1}] - [X_1^k]\|_F \\ & + c_2\|[X_2^{k+1}] - [X_2^k]\|_F + c_2\|[X_3^{k+1}] - [X_3^k]\|_F,\end{aligned}$$

where $c_2 = \max_k \|[\mathcal{G}^k]\|_F < \infty$ according to assertion (ii) in Theorem 3.1. Then it immediately follows from assertion (iii) in Theorem 3.1, $\lim_{k\to\infty} \|[\mathcal{Y}^{k+1}] - [\mathcal{Y}^k]\|_F = 0$.

Also, we have

$$\|\mathcal{L}^{k+1} - \mathcal{L}^k\|_F \leq \tilde{\delta} c_1^2 c_3 \|[\mathcal{Y}^{k+1}] - [\mathcal{Y}^k]\|_F + (1 - \tilde{\delta})\|\mathcal{S}^{k+1} - \mathcal{S}^k\|_F,$$

where $c_3 = \max(\sqrt{\mathcal{W}_{\mathbf{R}}})_{i_1 i_2 i_3}$. Then by assertion (iii) in Theorem 3.1, we have $\lim_{k\to\infty} \|\mathcal{L}^{k+1} - \mathcal{L}^k\|_F = 0$. □

3.2.3 Substationarity, Symmetry, and Feasibility for $[X_i]$

We apply the results in [25] to the updates of $[X_i]$ in our proposed algorithm. Those results are useful for proving three equalities of substationarity, symmetry, and feasibility for $[X_i]$. According to Lemma 3.3 in [25], we can have the following lemma and then we prove the decrease of the function value H after each update of $[X_i]$.

Lemma 3.2. *[25] Let $h : \mathbb{R}^{m\times n} \to \mathbb{R}$ be defined by $h(X) = \frac{1}{2}\|X - PQ^\top\|_F^2$, where $Q \in \mathbb{R}^{n\times m}$, $P \in \mathbb{R}^{m\times m}$ and $m \geq n$. If $X \in \mathbb{S}_{m,n}$ and*

$$\bar{X} = \mathrm{Proj}_{\mathbb{S}_{m,n}}(X - \tau(X - PQ^\top)), \tag{3.12}$$

where $\tau \in (0, 1)$, then we have $\bar{X} \in \mathbb{S}_{m,n}$ and

$$h(X) - h(\bar{X}) \geq \frac{\tau^{-1} - 1}{2(\tau^{-1} + 1 + \theta)^2}\|(\mathbf{I}_m - XX^\top)\nabla h(X)\|_F^2, \tag{3.13}$$

where $\theta = \|PQ^\top\|_2$.

Proposition 3.1. *Let H be defined as in* (3.1). *Let $\{\mathcal{Z}^k\}$ be the sequence generated by Algorithm 1 and $[X_i^{k+1}]$ be computed by* (3.4). *Then $[X_i^{k+1}] \in [\mathbb{S}_{m_i,n_i}]$ and the following inequality holds:*

$$\begin{aligned} &H([X_1^k],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k) - H([X_1^{k+1}],[X_2^{k+1}],[X_3^{k+1}],[\mathcal{G}^k],\mathcal{L}^k) \\ \geq & c_4 \sum_{i=1}^{3} \left\| ([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top)[H_i^k] \right\|_F^2 \end{aligned} \tag{3.14}$$

and

$$\left\| [H_i^k]^\top [X_i^k] - [X_i^k]^\top [H_i^k] \right\|_F \leq c_5 \| [X_{i+1}^k] - [X_i^k] \|_F, \tag{3.15}$$

where $c_4 = \frac{\alpha_X}{2(\alpha_X+2+\theta_{\max})^2}$, $\theta_{\max} = \max_{ijk} \|[P_i^k]^{(j)}([Q_i^k]^{(j)})^\top\|$, $c_5 = 2(\alpha_X\sqrt{n_{\max}}+\theta_{\max})$, $n_{\max} = \max\{n_1,n_2,n_3\}$, *and*

$$\begin{aligned} [H_1^k] &= \nabla_{[X_1]} H([X_1^k],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k), \\ [H_2^k] &= \nabla_{[X_2]} H([X_1^{k+1}],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k), \\ [H_3^k] &= \nabla_{[X_3]} H([X_1^{k+1}],[X_2^{k+1}],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k), \end{aligned}$$

which give

$$[H_i^k] = ([X_i^k][Q_i^k] - [P_i^k])[Q_i^k]^\top, \quad i = 1,2,3.$$

Proof. It follows from Lemma 3.2 that $[X_i^{k+1}] \in \mathbb{S}_{m_i,n_i}$.

To show the first inequality holds, the update of $[X_i^{k+1}]$ in (3.4) can be viewed as the iteration in (3.12) with $h([X]) = \frac{1}{2}\|[X] - [P_i^k][Q_i^k]^\top\|_F^2$, $[X] = [X_i^k]$, $[\bar{X}] = [X_i^{k+1}]$

and $\tau = \tilde{\alpha}_X = \frac{1}{1+\alpha_X}$. Then for $i = 1$ we have

$$\begin{aligned}
&H([X_1^k],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k) - H([X_1^{k+1}],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k)\\
=&\frac{1}{2}\|[X_i^{k+1}][Q_i^k] - [P_i^k]\|_F^2 - \frac{1}{2}\|[X_i^k][Q_i^k] - [P_i^k]\|_F^2\\
=&\frac{1}{2}\|[X_i^{k+1}] - [P_i^k][Q_i^k]^\top\|_F^2 - \frac{1}{2}\|[X_i^k] - [P_i^k][Q_i^k]^\top\|_F^2\\
\geq& c_{ij}^k \left\| \left([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top\right)\left([X_i^k] - [P_i^k][Q_i^k]^\top\right)\right\|_F^2,
\end{aligned}$$

where $c_{ij}^k = \frac{\tilde{\alpha}_X^{-1}-1}{2(\tilde{\alpha}_X^{-1}+1+\|[P_i^k]^{(j)}([Q_i^k]^{(j)})^\top\|)^2} \geq c_4$ and $\theta_{\max}$ is bounded, since the sequence $\{\mathcal{Z}^k\}$ is bounded. Using $[X_i^k] \in [\mathbb{S}_{m_i,n_i}]$ and $[X_i^{k+1}] \in [\mathbb{S}_{m_i,n_i}]$, we can obtain the fourth line above and rewrite part of the last line as follows

$$\begin{aligned}
&\left([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top\right)\left([X_i^k] - [P_i^k][Q_i^k]^\top\right)\\
=&\left([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top\right)\left([X_i^k][Q_i^k] - [P_i^k]\right)[Q_i^k]^\top\\
=&\left([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top\right)[H_i^k].
\end{aligned}$$

That is, we have

$$\begin{aligned}
&H([X_1^k],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k) - H([X_1^{k+1}],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k)\\
\geq& c_4 \left\| \left([\mathbf{I}_{m_1}] - [X_1^k][X_1^k]^\top\right)[H_1^k]\right\|_F^2.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&H([X_1^{k+1}],[X_2^k],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k) - H([X_1^{k+1}],[X_2^{k+1}],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k)\\
\geq& c_4 \left\| \left([\mathbf{I}_{m_2}] - [X_2^k][X_2^k]^\top\right)[H_2^k]\right\|_F^2
\end{aligned}$$

and

$$\begin{aligned}
&H([X_1^{k+1}],[X_2^{k+1}],[X_3^k],[\mathcal{G}^k],\mathcal{L}^k) - H([X_1^{k+1}],[X_2^{k+1}],[X_3^{k+1}],[\mathcal{G}^k],\mathcal{L}^k)\\
\geq& c_4 \left\| \left([\mathbf{I}_{m_3}] - [X_3^k][X_3^k]^\top\right)[H_3^k]\right\|_F^2.
\end{aligned}$$

Summing the inequalities above, (3.14) immediately holds.

Next, we show the second inequality holds. According to the update of $[X_i^{k+1}]$ in (3.4), if we let

$$[U^k][\Sigma^k][V^k]^\top = [X_i^k] - \tilde{\alpha}_X \left([X_i^k] - [P_i^k][Q_i^k]^\top\right) = \frac{1}{\alpha_X + 1}\left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right),$$

then the update $[X_i^{k+1}] = [U^k][V^k]^\top$. Hence, we have

$$\begin{aligned}[X_i^{k+1}]^\top \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right) &= \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right)^\top [X_i^{k+1}] \\ &= (\alpha_X + 1)[V]^k[\Sigma^k][V^k]^\top.\end{aligned} \tag{3.16}$$

Then using $[X_i^k] \in [\mathbb{S}_{m_i,n_i}]$ and $[X_i^{k+1}] \in [\mathbb{S}_{m_i,n_i}]$, we can rewrite

$$\begin{aligned}&[H_i^k]^\top[X_i^k] - [X_i^k]^\top[H_i^k] \\ =&\left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right)^\top [X_i^k] - [X_i^k]^\top \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right) \\ =&\left([X_i^{k+1}] - [X_i^k]\right)^\top \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right) \\ &- \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right)^\top \left([X_i^{k+1}] - [X_i^k]\right),\end{aligned}$$

where the last equation is obtained by (3.16). Taking the Frobenius norm of both sides, we obtain

$$\begin{aligned}&\|[H_i^k]^\top[X_i^k] - [X_i^k]^\top[H_i^k]\|_F \\ =&\| \left([X_i^{k+1}] - [X_i^k]\right)^\top \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right) \\ &- \left(\alpha_X[X_i^k] + [P_i^k][Q_i^k]^\top\right)^\top \left([X_i^{k+1}] - [X_i^k]\right) \|_F \\ \leq& 2\|\alpha_X[X_i^k]^{(j)} + [P_i^k]^{(j)}([Q_i^k]^{(j)})^\top\|_F \|[X_i^{k+1}] - [X_i^k]\|_F \\ \leq& c_5\|[X_i^{k+1}] - [X_i^k]\|_F,\end{aligned}$$

where $2\|\alpha_X[X_i^k]^{(j)} + [P_i^k]^{(j)}([Q_i^k]^{(j)})^\top\|_F \leq c_5$. Then (3.15) holds. □

3.2.4 Subsequence Convergence

Lastly, we show that every convergent subsequence converges to a first order stationary point of problem (2.1).

Theorem 3.2. *Let $\{\mathcal{Z}^k\}$ be the sequence generated by Algorithm 1. Then every accumulation point of $\{\mathcal{Z}^k\}$ is a first order stationary point of problem* (2.1)*.*

Proof. Suppose that $\{\mathcal{Z}^k\}_{k\in\mathcal{K}}$ is a convergent subsequence of $\{\mathcal{Z}^k\}$ and converges to $\bar{\mathcal{Z}}$ as $k \in \mathcal{K}$ approaches ∞. By the updates of $\mathcal{S}^k$, $[\mathcal{G}^k]$ and $\mathcal{L}^k$, we have for any $k = 0, 1, \ldots$

$$\begin{aligned}
&\alpha_{\mathcal{S}}(\mathcal{S}^{k-1} - \mathcal{S}^k) + \delta(\mathcal{L}^k - \mathcal{L}^{k-1}) \\
&\in \delta(\mathcal{L}^k + \mathcal{S}^k - \mathcal{D}) + \gamma \left(\sqrt{\mathcal{W}_{\mathbf{R}}}\right)^{-1} \odot \partial\|\cdot\|_{2,\psi}(\sqrt{\mathcal{W}_{\mathbf{R}}} \odot \mathcal{S}^k), \\
&\alpha_{\mathcal{G}}([\mathcal{G}^{k-1}] - [\mathcal{G}^k]) + ([\mathcal{O}^k] - [\tilde{\mathcal{O}}^k]) \\
&\in [\mathcal{G}^k] - [\tilde{\mathcal{O}}^k] + \partial\Phi([\mathcal{G}^k]),
\end{aligned}$$

and

$$0 = \delta(\mathcal{L}^k + \mathcal{S}^k - \mathcal{D}) + \mathcal{L}^k - (\mathcal{W}_{\mathbf{R}})^{-1}\mathbf{R}^\top([\mathcal{Y}^k]),$$

where $[\tilde{\mathcal{O}}^k] = \mathbf{R}(\mathcal{L}^k) \times_1 [X_1^k]^\top \times_2 [X_2^k]^\top \times_3 [X_3^k]^\top$. Since we have

$$\begin{aligned}
\|[\mathcal{O}^k] - [\tilde{\mathcal{O}}^k]\|_F &\le c_3\|[X_1^k]\|_F\|[X_2^k]\|_F\|[X_3^k]\|_F\|\mathcal{L}^k - \mathcal{L}^{k-1}\|_F \\
&\le c_3\sqrt{n_1n_2n_3N^3}\|\mathcal{L}^k - \mathcal{L}^{k-1}\|_F,
\end{aligned}$$

it immediately follows from Corollary 3.1 that $\lim_{k\to\infty}\|[\mathcal{O}^k] - [\tilde{\mathcal{O}}^k]\|_F = 0$.

According to the definition of limiting subdifferential and the fact that $\|\cdot\|_{2,\psi}$ and Φ are continuous functions, we can take the limits of the relations above as $k \in \mathcal{K}$ approaches ∞. Note that $\mathcal{Z}^k \to \bar{\mathcal{Z}}$, $[\tilde{\mathcal{O}}^k] \to [\bar{\mathcal{O}}]$ and $[\mathcal{Y}^k] \to [\bar{\mathcal{Y}}]$, as $k \in \mathcal{K}$ approaches ∞. Together using Theorem 3.1 (iii) and Corollary 3.1, we have (3.10a), (3.10e) and (3.10f) hold.

Next, we show (3.10b)-(3.10d) hold. Using (3.14) and the inequalities in the proof for assertion (i) of Theorem 3.1, we have

$$F(\mathcal{Z}^k) - F(\mathcal{Z}^{k+1}) \geq c_4 \sum_{i=1}^{3} \left\| ([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top)[H_i^k] \right\|_F^2$$

and further obtain

$$\sum_{k=0}^{\infty} \sum_{i=1}^{3} \left\| ([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top)[H_i^k] \right\|_F^2 \leq \frac{1}{c_4} F(\mathcal{Z}^0).$$

This implies

$$\lim_{k\to\infty} \|([\mathbf{I}_{m_i}] - [X_i^k][X_i^k]^\top)[H_i^k]\|_F = 0. \tag{3.17}$$

And by taking the limit of both sides of (3.15) and using Theorem 3.1 (iii), we have

$$\lim_{k\to\infty} \left\| [H_i^k]^\top [X_i^k] - [X_i^k]^\top [H_i^k] \right\|_F = 0. \tag{3.18}$$

Since $\mathcal{Z}^k \to \bar{\mathcal{Z}}$ as $k \in \mathcal{K}$ approaches ∞, we have $[P_i^k] \to [\bar{P}_i]$, $[Q_1^k] \to [\bar{Q}_1]$, as $k \in \mathcal{K}$ approaches ∞. Also, we have

$$\begin{aligned} \|[Q_2^k] - [\bar{Q}_2]\|_F &\leq \|[Q_2^k] - [\tilde{Q}_2^k]\|_F + \|[\tilde{Q}_2^k] - [\bar{Q}_2]\|_F \\ &\leq c_2 \|[X_1^{k+1}] - [X_1^k]\|_F + \|[\tilde{Q}_2^k] - [\bar{Q}_2]\|_F \end{aligned}$$

and

$$\begin{aligned} \|[Q_3^k] - [\bar{Q}_3]\|_F &\leq \|[Q_3^k] - [\tilde{Q}_3^k]\|_F + \|[\tilde{Q}_3^k] - [\bar{Q}_3]\|_F \\ &\leq c_2 \|[X_1^{k+1}] - [X_1^k]\|_F + c_2 \|[X_2^{k+1}] - [X_2^k]\|_F + \|[\tilde{Q}_3^k] - [\bar{Q}_3]\|_F, \end{aligned}$$

where $[\tilde{Q}_2^k] = \left([\mathcal{G}^k] \times_1 [X_1^k] \times_3 [X_3^k]\right)_{[(2)]}$ and $[\tilde{Q}_3^k] = \left([\mathcal{G}^k] \times_1 [X_1^k] \times_2 [X_2^k]\right)_{[(3)]}$. Since $[\tilde{Q}_i^k] \to [\bar{Q}_i]$, $i = 2, 3$, by Theorem 3.1 (iii), we have $[Q_i^k] \to [\bar{Q}_i]$, $i = 2, 3$, as $k \in \mathcal{K}$ approaches ∞. Hence, $[H_i^k] \to [\bar{H}_i]$, $i = 1, 2, 3$, as $k \in \mathcal{K}$ approaches ∞. By (3.17) and (3.18), we have (3.10b) and (3.10c) hold. Since the Stiefel manifold is a compact set, we also have (3.10d) holds. □

Chapter 4

Application to HSI Denoising and Destriping

4.1 Problem Statement

Mixed noise often appears in HSIs due to instrumental defects and environmental factors. The mixed noise typically includes stripe noise, dead line noise, Gaussian noise, salt-and-pepper noise and so on [103, 66]. In the following, we present the features of each type of noise.

Stripe noise has a special linear structure and directionality. The primary causes of stripe noise in HSIs are the working principles of imaging spectrometers, imaging instruments, and various external imaging environment factors. The core component of remote sensing imaging instruments is the charge-coupled device (CCD) detector element. In this array of elements, there are many linearly arranged detector elements. Due to the limitations of the imaging instrument's own performance or the influence of the external imaging environment, different detector units often show inconsistent responses to the same radiation energy during the scanning of ground objects, resulting in stripe noise pollution in the collected images. Figure 4.1a shows a specific band of stripe noise in an HSI obtained by the Hyperion hyperspectral sensor.

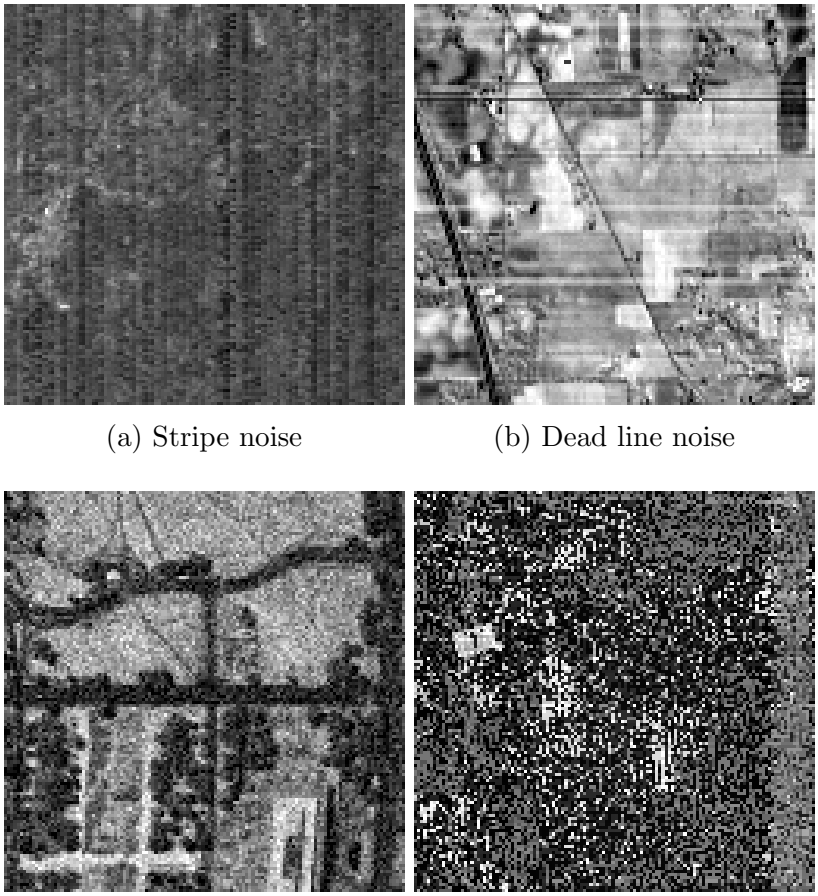
Dead line noise refers to a type of noise in imaging systems that occurs when there is a failure or malfunction in the detector elements or reading circuitry. This type of noise results in the formation of dead pixels or entire dead lines (either horizontal or vertical) in an image, where no data or incorrect data is recorded. Dead line noise typically appears as fully dark or fully bright lines, as shown in Figure 4.1b, where information is either missing or corrupted.

Gaussian noise generally refers to noise that follows a Gaussian distribution. During hyperspectral imaging, the CCD converts the captured electromagnetic energy into image information, which brings various forms of noise, including dark noise and readout noise. Since these types of noise frequently occur during hyperspectral imaging and usually satisfy a Gaussian distribution, researchers have defined this type of noise as Gaussian noise, assuming consistent noise levels across different bands. Figure 4.1c shows a specific band in an HSI obtained by the AVIRIS hyperspectral sensor degraded by Gaussian noise.

Salt-and-pepper noise, also known as impulse noise, refers to black noise points with minimum pixel values and white noise points with maximum pixel values in the image. During hyperspectral imaging, when there is significant electromagnetic interference causing abrupt signal changes or incorrect exposure of the photosensitive sheet, impulse noise appears in the obtained image. Figure 4.1d shows a specific band of mixed noise in an HSI obtained by the HYDICE hyperspectral sensor.

4.2 Model Formulation

In this section, we present an application of the proposed model (2.1) to HSI denoising and destriping and utilize the proposed P-BCD method given in Algorithm 1 for solving the model. In the following, we first introduce the nonlocal self-similarity of HSIs, that is, the choice of $\mathbf{R}$, present the sparsity-enhanced nonlocal low-rank



(a) Stripe noise (b) Dead line noise

(c) Gaussian noise (d) Salt-and-pepper noise

Figure 4.1: Four types of mixed noise in HSIs

tensor regularization for removing Gaussian noise, that is, the choice of Φ; and we mention the tensor $\ell_{2,p}$ norm for removing sparse noise with linear structures, that is, the choice of ψ for $\|\cdot\|_{2,\psi}$.

4.2.1 Nonlocal Self-similarity of HSIs

According to the image nonlocal self-similarity, image blocks share similar patterns. This prior is particularly evident in HSI, where multiple bands can be considered as multiple observations of the same scene under different wavelengths. The similarity across these observations provides powerful tools for denoising.

By utilizing the spectral correlation and the spatial nonlocal self-similarity of HSIs, a clean HSI can be approximated by nonlocal low-rank tensors [55, 52]. To denoise the HSI via the nonlocal low-rank tensor regularization, the first step is to extract nonlocal similar tensors that may have low-rank features and the second step is to characterize the low-rankness of the tensor naturally.

We first apply block matching to find similar blocks and then stack them into a fourth order nonlocal similar tensor. Given an HSI $\mathcal{L} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, we divide it into a total number of N overlapping FBBs of size $r \times r \times I_3$. For the j-th FBB, we search within a local window for a total of m_2 FBBs that are mostly similar to the reference block based on Euclidean distance. Then the j-th nonlocal similar sub-tensor of order 3 of $\mathcal{L}$, denoted as $\mathbf{R}_j(\mathcal{L})$, can be formed by unfolding all the nonlocal similar FBBs in the j-th group and then stacking them together. An illustrative figure on nonlocal low-rank tensor extraction is shown in Figure 4.2. As the nonlocal similar FBB sub-tensors are independent of each other, we can further stack them together into a fourth order nonlocal similar group tensor, denoted as $\mathbf{R}(\mathcal{L})$.

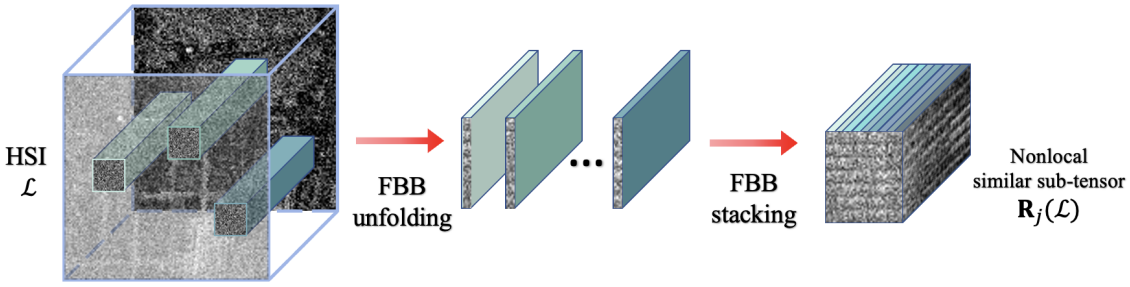


Figure 4.2: The procedure of block matching.

To be precise, we present the formulations for the j-th nonlocal similar sub-tensor extraction operator $\mathbf{R}_j$ and then for the nonlocal similar tensor extraction operator $\mathbf{R}$. We define $U_i^{(l)} \in \mathbb{R}^{r \times I_i}$ as a binary matrix such that $\mathcal{L} \times_1 U_1^{(l)} \times_2 U_2^{(l)}$ is exactly the l-th FBB of $\mathcal{L}$. And we define B_l as the Casorati matrix (a matrix whose columns

are vectorized bands of the HSI) of the l-th FBB as follows

$$B_l := \text{reshape}(\mathcal{L} \times_1 U_1^{(l)} \times_2 U_2^{(l)}, m_1, m_3),$$

where $m_1 = r^2$ and $m_3 = I_3$. Then the extraction operator of the j-th nonlocal similar sub-tensor $\mathbf{R}_j : \mathbb{R}^{I_1 \times I_2 \times I_3} \to \mathbb{R}^{m_1 \times m_2 \times m_3}$ can be defined by

$$\mathbf{R}_j(\mathcal{L}) := \text{reshape}((B_{l_{j1}}^\top, B_{l_{j2}}^\top, \ldots, B_{l_{jm_2}}^\top)^\top, m_1, m_2, m_3),$$

where the indices $l_{j1}, l_{j1}, \ldots, l_{jm_2}$ refer to the indices of FBBs that belong to the j-th nonlocal similar group and $j = 1, 2, \ldots, N$. Then the extraction operator of the nonlocal similar tensor $\mathbf{R} : \mathbb{R}^{I_1 \times I_2 \times I_3} \to \mathbb{R}^{m_1 \times m_2 \times m_3 \times N}$ is a linear map such that $[\mathbf{R}(\mathcal{L})]^{(j)} = \mathbf{R}_j(\mathcal{L})$.

Since the Frobenius inner product is invariant to reshaping, we have that for $\mathcal{Y} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$,

$$\langle \mathcal{Y}, \mathbf{R}_j(\mathcal{L}) \rangle_F = \sum_{i=1}^{m_2} \langle Y_{:i:}, B_{l_{ji}} \rangle_F = \langle \mathbf{R}_j^\top(\mathcal{Y}), \mathcal{L} \rangle_F,$$

where $\mathbf{R}_j^\top : \mathbb{R}^{m_1 \times m_2 \times m_3} \to \mathbb{R}^{I_1 \times I_2 \times I_3}$ is defined by

$$\mathbf{R}_j^\top(\mathcal{Y}) := \sum_{i=1}^{m_2} \text{reshape}(Y_{:i:}, r, r, I_2) \times_1 (U_1^{(l_{ji})})^\top \times_2 (U_2^{(l_{ji})})^\top. \tag{4.1}$$

And we further have that

$$\|\mathbf{R}(\mathcal{L})\|_F^2 = \sum_{j=1}^{N} \|\mathbf{R}_j(\mathcal{L})\|_F^2 = \sum_{j=1}^{N} \langle \mathcal{L}, \mathbf{R}_j^\top \mathbf{R}_j(\mathcal{L}) \rangle_F = \| \sqrt{\mathcal{W}_\mathbf{R}} \odot \mathcal{L} \|_F^2,$$

where $\odot$ represents the pointwise multiplication, and each entry of $\mathcal{W}_\mathbf{R} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ represents the number of nonlocal similar groups to which the corresponding pixel belongs. Since we assume that each pixel belongs to at least one nonlocal similar group, we have $\mathcal{W}_\mathbf{R} \in \mathbb{R}_{++}^{I_1 \times I_2 \times I_3}$.

4.2.2 Sparsity-enhanced Nonlocal Low-rank Tensor Regularization

Nonlocal low-rank tensor regularization is based on prior knowledge of the nonlocal similarity of HSIs. We impose the tensor low-rankness on the nonlocal similar tensors $\mathbf{R}(\mathcal{L})$ using (2.12), which is a low-rank tensor regularization with a sparsity enhancement $\|\cdot\|_{1,w}$ on the independent core tensors.

The update of $[\mathcal{G}]$ given in (3.6) can be efficiently computed via the proximal operator of $\|\cdot\|_{1,w}$. And the proximal operator of $\|\cdot\|_{1,w}$ can be computed componentwisely using the proximal operator of the ℓ_1 norm by the soft thresholding operator as follows

$$\begin{aligned}\left(\text{prox}_{\|\cdot\|_{1,w}}([\mathcal{G}])\right)^{(j)}_{i_1i_2i_3} &= \text{prox}_{w_j|\cdot|}\left([\mathcal{G}]^{(j)}_{i_1i_2i_3}\right)\\ &= \text{sign}\left([\mathcal{G}]^{(j)}_{i_1i_2i_3}\right)\max\left(\left|[\mathcal{G}]^{(j)}_{i_1i_2i_3}\right| - w_j, 0\right).\end{aligned}$$

4.2.3 Tensor $\ell_{2,p}$ Norm for Group Sparsity Regularization

The generalized group sparsity measure given in (2.2) can characterize the sparse component along a specific direction. We choose the tensor $\ell_{2,p}$ norm given in (2.4) as the group sparsity measure for $\mathcal{S}$. By applying Theorem 2.1, the update of $\mathcal{S}$ given in (3.2) can be efficiently computed. In particular, we calculate the (i_2, i_3)-th mode-1 fiber of $\mathcal{S}^{k+1}$ by

$$s^{k+1}_{:i_2i_3} = \Gamma_\mu\left(\left\|\sqrt{w_{:i_2i_3}} \odot \tilde{s}^k_{:i_2i_3}\right\|_2\right)\tilde{s}^k_{:i_2i_3},$$

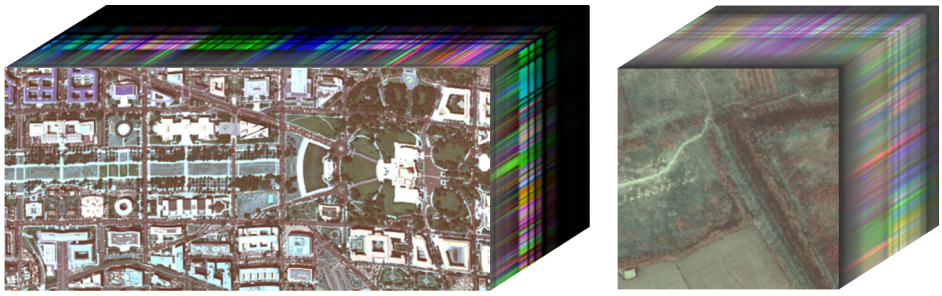
where $\Gamma_\mu(\cdot)$ is given in Theorem 2.1, $\tilde{\mathcal{S}}^k = \mathcal{S}^k - \tilde{\alpha}_{\mathcal{S}}\left(\mathcal{S}^k + \mathcal{L}^k - \mathcal{D}\right)$, and $\tilde{s}^k_{:i_2i_3}$ and $\sqrt{w_{:i_2i_3}}$ are the (i_2, i_3)-th mode-1 fibers of $\tilde{\mathcal{S}}^k$ and $\sqrt{\mathcal{W}_{\mathbf{R}}}$, respectively.

4.3 Numerical Experiments

In this section, we conduct numerical experiments for removing mixed noise in HSIs. We compare the proposed methods with five methods, which are BM4D [55] for removing Gaussian noise, and LRTD [14], SNLRSF [9], LRTFL0 [93] and QRNN3D [86] for removing mixed noise. All numerical experiments are implemented in Matlab R2018a and executed on a personal desktop (Intel Core i7 9750H at 2.60 GHz with 16 GB RAM).

4.3.1 Simulated Data Experiments

In this subsection, the proposed method and the competing methods are tested on simulated data. The test images are subimages of size $128 \times 128 \times 128$ randomly obtained from the Washington DC Mall[1] ($1280 \times 307 \times 191$) and the Xiong-An[2] ($256 \times 256 \times 256$). As shown in Figure 4.3, the Washington DC Mall is obtained from an urban area, where buildings are relatively dense; the Xiong-An is obtained from a hilly area, with mountains and shrubs.



(a) Washington DC Mall (b) Xiong-An

Figure 4.3: Original HSIs of two datasets. (a) Part of Washington DC Mall dataset (R:17, G:36, B:46); (b) Xiong-An dataset (R:71, G:110, B:120).

To simulate the noisy HSI data, Gaussian noise, stripes, or dead lines are added

[1] https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html

[2] http://www.hrs-cas.com/a/share/shujuchanpin/2019/0501/1049.html

to the normalized clean HSI data under the following cases:

- *Case 1:* Gaussian noise with a mean of zero and a standard deviation of 0.1 is added to all the bands. And then all the bands are selected and stripes with a density of 30% and a standard deviation of 0.2 are added to each band.

- *Case 2:* Gaussian noise with a mean of zero and a standard deviation of 0.1 is added to all the bands. And then $11-40$, $71-100$, $121-128$ bands are selected and stripes with a density of 20% and a standard deviation of 0.2 are added to each band.

- *Case 3:* Gaussian noise with a mean of zero and a standard deviation of 0.2 is added to all the bands. And then 25% of the bands are randomly selected and dead lines with a density of 5% are added to each band.

For comparing the quality of the restored images, four evaluation metrics are employed, which are the mean peak signal-to-noise ratio (MPSNR), the mean structural similarity index (MSSIM), the mean feature similarity index (MFSIM), the mean spectral angle mapping (MSAM), and the erreur relative globale adimensionnelle de synthese (ERGAS). Let $\mathcal{X}^*$ denote the restored HSI and $\hat{\mathcal{X}}$ denote the clean HSI. Then $X^*_{::i}$ and $\hat{X}_{::i}$ denote the i-th band of the restored HSI and clean HSI, respectively. The MPSNR value is defined as [34]

$$\text{MPSNR} = \frac{1}{I_3}\sum_{k=1}^{I_3} 10\log_{10}\left(\frac{\max^2(X^*_{::k})}{\text{mse}(X^*_{::k},\hat{X}_{::k})}\right),$$

which is the average PSNR value across the bands. Similarly, MSSIM and MFSIM values are defined as

$$\text{MSSIM} = \frac{1}{I_3}\sum_{k=1}^{I_3}\text{SSIM}(X^*_{::k},\hat{X}_{::k}) \quad \text{and MFSIM} = \frac{1}{I_3}\sum_{k=1}^{I_3}\text{FSIM}(X^*_{::k},\hat{X}_{::k}),$$

where SSIM is given in [109] and FSIM is given in [105]. The MSAM and ERGAS are defined as

$$\text{MSAM} = \frac{1}{I_1 I_2} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \arccos\left(\frac{{x_{ij:}^*}^\top \cdot \hat{x}_{ij:}}{\|x_{ij:}^*\| \|\hat{x}_{ij:}\|}\right)$$

and

$$\text{ERGAS} = 100 \sqrt{\frac{1}{I_3} \sum_{k=1}^{I_3} \frac{\text{mse}(X_{::k}^*, \hat{X}_{::k})}{\text{mean}(X_{::k}^*)}},$$

where MSAM is given in [44] and ERGAS is given in [78]. In addition, better denoising results are indicated by larger MPSNR, MSSIM, and MFSIM values, as well as smaller MSAM and ERGAS values.

Table 4.1: Numerical results tested on Washington DC Mall dataset

Case	Index	Noisy	NLTL2p	LRTFL0	SNLRSF	LRTD	BM4D	QRNN3D
1	MPSNR	14.90	**30.84**	30.29	25.92	26.47	16.35	25.27
	MSSIM	0.324	**0.930**	0.923	0.795	0.817	0.399	0.794
	MFSIM	0.653	0.952	**0.957**	0.902	0.907	0.724	0.885
	MSAM	0.474	**0.083**	0.083	0.138	0.132	0.407	0.150
	ERGAS	518.02	**88.21**	92.45	151.04	138.47	438.24	159.34
2	MPSNR	16.72	**31.75**	30.46	31.45	27.82	22.72	26.97
	MSSIM	0.428	**0.955**	0.920	0.912	0.854	0.639	0.844
	MFSIM	0.703	**0.969**	0.933	0.950	0.921	0.817	0.916
	MSAM	0.397	**0.056**	0.067	0.074	0.067	0.318	0.109
	ERGAS	419.52	**68.88**	78.65	83.10	105.07	330.58	119.73
3	MPSNR	13.90	**30.20**	29.83	28.01	25.44	20.80	25.62
	MSSIM	0.390	**0.943**	0.924	0.895	0.801	0.762	0.842
	MFSIM	0.680	**0.963**	0.929	0.943	0.897	0.879	0.913
	MSAM	0.606	**0.098**	0.099	0.153	0.503	0.467	0.164
	ERGAS	610.91	95.59	**93.19**	141.32	148.89	449.91	147.69

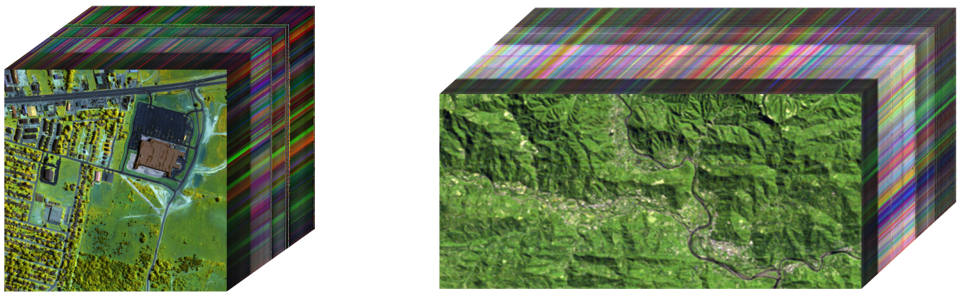
The numerical results of simulated data experiments for case 1, case 2, and case 3 are presented in Tables 4.1 and 4.2 for Washington DC Mall and Xiong-An datasets, respectively. Numerical results in bold font indicate the best performance of the indicator in the current case. It can be observed from Tables 4.1 and 4.2 that the

Table 4.2: Numerical results tested on Xiong-An dataset

Case	Index	Noisy	NLTL2p	LRTFL0	SNLRSF	LRTD	BM4D	QRNN3D
1	MPSNR	14.49	**32.94**	31.73	26.26	30.02	15.88	26.53
	MSSIM	0.102	**0.862**	0.810	0.562	0.710	0.151	0.633
	MFSIM	0.482	**0.923**	0.915	0.829	0.883	0.575	0.852
	MSAM	0.286	**0.032**	0.037	0.076	0.047	0.246	0.071
	ERGAS	382.96	**47.02**	55.28	112.07	66.10	328.21	97.50
2	MPSNR	17.19	**34.17**	33.21	30.25	30.71	24.29	28.27
	MSSIM	0.159	**0.877**	0.837	0.719	0.730	0.468	0.694
	MFSIM	0.562	**0.931**	0.924	0.885	0.894	0.722	0.880
	MSAM	0.230	**0.028**	0.033	0.056	0.042	0.176	0.061
	ERGAS	306.56	**40.98**	48.82	80.07	60.80	234.71	82.38
3	MPSNR	13.90	**31.68**	30.89	28.39	27.96	21.30	26.13
	MSSIM	0.238	**0.920**	0.858	0.828	0.765	0.692	0.755
	MFSIM	0.634	**0.959**	0.897	0.924	0.881	0.831	0.900
	MSAM	0.313	0.031	**0.029**	0.061	0.047	0.224	0.075
	ERGAS	435.95	60.74	**56.64**	103.41	83.80	321.25	101.38

proposed NLTL2p method outperforms other methods almost in terms of all the evaluation metrics. For example, in case 1 of Xiong-An dataset, the MPSNR value of the HSI restored by the NLTL2p method is 1.84 dB larger than the MPSNR value of the second best method, that is, the LRTFL0 method.

4.3.2 Real Data Experiments



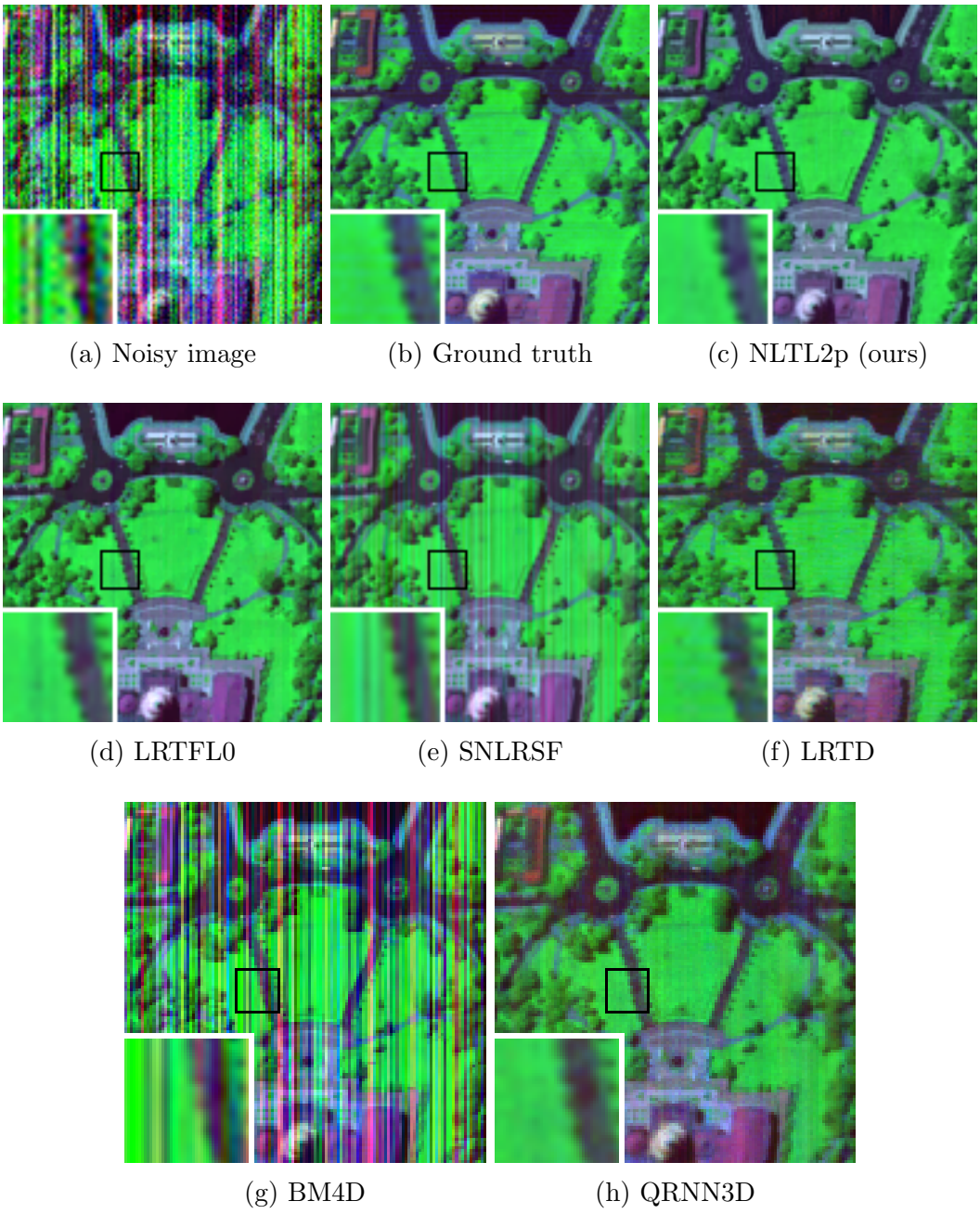
(a) HYDICE Urban (b) EO-1 Hyperion

Figure 4.4: Real HSIs of two datasets. (a) HYDICE Urban dataset (R:61, G:98, B:170); (b) EO-1 Hyperion dataset (R:101, G:114, B:160).

In this subsection, we test the proposed method and the competing methods on two real HSI datasets containing mixed noise. The test images are subimages of size $128 \times 128 \times 128$ randomly obtained from the HYDICE Urban[3] ($307 \times 307 \times 210$) and EO-1 Hyperion[4] ($400 \times 200 \times 166$), which are shown in Figure 4.4. A selected band of the HSI restored by each method is presented in Figure 4.11 and Figure 4.13 for HYDICE Urban dataset and EO-1 Hyperion dataset, respectively. It can be observed that the proposed NLTL2p method can remove the stripes while preserving the image details. However, the LRTFL0, LRTD, BM4D and QRNN3D methods are unable to eliminate the stripes when the band is contaminated by heavy mixed noise as shown in Figure 4.11; and the SNLRSF, LRTD, BM4D and QRNN3D methods remove not only the noise but also some structural details of the HSI as shown in Figure 4.13.

[3] http://www.erdc.usace.army.mil/Media/Fact-Sheets/Fact-Sheet-Article-View/Article/610433/hypercube/

[4] http://www.lmars.whu.edu.cn/prof_web/zhanghongyan/resource/noise_EOI.zip

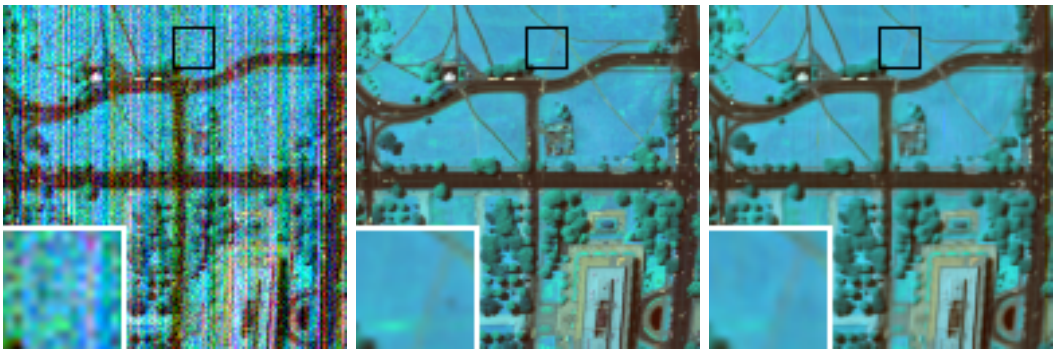


(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

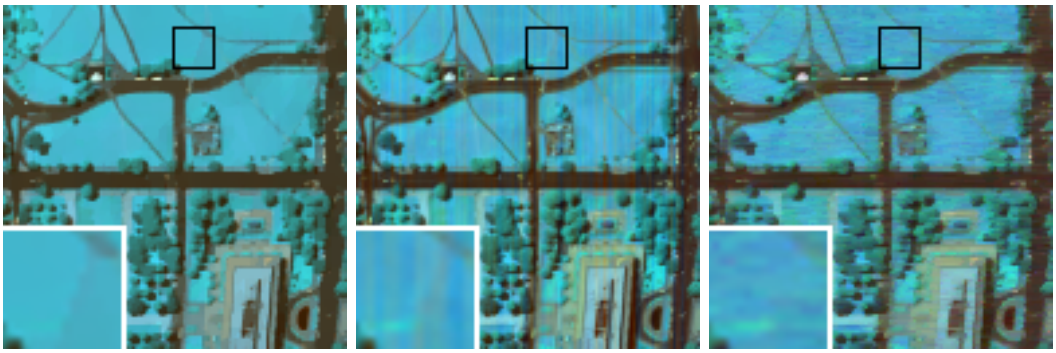
(d) LRTFL0 (e) SNLRSF (f) LRTD

(g) BM4D (h) QRNN3D

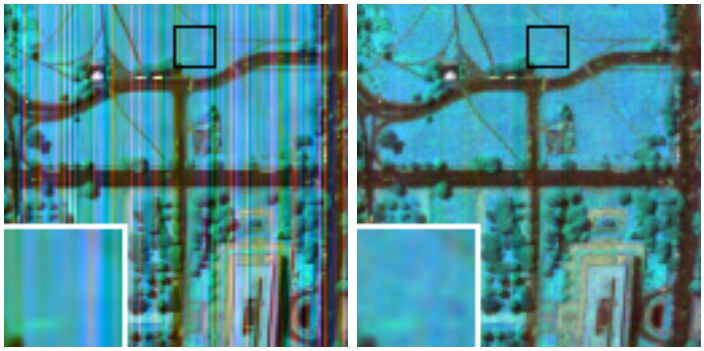
Figure 4.5: Comparison of HSIs (R:3, G:43, B:75) restored by different methods from Washington DC Mall in case 1. The PSNR value for each restored HSI: (a) Noisy image (14.90 dB); (c) NLTL2p (ours) (30.84 dB); (d) LRTFL0 (30.29 dB); (e) SNLRSF (25.92 dB); (f) LRTD (26.47 dB); (g) BM4D (16.35 dB); (h) QRNN3D (25.27 dB).



(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

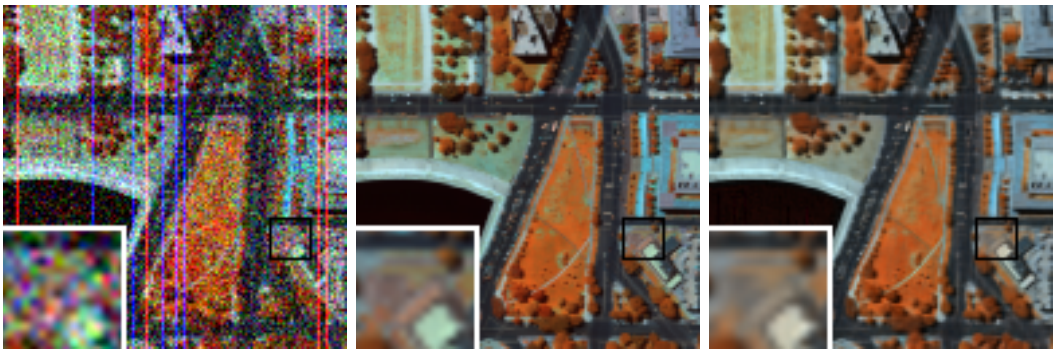


(d) LRTFL0 (e) SNLRSF (f) LRTD

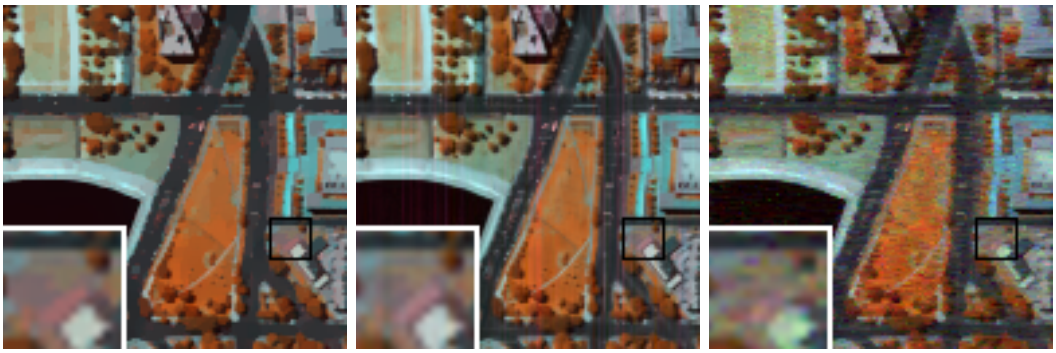


(g) BM4D (h) QRNN3D

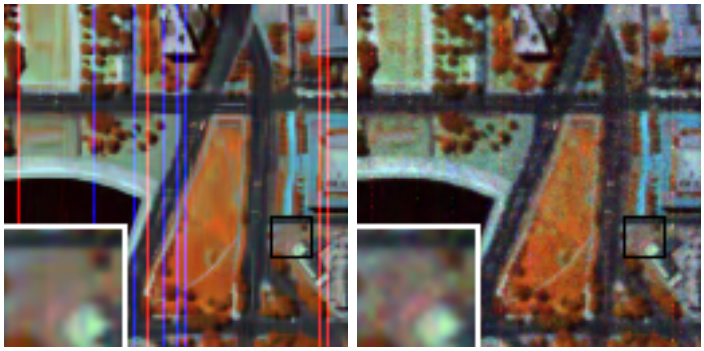
Figure 4.6: Comparison of HSIs (R:23, G:63, B:94) restored by different methods from Washington DC Mall in case 2. The PSNR value for each restored HSI: (a) Noisy image (16.72 dB); (c) NLTL2p (ours) (31.75 dB); (d) LRTFL0 (30.46 dB); (e) SNLRSF (31.45 dB); (f) LRTD (27.82 dB); (g) BM4D (22.72 dB); (h) QRNN3D (26.97 dB).



(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

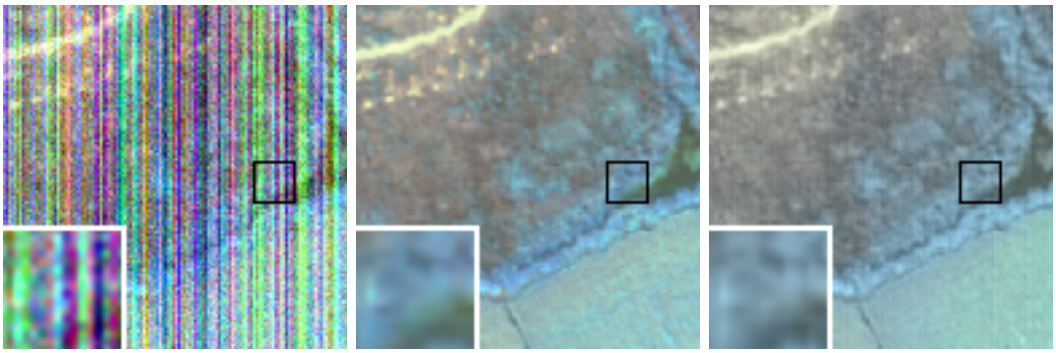


(d) LRTFL0 (e) SNLRSF (f) LRTD

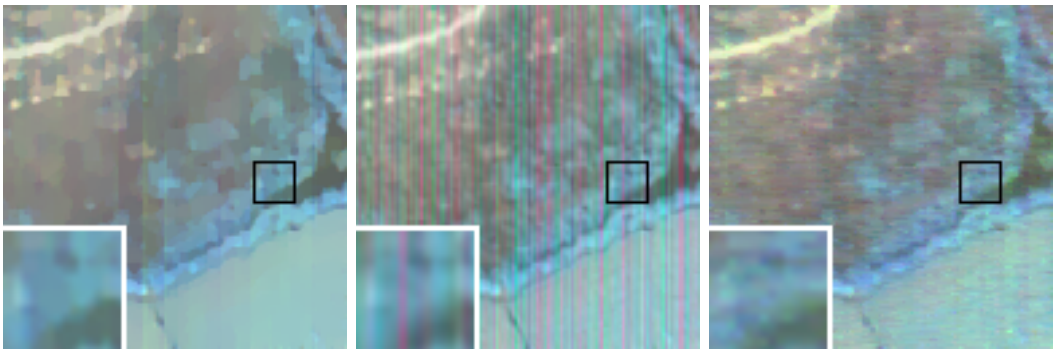


(g) BM4D (h) QRNN3D

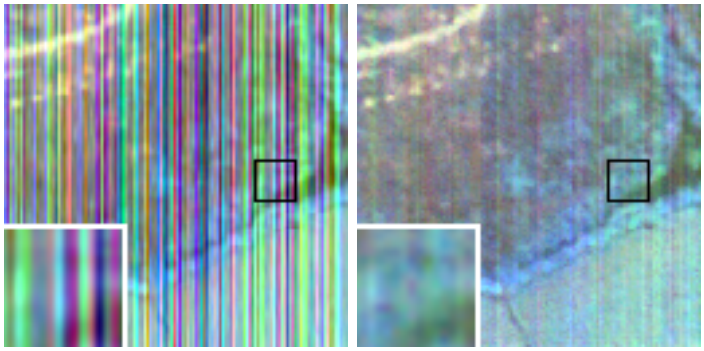
Figure 4.7: Comparison of HSIs (R:16, G:70, B:100) restored by different methods from Washington DC Mall in case 3. The PSNR value for each restored HSI: (a) Noisy image (13.90 dB); (c) NLTL2p (ours) (30.20 dB); (d) LRTFL0 (29.83 dB); (e) SNLRSF (28.01 dB); (f) LRTD (25.44 dB); (g) BM4D (20.80 dB); (h) QRNN3D (25.62 dB).



(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

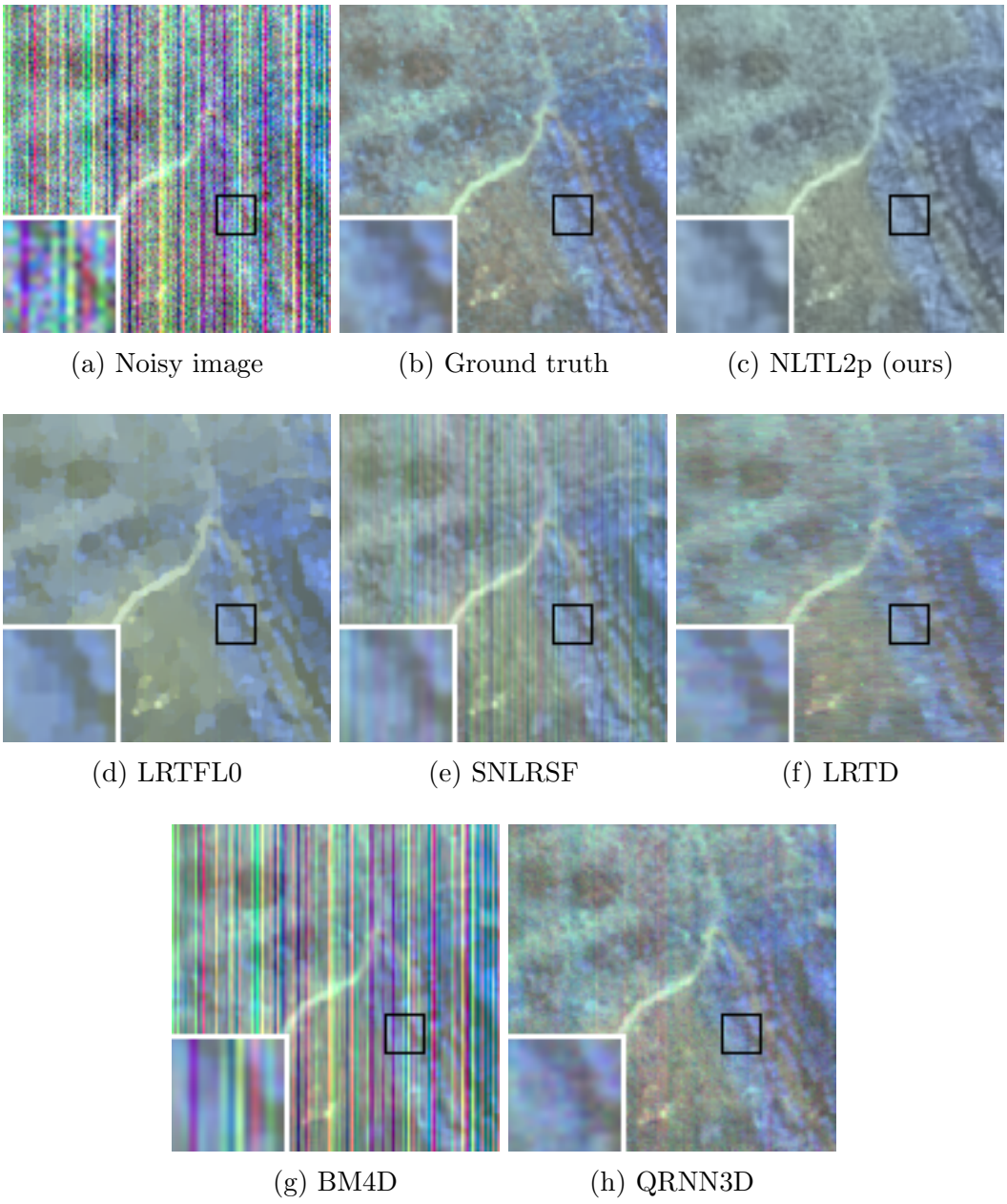


(d) LRTFL0 (e) SNLRSF (f) LRTD



(g) BM4D (h) QRNN3D

Figure 4.8: Comparison of HSIs (R:30, G:55, B:115) restored by different methods from Xiong-An in case 1. The PSNR value for each restored HSI: (a) Noisy image (14.49 dB); (c) NLTL2p (ours) (32.94 dB); (d) LRTFL0 (31.73 dB); (e) SNLRSF (26.26 dB); (f) LRTD (30.02 dB); (g) BM4D (15.88 dB); (h) QRNN3D (26.53 dB).

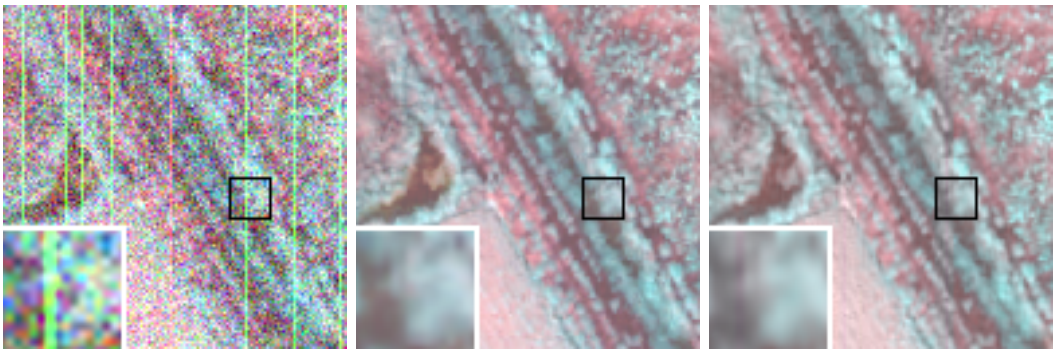


(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

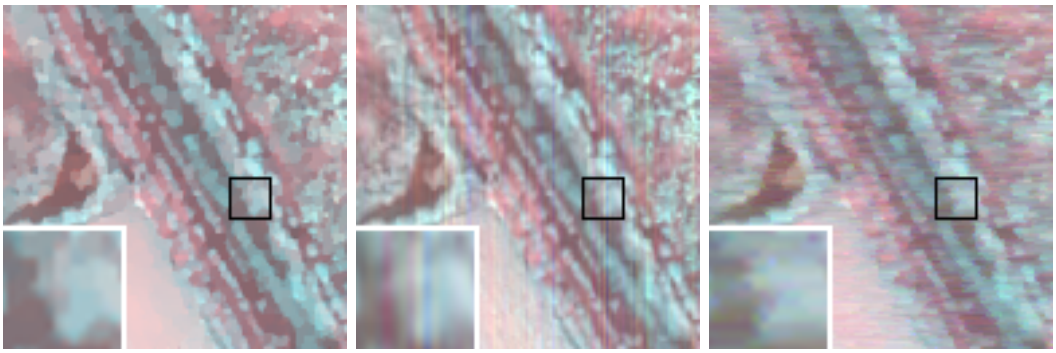
(d) LRTFL0 (e) SNLRSF (f) LRTD

(g) BM4D (h) QRNN3D

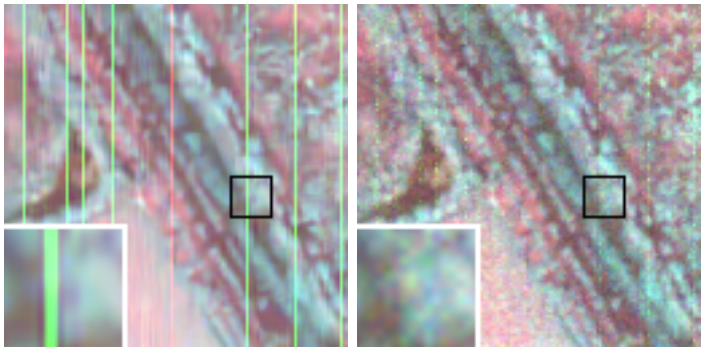
Figure 4.9: Comparison of HSIs (R:20, G:45, B:71) restored by different methods from Xiong-An in case 2. The PSNR value for each restored HSI: (a) Noisy image (17.19 dB); (c) NLTL2p (ours) (34.17 dB); (d) LRTFL0 (33.21 dB); (e) SNLRSF (30.25 dB); (f) LRTD (30.71 dB); (g) BM4D (24.29 dB); (h) QRNN3D (28.27 dB).



(a) Noisy image (b) Ground truth (c) NLTL2p (ours)

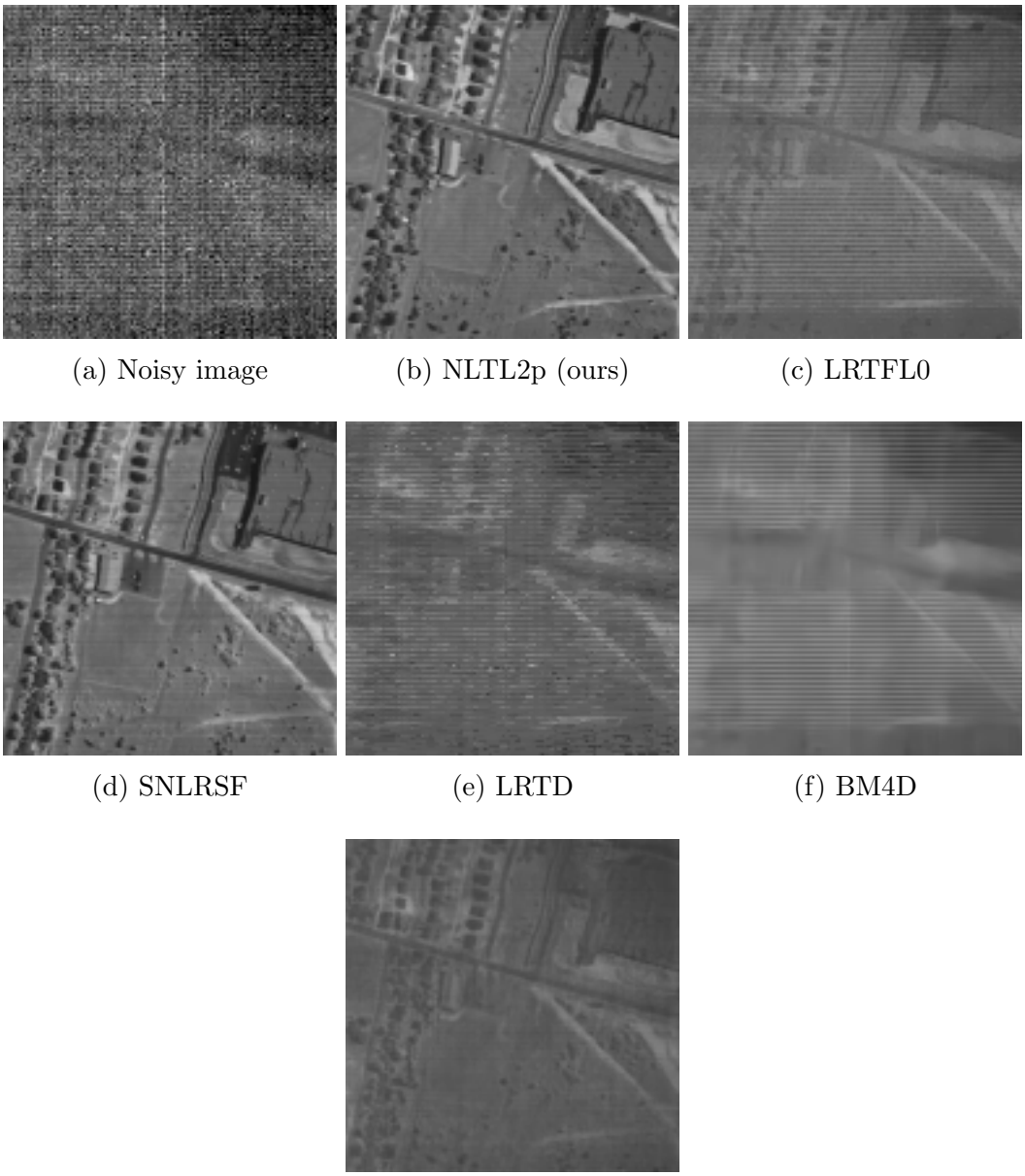


(d) LRTFL0 (e) SNLRSF (f) LRTD



(g) BM4D (h) QRNN3D

Figure 4.10: Comparison of HSIs (R:38, G:67, B:90) restored by different methods from Xiong-An in case 3. The PSNR value for each restored HSI: (a) Noisy image (13.90 dB); (c) NLTL2p (ours) (31.68 dB); (d) LRTFL0 (30.89 dB); (e) SNLRSF (28.39 dB); (f) LRTD (27.96 dB); (g) BM4D (21.30 dB); (h) QRNN3D (26.13 dB).

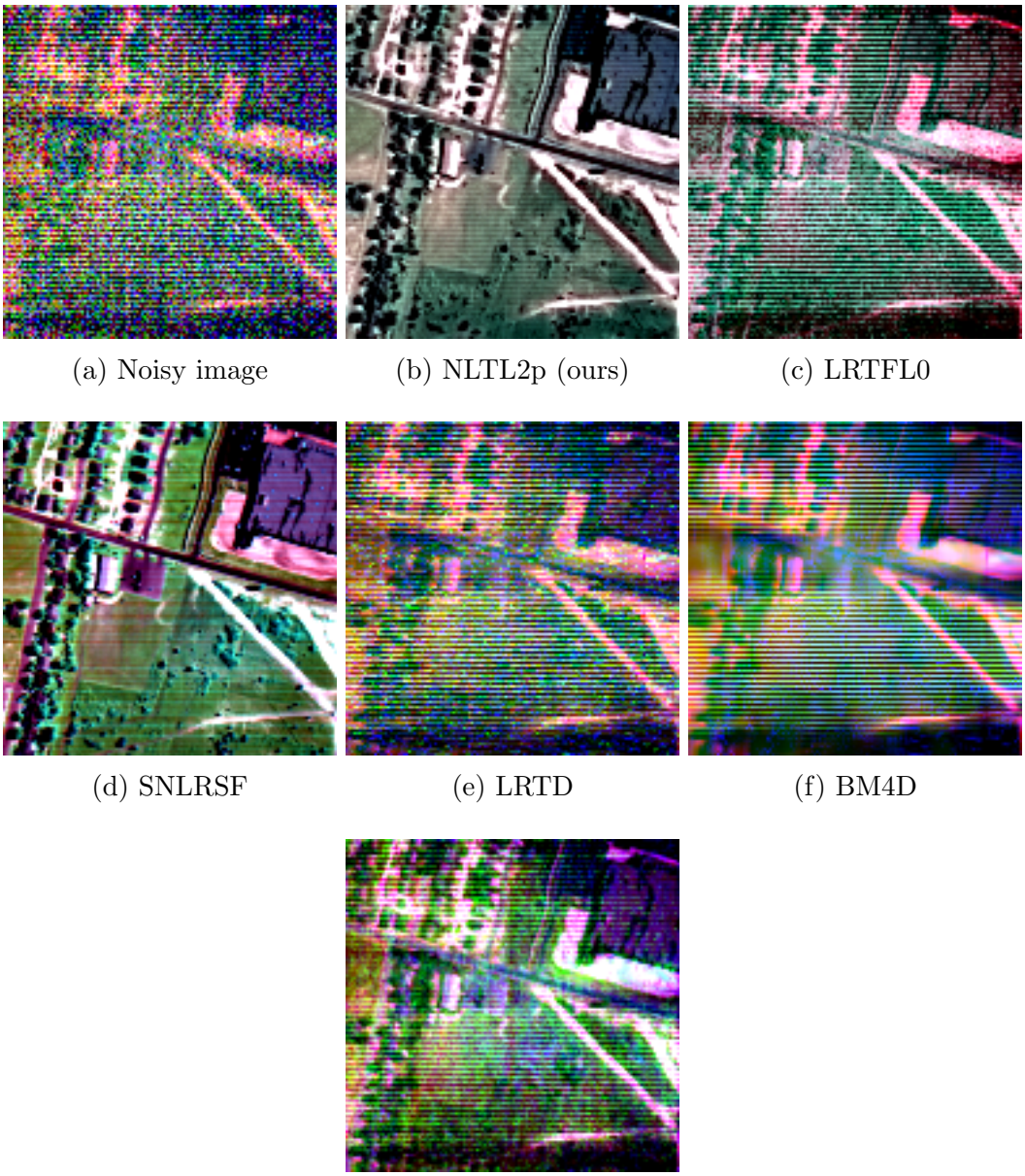


(a) Noisy image (b) NLTL2p (ours) (c) LRTFL0

(d) SNLRSF (e) LRTD (f) BM4D

(g) QRNN3D

Figure 4.11: Comparison of the 25-th band of the HSI restored by different methods from HYDICE Urban.

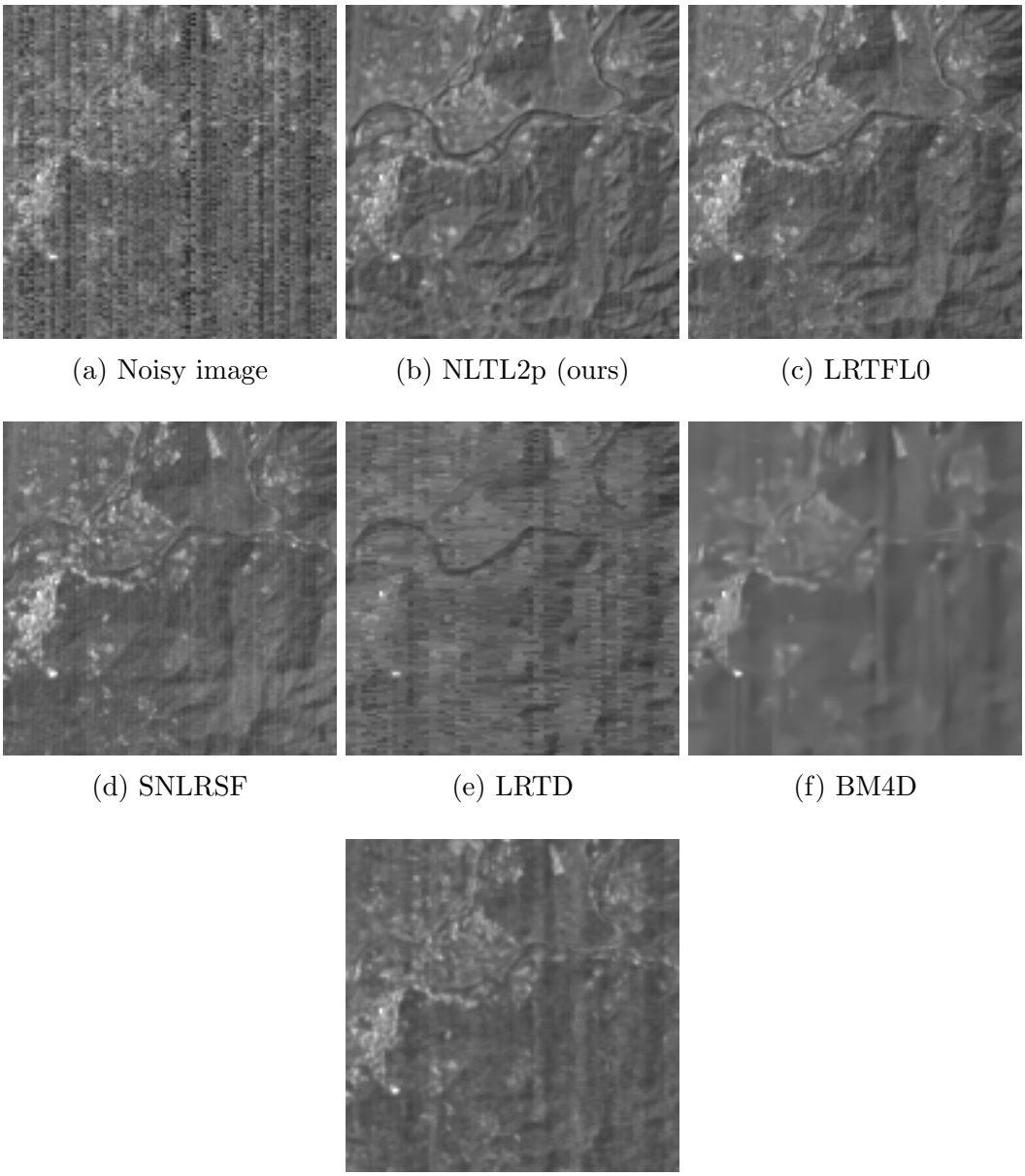


(a) Noisy image (b) NLTL2p (ours) (c) LRTFL0

(d) SNLRSF (e) LRTD (f) BM4D

(g) QRNN3D

Figure 4.12: Comparison of the fake color image of the HSI restored by different methods from HYDICE Urban.

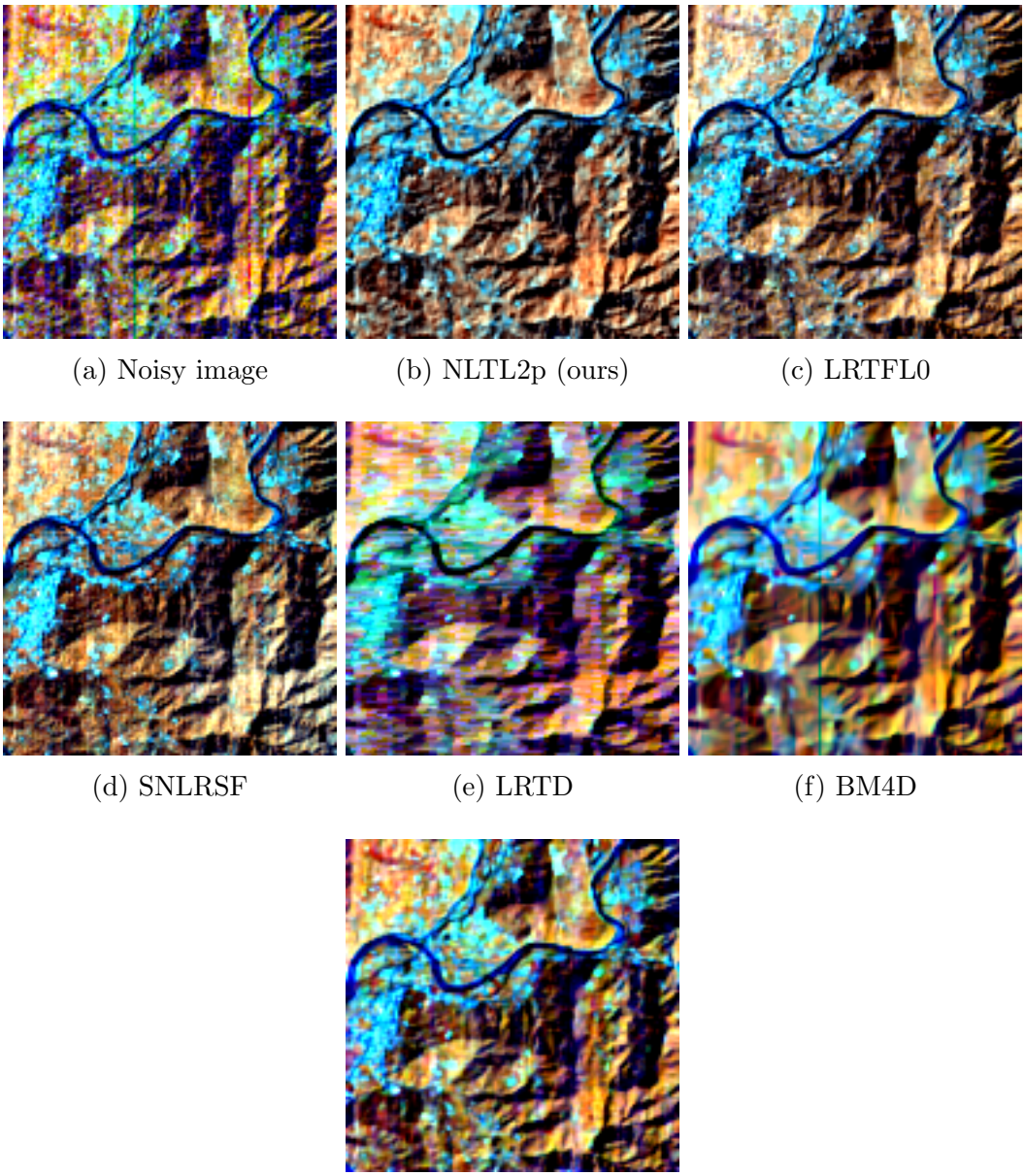


(a) Noisy image (b) NLTL2p (ours) (c) LRTFL0

(d) SNLRSF (e) LRTD (f) BM4D

(g) QRNN3D

Figure 4.13: Comparison of the 128-th band of the HSI restored by different methods from EO-1 Hyperion.



(a) Noisy image (b) NLTL2p (ours) (c) LRTFL0

(d) SNLRSF (e) LRTD (f) BM4D

(g) QRNN3D

Figure 4.14: Comparison of the fake color image of the HSI restored by different methods from EO-1 Hyperion.

Chapter 5

Application to Hyperspectral Anomaly Detection

5.1 Problem Statement

Hyperspectral anomaly detection involves identifying pixels or regions within an image that significantly differ from the background [73]. Unlike conventional imaging, hyperspectral sensors capture detailed spectral data across hundreds of bands, allowing for a fine-grained analysis of the material composition of each pixel. This spectral richness makes HSIs particularly suitable for detecting small-scale anomalies such as vehicles, aircraft, and ships. Hence, hyperspectral anomaly detection is critical in applications such as environmental monitoring [72], agriculture[53], and defense[70].

Detecting these anomalies is challenging due to several factors. The high dimensionality of HSI data means that each pixel is represented by a complex spectral vector, making it difficult to differentiate anomalous pixels from natural variability [73]. Anomalies often blend into the background due to subtle spectral differences, which require well-designed algorithms to identify. Additionally, HSIs commonly contain various types of noise, such as Gaussian noise, stripes, and dead lines, which can mimic anomalies, complicating the detection process [80]. The sparsity of anomalies further increases the difficulty, as these unusual features typically occupy only a

small fraction of the image, standing out against a dominant background.

5.2 Model Formulation

According to the high spectral correlation of HSIs, a clean HSI $\mathcal{L} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ can be expressed in a low-rank tensor representation as follows

$$\mathcal{L} = \mathcal{Z} \times_3 E, \tag{5.1}$$

where $E \in \mathbb{R}^{n_3 \times r}$ represents a basis of the spectral subspace and the tensor $\mathcal{Z} \in \mathbb{R}^{n_1 \times n_2 \times r}$ denotes the representation coefficient of $\mathcal{L}$ with respect to E. In particular, we choose E as an orthogonal basis, that is, $E^\top E = \mathbf{I}_r$ with $\mathbf{I}_r$ denoting the identity matrix of size r. And each band of $\mathcal{Z}$ is called an eigenimage. Then a noisy HSI $\mathcal{O} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ can be formulated mathematically as

$$\mathcal{O} = \mathcal{Z} \times_3 E + \mathcal{S} + \mathcal{N},$$

where $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ represents the sparse components such as anomalous objects, and $\mathcal{N} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ represents Gaussian noise.

To remove Gaussian noise and detect anomalous objects simultaneously, we propose an optimization model as follows

$$\begin{aligned} &\min_{\mathcal{Z},E,\mathcal{S}} \frac{\delta}{2}\|\mathcal{Z} \times_3 E + \mathcal{S} - \mathcal{O}\|_F^2 + \tau\|\mathcal{S}\|_{2,\psi} + \Phi_\Sigma(\mathcal{Z}) \\ &\text{s.t.} \quad E^\top E = \mathbf{I}_r, \end{aligned} \tag{5.2}$$

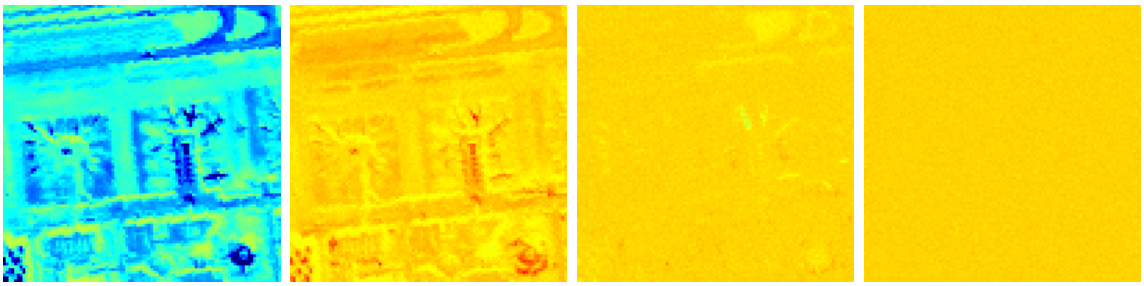
where $\|\cdot\|_{2,\psi}$ represents a generalized group sparsity measure for detecting anomalous objects, $\Phi_\Sigma(\cdot)$ represents a proximal denoiser prior for removing Gaussian noise, and E is a learnable orthogonal basis. The resulting model is a nonconvex nonsmooth minimization problem with an orthogonal constraint. More details of this model will be provided in the following subsections.

5.2.1 Proximal Denoiser Prior in a PnP Framework

If the observed HSI $\mathcal{O}$ is degraded by Gaussian noise, the components of its tensor decomposition will also contain some noise, i.e.,

$$\mathcal{O} = \widetilde{\mathcal{Z}} \times_3 \widetilde{E}, \tag{5.3}$$

where $\widetilde{\mathcal{Z}}$ denotes the eigenimages degraded by noise, and $\widetilde{E}$ denotes the orthogonal basis with bias. For illustration, in Figure 5.1 we present some selected eigenimages of a noisy HSI with a noise level of 0.03 using the subspace obtained by the HySime algorithm [4]. As shown in Figure 5.1, the first eigenimage is clean, while the noise level of the other eigenimages increases as the band index increases. To remove the noise in the HSI, we utilize a deep denoiser $\mathcal{D}_\sigma$ on each eigenimage with an adaptive noise level σ.



(a) 1st eigenimage (b) 2nd eigenimage (c) 5th eigenimage (d) 10th eigenimage

Figure 5.1: Illustration of eigenimages obtained from a noisy HSI.

The deep denoiser $\mathcal{D}_\sigma$ that we will use is a proximal denoiser proposed by Hurault et al. in [39], which has the form of a gradient descent step:

$$\mathcal{D}_\sigma := \mathrm{Id} - \nabla g_\sigma, \tag{5.4}$$

where g_σ denotes a smooth parameterized neural network. In particular, g_σ is defined as follows

$$g_\sigma(X) = \frac{1}{2}\|X - \mathcal{N}_\sigma(X)\|_F^2, \tag{5.5}$$

with $\mathcal{N}_\sigma(X)$ being a $\mathcal{C}^2$ neural network, specifically DRUNet [104], pre-trained for denoising grayscale or color images. And the denoiser is carefully trained to ensure that g_σ approximately has an L_{g_σ}-Lipschitz gradient with $L_{g_\sigma} < 1$. The overall denoiser $\mathcal{D}_\sigma$ is called a proximal denoiser, because it behaves like a proximal operator as shown in Proposition 5.1. More discussions on this proximal denoiser can also be found in [35, 89, 33].

Proposition 5.1. *(See [38, Prop. 1]) Let $g_\sigma : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ be a $\mathcal{C}^2$ function with ∇g_σ being L_{g_σ}-Lipschitz and $L_{g_\sigma} < 1$. Then, for $\mathcal{D}_\sigma$ defined as in* (5.4)*, there exists a potential $\phi_\sigma : \mathbb{R}^{n_1 \times n_2} \to [0, +\infty)$ such that $\operatorname{prox}_{\phi_\sigma}$ is one-to-one and*

$$\mathcal{D}_\sigma = \operatorname{prox}_{\phi_\sigma},$$

where

$$\phi_\sigma(X) = \begin{cases} g_\sigma(\mathcal{D}_\sigma^{-1}(X)) - \frac{1}{2}\|\mathcal{D}_\sigma^{-1}(X) - X\|_F^2, & \text{if } X \in \operatorname{Im}(\mathcal{D}_\sigma), \\ +\infty, & \text{otherwise.} \end{cases} \tag{5.6}$$

Moreover, ϕ_σ is $\frac{L_{g_\sigma}}{L_{g_\sigma}+1}$-weakly convex and $\phi_\sigma(X) \geq g_\sigma(X)$ for any $X \in \mathbb{R}^{n_1 \times n_2}$.

To better adapt the proximal denoiser $\mathcal{D}_\sigma$ to denoise each eigenimage $Z_{::n}$, we first consider a relaxed version of the proximal denoiser discussed in [38] as follows

$$\mathcal{D}_\sigma^\gamma = \gamma \mathcal{D}_\sigma + (1 - \gamma)\operatorname{Id} = \operatorname{Id} - \gamma \nabla g_\sigma$$

with parameter $\gamma \in [0, 1]$. Then by applying Proposition 5.1 with $g_\sigma^\gamma = \gamma g_\sigma$, we get that there exists a $\frac{\gamma L_{g_\sigma}}{\gamma L_{g_\sigma}+1}$-weakly convex function ϕ_σ^γ such that $\mathcal{D}_\sigma^\gamma = \operatorname{prox}_{\phi_\sigma^\gamma}$ if $\gamma L_{g_\sigma} < 1$. This allows us to control the weak convexity of the regularization function ϕ_σ^γ, leading to a wide range for the selection of step size in the algorithm that we will propose in the next section.

Second, as eigenimages may not fall in $[0, 1]$, we consider a shifted denoiser for an eigenimage Z as follows

$$\widetilde{\mathcal{D}}_\sigma^\gamma(Z) = \frac{1}{a}\left(\mathcal{D}_\sigma^\gamma(aZ + b) - b\right),$$

where $a > 0$ and b is a constant. Then

$$\widetilde{\mathcal{D}}_\sigma^\gamma(Z) = \frac{1}{a}\left(\text{prox}_{\phi_\sigma^\gamma}(aZ+b) - b\right) = \text{prox}_{\widetilde{\phi}_\sigma^\gamma}(Z),$$

where $\widetilde{\phi}_\sigma^\gamma(Z) = \frac{1}{a^2}\phi_\sigma^\gamma(aZ+b)$ and $\widetilde{\phi}_\sigma^\gamma$ is $\frac{\gamma L_{g_\sigma}}{\gamma L_{g_\sigma}+1}$-weakly convex.

Lastly, we define the proximal denoiser prior $\Phi_\Sigma(\cdot)$ used in our proposed model (5.2) for removing noise in eigenimages as follows

$$\Phi_\Sigma(\mathcal{Z}) = \lambda \sum_{n=1}^{r} \widetilde{\phi}_{\sigma_n}^\gamma(Z_{::n}), \tag{5.7}$$

where $\lambda > 0$ and $\Sigma := \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with σ_n denoting the noise level of the n-th eigenimage $Z_{::n}$.

5.2.2 Learnable Orthogonal Basis

The orthogonal constrained set on E is the Stiefel manifold $\mathbb{S}_{n_3,r}$ with $n_3 \geq r$. By choosing an orthogonal basis E, the eigenimages $Z_{::n}$ are linearly independent to each other. This allows us to apply the denoiser to each eigenimage $Z_{::n}$ independently and the noise covariance matrix is a diagonal matrix.

In some existing works, some subspace methods consider a fixed basis for the low-rank tensor decomposition. However, according to (5.3), this may result in false labels. Hence, we consider the learnable basis E, which will be updated iteratively.

5.2.3 Tensor Relaxed $\ell_{2,p}$ Norm for Group Sparsity Regularization

The generalized group sparsity measure given in (2.2) can characterize the sparse component along a specific direction. We choose ψ as the relaxed ℓ_p norm with $p \in (0,1)$, i.e., $\psi(t) = (|t| + \varepsilon)^p - \varepsilon^p$, $\varepsilon > 0$ and measure the group sparsity along the mode-3 direction. Then the corresponding generalized tensor group sparsity measure

of a third order tensor $\mathcal{S}$, denoted as $\|\mathcal{S}\|_{2,p,\varepsilon}^p$, is defined by

$$\|\mathcal{S}\|_{2,p,\varepsilon}^p = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\|s_{ij:}\|_2+\varepsilon)^p = \sum_{i_3=1}^{I_3}\sum_{i_2=1}^{I_2}\left(\sum_{i_1=1}^{I_1} s_{i_1 i_2 i_3}^2\right)^{\frac{p}{2}}. \tag{5.8}$$

And the proximal operator of $\|\cdot\|_{2,p,\varepsilon}^p$ can be efficiently computed.

5.3 The PnP-PBCD Algorithm

In this section, we propose a PnP-PBCD method for solving model (5.2), which is a nonconvex and nonsmooth optimization problem over a Stiefel manifold. In particular, in model (5.2), both $\|\cdot\|_{2,\psi}$ and Φ_Σ are weakly convex functions.

Let F denote the objective function of the proposed model (5.2) and

$$H(\mathcal{Z}, E, \mathcal{S}) := \frac{\delta}{2}\|\mathcal{Z}\times_3 E + \mathcal{S} - \mathcal{O}\|_F^2. \tag{5.9}$$

Then a PnP-PBCD algorithm for problem (5.2) is summarized as follows

$$\mathcal{S}^{k+1} \in \operatorname*{argmin}_{\mathcal{S}} H(\mathcal{Z}^k, E^k, \mathcal{S}) + \tau\|\mathcal{S}\|_{2,\psi} + \frac{\alpha_{\mathcal{S}}}{2}\|\mathcal{S}-\mathcal{S}^k\|_F^2, \tag{5.10}$$

$$E^{k+1} \in \operatorname*{argmin}_{E\in\mathbb{S}_{n_3,r}} H(\mathcal{Z}^k, E, \mathcal{S}^{k+1}) + \frac{\alpha_E}{2}\|E-E^k\|_F^2, \tag{5.11}$$

$$\mathcal{Z}^{k+1} = \operatorname*{argmin}_{\mathcal{Z}} H(\mathcal{Z}, E^{k+1}, \mathcal{S}^{k+1}) + \Phi_\Sigma(\mathcal{Z}) + \frac{\alpha_{\mathcal{Z}}}{2}\|\mathcal{Z}-\mathcal{Z}^k\|_F^2, \tag{5.12}$$

where the step sizes $\alpha_{\mathcal{S}}, \alpha_{\mathcal{Z}} \geq 0$ and $\alpha_E > 0$.

In the following, we present the details for computing each update. We will conduct a convergence analysis for the proposed PnP-PBCD algorithm in the next section.

5.3.1 The Update of $\mathcal{S}$

Combining the function H and the proximal term in (5.10), the update $\mathcal{S}^{k+1}$ can be written in terms of the proximal operator of $\|\cdot\|_{2,\psi}$ as follows

$$\mathcal{S}^{k+1} \in \operatorname{prox}_{\tilde{\tau}\|\cdot\|_{2,\psi}}\left(\mathcal{S}^k - \tilde{\alpha}_{\mathcal{S}}\left(\mathcal{S}^k + \mathcal{Z}^k \times_3 E^k - \mathcal{O}\right)\right), \tag{5.13}$$

where $\tilde{\tau} = \frac{\tau}{\delta+\alpha_{\mathcal{S}}}$ and $\tilde{\alpha}_{\mathcal{S}} = \frac{\delta}{\delta+\alpha_{\mathcal{S}}}$. Since $\|\cdot\|_{2,\psi}$ is separable, the (i,j)-th mode-3 fiber of $\mathcal{S}^{k+1}$ can be computed via

$$s_{ij:}^{k+1} \in \operatorname{prox}_{\tilde{\tau}\psi\circ\|\cdot\|_2}(\hat{s}_{ij:}^k),$$

where $\hat{s}_{ij:}^k$ is the (i,j)-th mode-3 fiber of $\widehat{\mathcal{S}}^k = \mathcal{S}^k - \tilde{\alpha}_{\mathcal{S}}\left(\mathcal{S}^k + \mathcal{Z}^k \times_3 E^k - \mathcal{O}\right)$.

5.3.2 The Update of E

Before we compute the update of E given in (5.11), we first introduce a lemma for the optimization problems over a Stiefel manifold as follows.

Lemma 5.1. *Let $X \in \mathbb{R}^{n\times r}$, $A \in \mathbb{R}^{r\times m}$ and $B \in \mathbb{R}^{n\times m}$. Then the solutions to the following problems are the same:*

$$\min_{X\in\mathbb{S}_{n,r}} \frac{1}{2}\|XA - B\|_F^2, \tag{5.14a}$$

$$\min_{X\in\mathbb{S}_{n,r}} -\langle X, BA^\top\rangle, \tag{5.14b}$$

and

$$\min_{X\in\mathbb{S}_{n,r}} \frac{1}{2}\|X - BA^\top\|_F^2. \tag{5.14c}$$

Proof. If $X \in \mathbb{S}_{n,r}$, i.e., $X^\top X = \mathbf{I}_r$, we have

$$\begin{aligned}\|XA - B\|_F^2 =& \|A\|_F^2 - 2\langle X, BA^\top\rangle + \|B\|_F^2\\ =& \|X - BA^\top\|_F^2 - \|BA^\top\|_F^2 - r^2 + \|A\|_F^2 + \|B\|_F^2.\end{aligned} \tag{5.15}$$

Then the equivalence is immediately achieved. □

Recall that for unfolding of tensors

$$\mathcal{L} = \mathcal{Z} \times_3 E \quad \text{if and only if} \quad \mathcal{L}_{(3)} = E\mathcal{Z}_{(3)}. \tag{5.16}$$

Then the function H can be rewritten as $H(\mathcal{Z}, E, \mathcal{S}) = \frac{\delta}{2}\|E\mathcal{Z}_{(3)} + (\mathcal{S} - \mathcal{O})_{(3)}\|_F^2$, which has the same form as (5.14a). It follows from Lemma 5.1 that minimizing (5.14a) is equivalent to minimizing more simple forms as (5.14b) and (5.14c). In particular, we consider $\widetilde{H}(\mathcal{Z}, E, \mathcal{S}) = -\delta\langle E, (\mathcal{O} - \mathcal{S})_{(3)}(\mathcal{Z}_{(3)})^\top\rangle$ of the form as (5.14b). Then the update of E can be computed as follows

$$\begin{aligned} E^{k+1} &\in \operatorname*{argmin}_{E \in \mathbb{S}_{n_3,r}} \widetilde{H}(\mathcal{Z}^k, E, \mathcal{S}^{k+1}) + \frac{\alpha_E}{2}\|E - E^k\|_F^2, \\ &= \operatorname{Proj}_{\mathbb{S}_{n_3,r}}\left(E^k + \tilde{\alpha}_E(\mathcal{S}^{k+1} - \mathcal{O})_{(3)}(\mathcal{Z}_{(3)}^k)^\top\right), \end{aligned} \tag{5.17}$$

where $\tilde{\alpha}_E = \frac{\delta}{\alpha_E}$ and $\operatorname{Proj}_{\mathbb{S}_{n_3,r}}$ denotes the projection onto the Stiefel manifold $\mathbb{S}_{n_3,r}$. And it follows from Lemma 3.1 that the projection $\operatorname{Proj}_{\mathbb{S}_{n_3,r}}$ has a closed form and E^{k+1} can be computed as follows

$$E^{k+1} = U^{k+1}(V^{k+1})^\top, \quad \text{with } U^{k+1}\widehat{\Sigma}^{k+1}(V^{k+1})^\top = \widehat{E}^k, \tag{5.18}$$

where $\widehat{E}^k = E^k + \tilde{\alpha}_E(\mathcal{S}^{k+1} - \mathcal{O})_{(3)}(\mathcal{Z}_{(3)}^k)^\top$, $U^{k+1}\widehat{\Sigma}^{k+1}(V^{k+1})^\top$ is a reduced SVD of $\widehat{E}^k$, $U^{k+1} \in \mathbb{R}^{n_3 \times r}$, $V^{k+1} \in \mathbb{R}^{r \times r}$, and $\widehat{\Sigma}^{k+1} \in \mathbb{R}^{r \times r}$.

5.3.3 The Update of $\mathcal{Z}$

By applying $E^\top E = \mathbf{I}_r$, we have

$$\|\mathcal{Z} \times_3 E - \mathcal{L}\|_F^2 = \|\mathcal{Z} - \mathcal{L} \times_3 E^\top\|_F^2.$$

Then the subproblem for updating $\mathcal{Z}$ can be reformulated as

$$\begin{aligned} \mathcal{Z}^{k+1} &\in \operatorname*{argmin}_{\mathcal{Z}} \Phi_\Sigma(\mathcal{Z}) + \frac{\delta}{2}\|\mathcal{Z} + (\mathcal{S}^{k+1} - \mathcal{O}) \times_3 (E^{k+1})^\top\|_F^2 + \frac{\alpha_{\mathcal{Z}}}{2}\|\mathcal{Z} - \mathcal{Z}^k\|_F^2, \\ &= \operatorname{prox}_{\tilde{\alpha}_{\mathcal{Z}}\Phi_\Sigma}\left(\mathcal{Z}^k - \tilde{\alpha}_{\mathcal{Z}}(\mathcal{Z}^k - (\mathcal{O} - \mathcal{S}^{k+1}) \times_3 (E^{k+1})^\top\right), \end{aligned} \tag{5.19}$$

where $\tilde{\alpha}_{\mathcal{Z}} = \frac{\delta}{\delta+\alpha_{\mathcal{Z}}}$. By choosing the parameter $\lambda = \frac{1}{\tilde{\alpha}_{\mathcal{Z}}}$ in Φ_Σ defined in (5.7), we have $\tilde{\alpha}_{\mathcal{Z}}\Phi_\Sigma(\mathcal{Z}) = \sum_{n=1}^{r} \widetilde{\phi}^{\gamma}_{\sigma_n}(Z_{::n})$. Then each eigenimage, i.e., $Z^{k+1}_{::n}$ can be computed via the shifted and relaxed proximal denoiser as follows

$$Z^{k+1}_{::n} = \widetilde{\mathcal{D}}^{\gamma}_{\sigma_n}(\widehat{Z}^{k}_{::n}), \tag{5.20}$$

where $\widehat{\mathcal{Z}}^k = \mathcal{Z}^k - \tilde{\alpha}_{\mathcal{Z}}(\mathcal{Z}^k - (\mathcal{O} - \mathcal{S}^{k+1}) \times_3 (E^{k+1})^\top)$.

All in all, the proposed PnP-PBCD algorithm for model (5.2) is summarized in Algorithm 2.

Algorithm 2 The PnP-PBCD algorithm for model (5.2)

1: Initialize $(\mathcal{Z}^0, E^0, \mathcal{S}^0)$ with $(E^0)^\top E^0 = \mathbf{I}_r$.
2: Set parameters $\alpha_{\mathcal{S}}, \alpha_{\mathcal{Z}} \geq 0$ and $\alpha_E > 0$.
3: Set $k = 0$.
4: **repeat**
5: Compute $\mathcal{S}^{k+1}$ by (5.13).
6: Compute E^{k+1} by (5.18).
7: Compute $\mathcal{Z}^{k+1}$ by (5.19).
8: $k \leftarrow k + 1$.
9: **until** the stopping criterion is met.

Output: $(\mathcal{Z}^k, E^k, \mathcal{S}^k)$.

5.4 Convergence Analysis

In this section, we conduct a convergence analysis of the proposed PnP-PBCD method. We first define the first order optimality condition of problem (5.2) and then prove that any accumulation point of the sequence generated by the PnP-PBCD method given in Algorithm 2 is a stationary point of problem (5.2).

5.4.1 The First Order Optimality Condition

We define the first order optimality condition of the orthogonal constrained optimization problem as in (5.2). The point $(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$ is a first order stationary point of

problem (5.2) if $0 \in \partial F(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$, that is,

$$0 \in \nabla_{\mathcal{S}} H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}) + \tau \partial \|\cdot\|_{2,\psi}(\bar{\mathcal{S}}), \tag{5.21a}$$

$$0 = \operatorname{grad}_E H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}), \quad \bar{E}^\top \bar{E} = \mathbf{I}_r, \tag{5.21b}$$

$$0 \in \nabla_{\mathcal{Z}} H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}) + \partial \Phi_\Sigma(\bar{\mathcal{Z}}), \tag{5.21c}$$

where $\operatorname{grad}_E H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$ denotes the Riemannian gradient of H with respect to E evaluated at $(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$, and $\partial\|\cdot\|_{2,\psi}$ and $\partial\Phi_\Sigma$ denote the subdifferentials of $\|\cdot\|_{2,\psi}$ and Φ_Σ, respectively.

5.4.2 Subsequence Convergence

We present some assumptions for problem (5.2) as follows.

(A1) $\tau\|\cdot\|_{2,\psi}$ is ρ_1-weakly convex.

(A2) Φ_Σ is coercive and ρ_2-weakly convex.

Note that the coercivity on Φ_Σ required in Assumption (A2) can be achieved by the coercivity of g_σ, according to [38].

We first show two lemmas that we will use in the convergence analysis. The first lemma is for analyzing the updates of $\mathcal{Z}$ and $\mathcal{S}$, and the second lemma is for understanding the update of E.

Lemma 5.2. *Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a ρ-weakly convex function, and let $h : \mathbb{R}^d \to \mathbb{R}$ be a differentiable and μ-strongly convex function. If we have*

$$\hat{x} = \operatorname*{argmin}_x f(x) + h(x) + \frac{\alpha}{2}\|x - x_0\|^2, \tag{5.22}$$

where $\mu - \rho + \alpha > 0$, then

$$f(x_0) + h(x_0) \geq f(\hat{x}) + h(\hat{x}) + \frac{\mu - \rho + \alpha}{2}\|x_0 - \hat{x}\|^2. \tag{5.23}$$

Proof. Since $f(x) + \frac{\rho}{2}\|x - x_0\|^2$ and $h(x) - \frac{\mu}{2}\|x - x_0\|^2$ are convex, we have $f(x) + h(x) + \frac{\alpha}{2}\|x - x_0\|^2$ is $(\mu - \rho + \alpha)$-strongly convex. Then (5.22) is well defined and $\hat{x}$ is uniquely obtained.

Next, for the ρ-weakly convex function f, it follows from Lemma 2.1 in [16] that

$$f(x_0) \geq f(\hat{x}) + \langle u, x_0 - \hat{x}\rangle - \frac{\rho}{2}\|x_0 - \hat{x}\|^2,$$

for $\forall u \in \partial f(\hat{x})$. And for the differentiable and μ-strongly convex function h, it follows from [62] that

$$h(x_0) \geq h(\hat{x}) + \langle \nabla h(\hat{x}), x_0 - \hat{x}\rangle + \frac{\mu}{2}\|x_0 - \hat{x}\|^2.$$

Summing the inequalities above, we obtain

$$f(x_0) + h(x_0) \geq f(\hat{x}) + h(\hat{x}) + \langle u + \nabla h(\hat{x}), x_0 - \hat{x}\rangle + \frac{\mu - \rho}{2}\|x_0 - \hat{x}\|^2.$$

By the update of $\hat{x}$ in (5.22), we have $0 \in \partial f(\hat{x}) + \nabla h(\hat{x}) + \alpha(\hat{x} - x_0)$. That is, $-\nabla h(\hat{x}) + \alpha(x_0 - \hat{x}) \in \partial f(\hat{x})$. Substituting $u = -\nabla h(\hat{x}) + \alpha(x_0 - \hat{x})$ into the above inequality, we obtain (5.23). □

Proposition 5.2. *Let $X \in \mathbb{R}^{n\times r}$, $A \in \mathbb{R}^{r\times m}$ and $B \in \mathbb{R}^{n\times m}$. And let H_1, H_2, and H_3 denote the objective functions of* (5.14a), (5.14b), *and* (5.14c), *respectively. Then we have*

$$\operatorname{grad} H_1(X) = \operatorname{grad} H_2(X) = \operatorname{grad} H_3(X). \tag{5.24}$$

Proof. We first compute the gradients of H_1, H_2, and H_3 as follows

$$\nabla H_1 = (XA - B)A^\top = XAA^\top + \nabla H_2,$$

$$\nabla H_2 = -BA^\top,$$

$$\nabla H_3 = X - BA^\top = X + \nabla H_2.$$

Then by projecting the gradients above onto the tangent space of the Stiefel manifold at X, we can compute the Riemannian gradients of H_1, H_2, and H_3. Since $\mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}$ is linear, we have

$$\mathrm{grad}\, H_1(X) = \mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(XAA^\top) + \mathrm{grad}\, H_2(X)$$

and

$$\mathrm{grad}\, H_3(X) = \mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(X) + \mathrm{grad}\, H_3(X).$$

It is easy to verify that

$$\mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(XAA^\top) = XAA^\top - \frac{1}{2}X(X^\top XAA^\top + (XAA^\top)^\top X) = 0$$

and $\mathrm{Proj}_{\mathcal{T}_X\mathbb{S}_{m,n}}(X) = X - \frac{1}{2}X(X^\top X + X^\top X) = 0$. Therefore, we obtain (5.24). $\square$

Next, we prove the non-increasing monotonicity of the objective sequence $\{F(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ and the boundedness of the sequence $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ generated by Algorithm 2.

Theorem 5.1. *Assume that assumptions (A1) and (A2) are satisfied. Let $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ be the sequence generated by Algorithm 2 with $\alpha_{\mathcal{S}} > \rho_1 - \delta$ and $\alpha_{\mathcal{Z}} > \rho_2 - \delta$. Then the following statements hold.*

(i) The sequence $\{F(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ of function values at the iteration points decreases monotonically, and

$$\begin{aligned} &F(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^{k-1}) - F(\mathcal{Z}^k, E^k, \mathcal{S}^k) \\ \geq &\frac{c_1}{2}\left(\|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F^2 + \|E^k - E^{k-1}\|_F^2 + \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F^2\right), \end{aligned} \tag{5.25}$$

for some $c_1 > 0$.

(ii) The sequence $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ is bounded.

(iii) $\lim_{k\to\infty} \|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F = 0$, $\lim_{k\to\infty} \|E^k - E^{k-1}\|_F = 0$, *and* $\lim_{k\to\infty} \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F = 0$, *for any* $i = 1, 2, 3$.

Proof. (i) The updates of $\mathcal{S}$ and $\mathcal{Z}$ have the form of (5.22). Then Lemma 5.2 implies that

$$F(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^{k-1}) - F(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^k) \geq \frac{\delta - \rho_1 + \alpha_{\mathcal{S}}}{2} \|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F^2$$

and

$$F(\mathcal{Z}^{k-1}, E^k, \mathcal{S}^k) - F(\mathcal{Z}^k, E^k, \mathcal{S}^k) \geq \frac{\delta - \rho_2 + \alpha_{\mathcal{Z}}}{2} \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F^2.$$

Next, it follows from (5.15) and (5.17), we have

$$\begin{aligned} & F(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^k) - F(\mathcal{Z}^{k-1}, E^k, \mathcal{S}^k) \\ =& \widetilde{H}(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^k) - \widetilde{H}(\mathcal{Z}^{k-1}, E^k, \mathcal{S}^k) \\ \geq& \frac{\alpha_E}{2} \|E^k - E^{k-1}\|_F^2. \end{aligned}$$

Combining the inequalities above, we obtain (5.25) with $c_1 = \min\{\delta - \rho_1 + \alpha_{\mathcal{S}}, \alpha_E, \delta - \rho_2 + \alpha_{\mathcal{Z}}\}$.

(ii) Since $(E^k)^\top E^k = \mathbf{I}_r$, we have the sequence $\{E^k\}$ is bounded. By (i), we have $F(\mathcal{Z}^k, E^k, \mathcal{S}^k) \leq F(\mathcal{Z}^0, E^0, \mathcal{S}^0)$. Also, we observe that $F(\mathcal{Z}^k, E^k, \mathcal{S}^k) \geq \tau \|\mathcal{S}^k\|_{2,\psi} + \Phi_\Sigma(\mathcal{Z}^k) \geq 0$. Since both $\|\cdot\|_{2,\psi}$ and Φ_Σ are coercive, that is,

$$\lim_{\|\mathcal{S}\|_F \to \infty} \|\mathcal{S}\|_{2,\psi} = \infty \quad \text{and} \quad \lim_{\|\mathcal{Z}\|_F \to \infty} \Phi_\Sigma(\mathcal{Z}) = \infty,$$

we have the sequences $\{\mathcal{S}^k\}$ and $\{\mathcal{Z}^k\}$ are bounded.

(iii) Let K be an arbitrary integer. Summing (5.25) from $k = 1$ to $K - 1$, we

have

$$\begin{aligned}
&\sum_{k=1}^{K} \left(\|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F^2 + \|E^k - E^{k-1}\|_F^2 + \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F^2 \right) \\
\leq &\frac{2}{c_1} \left(F(\mathcal{Z}^0, E^0, \mathcal{S}^0) - F(\mathcal{Z}^K, E^K, \mathcal{S}^K) \right) \\
\leq &\frac{2}{c_1} F(\mathcal{Z}^0, E^0, \mathcal{S}^0).
\end{aligned}$$

Taking the limits of both sides of the inequality as $K \to \infty$, we have

$$\sum_{k=1}^{\infty} \left(\|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F^2 + \|E^k - E^{k-1}\|_F^2 + \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F^2 \right) < \infty.$$

Then assertion (iii) immediately holds. □

Corollary 5.1. *Assume that assumptions (A1) and (A2) are satisfied. Let* $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ *be the sequence generated by Algorithm 2 with* $\alpha_{\mathcal{S}} > \rho_1 - \delta$ *and* $\alpha_{\mathcal{Z}} > \rho_2 - \delta$. *Then* $\lim_{k\to\infty} \|\mathcal{Z}^k \times_3 E^k - \mathcal{Z}^{k-1} \times_3 E^{k-1}\|_F = 0$.

Proof. We have

$$\begin{aligned}
&\|\mathcal{Z}^k \times_3 E^k - \mathcal{Z}^{k-1} \times_3 E^{k-1}\|_F \\
\leq &\|(\mathcal{Z}^k - \mathcal{Z}^{k-1}) \times_3 E^k\|_F + \|\mathcal{Z}^{k-1} \times_3 (E^k - E^{k-1})\|_F \\
\leq &\sqrt{r}\|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F + c_2\|E^k - E^{k-1}\|_F, \qquad (5.26)
\end{aligned}$$

where $c_2 = \max_k \|\mathcal{Z}^k\|_F < \infty$ according to assertion (ii) in Theorem 5.1. Then by assertion (iii) in Theorem 5.1, the result holds.

□

Lastly, we show that every convergent subsequence converges to the first order stationary point of problem (5.2).

Theorem 5.2. *Assume that assumptions (A1) and (A2) are satisfied. Let* $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ *be the sequence generated by Algorithm 2 with* $\alpha_{\mathcal{S}} > \rho_1 - \delta$ *and* $\alpha_{\mathcal{Z}} > \rho_2 - \delta$*. Then every accumulation point of* $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ *is a first order stationary point of problem* (5.2)*.*

Proof. By the updates of $\mathcal{S}^k$, E^k and $\mathcal{Z}^k$ given in (5.10), (5.11), and (5.12), respectively, we have for any $k = 1, 2, \ldots$

$$
\begin{aligned}
&0 \in \nabla_{\mathcal{S}} H(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^k) + \tau \partial \|\cdot\|_{2,\psi}(\mathcal{S}^k) + \alpha_{\mathcal{S}}(\mathcal{S}^k - \mathcal{S}^{k-1}),\\
&0 = \operatorname{grad}_E \widetilde{H}(\mathcal{Z}^{k-1}, E^k, \mathcal{S}^k), \quad (E^k)^\top E^k = \mathbf{I}_r,\\
&0 \in \nabla_{\mathcal{Z}} H(\mathcal{Z}^k, E^k, \mathcal{S}^k) + \partial \Phi_\Sigma(\mathcal{Z}^k) + \alpha_{\mathcal{Z}}(\mathcal{Z}^k - \mathcal{Z}^{k-1}).
\end{aligned}
$$

Then we have

$$
\begin{aligned}
A_{\mathcal{S}}^k :&= -\alpha_{\mathcal{S}}(\mathcal{S}^k - \mathcal{S}^{k-1}) + \nabla_{\mathcal{S}} H(\mathcal{Z}^k, E^k, \mathcal{S}^k) - \nabla_{\mathcal{S}} H(\mathcal{Z}^{k-1}, E^{k-1}, \mathcal{S}^k) \qquad (5.27a)\\
&\in \nabla_{\mathcal{S}} H(\mathcal{Z}^k, E^k, \mathcal{S}^k) + \tau \partial \|\cdot\|_{2,\psi}(\mathcal{S}^k),\\
A_E^k :&= \operatorname{grad}_E \widetilde{H}(\mathcal{Z}^k, E^k, \mathcal{S}^k) - \operatorname{grad}_E \widetilde{H}(\mathcal{Z}^{k-1}, E^k, \mathcal{S}^k) \qquad (5.27b)\\
&= \operatorname{grad}_E H(\mathcal{Z}^k, E^k, \mathcal{S}^k),
\end{aligned}
$$

and

$$
\begin{aligned}
A_{\mathcal{Z}}^k :&= -\alpha_{\mathcal{Z}}(\mathcal{Z}^k - \mathcal{Z}^{k-1}) \qquad (5.27c)\\
&\in \nabla_{\mathcal{Z}} H(\mathcal{Z}^k, E^k, \mathcal{S}^k) + \partial \Phi_\Sigma(\mathcal{Z}^k).
\end{aligned}
$$

Note that (5.27b) is obtained by Proposition 5.2. Furthermore, since $\nabla_{\mathcal{S}} H(\mathcal{Z}, E, \mathcal{S}) = \delta(\mathcal{Z} \times_3 E + \mathcal{S} - \mathcal{O})$, by (5.26) we have

$$
\begin{aligned}
\|A_{\mathcal{S}}^k\|_F &\le \alpha_{\mathcal{S}} \|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F + \delta \|\mathcal{Z}^k \times_3 E^k - \mathcal{Z}^{k-1} \times_3 E^{k-1}\|_F\\
&\le \alpha_{\mathcal{S}} \|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F + \delta\sqrt{r} \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F + c_2 \delta \|E^k - E^{k-1}\|_F. \qquad (5.28)
\end{aligned}
$$

Since $\operatorname{grad}_E \widetilde{H}(\mathcal{Z}, E, \mathcal{S}) = \operatorname{Proj}_{\mathcal{T}_{E^{\mathbb{S}_{n_3,r}}}} \left(-\delta(\mathcal{O}-\mathcal{S})_{(3)}(\mathcal{Z}_{(3)})^\top\right)$, we have

$$\begin{aligned}\|A_E^k\|_F &\le \left\|\operatorname{Proj}_{\mathcal{T}_{E^k}\mathbb{S}_{n_3,r}}\left(-\delta(\mathcal{O}-\mathcal{S}^k)_{(3)}(\mathcal{Z}_{(3)}^k - \mathcal{Z}_{(3)}^{k-1})^\top\right)\right\|_F \\ &\le (1+r^2)\|\delta(\mathcal{O}-\mathcal{S}^k)_{(3)}(\mathcal{Z}_{(3)}^k - \mathcal{Z}_{(3)}^{k-1})^\top\|_F \\ &\le (1+r^2)\delta c_3\|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F,\end{aligned}$$

where $c_3 = \max_k \|\mathcal{O}-\mathcal{S}^k\|_F$. Also, we have

$$\|A_{\mathcal{Z}}^k\|_F \le \alpha_{\mathcal{Z}}\|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F.$$

Suppose that $\{(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l})\}$ is a subsequence of $\{(\mathcal{Z}^k, E^k, \mathcal{S}^k)\}$ which converges to $(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$ as $l \to \infty$. It immediately follows from Theorem 5.2 (iii) that $A_{\mathcal{S}}^{k_l} \to 0$, $A_E^{k_l} \to 0$, and $A_{\mathcal{Z}}^{k_l} \to 0$, as $l \to \infty$. Also, due to the continuity of $\nabla_{\mathcal{S}} H$, $\operatorname{grad}_E H$ and $\nabla_{\mathcal{Z}} H$, we have

$$\begin{aligned}\nabla_{\mathcal{S}} H(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l}) &\to \nabla_{\mathcal{S}} H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}), \\ \operatorname{grad}_E H(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l}) &\to \operatorname{grad}_E H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}), \\ \nabla_{\mathcal{Z}} H(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l}) &\to \nabla_{\mathcal{Z}} H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}}),\end{aligned}$$

as $l \to \infty$. Then $(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$ satisfies (5.21b) for the first order optimality condition of problem (5.2).

Next, we show $(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$ satisfies (5.21a) and (5.21c). Since $\|\cdot\|_{2,\psi}$ is lower semicontinuous, we have

$$\liminf_{l\to\infty} \|\mathcal{S}^{k_l}\|_{2,\psi} \ge \|\bar{\mathcal{S}}\|_{2,\psi}.$$

Then according to (5.10), we have

$$\begin{aligned}&H(\mathcal{Z}^{k_l-1}, E^{k_l-1}, \mathcal{S}^{k_l}) + \tau\|\mathcal{S}^{k_l}\|_{2,\psi} + \frac{\alpha_{\mathcal{S}}}{2}\|\mathcal{S}^{k_l} - \mathcal{S}^{k_l-1}\|_F^2 \\ \le &H(\mathcal{Z}^{k_l-1}, E^{k_l-1}, \bar{\mathcal{S}}) + \tau\|\bar{\mathcal{S}}\|_{2,\psi} + \frac{\alpha_{\mathcal{S}}}{2}\|\bar{\mathcal{S}} - \mathcal{S}^{k_l-1}\|_F^2.\end{aligned}$$

The continuity of H implies that $\lim_{l\to\infty} H(\mathcal{Z}^{k_l-1}, E^{k_l-1}, \bar{\mathcal{S}}) - H(\mathcal{Z}^{k_l-1}, E^{k_l-1}, \bar{\mathcal{S}}) = 0$. By rewriting the above inequality and taking the limit superior of both sides as $l \to \infty$, we have

$$\limsup_{l\to\infty} \|\mathcal{S}^{k_l}\|_{2,\psi} \leq \|\bar{\mathcal{S}}\|_{2,\psi}.$$

Recall the definition of (limiting) subdifferential. Since we have $\mathcal{S}^{k_l} \to \bar{\mathcal{S}}$, $\|\mathcal{S}^{k_l}\|_{2,\psi} \to \|\bar{\mathcal{S}}\|_{2,\psi}$, $A_{\mathcal{S}}^{k_l} - \nabla_{\mathcal{S}} H(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l}) \in \tau\partial\|\cdot\|_{2,\psi}(\mathcal{S}^{k_l})$ and $A_{\mathcal{S}}^{k_l} - \nabla_{\mathcal{S}} H(\mathcal{Z}^{k_l}, E^{k_l}, \mathcal{S}^{k_l}) \to \nabla_{\mathcal{S}} H(\bar{\mathcal{Z}}, \bar{E}, \bar{\mathcal{S}})$, as $l \to \infty$, we have (5.21a) is satisfied.

Similarly, we can show (5.21c) is satisfied. □

Remark 5.1. *The parameter α_E can be chosen as 0, only if $(\mathcal{O}-\bar{\mathcal{S}})_{(3)}(\bar{\mathcal{Z}}_{(3)})^\top$ is full-rank. In this case, we can achieve $\lim_{k\to\infty} \|E^k - E^{k-1}\|_F = 0$ using (5.18), $\lim_{k\to\infty} \|\mathcal{Z}^k - \mathcal{Z}^{k-1}\|_F = 0$, and $\lim_{k\to\infty} \|\mathcal{S}^k - \mathcal{S}^{k-1}\|_F = 0$, without using $\alpha_E > 0$.*

5.5 Numerical Experiments

In this section, we conduct numerical experiments for anomaly detection in noisy HSIs. We compare the proposed method with RX [67], LRASR [94], LSCTV [24], TLRSR [81], Auto-AD [82] methods. All the numerical experiments are executed on a personal desktop with an Intel Core i7 9750H at 2.60 GHz with 16 GB RAM.

The proposed anomaly detection method and the competing methods are tested on four HSIs from the classic "airport-beach-urban" dataset [40], where items such as boats, cars, and airplanes are labeled as anomaly objects. The details of the test HSIs are shown as follows.

- "Airport" was captured by the airborne visible/infrared imaging spectrometer (AVIRIS) [28] sensor over Los Angeles, with a spatial resolution of 7.1 m. The HSI size is $100 \times 100 \times 205$.

- "Beach" was captured by the ROSIS-03 sensor in Pavia, with a spatial resolution of 1.3 m. The HSI size is $150 \times 150 \times 102$.

- "Urban 1" and "Urban 2" were also captured by the AVIRIS over Los Angeles, with a spatial resolution of 7.1 m. The HSIs are of size $100 \times 100 \times 205$.

We test all the methods for anomaly detection on the original HSIs and on simulated noisy HSIs degraded by Gaussian noise with a noise level of 0.03.

To evaluate the accuracy of object detection, we plot the receiver operating characteristic (ROC) curve [41] and calculate the area under the ROC curve (AUC) [42]. The ROC curve plots the probability of detection vs the false alarm rate for various possible thresholds. The closer the ROC curve is to the upper-left corner and the larger AUC score, the better the detection performance.

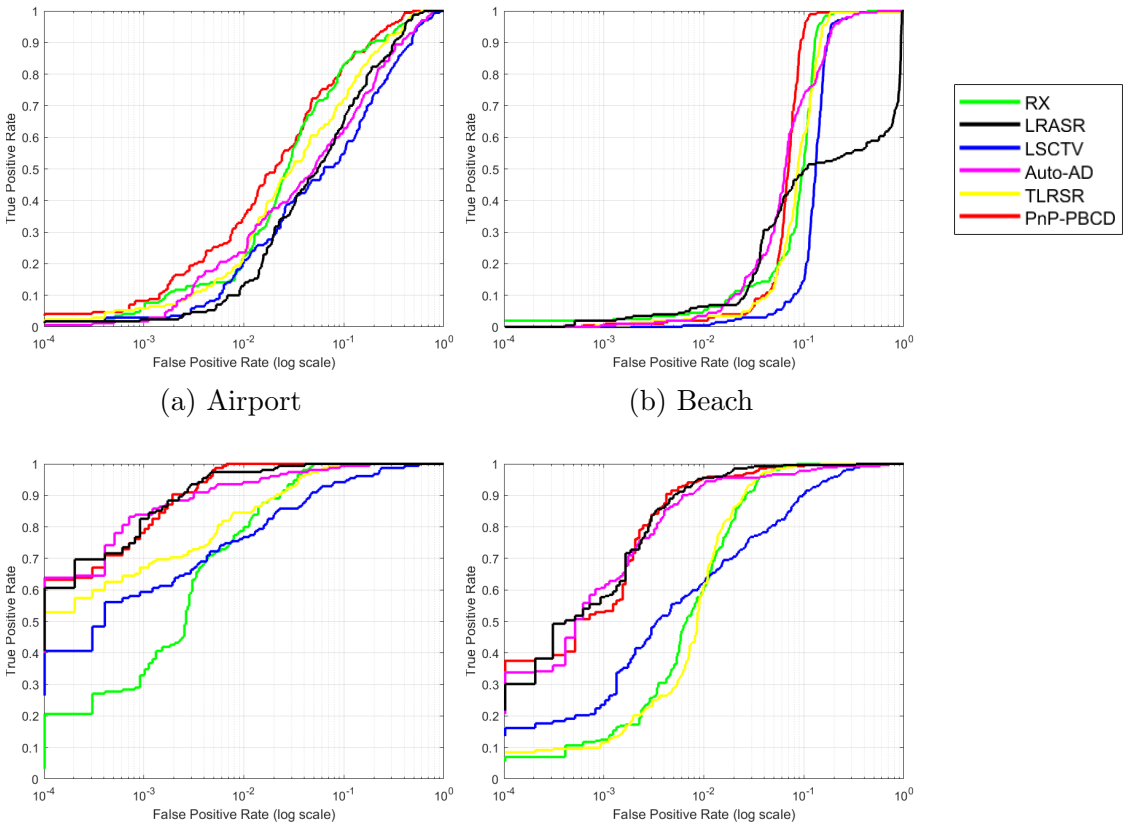
Table 5.1: Comparison of average AUC scores and average computational time obtained by different methods.

Noise level	**Index**	**RX**	**LRASR**	**LSCTV**	**Auto-AD**	**TLRSR**	**PnP-PBCD (ours)**
0	Ave AUC	0.9550	0.8702	0.9137	0.9437	0.9500	**0.9663**
	Ave Time	**0.21**	4.06	364.13	3.75	4.79	14.86
0.03	Ave AUC	0.8948	0.8831	0.9089	0.9478	0.7939	**0.9607**
	Ave Time	**0.21**	5.02	382.90	4.47	5.15	13.62

Table 5.1 presents the average AUC scores and computational times of several methods obtained from HSIs with no noise and with a noise level of 0.03. The results demonstrate that PnP-PBCD consistently achieves superior anomaly detection performance while maintaining reasonable computational efficiency. At a noise level of 0, PnP-PBCD achieves the highest average AUC score of 0.9663, outperforming all other methods. RX and TLRSR also perform well, with scores of 0.9550 and 0.9500, respectively. In terms of computational time, RX is the fastest, requiring only 0.21 seconds. While PnP-PBCD takes 14.86 seconds, it balances this computational cost

with superior detection accuracy. At a noise level of 0.03, PnP-PBCD maintains its leading position with an average AUC score of 0.9607, showcasing its robustness to noise. Auto-AD follows closely with a score of 0.9478, while RX achieves 0.8948. Computational times remain consistent, with RX retaining the shortest time of 0.21 seconds, and PnP-PBCD achieving a good balance between accuracy and computational cost at 13.62 seconds.

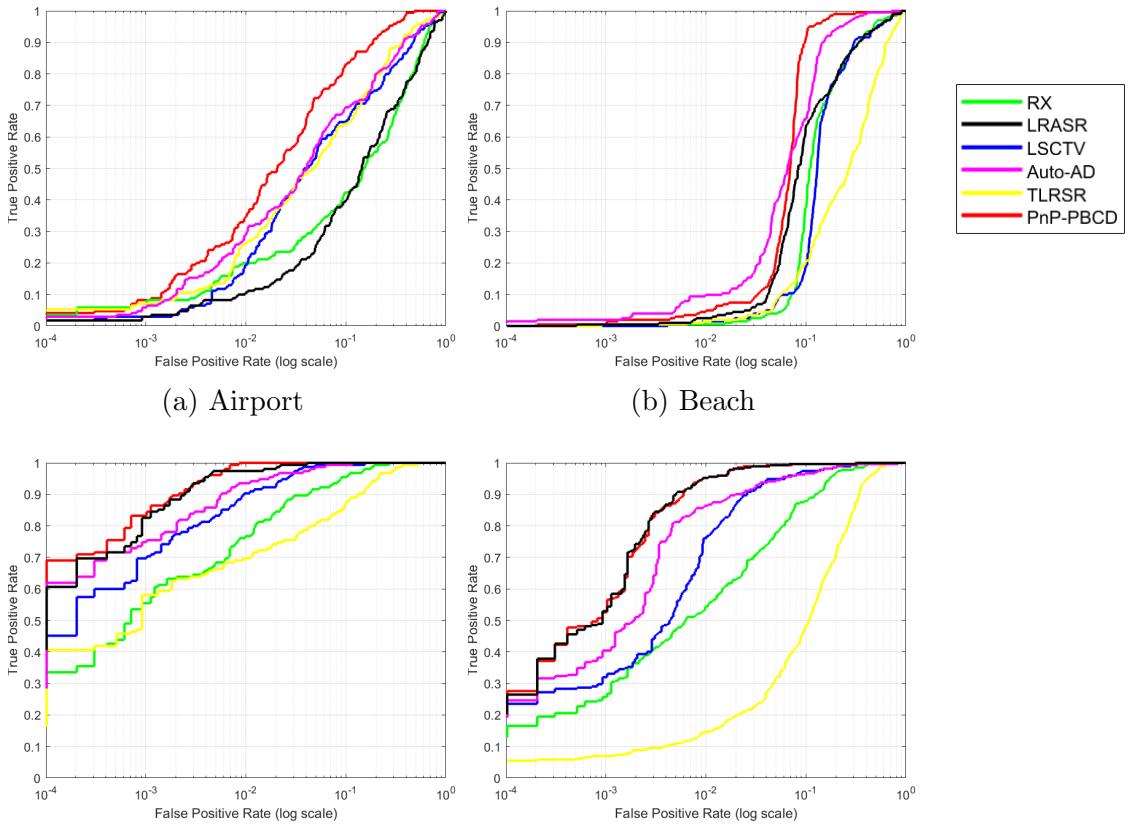
Figure 5.2 and Figure 5.3 present the ROC curves obtained from HSIs with no noise and with a noise level of 0.03, respectively. Each figure evaluates four scenes: Airport, Beach, Urban 1, and Urban 2. The x-axis represents the false positive rate (log scale), and the y-axis represents the true positive rate.



(a) Airport (b) Beach

(c) Urban 1 (d) Urban 2

Figure 5.2: Comparison of ROC curves obtained by different methods from HSIs with no noise.



(a) Airport (b) Beach

(c) Urban 1 (d) Urban 2

Figure 5.3: Comparison of ROC curves obtained by different methods from HSIs with a noise level of 0.03.

In Figure 5.2, which corresponds to noise-free data, PnP-PBCD (red curve) consistently outperforms all other methods, achieving higher true positive rates at lower false positive rates across all scenes. The method shows a particularly clear advantage in Urban 1 and Urban 2. While Auto-AD (pink curve) and RX (green curve) show competitive performance in certain regions, they generally underperform compared to PnP-PBCD.

In Figure 5.3, with a noise level of 0.03, PnP-PBCD maintains its strong performance across all scenarios, highlighting its robustness. Its ROC curves remain above those of competing methods in most cases, particularly at lower false positive rates. Auto-AD performs well under noisy conditions, achieving competitive results

in several scenes. However, methods such as TLRSR (yellow curve) and RX are more sensitive to noise, showing reduced performance, particularly in Urban 1. LRASR demonstrates strong performance in some scenes, such as Urban 1 and Urban 2, but struggles in others, such as Airport and Beach.

Overall, these results highlight the robustness and reliability of the proposed PnP-PBCD method. It not only delivers superior detection performance in noise-free settings but also demonstrates resilience to noise, maintaining consistently high AUC scores and favorable ROC curves across diverse test scenarios. While RX excels in computational efficiency and Auto-AD performs well in noisy conditions, PnP-PBCD stands out as a comprehensive solution that effectively balances accuracy, robustness, and efficiency for anomaly detection in HSIs.

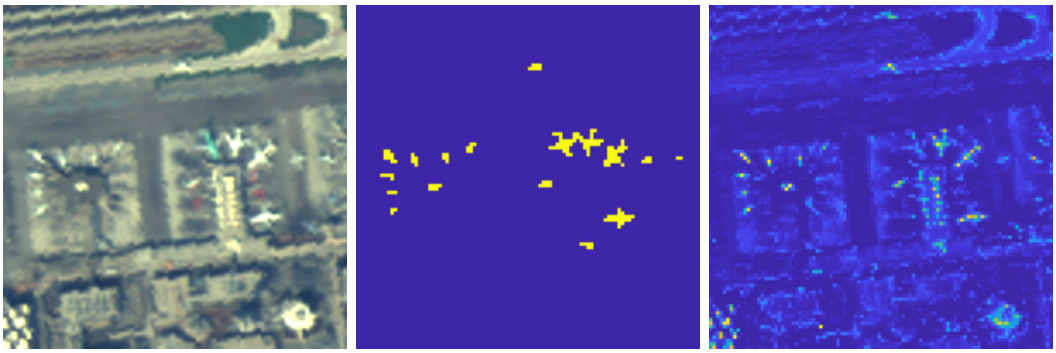
Lastly, Figures 5.4-5.9 illustrate the visual comparisons of anomaly detection using different methods, showcasing various scenes under both noise-free and noisy conditions. These comparisons highlight the performance of RX, LRASR, LSCTV, Auto-AD, TLRSR, and the proposed PnP-PBCD method in terms of their ability to accurately detect anomalies while avoiding noise interference.

In noise-free conditions, as shown in Figures 5.4, 5.6, and 5.5, the proposed PnP-PBCD method consistently provides the most accurate and complete detection results. In all scenes, PnP-PBCD achieves strong alignment with the ground truth, detecting anomalies with well-defined boundaries and high density. Auto-AD also performs effectively, capturing the anomalies with a slightly lower intensity compared to PnP-PBCD, but without significant omissions. In contrast, RX, LRASR, and LSCTV exhibit incomplete or scattered detections, often missing key anomalies or introducing artifacts. TLRSR, while capturing many anomalies, tends to include additional noise or false positives, reducing its overall precision in comparison to PnP-PBCD and Auto-AD.

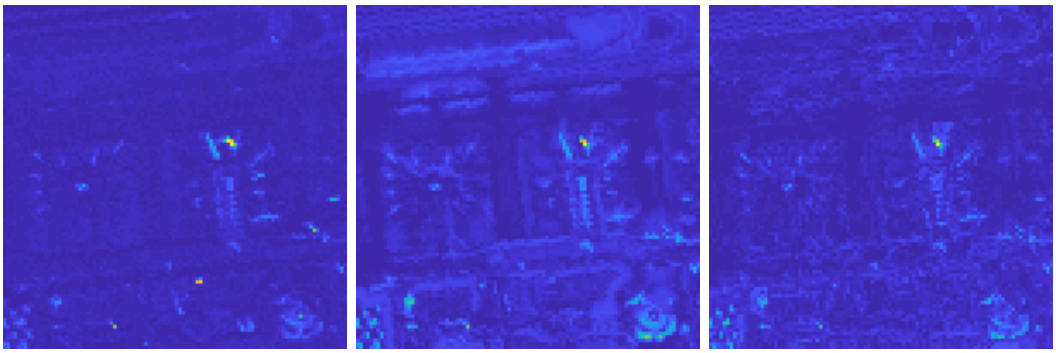
When noise is presented, as shown in Figures 5.7, 5.8, and 5.9, the robustness of

each method becomes more apparent. The proposed PnP-PBCD method maintains its great performance, effectively distinguishing anomalies from noise and preserving high detection density and accuracy. Auto-AD also remains reliable under noisy conditions, producing comparable results to its noise-free performance. In contrast, RX and TLRSR are significantly affected by the noise, frequently misclassifying noise as anomalies, leading to high false positive rates. LRASR and LSCTV also struggle in noisy environments, showing reduced detection accuracy and failing to adapt effectively to the added noise.

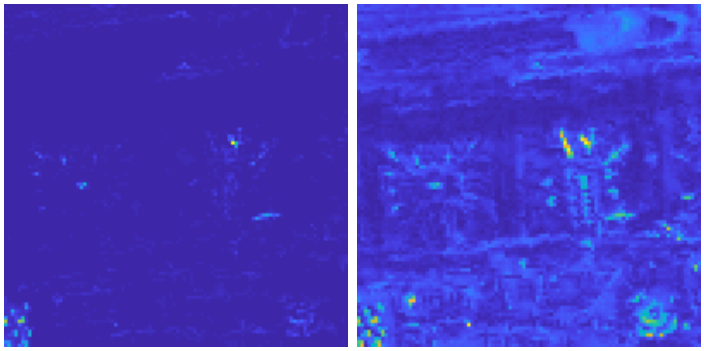
All in all, the proposed PnP-PBCD method is an efficient method for anomaly detection, outperforming the competing methods both numerically and visually, especially in noisy HSIs.



(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

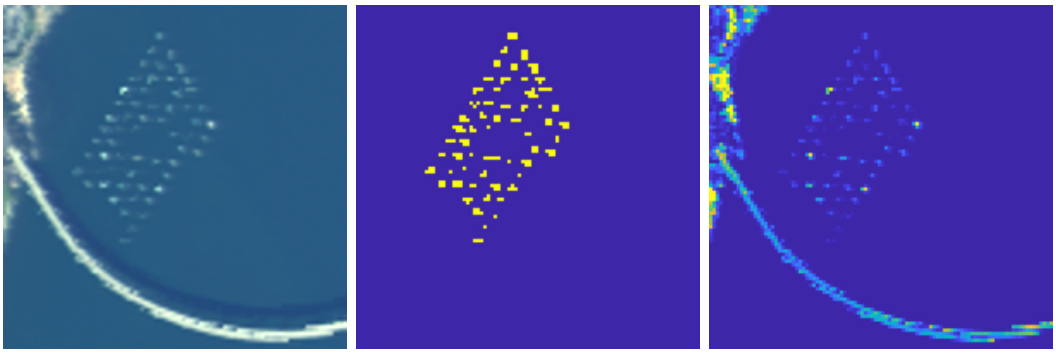


(d) RX (e) LRASR (f) LSCTV

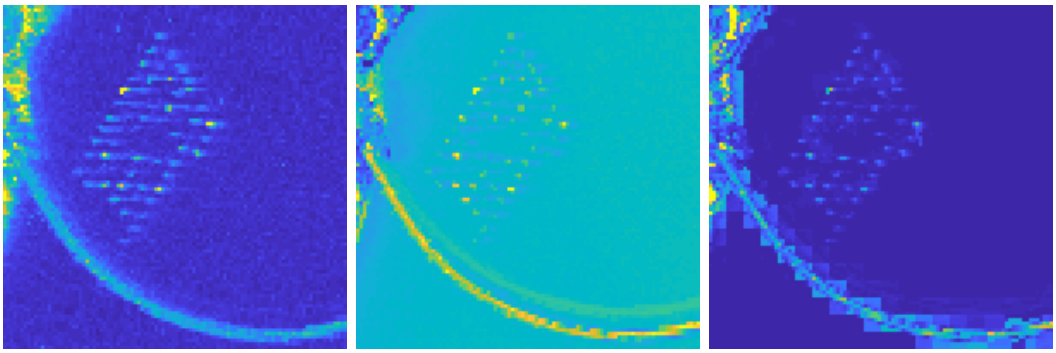


(g) Auto-AD (h) TLRSR

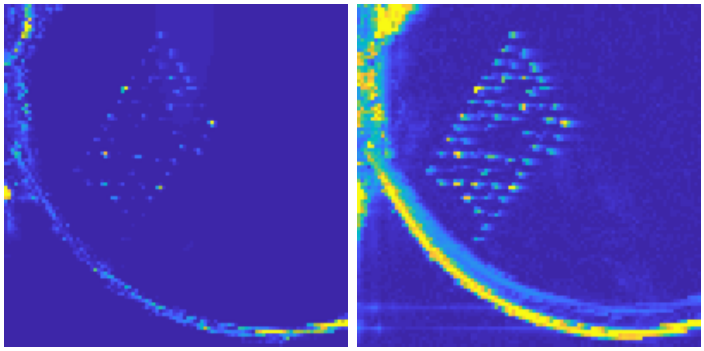
Figure 5.4: Comparison of anomaly objects detected by different methods from "Airport" with no noise.



(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

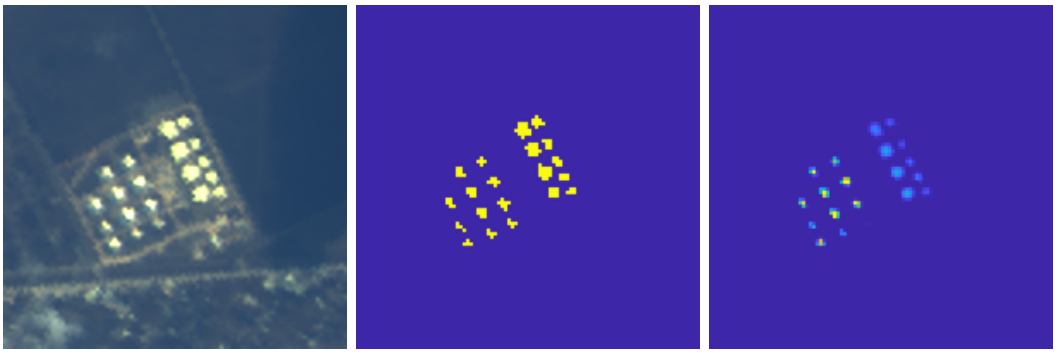


(d) RX (e) LRASR (f) LSCTV

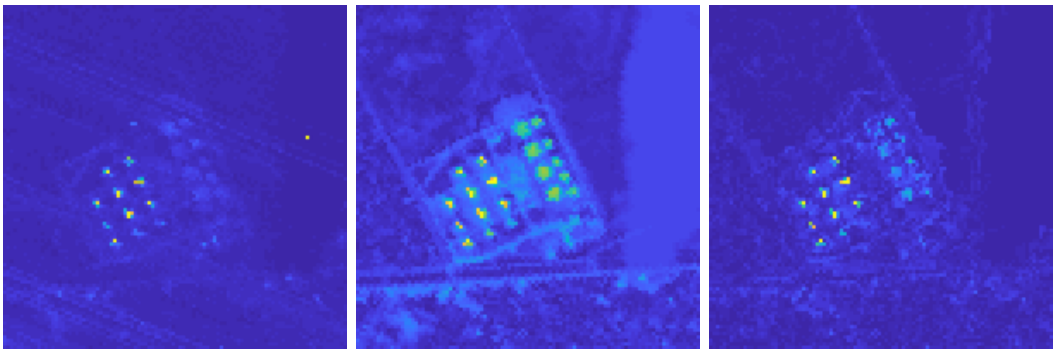


(g) Auto-AD (h) TLRSR

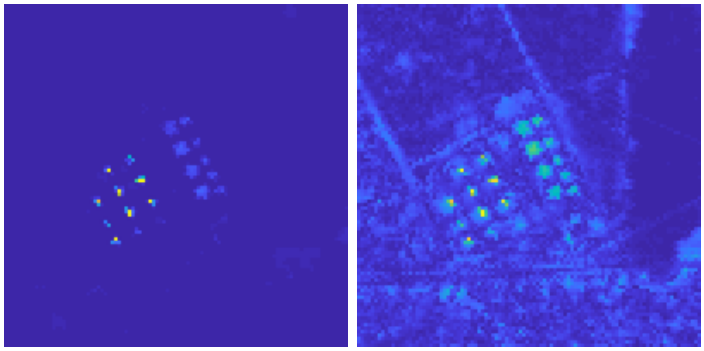
Figure 5.5: Comparison of anomaly objects detected by different methods from "Beach" with no noise.



(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

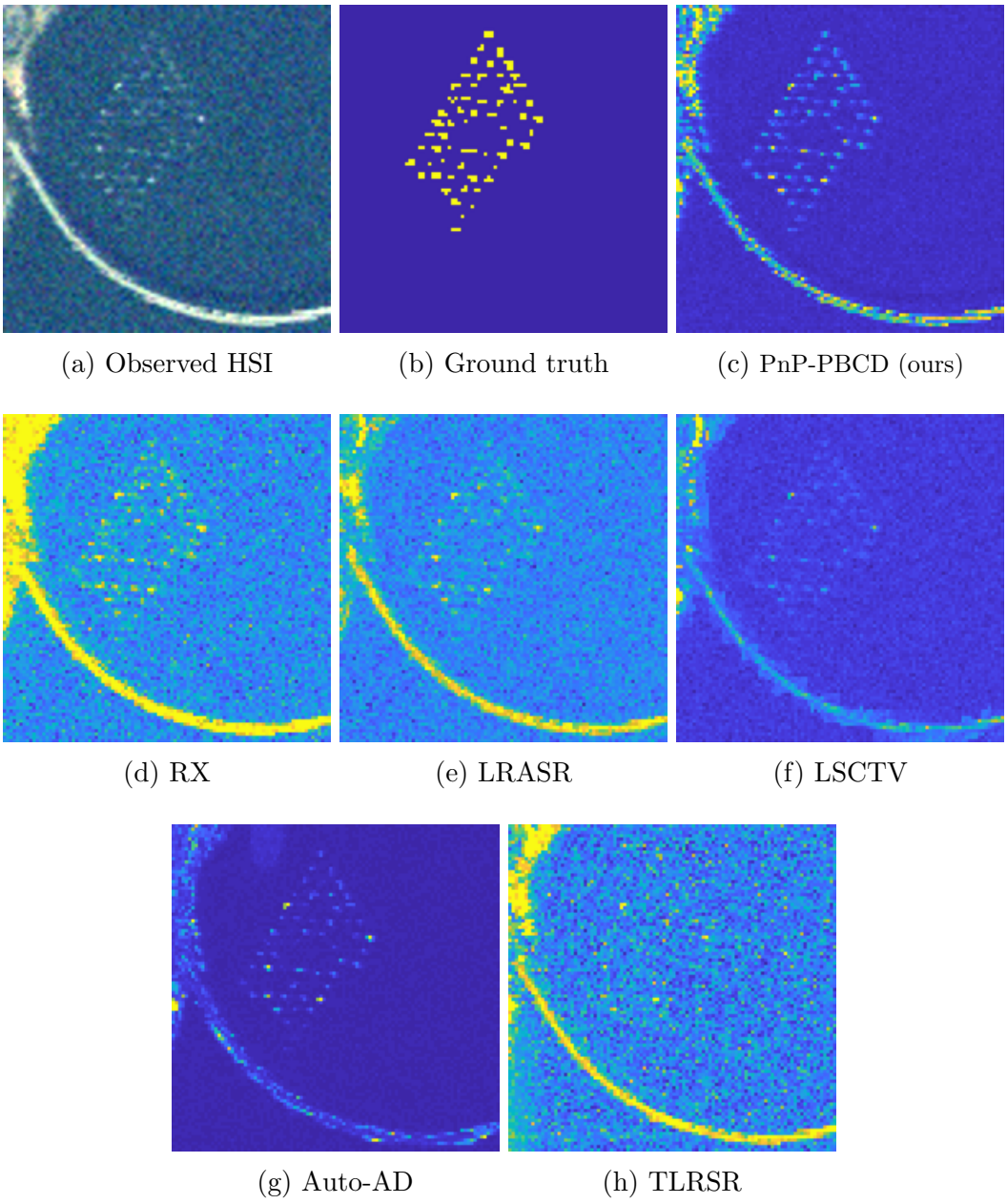


(d) RX (e) LRASR (f) LSCTV



(g) Auto-AD (h) TLRSR

Figure 5.6: Comparison of anomaly objects detected by different methods from "Urban 1" with no noise.

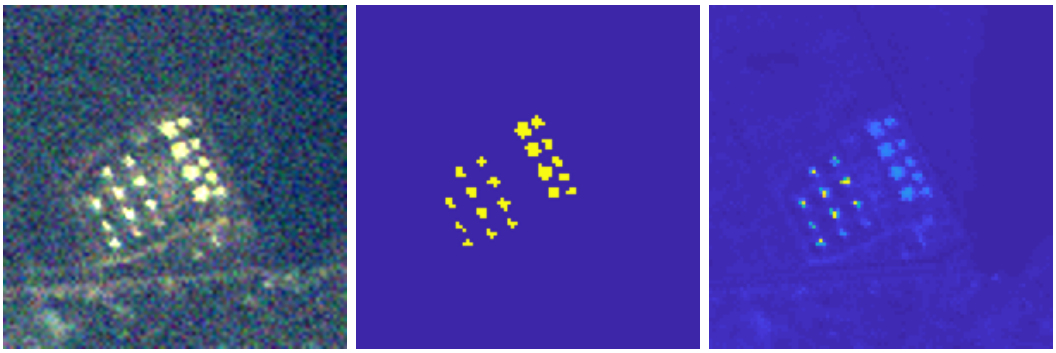


(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

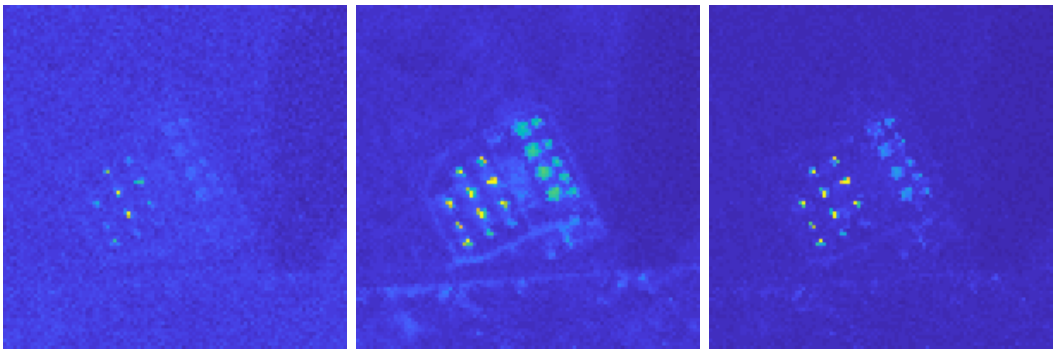
(d) RX (e) LRASR (f) LSCTV

(g) Auto-AD (h) TLRSR

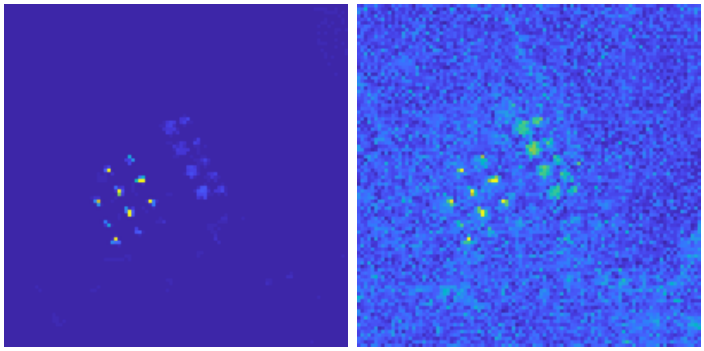
Figure 5.7: Comparison of anomaly objects detected by different methods from "Beach" with a noise level of 0.03.



(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

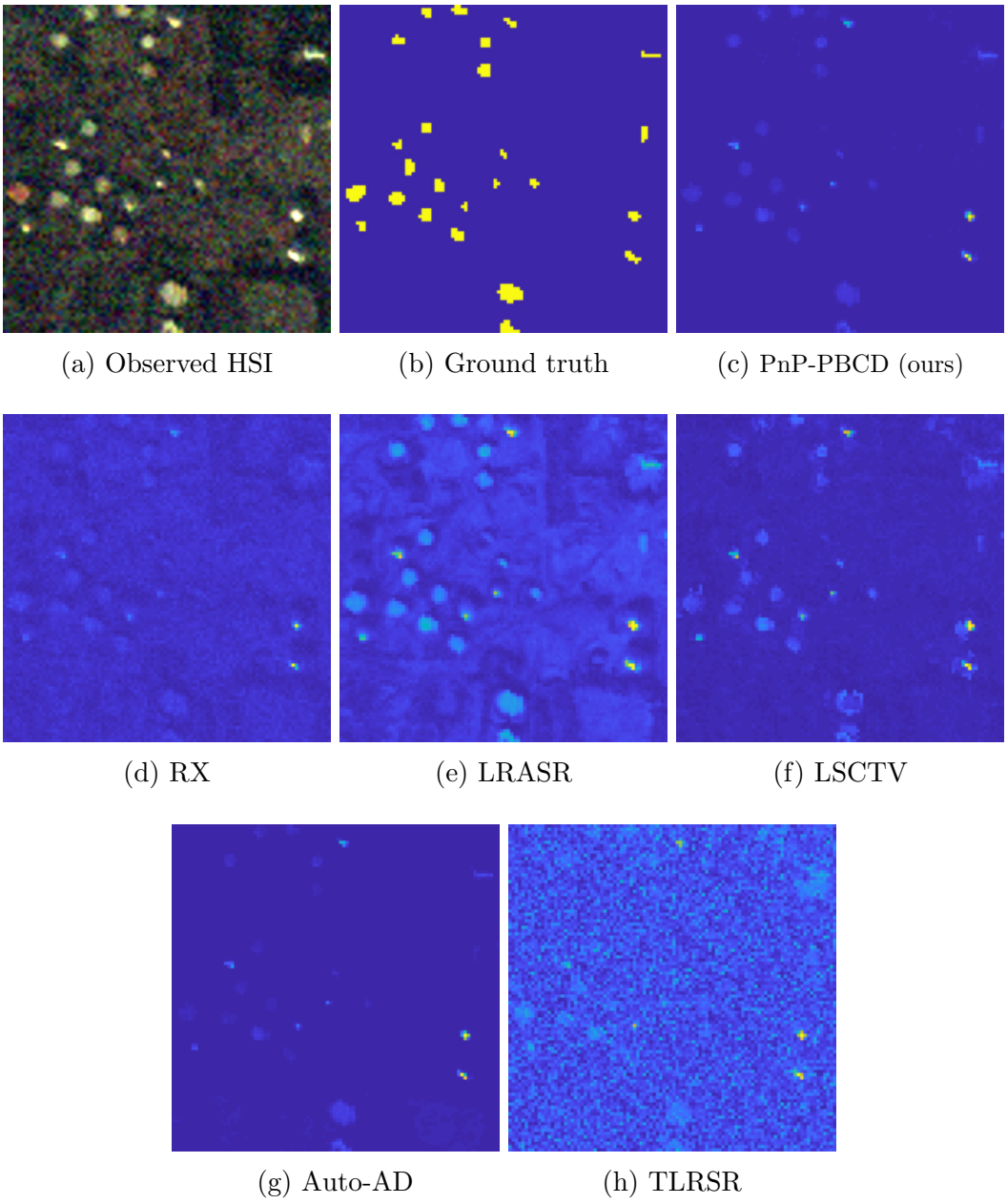


(d) RX (e) LRASR (f) LSCTV



(g) Auto-AD (h) TLRSR

Figure 5.8: Comparison of anomaly objects detected by different methods from "Urban 1" with a noise level of 0.03.



(a) Observed HSI (b) Ground truth (c) PnP-PBCD (ours)

(d) RX (e) LRASR (f) LSCTV

(g) Auto-AD (h) TLRSR

Figure 5.9: Comparison of anomaly objects detected by different methods from "Urban 2" with a noise level of 0.03.

Chapter 6

Conclusion and Future Work

This chapter concludes the thesis and points out some possible research directions related to the work done in this thesis.

6.1 Conclusion

This thesis presented a comprehensive study on HSI restoration and anomaly detection, addressing key challenges associated with noise, and high-dimensional data. By utilizing nonlocal low-rank tensor regularization and tensor group sparsity measure $\|\cdot\|_{2,\psi}$, along with spatial and spectral priors of HSIs or deep learning based proximal denoiser priors, we developed two optimization models and iterative algorithms with convergence guarantees. These models effectively addressed tasks in HSI processing such as mixed noise removal and hyperspectral anomaly detection.

6.2 Future Work

Some possible future research directions extended from this thesis are presented as follows.

- The proposed P-BCD algorithm presented in Algorithm 1 updates the variables associated with orthogonal constraints using a projection given in Lemma 3.1.

However, conducting the SVD on a matrix with large sizes may be time-consuming. We may propose new algorithms that conduct projections onto the tangent space of the Stiefel manifold to avoid implementing SVD. For example, we may extend the manifold proximal gradient algorithm [12] or the alternating manifold proximal gradient algorithm [13] from solving objective functions with one or two variables to solving objective functions with multiple variables, specifically for our model (2.1).

- The proposed PnP-PBCD method presented in Algorithm 2 utilizes the proximal denoiser proposed in [39, 38]. However, the existing proximal-type denoisers are limited to the formulation proposed in [39, 38] and a specific pre-trained neural network for denoising grayscale or color images, called DRUNet. Hence, we may develop new proximal-type denoisers which can denoise HSIs directly without using eignimages band by band.

- In addition to the subsequence convergence that we have proved for both the P-BCD algorithm and the PnP-PBCD method, we may show the whole sequence convergence of both methods by assuming the KL property on the objective functions.

Bibliography

[1] T. Abrudan, J. Eriksson, and V. Koivunen, *Conjugate gradient algorithm for optimization under unitary matrix constraint*, Signal Processing, 89 (2009), pp. 1704–1714.

[2] T. E. Abrudan, J. Eriksson, and V. Koivunen, *Steepest descent algorithms for optimization under unitary matrix constraint*, IEEE Transactions on Signal Processing, 56 (2008), pp. 1134–1147.

[3] P. A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.

[4] J. M. Bioucas-Dias and J. M. Nascimento, *Hyperspectral subspace identification*, IEEE Transactions on Geoscience and Remote Sensing, 46 (2008), pp. 2435–2445.

[5] A. Böhm and S. J. Wright, *Variable smoothing for weakly convex composite functions*, Journal of Optimization Theory and Applications, 188 (2021), pp. 628–649.

[6] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[7] J. Bradbury, S. Merity, C. Xiong, and R. Socher, *Quasi-recurrent neural networks*, arXiv preprint arXiv:1611.01576, (2016).

[8] A. Buades, B. Coll, and J.-M. Morel, *A non-local algorithm for image denoising*, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, IEEE, 2005, pp. 60–65.

[9] C. Cao, J. Yu, C. Zhou, K. Hu, F. Xiao, and X. Gao, *Hyperspectral image denoising via subspace-based nonlocal low-rank and sparse factorization*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12 (2019), pp. 973–988.

[10] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, vol. 1, Springer Science and Business Media, 2003.

[11] J. Chen and Y. Saad, *On the tensor SVD and the optimal low rank orthogonal approximation of tensors*, SIAM Journal on Matrix Analysis and Applications, 30 (2009), pp. 1709–1734.

[12] S. Chen, S. Ma, A. Man-Cho So, and T. Zhang, *Proximal gradient method for nonsmooth optimization over the Stiefel manifold*, SIAM Journal on Optimization, 30 (2020), pp. 210–239.

[13] S. Chen, S. Ma, L. Xue, and H. Zou, *An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis*, INFORMS Journal on Optimization, 2 (2020), pp. 192–208.

[14] Y. Chen, T.-Z. Huang, and X.-L. Zhao, *Destriping of multispectral remote sensing image using low-rank tensor decomposition*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11 (2018), pp. 4950–4967.

[15] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising by sparse 3D transform-domain collaborative filtering*, IEEE Transactions on Image Processing, 16 (2007), pp. 2080–2095.

[16] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 207–239.

[17] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A multilinear singular value decomposition*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 1253–1278.

[18] J. H. de Morais Goulart and G. Favier, *Low-rank tensor recovery using sequentially optimal modal projections in iterative hard thresholding (SeMPIHT)*, SIAM Journal on Scientific Computing, 39 (2017), pp. A860–A889.

[19] Y. Ding, W. He, J. Tang, Q. Zou, and F. Guo, *Laplacian regularized sparse representation based classifier for identifying DNA N4-methylcytosine sites via $l_{2,1/2}$-matrix norm*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20 (2021), pp. 500–511.

[20] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM Journal on Matrix Analysis and Applications, 20 (1998), pp. 303–353.

[21] M. Elad and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, IEEE Transactions on Image Processing, 15 (2006), pp. 3736–3745.

[22] J. Fan, *Comments on "wavelets in statistics: A review" by a. antoniadis*, Journal of the Italian Statistical Society, 6 (1997), p. 131.

[23] M. Feng, W. Chen, Y. Yang, Q. Shu, H. Li, and Y. Huang, *Hyperspectral anomaly detection based on tensor ring decomposition with factors TV regularization*, IEEE Transactions on Geoscience and Remote Sensing, 61 (2023), pp. 1–14.

[24] R. Feng, H. Li, L. Wang, Y. Zhong, L. Zhang, and T. Zeng, *Local spatial constraint and total variation for hyperspectral anomaly detection*, IEEE Transactions on Geoscience and Remote Sensing, 60 (2021), pp. 1–16.

[25] B. Gao, X. Liu, X. Chen, and Y.-X. Yuan, *A new first-order algorithmic framework for optimization problems with orthogonality constraints*, SIAM Journal on Optimization, 28 (2018), pp. 302–332.

[26] K. Gao and Z.-H. Huang, *Tensor robust principal component analysis via tensor fibered rank and minimization*, SIAM Journal on Imaging Sciences, 16 (2023), pp. 423–460.

[27] H. Grahn and P. Geladi, *Techniques and Applications of Hyperspectral Image Analysis*, John Wiley and Sons, 2007.

[28] R. O. Green, M. L. Eastwood, C. M. Sarture, T. G. Chrien, M. Aronsson, B. J. Chippendale, J. A. Faust, B. E. Pavri, C. J. Chovit, M. Solis, et al., *Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (aviris)*, Remote Sensing of Environment, 65 (1998), pp. 227–248.

[29] J. Guo, Y. Guo, Q. Jin, M. Kwok-Po Ng, and S. Wang, *Gaussian patch mixture model guided low-rank covariance matrix minimization for image denoising*, SIAM Journal on Imaging Sciences, 15 (2022), pp. 1601–1622.

[30] Q. Guo, B. Zhang, Q. Ran, L. Gao, J. Li, and A. Plaza, *Weighted-RXD and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7 (2014), pp. 2351–2366.

[31] N. J. Higham, *Matrix nearness problems and applications*, Applications of Matrix Theory, 22 (1989).

[32] C. Hildreth, *A quadratic programming procedure*, Naval Research Logistics Quarterly, 4 (1957), pp. 79–85.

[33] J. T. Hong Ye Tan, Subhadip Mukherjee and C.-B. Schönlieb, *Provably convergent plug-and-play quasi-Newton methods*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 785–819.

[34] A. Hore and D. Ziou, *Image quality metrics: PSNR vs. SSIM*, in 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 2366–2369.

[35] C. Huang, Z. Wu, Y. Cheng, T. Zeng, C.-B. Schönlieb, and A. I. Aviles-Rivero, *Deep block proximal linearised minimisation algorithm for non-convex inverse problems*, arXiv preprint arXiv:2406.02458, (2024).

[36] J. Huang and C. X. Ling, *Using AUC and accuracy in evaluating learning algorithms*, IEEE Transactions on Knowledge and Data Engineering, 17 (2005), pp. 299–310.

[37] W. Huang, K. A. Gallivan, and P.-A. Absil, *A broyden class of quasi-Newton methods for Riemannian optimization*, SIAM Journal on Optimization, 25 (2015), pp. 1660–1685.

[38] S. Hurault, A. Chambolle, A. Leclaire, and N. Papadakis, *A relaxed proximal gradient descent algorithm for convergent plug-and-play with proximal denoiser*, in International Conference on Scale Space and Variational Methods in Computer Vision, Springer, 2023, pp. 379–392.

[39] S. Hurault, A. Leclaire, and N. Papadakis, *Proximal denoiser for convergent plug-and-play optimization with nonconvex regularization*, in International Conference on Machine Learning, PMLR, 2022, pp. 9483–9505.

[40] X. Kang, X. Zhang, S. Li, K. Li, J. Li, and J. A. Benediktsson, *Hyperspectral anomaly detection with attribute and edge-preserving filters*, IEEE Transactions on Geoscience and Remote Sensing, 55 (2017), pp. 5600–5611.

[41] J. Kerekes, *Receiver operating characteristic curve confidence intervals and regions*, IEEE Geoscience and Remote Sensing Letters, 5 (2008), pp. 251–255.

[42] S. Khazai, S. Homayouni, A. Safari, and B. Mojaradi, *Anomaly detection in hyperspectral images based on an adaptive support vector method*, IEEE Geoscience and Remote Sensing Letters, 8 (2011), pp. 646–650.

[43] Y.-D. Kim and S. Choi, *Nonnegative Tucker decomposition*, in 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[44] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. Shapiro, P. Barloon, and A. F. Goetz, *The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data*, Remote sensing of environment, 44 (1993), pp. 145–163.

[45] H. Kwon and N. M. Nasrabadi, *Kernel RX-algorithm: A nonlinear anomaly detector for hyperspectral imagery*, IEEE Transactions on Geoscience and Remote Sensing, 43 (2005), pp. 388–397.

[46] R. Lai and S. Osher, *A splitting method for orthogonality constrained problems*, Journal of Scientific Computing, 58 (2014), pp. 431–449.

[47] B.-Z. Li, X.-L. Zhao, X. Zhang, T.-Y. Ji, X. Chen, and M. K. Ng, *A learnable group-tube transform induced tensor nuclear norm and its application for tensor completion*, SIAM Journal on Imaging Sciences, 16 (2023), pp. 1370–1397.

[48] J. Li, H. Zhang, L. Zhang, and L. Ma, *Hyperspectral anomaly detection by the use of background joint sparse representation*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8 (2015), pp. 2523–2533.

[49] Z. Li, F. Nie, J. Bian, D. Wu, and X. Li, *Sparse PCA via $\ell_{2,p}$-norm regularization for unsupervised feature selection*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (2021), pp. 5322–5328.

[50] Z. Li, J. Tang, and X. He, *Robust structured nonnegative matrix factorization for image representation*, IEEE Transactions on Neural Networks and Learning Systems, 29 (2017), pp. 1947–1960.

[51] T. Liu, Z. Lu, X. Chen, and Y.-H. Dai, *An exact penalty method for semidefinite-box-constrained low-rank matrix optimization problems*, IMA Journal of Numerical Analysis, 40 (2018), pp. 563–586.

[52] X. Liu, J. Lu, L. Shen, C. Xu, and Y. Xu, *Multiplicative noise removal: Nonlocal low-rank model and its proximal alternating reweighted minimization algorithm*, SIAM Journal on Imaging Sciences, 13 (2020), pp. 1595–1629.

[53] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, *Recent advances of hyperspectral imaging technology and applications in agriculture*, Remote Sensing, 12 (2020), p. 2659.

[54] X. Ma, Q. Ye, and H. Yan, *L2P-norm distance twin support vector machine*, IEEE Access, 5 (2017), pp. 23473–23483.

[55] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi, *Nonlocal transform-domain filter for volumetric data denoising and reconstruction*, IEEE Transactions on Image Processing, 22 (2012), pp. 119–133.

[56] J. H. Manton, *Optimization algorithms exploiting unitary constraints*, IEEE Transactions on Signal Processing, 50 (2002), pp. 635–650.

[57] G. Marjanovic and V. Solo, *On l_q optimization and matrix completion*, IEEE Transactions on Signal Processing, 60 (2012), pp. 5714–5724.

[58] J. M. Molero, E. M. Garzon, I. Garcia, and A. Plaza, *Analysis and optimizations of global and local versions of the RX algorithm for anomaly detection in hyperspectral data*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6 (2013), pp. 801–814.

[59] Y. Nishimori and S. Akaho, *Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold*, Neurocomputing, 67 (2005), pp. 106–135.

[60] H. Oviedo, O. Dalmau, and H. Lara, *Two adaptive scaled gradient projection methods for Stiefel manifold constrained optimization*, Numerical Algorithms, 87 (2021), pp. 1107–1127.

[61] J. Pan, M. K. Ng, Y. Liu, X. Zhang, and H. Yan, *Orthogonal nonnegative Tucker decomposition*, SIAM Journal on Scientific Computing, 43 (2021), pp. B55–B81.

[62] B. T. Polyak, *Existence theorems and convergence of minimizing sequences for extremal problems with constraints*, Doklady Akademii Nauk SSSR, 166 (1966), pp. 287–290.

[63] A. B. Pour, B. Zoheir, B. Pradhan, and M. Hashim, *Multispectral and Hyperspectral Remote Sensing Data for Mineral Exploration and Environmental Monitoring of Mined Areas*, MDPI, 2021.

[64] C. Prévost, R. A. Borsoi, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, *Hyperspectral super-resolution accounting for spectral variability: Coupled tensor ll1-based recovery and blind unmixing of the unknown super-resolution image*, SIAM Journal on Imaging Sciences, 15 (2022), pp. 110–138.

[65] A. Rajwade, D. Kittle, T.-H. Tsai, D. Brady, and L. Carin, *Coded hyperspectral imaging and blind compressive sensing*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 782–812.

[66] B. Rasti, P. Scheunders, P. Ghamisi, G. Licciardi, and J. Chanussot, *Noise reduction in hyperspectral imagery: overview and application*, Remote Sensing, 10 (2018), p. 482.

[67] I. S. Reed and X. Yu, *Adaptive multiple-band constant false alarm rate detection of an optical pattern with unknown spectral distribution*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 38 (1990), pp. 1760–1770.

[68] H. Sato and K. Aihara, *Cholesky QR-based retraction on the generalized Stiefel manifold*, Computational Optimization and Applications, 72 (2019), pp. 293–308.

[69] L. Shen, B. W. Suter, and E. E. Tripp, *Structured sparsity promoting functions*, Journal of Optimization Theory and Applications, 183 (2019), pp. 386–421.

[70] M. Shimoni, R. Haelterman, and C. Perneel, *Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques*, IEEE Geoscience and Remote Sensing Magazine, 7 (2019), pp. 101–117.

[71] Y. Song, E.-H. Djermoune, J. Chen, C. Richard, and D. Brie, *Online deconvolution for industrial hyperspectral imaging systems*, SIAM Journal on Imaging Sciences, 12 (2019), pp. 54–86.

[72] M. B. Stuart, A. J. S. McGonigle, and J. R. Willmott, *Hyperspectral imaging in environmental monitoring: A review of recent developments and technological advances in compact field deployable systems*, Sensors, 19 (2019), p. 3071.

[73] H. Su, Z. Wu, H. Zhang, and Q. Du, *Hyperspectral anomaly detection: A survey*, IEEE Geoscience and Remote Sensing Magazine, 10 (2021), pp. 64–90.

[74] H. Sun, M. Liu, K. Zheng, D. Yang, J. Li, and L. Gao, *Hyperspectral image denoising via low-rank representation and CNN denoiser*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15 (2022), pp. 716–728.

[75] Y. Teng, L. Yang, B. Yu, and X. Song, *A penalty PALM method for sparse portfolio selection problems*, Optimization Methods and Software, 32 (2017), pp. 126–147.

[76] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of Optimization Theory and Applications, 109 (2001), pp. 475–494.

[77] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.

[78] L. Wald, *Quality of high resolution synthesised images: Is there a simple criterion?*, in Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images", SEE/URISCA, 2000, pp. 99–103.

[79] L. Wang and S. Chen, *$l_{2,p}$ matrix norm and its application in feature selection*, arXiv preprint arXiv:1303.3987, (2013).

[80] M. Wang, D. Hong, B. Zhang, L. Ren, J. Yao, and J. Chanussot, *Learning double subspace representation for joint hyperspectral anomaly detection and noise removal*, IEEE Transactions on Geoscience and Remote Sensing, 61 (2023), pp. 1–17.

[81] M. Wang, Q. Wang, D. Hong, S. K. Roy, and J. Chanussot, *Learning tensor low-rank representation for hyperspectral anomaly detection*, IEEE Transactions on Cybernetics, 53 (2022), pp. 679–691.

[82] S. Wang, X. Wang, L. Zhang, and Y. Zhong, *Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder*, IEEE Transactions on Geoscience and Remote Sensing, 60 (2021), pp. 1–14.

[83] Y. Wang, J. Peng, Q. Zhao, Y. Leung, X.-L. Zhao, and D. Meng, *Hyperspectral image restoration via total variation regularized low-rank tensor decomposition*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11 (2017), pp. 1227–1243.

[84] Z. Wang, Q. Lin, Y. Chen, and P. Zhong, *Block-based multi-view classification via view-based $l_{2,p}$ sparse representation and adaptive view fusion*, Engineering Applications of Artificial Intelligence, 116 (2022), p. 105337.

[85] J. Warga, *Minimizing certain convex functions*, Journal of the Society for Industrial and Applied Mathematics, 11 (1963), pp. 588–593.

[86] K. Wei, Y. Fu, and H. Huang, *3-D quasi-recurrent neural network for hyperspectral image denoising*, IEEE Transactions on Neural Networks and Learning Systems, 32 (2020), pp. 363–375.

[87] Z. Wen, C. Yang, X. Liu, and Y. Zhang, *Trace-penalty minimization for large-scale eigenspace computation*, Journal of Scientific Computing, 66 (2016), pp. 1175–1203.

[88] D. Wu and D.-W. Sun, *Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review—part i: Fundamentals*, Innovative Food Science and Emerging Technologies, 19 (2013), pp. 1–14.

[89] Z. Wu, C. Huang, and T. Zeng, *Extrapolated plug-and-play three-operator splitting methods for nonconvex optimization with applications to image restoration*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 1145–1181.

[90] N. Xiao, X. Liu, and Y.-X. Yuan, *A class of smooth exact penalty function methods for optimization problems with orthogonality constraints*, Optimization Methods and Software, 37 (2022), pp. 1205–1241.

[91] W. Xie, J. Lei, B. Liu, Y. Li, and X. Jia, *Spectral constraint adversarial autoencoders approach to feature representation in hyperspectral anomaly detection*, Neural Networks, 119 (2019), pp. 222–234.

[92] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin, *Dictionary learning for noisy and incomplete hyperspectral images*, SIAM Journal on Imaging Sciences, 5 (2012), pp. 33–56.

[93] F. Xiong, J. Zhou, and Y. Qian, *Hyperspectral restoration via l_0 gradient regularized low-rank tensor factorization*, IEEE Transactions on Geoscience and Remote Sensing, 57 (2019), pp. 10410–10425.

[94] Y. Xu, Z. Wu, J. Li, A. Plaza, and Z. Wei, *Anomaly detection in hyperspectral images based on low-rank and sparse representation*, IEEE Transactions on Geoscience and Remote Sensing, 54 (2015), pp. 1990–2000.

[95] Y. Xu and W. Yin, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, Journal of Scientific Computing, 72 (2017), pp. 700–734.

[96] C. Yang, J. C. Meza, and L.-W. Wang, *A trust region direct constrained minimization algorithm for the Kohn–Sham equation*, SIAM Journal on Scientific Computing, 29 (2007), pp. 1854–1875.

[97] F. Yang, X. Chen, and L. Chai, *Hyperspectral image destriping and denoising using stripe and spectral low-rank matrix recovery and global spatial-spectral total variation*, Remote Sensing, 13 (2021), p. 827.

[98] Q. Yu and M. Bai, *Generalized nonconvex hyperspectral anomaly detection via background representation learning with dictionary constraint*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 917–950.

[99] Q. Yuan, L. Zhang, and H. Shen, *Hyperspectral image denoising employing a spectral–spatial adaptive total variation model*, IEEE Transactions on Geoscience and Remote Sensing, 50 (2012), pp. 3660–3677.

[100] Q. Yuan, Q. Zhang, J. Li, H. Shen, and L. Zhang, *Hyperspectral image denoising employing a spatial-spectral deep residual convolutional neural network*, IEEE Transactions on Geoscience and Remote Sensing, 57 (2018), pp. 1205–1218.

[101] C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of Statistics, 38 (2010), pp. 894 – 942.

[102] H. Zhang, J. Cai, W. He, H. Shen, and L. Zhang, *Double low-rank matrix decomposition for hyperspectral image denoising and destriping*, IEEE Transactions on Geoscience and Remote Sensing, 60 (2022), pp. 1–19.

[103] H. Zhang, W. He, L. Zhang, H. Shen, and Q. Yuan, *Hyperspectral image restoration using low-rank matrix recovery*, IEEE Transactions on Geoscience and Remote Sensing, 52 (2013), pp. 4729–4743.

[104] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. V. Gool, and R. Timofte, *Plug-and-play image restoration with deep denoiser prior*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2021), pp. 6360–6376.

[105] L. Zhang, L. Zhang, X. Mou, and D. Zhang, *FSIM: A feature similarity index for image quality assessment*, IEEE Transactions on Image Processing, 20 (2011), pp. 2378–2386.

[106] C. Zhao and L. Zhang, *Spectral-spatial stacked autoencoders based on low-rank and sparse matrix decomposition for hyperspectral anomaly detection*, Infrared Physics & Technology, 92 (2018), pp. 166–176.

[107] Y.-Q. Zhao and J. Yang, *Hyperspectral image denoising via sparse representation and low-rank constraint*, IEEE Transactions on Geoscience and Remote Sensing, 53 (2014), pp. 296–308.

[108] H. Zheng, Y. Lou, G. Tian, and C. Wang, *A scale-invariant relaxation in low-rank tensor recovery with an application to tensor completion*, SIAM Journal on Imaging Sciences, 17 (2024), pp. 756–783.

[109] W. Zhou, B. Alan Conrad, S. Hamid Rahim, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE Transactions on Image Processing, 13 (2004), pp. 600–612.