Hyperspectral Super-Resolution via Interpretable Block-Term Tensor Modeling

Meng Ding, Xiao Fu, Ting-Zhu Huang, Jun Wang, and Xi-Le Zhao

***Abstract*—Hyperspectral super-resolution (HSR) aims at fusing a pair of hyperspectral and multispectral images to recover a super-resolution image (SRI). This work revisits coupled tensor decomposition (CTD)-based HSR. The vast majority of the HSR approaches take a low-rank matrix recovery perspective. The challenge is that theoretical guarantees for recovering the SRI using low-rank matrix models are either elusive or derived under stringent conditions. A couple of recent CTD-based methods ensure recoverability for the SRI under relatively mild conditions, leveraging algebraic properties of the *canonical polyadic decomposition* (CPD) and the *Tucker* decomposition models, respectively. However, the latent factors of both the CPD and Tucker models have no physical interpretations in the context of spectral image analysis, which makes incorporating prior information challenging—but using priors is often essential for enhancing performance in noisy environments. This work employs an idea that models spectral images as tensors following the block-term decomposition model with multilinear rank-$(L_r, L_r, 1)$ terms (i.e., the LL1 model) and formulates the HSR problem as a coupled LL1 tensor decomposition problem. Similar to the existing CTD approaches, recoverability of the SRI is shown under mild conditions. More importantly, the latent factors of the LL1 model can be interpreted as the key constituents of spectral images, i.e., the endmembers' spectral signatures and abundance maps. This connection allows us to incorporate prior information for performance enhancement. A flexible algorithmic framework that can work with a series of structural information is proposed to take advantages of the model interpretability. The effectiveness is showcased using simulated and real data.**

***Index Terms*—hyperspectral super-resolution, block-term tensor decomposition, recoverability, regularization.**

M. Ding, T. Huang, and X. Zhao are supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61772003 and 61876203, Key Projects of Applied Basic Research in Sichuan Province (No. 2020YJ0216), and the Applied Basic Research Project of Sichuan Province (No. 21YYJC3042). X. Fu is supported in part by National Science Foundation (NSF) under Projects ECCS 1608961, ECCS 1808159 and ECCS-2024058, and the Army Research Office (ARO) under Project ARO W911NF-19-1-0247. J. Wang is supported in part by the NSFC under Grants 62071096, U19B2014 and 62001089, the Foundation of National Key Laboratory of Science and Technology on Communications, and the Innovation Fund of NCL (IFN) under IFN2019102. (*Corresponding author: Xiao Fu.*)

M. Ding, T. Huang, and X. Zhao are with the School of Mathematical Sciences at University of Electronic Science and Technology of China, Chengdu, China. Emails: dingmeng56@163.com, tingzhuhuang@126.com, and xlzhao122003@163.com. The work is developed during M. Ding's visit to Oregon State University.

X. Fu is with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, United States. Email: xiao.fu@oregonstate.edu.

J. Wang is with the National Key Laboratory of Science and Technology on Communications at University of Electronic Science and Technology of China, Chengdu, China. Email: junwang@uestc.edu.cn.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a number of additional proofs and detailed derivations. Contact xiao.fu@oregonstate.edu for further questions about this work.

I. Introduction

The spatial and spectral resolution tradeoff in spectral image sensing has been a well-known effect [1]. To be specific, hyperspectral sensors acquire data with high spectral resolution but low spatial resolution. Roughly speaking, in spectral image sensing (especially remote sensing), images are captured by sensing the sunlight reflected from objects/scenes. The total energy of sunlight reflected from the scene is limited. Producing a pixel at a spectral band is only possible when the signal-to-noise ratio reaches a certain level. Hence, the limited amount of energy carried by the reflected sunlight has to be allocated to different pixels (spatial locations) and frequency bands—which naturally presents a tradeoff [1], [2]. However, multispectral sensors produce images that exhibit high spatial resolution—but the spectral resolution is often coarse. Since both types of information are of great interest, many hyperspectral super-resolution (HSR) approaches have been proposed [1]. The goal of HSR is to fuse a pair of co-registered hyperspectral image (HSI) and multispectral image (MSI) to produce a super-resolution image (SRI) that has the "best of the two"—i.e., high resolutions in both the spectral and spatial domains.

Many early HSR techniques (e.g., [3]–[5]) view the SRI as a low-rank matrix, which is reminiscent of the *linear mixture model* (LMM) for spectral images [6]. The HSI and MSI are modeled as "compressed" data of the SRI by degrading from the spatial and spectral modes, respectively. Consequently, recovering the SRI can be understood as low-rank matrix estimation from downsampled data—which is an ill-posed inverse problem. Recoverability for the SRI using low-rank estimation techniques had been elusive until a couple of recent works [7], [8]. Nevertheless, as pointed out in [7], [8], matrix estimation-based approaches have recoverability guarantees for the SRI under stringent conditions, e.g., when every pixel only contains a small number of endmembers.

In 2018, Kanatsoulis *et al.* proposed a *coupled tensor decomposition* (CTD) framework for HSR [9]. There, the spectral images are modeled as tensors with low-rank canonical polyadic decomposition (CPD) representations. Utilizing the algebraic properties of CPD, recoverability of the SRI is shown to hold under mild conditions. Later on, the work in [10] adopted a similar idea but used the Tucker tensor model, which facilitates a fast algorithm leveraging on the higher-order singular value decomposition (SVD); also see a pre-print [11] that extends the coupled Tucker model to handle more complex cases where the spectral images exhibit certain variability. Note that a number of works from the

computer vision and remote sensing communities also take tensor perspectives for HSR; see, e.g., [12], [13]. Nonetheless, these works put more emphasis on the computation side, but no recoverability support was offered.

Compared to the low-rank matrix-based methods, e.g., those in [3]–[5], the tensor methods in [9], [10] are advantageous in terms of recoverability guarantees. Nevertheless, the performance gain obtained by the tensor methods is not always obvious. One possible reason is that the model parameters (or, the latent factors) of the tensor models used in [9], [10] do not have physical interpretations in the context of HSR. Consequently, it is hard to incorporate prior information into the CTD frameworks for performance enhancement. On the other hand, since the matrix-based approaches have clear connections to the LMM, the latent factors there can be interpreted as two key constituents of spectral images, i.e., the spectral signatures of endmembers (i.e., materials captured in the image) and the corresponding abundance maps—which makes using prior information (e.g., nonnegativity and spatial smoothness of the abundance maps) fairly natural through adding constraints and regularization terms to their optimization criteria—see examples in [14]–[16]. Note that for real data, since noise and modeling errors always exist, taking advantage of prior information oftentimes plays an essential role in producing high-quality image fusion results.

In this work, our objective is an alternative CTD approach offers similar recoverability guarantees as in the CPD and Tucker-based methods [9], [10]—and at the same time has physical interpretations for its model parameters. This way, prior information can be swiftly incorporated in the framework to enhance the HSR performance under challenging scenarios. To this end, our idea is to employ the tensor decomposition model with multilinear rank-$(L_r, L_r, 1)$ block-terms (or, the LL1 model for short) [17] for modeling the spectral images. The LL1 model was recently connected to HSIs in [18], where the latent factors of the LL1 model were linked to the LMM model. This connection was exploited for hyperspectral unmixing in [18]. Nonetheless, using the LL1 model for HSR has not been considered, to our best knowledge.

Contributions: Our detailed contributions are as follows:

• **Recoverability Guarantees.** We formulate the HSR problem as a coupled LL1 tensor decomposition problem, and show that the formulated criteria can provably recover the SRI under mild conditions. Similar properties have been observed in existing tensor approaches that are based on the CPD or Tucker models. Nonetheless, our analysis and recoverability conditions are new and specialized for the LL1 modeling. As in [9], [10], we also show that the LL1 model guarantees recoverability under a realistic yet challenging scenario, i.e., the case where the spatial degradation process is unknown. The recoverability analysis ensures that one does not lose theoretical guarantees when using the advocated LL1 model.

• **Flexible Algorithmic Framework.** We recast the formulated criteria to optimization forms that are convenient to incorporate regularization and constraints. In particular, we propose a change of variable scheme together with a Shatten-p function-based low-rank promotion regularization, for approximating the LL1 model. This way, the working optimization problems admit continuously differentiable objectives, and thus many off-the-shelf optimization tools for smooth cost functions can be used for tackling them. More importantly, this reformulation strategy makes the spectral signatures and abundance maps explicitly present in the optimization objectives, and thus constraints and regularization can be naturally imposed. We showcase the flexibility of this framework via incorporating nonnegativity of the endmembers and the spatial smoothness of the abundance maps. We propose to handle the formulated problems using a unified inexact and accelerated *block coordinate descent* (BCD) algorithmic framework.

• **Extensive Experiments.** We test the proposed algorithms over four different semi-real datasets (namely, semi-real data derived from the Salinas, Pavia University, Indian Pines, and Jasper Ridge datasets[1]) and different performance metrics. In addition, we test the algorithms on a pair of co-registered real HSI and MSI that were provided in the recent paper [19].

A conference version that contains part of the work was published in Proc. IEEE CAMSAP 2019 [20]. In the journal version, we additionally included more recoverability analysis results, the flexible algorithmic framework, extensive semi-real experiments, and the real experiment.

Notation. A scalar, a vector, a matrix, and a tensor are denoted as x, $\boldsymbol{x}$, $\boldsymbol{X}$, and $\underline{\boldsymbol{X}}$, respectively. $[\boldsymbol{x}]_i$, $[\boldsymbol{X}]_{i,j}$, and $[\underline{\boldsymbol{X}}]_{i,j,k}$ denote the i-th, (i,j)-th, and (i,j,k)-th element of $\boldsymbol{x} \in \mathbb{R}^I$, $\boldsymbol{X} \in \mathbb{R}^{I\times J}$, and $\underline{\boldsymbol{X}} \in \mathbb{R}^{I\times J\times K}$, respectively. The notation $\boldsymbol{X}(:,j)$ or $\boldsymbol{x}_j$ represents the j-th column of a matrix $\boldsymbol{X} \in \mathbb{R}^{I\times J}$ in an exchangeable manner. $\underline{\boldsymbol{X}}(i,j,:)$ and $\underline{\boldsymbol{X}}(:,:,k)$ denote the (i,j)-th tube and the k-th slab of $\underline{\boldsymbol{X}}$, respectively. The Frobenius norms of $\boldsymbol{X}$ and $\underline{\boldsymbol{X}}$ are denoted as $\|\boldsymbol{X}\|_F = \sqrt{\sum_{i,j}[\boldsymbol{X}]_{i,j}^2}$ and $\|\underline{\boldsymbol{X}}\|_F = \sqrt{\sum_{i,j,k}[\underline{\boldsymbol{X}}]_{i,j,k}^2}$, respectively. Given two vectors $\boldsymbol{x} \in \mathbb{R}^M$, $\boldsymbol{y} \in \mathbb{R}^N$ and two matrices $\boldsymbol{X} \in \mathbb{R}^{I\times J}$, $\boldsymbol{Y} \in \mathbb{R}^{P\times Q}$, the outer product $\boldsymbol{x} \circ \boldsymbol{y}$, $\boldsymbol{X} \circ \boldsymbol{y}$, and the Kronecker product $\boldsymbol{X} \otimes \boldsymbol{Y}$ are a $M \times N$ matrix, a $I \times J \times N$ tensor, and a $IP \times JQ$ matrix, respectively. The Khatri-Rao product $\odot_c$ is defined as $\boldsymbol{X} \odot_c \boldsymbol{Y} = [\boldsymbol{x}_1 \otimes \boldsymbol{y}_1, \cdots, \boldsymbol{x}_J \otimes \boldsymbol{y}_J] \in \mathbb{R}^{IP\times J}$ when $J = Q$. Let $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_R] \in \mathbb{R}^{I\times\sum_r J_r}$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_R] \in \mathbb{R}^{P\times\sum_r Q_r}$ be two partitioned matrices, the partitioned Khatri-Rao product $\odot$ is defined as $\boldsymbol{X} \odot \boldsymbol{Y} = [\boldsymbol{X}_1 \otimes \boldsymbol{Y}_1, \cdots, \boldsymbol{X}_R \otimes \boldsymbol{Y}_R] \in \mathbb{R}^{IP\times\sum_r J_r Q_r}$ [17]. The i-th singular value of $\boldsymbol{X}$ is denoted as $\sigma_i(\boldsymbol{X})$ and the largest one as $\sigma_{\max}(\boldsymbol{X})$. $\widehat{\boldsymbol{X}}$ is used as the estimator of $\boldsymbol{X}$.

II. Background

In this section, we first briefly introduce some preliminaries pertaining to our framework.

A. Tensor Modeling for Spectral Images

Third-order tensors (three-dimensional arrays) $\underline{\boldsymbol{X}} \in \mathbb{R}^{I\times J\times K}$ can be understood as generalization for vectors and matrices (see Fig. 1). Apparently, spectral images are tensors, which admit two spatial dimensions and one spectral dimension—also see Fig. 1.

[1] https://rslab.ut.ac.ir/data

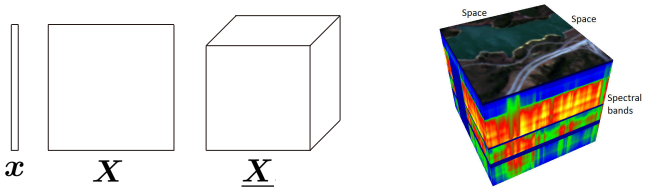


Fig. 1. Left: A vector, matrix, and a tensor. Right: A spectral image can be understood as a third-order tensor. Every spectral pixel is measured over a large number of spectral bands.

One of the key differences between tensors and matrices is that the rank decomposition for tensors have various forms. The arguably most popular tensor rank decomposition model is the so-called *canonical polyadic decomposition* (CPD) model [21], which is also known as the parallel factor analysis (PARAFAC) model. Under the CPD model, a tensor with canonical polyadic (CP) rank R can be written as

$$\underline{\boldsymbol{X}} = \sum_{r=1}^{R} \boldsymbol{A}(:,r) \circ \boldsymbol{B}(:,r) \circ \boldsymbol{C}(:,r).$$

Besides, the Tucker model [22] has also been widely used in the imaging society. Under the Tucker model, a tensor can be decomposed as

$$\underline{\boldsymbol{X}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \underline{\boldsymbol{G}}(i,j,k) \boldsymbol{A}(:,i) \circ \boldsymbol{B}(:,j) \circ \boldsymbol{C}(:,k).$$

Both the CPD and Tucker models are able to "encode" dependence across different dimensions of the data, and thus are meaningful low-rank models (see illustrations in Fig. 2). In fact, both the CPD and Tucker models have been utilized for modeling spectral images, and the recoverability of SRI was shown under tensor modeling with mild conditions [9], [10]. Note that most low-rank matrix modeling based HSR approaches, e.g., [3], [4], [14], [16], do not have recoverability guarantees (except for some recent ones under relatively stringent conditions [7], [8]).

B. Linear Mixture Model (LMM) and Matrix-based HSR

The CPD and Tucker based HSR frameworks are appealing due to the SRI recoverability guarantees under mild conditions.

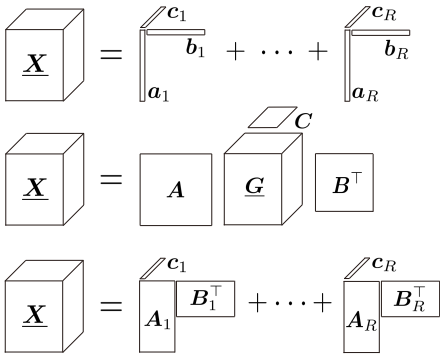


Fig. 2. Illustration of three tensor decompositions. Top: CPD. Middle: Tucker. Bottom: LL1.

However, the performance in practice is not always better than the low-rank matrix estimation-based approaches that do not have recoverability support. One reason may be that the matrix models often exploit the physical interpretations of their latent factors. Imposing structural information of the latent factors helps regularize the HSR criteria with prior knowledge. This is often essential for performance enhancement when noise and modeling errors are present. On the other hand, both the CPD and the Tucker models do not have physical interpretations for their latent factors. Therefore, in spite of the recoverability appeal, the challenge for incorporating prior information makes it hard to further improve the performance of the CPD/Tucker-based HSR approaches.

To see how the matrix factorization-based approaches exploit the physical interpretations, consider the LMM [6] of spectral pixels. In the noise-free case, a spectral pixel of an $I \times J \times K$ image can be written as follows:

$$\boldsymbol{y}_\ell = \boldsymbol{C}\boldsymbol{s}_\ell \in \mathbb{R}^K, \ \ell = 1, 2, \ldots, N,$$

where K denotes the number of spectral bands, $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_R] \in \mathbb{R}^{K \times R}$ collects R different spectral signatures of materials (endmembers) contained in the pixel, N is the total number of pixels, and $\boldsymbol{s}_\ell(r)$ is the proportion of endmember r in pixel ℓ (i.e., the abundance of endmember r). Putting all the pixels together one has

$$\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{S}^\top, \tag{1}$$

where

$$\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N]^\top \tag{2}$$

and $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$. Note that $\boldsymbol{S}(:,r)$ can be reshaped to an $I \times J$ matrix, which is often referred to as the abundance map of endmember r (see Fig. 3). Both the HSI and MSI can be expressed following the above LMM. If the HSI and MSI are co-registered, the two LMMs' latent factors are coupled together. Then, the HSR problem can be formulated as a coupled matrix factorization problem; see details in [5], [14], [16]. In addition, the latent factors in the LMM (i.e., $\boldsymbol{C}$ and $\boldsymbol{S}$) both have strong physical interpretations, and thus a variety of prior knowledge can be readily utilized. For example, both $\boldsymbol{C}$ and $\boldsymbol{S}$ should be nonnegative [5], [14], and the abundance map $\mathrm{mat}(\boldsymbol{S}(:,r))$ should exhibit low total variation (TV) across both the row and the column dimensions [4]. As a consequence, these priors can be incorporated into the computational frameworks to enhance the HSR performance.

C. Interpretable Tensor Modeling for Spectral Images

Recently, tensor model, namely, the block-term decomposition into multilinear rank-$(L_r, L_r, 1)$ terms model (or simply the LL1 model) [17], was connected to spectral images [18]. The model is very similar to the LMM, but has an extra assumption—i.e., the abundance maps are low-rank matrices. To be specific, instead of looking at the LMM model from a matrix view as in (1), one can re-write the LMM as follows:

$$\underline{\boldsymbol{Y}} = \sum_{r=1}^{R} \boldsymbol{S}_r \circ \boldsymbol{C}(:,r), \tag{3}$$

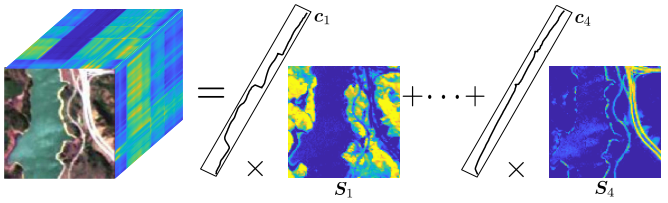


Fig. 3. The LMM of spectral images. The vector $\boldsymbol{c}_r$ denotes the spectral signature of material r; the matrix $\boldsymbol{S}_r$ denotes the corresponding abundance map.

where

$$\boldsymbol{S}_r = \text{mat}(\boldsymbol{S}(:,r)) \in \mathbb{R}^{I\times J} \tag{4}$$

represents the abundance map in its (more natural) matrix form, and $\boldsymbol{C}(:,r) = \boldsymbol{c}_r \in \mathbb{R}^K$ the r-th endmember; see Fig. 3 for illustration. Let $\text{rank}(\boldsymbol{S}_r) = L_r \leq \min\{I, J\}$, the above can be re-written as

$$\underline{\boldsymbol{Y}} = \sum_{r=1}^{R} (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{C}(:,r), \tag{5}$$

where $\boldsymbol{A}_r \in \mathbb{R}^{I\times L_r}$ and $\boldsymbol{B}_r \in \mathbb{R}^{J\times L_r}$ are full column-rank matrices and $\boldsymbol{S}_r = \boldsymbol{A}_r \boldsymbol{B}_r^\top$. The tensor model in (5) is exactly the LL1 model. The LL1 model admits the following important model uniqueness property:

Lemma 1: [17] Let $(\{\boldsymbol{A}_r \in \mathbb{R}^{I\times L_r}, \boldsymbol{B}_r \in \mathbb{R}^{J\times L_r}\}_{r=1}^R, \boldsymbol{C} \in \mathbb{R}^{K\times R})$ be the latent factors of the LL1 tensor $\underline{\boldsymbol{Y}} = \sum_{r=1}^R (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{C}(:,r)$ where $L_r = L$. Assume that $\boldsymbol{A}_r$, $\boldsymbol{B}_r$, and $\boldsymbol{C}$ are drawn from any absolutely continuous distributions. Then, the LL1 decomposition of $\underline{\boldsymbol{Y}}$ is essentially unique almost surely, if $IJ \geq L^2 R$ and $\min\left(\left\lfloor \frac{I}{L} \right\rfloor, R\right) + \min\left(\left\lfloor \frac{J}{L} \right\rfloor, R\right) + \min(K, R) \geq 2R + 2$. Here, essentially uniqueness means that if $\underline{\boldsymbol{Y}} = \sum_{r=1}^R (\boldsymbol{A}_r^* (\boldsymbol{B}_r^*)^\top) \circ \boldsymbol{C}^*(:,r)$, then, we must have $\boldsymbol{S}^* = \boldsymbol{S}\boldsymbol{\Pi}\boldsymbol{\Lambda}$, $\boldsymbol{C}^* = \boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\Lambda}^{-1}$, where $\boldsymbol{\Pi}$ and $\boldsymbol{\Lambda}$ denote a permutation matrix and a nonsingular scaling matrix, respectively, $\boldsymbol{S}^\star = [\text{vec}(\boldsymbol{S}_1^\star), \ldots, \text{vec}(\boldsymbol{S}_R^\star)]$, $\boldsymbol{S}_r^\star = \boldsymbol{A}_r^\star (\boldsymbol{B}_r^\star)^\top$ and $\boldsymbol{S} = [\text{vec}(\boldsymbol{S}_1), \ldots, \text{vec}(\boldsymbol{S}_R)]$, $\boldsymbol{S}_r = \boldsymbol{A}_r \boldsymbol{B}_r^\top$.
More details and properties of the LL1 model can be found in [17].

The LL1 model was used in hyperspectral imaging for unmixing the endmembers [18], [23]. This is well motivated. First, since the abundance maps are correlated across the two spatial dimensions, assuming $\boldsymbol{S}_r$ to be low-rank is plausible. Second, since the LL1 model is unique under mild conditions, the abundance maps and the endmembers can be identified up to certain trivial ambiguities with provable guarantees.

In Fig. 4, we present the singular values of the abundance maps of two datasets, i.e., Jasper Ridge and Samon, whose ground-truth abundance maps are available online (https://rslab.ut.ac.ir/data). One can see that for Jasper, the first 20 principal components of $\boldsymbol{S}_r$ (i.e., first 20 left, right singular vectors and singular values) contain more than 90% of its energy. For Samon, the first 10 principal components contain more than 95% of $\boldsymbol{S}_r$'s energy. This result supports the postulate that $\boldsymbol{S}_r$'s are approximately low-rank matrices.

Using the LL1 model for HSR seems to be appealing, since physical interpretations of the key constituents in spectral images are reflected in the model. Nevertheless, although the

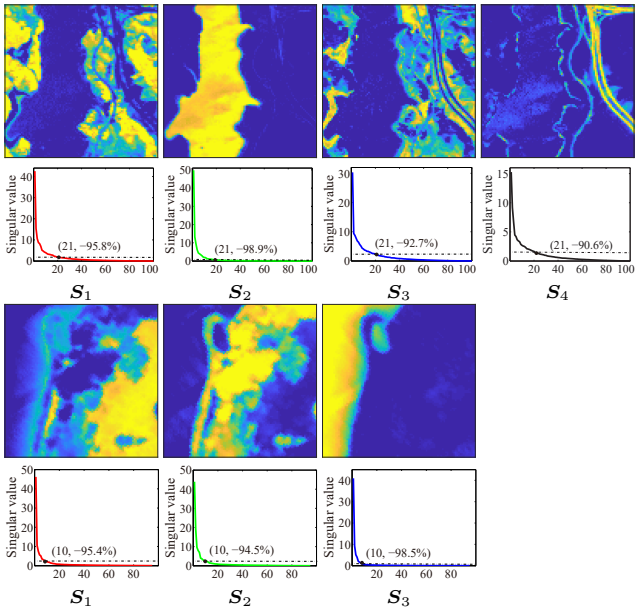


Fig. 4. The abundance maps and the corresponding singular values of Jasper Ridge (first two rows) and Samon (last two rows), respectively.

LL1 model exhibits promising advantages in hyperspectral unmixing (HU), the HU problem has very different settings and objectives relative to HSR. Particularly, there are a number of challenges of utilizing the LL1 model for HSR. First, it is unclear if certain SRI recoverability guarantees, e.g., those were shown under the CPD and Tucker models, still exist under the LL1 model. Second, optimization involving the LL1 model while considering structural constraints on the abundance maps and endmembers is a challenging problem.

III. Coupled LL1 Tensor Decomposition for HSR

In this section, we propose an LL1 model-based HSR approach, and discuss its recoverability properties.

A. Problem Formulation

We follow the setup in [9] for tensor-based HSR. Specifically, we use $\underline{\boldsymbol{Y}}_H \in \mathbb{R}^{I_H\times J_H\times K_H}$, $\underline{\boldsymbol{Y}}_M \in \mathbb{R}^{I_M\times J_M\times K_M}$ and $\underline{\boldsymbol{Y}}_S \in \mathbb{R}^{I_M\times J_M\times K_H}$ to denote the HSI, MSI, and SRI, respectively (note that $I_M \gg I_H$, $J_M \gg J_H$, and $K_H \gg K_M$ usually hold). All the spectral images are represented in a "space$\times$ space$\times$ spectrum" format. We assume that the HSI and MSI are downsampled/degraded from a SRI. Note that the SRI has the spatial resolution of the MSI (i.e., $I_M \times J_M$) and the spectral resolution of the HSI (i.e., K_H).

We assume that the abundance maps are low-rank matrices. Consequently, the SRI follows the LL1 model:

$$\underline{\boldsymbol{Y}}_S = \sum_{r=1}^{R} (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{C}(:,r), \tag{6}$$

where $\boldsymbol{A}_r \in \mathbb{R}^{I_M\times L_r}$, $\boldsymbol{B}_r \in \mathbb{R}^{J_M\times L_r}$ for $r = 1, \ldots, R$, $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_R] \in \mathbb{R}^{K_H\times R}$. The HSI can be understood as a spatially blurred and downsampled version of the SRI [14], [16]. This can be modeled by multiplying two blurring

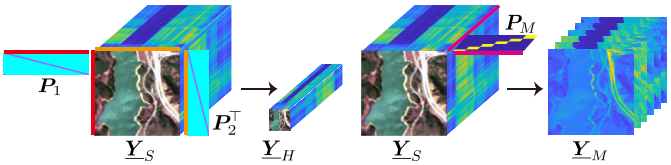


Fig. 5. Illustration of spatial and spectral degradation procedures from SRI to HSI and MSI, respectively.

and compressing matrices to the row and column dimensions to every slab $\underline{\boldsymbol{Y}}_S(:,:,k)$, i.e.,

$$\underline{\boldsymbol{Y}}_H(:,:,k) = \boldsymbol{P}_1 \underline{\boldsymbol{Y}}_S(:,:,k) \boldsymbol{P}_2^\top, \quad k = 1, \ldots, K_H, \tag{7}$$

where $\boldsymbol{P}_1 \in \mathbb{R}^{I_H \times I_M}$ and $\boldsymbol{P}_2 \in \mathbb{R}^{J_H \times J_M}$ represent two spatial blurring and downsampling matrices along the two spatial dimensions, respectively. In hyperspectral imaging, the blurring operator can be modeled by some kernel functions, e.g., the widely used Gaussian kernel—see detailed discussions in [9].

If $\underline{\boldsymbol{Y}}_S$ follows the LL1 model and the spatial degradation model in (7) holds, we have the following:

$$\begin{aligned} \underline{\boldsymbol{Y}}_H &= \sum_{r=1}^{R} (\boldsymbol{P}_1 \boldsymbol{S}_r \boldsymbol{P}_2^\top) \circ \boldsymbol{C}(:,r) \\ &= \sum_{r=1}^{R} (\boldsymbol{P}_1 \boldsymbol{A}_r (\boldsymbol{P}_2 \boldsymbol{B}_r)^\top) \circ \boldsymbol{C}(:,r). \end{aligned} \tag{8}$$

Similarly, the MSI can be modeled as spectral band-aggregated version of the SRI, i.e.,

$$\underline{\boldsymbol{Y}}_M(i,j,:) = \boldsymbol{P}_M \underline{\boldsymbol{Y}}_S(i,j,:), \quad \forall i,j, \tag{9}$$

where $\boldsymbol{P}_M \in \mathbb{R}^{K_M \times K_H}$ is a band aggregation matrix that is related to sensor specifications; see [9]. The above leads to the following compact representation:

$$\underline{\boldsymbol{Y}}_M = \sum_{r=1}^{R} (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{P}_M \boldsymbol{C}(:,r). \tag{10}$$

The spatial and spectral degradation procedures are illustrated in Fig. 5. A side comment is that the operations in (7) and (9) are the so-called "mode products" in tensor algebra; see more details in [9] and references therein.

Under the degradation model, the main task of recovering $\underline{\boldsymbol{Y}}_S$ boils down to estimating the latent factors $\{\boldsymbol{A}_r \boldsymbol{B}_r^\top\}_{r=1}^R$ (high-resolution abandance maps) and $\boldsymbol{C}$ (high-resolution endmembers) from $\underline{\boldsymbol{Y}}_H$ and $\underline{\boldsymbol{Y}}_M$. Once $\{\boldsymbol{A}_r \boldsymbol{B}_r^\top\}_{r=1}^R$ and $\boldsymbol{C}$ are identified, one can readily reconstruct the SRI using $\underline{\boldsymbol{Y}}_S = \sum_{r=1}^R (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{C}(:,r)$.

B. Recoverability Analysis

A key benefit for tensor modeling is that the recoverability of the SRI can be established via exploiting the algebraic properties of various tensor structures, e.g., the CPD and Tucker decompositions, as shown in [9], [10]. It is also of interest to consider the recoverability properties under the LL1 model. Note that the characteristics of the LL1 model are different from those of the CPD and Tucker models. Hence, custom analysis for the LL1 model is needed.

To answer the recoverability inquiry, we consider the following formulation:

$$\begin{aligned} &\text{find } \{\boldsymbol{A}_r \in \mathbb{R}^{I_M \times L_r}, \boldsymbol{B}_r \in \mathbb{R}^{J_M \times L_r}\}_{r=1}^R, \boldsymbol{C} \in \mathbb{R}^{K_H \times R} \\ &\text{s.t. } \underline{\boldsymbol{Y}}_H = \sum_{r=1}^{R} (\boldsymbol{P}_1 \boldsymbol{A}_r (\boldsymbol{P}_2 \boldsymbol{B}_r)^\top) \circ \boldsymbol{C}(:,r), \\ &\qquad \underline{\boldsymbol{Y}}_M = \sum_{r=1}^{R} (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{P}_M \boldsymbol{C}(:,r). \end{aligned} \tag{11}$$

The above can be understood as a coupled LL1 tensor decomposition criterion: we aim at finding an LL1 tensor (i.e., the SRI) that fits both of the degraded tensors (the HSI and MSI). The recoverability problem amounts to answering if the solution to (11) always reconstructs $\underline{\boldsymbol{Y}}_S$. Our analysis shows the following:

Theorem 1: Assume that the SRI $\underline{\boldsymbol{Y}}_S$ follows the LL1 model with $L_r = L$ for all r. Let $(\{\boldsymbol{A}_r, \boldsymbol{B}_r\}_{r=1}^R, \boldsymbol{C})$ be the ground-truth latent factors of $\underline{\boldsymbol{Y}}_S$. Assume that $\underline{\boldsymbol{Y}}_H$ and $\underline{\boldsymbol{Y}}_M$ follow the degradation models in (8) and (10), respectively. Suppose that each of $\boldsymbol{A}_r \in \mathbb{R}^{I_M \times L}$, $\boldsymbol{B}_r \in \mathbb{R}^{J_M \times L}$, and $\boldsymbol{C} \in \mathbb{R}^{K_H \times R}$ is drawn from any absolutely continuous distribution, that $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, and $\boldsymbol{P}_M$ have full row rank, and that $(\{\boldsymbol{A}_r^\star, \boldsymbol{B}_r^\star\}_{r=1}^R, \boldsymbol{C}^\star)$ represent any solution to Problem (11). Then, the ground-truth $\underline{\boldsymbol{Y}}_S$ is uniquely recovered with probability one by

$$\underline{\boldsymbol{Y}}_S = \sum_{r=1}^{R} (\boldsymbol{A}_r^\star (\boldsymbol{B}_r^\star)^\top) \circ \boldsymbol{C}^\star(:,r),$$

if $I_M J_M \geq L^2 R$, $I_H J_H \geq LR$, and

$$\min\left(\left\lfloor \frac{I_M}{L} \right\rfloor, R\right) + \min\left(\left\lfloor \frac{J_M}{L} \right\rfloor, R\right) + \min(K_M, R) \geq 2R+2.$$

The proof of Theorem 1 is relegated to Appendix A. The criterion and Theorem 1 consider the case where the spatial and spectral degradation operators are both known. In [9], [10], it was argued that in some cases the spatial degradation operators may not be easy to estimate. Indeed, the spatial degradation process may be complex and may vary from case to case, caused by various reasons such as the sensor specifications and the sensing environment—modeling this process *per se* may be a challenging task. A side benefit of using tensor modeling in [9], [10] is that the SRI recoverability can still be ensured even if the spatial degradation operators $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are unknown. In this work, we show that this benefit remains under the LL1 model. To see this, we consider the following criterion:

$$\begin{aligned} &\text{find } \{\widetilde{\boldsymbol{A}}_r \in \mathbb{R}^{I_H \times L_r}, \widetilde{\boldsymbol{B}}_r \in \mathbb{R}^{J_H \times L_r}, \boldsymbol{A}_r, \boldsymbol{B}_r\}_{r=1}^R, \boldsymbol{C} \\ &\text{s.t. } \underline{\boldsymbol{Y}}_H = \sum_{r=1}^{R} (\widetilde{\boldsymbol{A}}_r \widetilde{\boldsymbol{B}}_r^\top) \circ \boldsymbol{C}(:,r), \qquad \text{(12a)} \\ &\qquad \underline{\boldsymbol{Y}}_M = \sum_{r=1}^{R} (\boldsymbol{A}_r \boldsymbol{B}_r^\top) \circ \boldsymbol{P}_M \boldsymbol{C}(:,r) \qquad \text{(12b)} \end{aligned}$$

In the above, $\widetilde{\boldsymbol{A}}_r$ and $\widetilde{\boldsymbol{B}}_r$ "absorb" the spatial degradation operators (i.e., we use $\widetilde{\boldsymbol{A}}_r$ and $\widetilde{\boldsymbol{B}}_r$ to model $\boldsymbol{P}_1 \boldsymbol{A}_r$ and $\boldsymbol{P}_2 \boldsymbol{B}_r$, respectively). For this more challenging case, we also show recoverability as follows:

Theorem 2: Under the same settings as in Theorem 1, assume that $K_M \geq 2$ and that $(\{\widehat{\boldsymbol{A}}_r^\star, \widetilde{\boldsymbol{B}}_r^\star\}_{r=1}^R, \{\boldsymbol{A}_r^\star, \boldsymbol{B}_r^\star\}_{r=1}^R, \boldsymbol{C}^\star)$ is any solution to Problem (12). Then, if $I_H J_H \geq L^2 R$ and $\min\left(\left\lfloor \frac{I_H}{L} \right\rfloor, R\right) + \min\left(\left\lfloor \frac{J_H}{L} \right\rfloor, R\right) + \min(K_M, R) \geq 2R + 2$, the ground-truth $\underline{\boldsymbol{Y}}_S$ is uniquely recovered by $\underline{\boldsymbol{Y}}_S = \sum_{r=1}^R (\boldsymbol{A}_r^\star (\boldsymbol{B}_r^\star)^\top) \circ \boldsymbol{C}^\star(:, r)$ with probability one.

The proof of Theorem 2 is given in Appendix B. A remark is that the conditions in Theorem 2 are more restrictive relative to those in Theorem 1. In particular, smaller L and R are needed for the conditions in Theorem 2 to hold—which stands for the price to pay for not knowing $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.

Remark 1: The proofs of Theorems 1-2 are reminiscent of a number of prior works that consider coupled tensor decomposition [9], [24]–[26]. In particular, the work in [24], [25] considered coupled LL1 decomposition under various settings. Nonetheless, the recoverability results in [24], [25] cannot be directly applied to our case. First, the work in [25] does not consider "degradation" (e.g., compression by $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_M$ matrices in our case). Second, the work in [24] considers special degradation matrices. There, the degradation matrices are all slab/fiber selection matrices, which makes the proof relatively simple. In our case, $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_M$ are general compression matrices, which requires more careful and tailored derivations (e.g., invoking Lemma 2 in Appendix A) to help establish recoverability. Our proof and theorems therefore cover more cases that could not be covered by the recoverability theorems in [24]. Another remark is that our proofs consider the *generic case*. One may also consider the *worst-case* recoverability conditions via characterizing the so-called k'-rank of the latent factors [17] in the LL1 model, which may help remove the continuous distribution assumption on $\boldsymbol{A}_r, \boldsymbol{B}_r$ and $\boldsymbol{C}$.

Remark 2: Our recoverability analysis is built upon the LL1 identifiability (or, model essential uniqueness) of $\underline{\boldsymbol{Y}}_H$ and/or $\underline{\boldsymbol{Y}}_M$. Identifiability is critical to tie the physical interpretation with the latent factors, i.e., $\boldsymbol{A}_r \boldsymbol{B}_r^\top$ (abundance map) and $\boldsymbol{C}(:, r)$ (endmember), which is problematic if the factorization model is not unique. Although the LL1 identifiability does not rely on prior information such as nonnegativity of the abundance maps, such physical meaning can be used in our algorithm design to combat noise for performance enhancement in practice. A side note is that the recoverability condition could be improved, since latent factor identification is not a necessary condition for recovering the SRI; see, e.g., [10]. Also note that our recoverability analysis is based on a noise-free model, as was in [9], [10], [24]. In practice, when noise is present, efficient and effective HSR relies on custom algorithm design and effective regularization/constraints to incorporate prior knowledge, which will be addressed in the next section.

IV. Algorithm Design: Exploiting Physical Interpretations

Besides the exact recoverability guarantees (i.e., recoverability guarantees under noiseless cases such as those in Theorems 1-2), another important fact is that the LL1 tensor model is consistent with the LMM of spectral images. Therefore, the latent factors of the LL1 model admit physical interpretations. This connection allows us to incorporate prior information about the endmembers and the abundance maps of the materials contained in the HSI and MSI. Note that utilizing prior information is often important for parameter estimation, especially in the presence of heavy noise or modeling errors.

A. Reformulation

To proceed, let us first recast the criterion in (11) using the following form:

$$\begin{aligned}\min_{\{\boldsymbol{S}_r\}_{r=1}^R, \boldsymbol{C}} \quad & \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_H - \sum_{r=1}^R \left(\boldsymbol{P}_1 \boldsymbol{S}_r \boldsymbol{P}_2^\top\right) \circ \boldsymbol{C}(:, r)\right\|_F^2 \\ & + \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_M - \sum_{r=1}^R \boldsymbol{S}_r \circ (\boldsymbol{P}_M \boldsymbol{C}(:, r))\right\|_F^2 \\ & + \sum_{r=1}^R \theta_r \varphi(\boldsymbol{S}_r) + \frac{\lambda}{2}\|\boldsymbol{C}\|_F^2 \\ \text{s.t. } & \boldsymbol{S}_r \geq 0, \text{rank}(\boldsymbol{S}_r) = L_r, r = 1, \ldots, R, \boldsymbol{C} \geq 0,\end{aligned} \tag{13}$$

where $\theta_r, \lambda \geq 0$ are regularization parameters, and $\varphi(\cdot)$ is a regularization function that promotes small 2D total variation.

To explain the above reformulation, first notice that we have lifted the equality constraints in (11) to the cost function using two fitting terms. The reason is that the criterion in (11) is derived for the exact recovery of the HSI under the noiseless case. If there is noise, the problem may not have any feasible solution. Using a model fitting based approximation may circumvent such situations. In addition, we have replaced $\boldsymbol{A}_r \boldsymbol{B}_r^\top$ by $\boldsymbol{S}_r$ under the constraint $\text{rank}(\boldsymbol{S}_r) = L_r$. The reason is that we hope to "bring out" $\boldsymbol{S}_r$ in our formulation, since $\boldsymbol{S}_r$ is the abundance map of endmember r, and in many cases structural constraints/regularization on $\boldsymbol{S}_r$ can be added to enhance performance, especially in noisy cases. Third, we have imposed nonnegativity constraints on $\boldsymbol{S}_r$ and $\boldsymbol{C}$, per their physical interpretations. For example, we have added a spatial 2D TV regularizer $\varphi(\cdot)$ on $\boldsymbol{S}_r$ (which will be detailed in Sec. IV-A1). This is motivated by the fact that an endmember oftentimes exhibits very similar abundances in neighboring pixels [27], [28]. The regularization on $\boldsymbol{C}$ is to combat the scaling/counter-scaling effect in factorization models; see discussions in [29].

The reformulation is still very hard to handle. One reason is that (13) has a nonconvex rank constraint on $\boldsymbol{S}_r$. In addition, the commonly used 2D TV regularization for $\boldsymbol{S}_r$ is ℓ_1-norm based—thereby being nonsmooth; see, e.g., [28], [30], [31]. Hence, the problem is a nonconvex set-constrained nonsmooth optimization criterion, making designing effective and convergence-guaranteed algorithms challenging. To circumvent these difficulties, we propose to approximate the

formulation in (13) using the following optimization surrogate:

$$\begin{aligned}\min_{\{\boldsymbol{S}_r\}_{r=1}^R,\boldsymbol{C}} \quad & \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_H - \sum_{r=1}^R (\boldsymbol{P}_1\boldsymbol{S}_r\boldsymbol{P}_2^\top)\circ\boldsymbol{C}(:,r)\right\|_F^2 \\ & + \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_M - \sum_{r=1}^R \boldsymbol{S}_r\circ(\boldsymbol{P}_M\boldsymbol{C}(:,r))\right\|_F^2 \\ & + \sum_{r=1}^R \theta_r\varphi(\boldsymbol{S}_r) + \sum_{r=1}^R \eta_r\phi_{p,\tau}(\boldsymbol{S}_r) + \frac{\lambda}{2}\|\boldsymbol{C}\|_F^2 \\ \text{s.t. } & \boldsymbol{S}_r\geq\boldsymbol{0}, r=1,\ldots,R, \quad \boldsymbol{C}\geq\boldsymbol{0}.\end{aligned} \tag{14}$$

Here, $\eta_r \geq 0$, $\varphi(\boldsymbol{S}_r)$ is a smoothed 2D TV regularizer as before, and $\phi_{p,\tau}(\boldsymbol{S}_r)$ is low-rank promoting regularizer—which is introduced to serve as a surrogate for the hard constraint $\mathrm{rank}(\boldsymbol{S}_r) = L_r$, since dealing with the low-rank and nonnegativity constraints simultaneously is computationally challenging [32]. As we will see, if $\phi_{p,\tau}(\boldsymbol{S}_r)$ is properly designed, an efficient optimization algorithm for handling (14) can be devised. In the following, we provide design details of the requlatization terms $\varphi(\boldsymbol{S}_r)$ and $\phi_{p,\tau}(\boldsymbol{S}_r)$, respectively.

1) ℓ_q Function-Based Total Variation Regularization: To explain, we use the following TV surrogate:

$$\varphi(\boldsymbol{S}_r) = \varphi_{q,\varepsilon}(\boldsymbol{H}_x\boldsymbol{q}_r) + \varphi_{q,\varepsilon}(\boldsymbol{H}_y\boldsymbol{q}_r), \tag{15}$$

where $\boldsymbol{q}_r = \mathrm{vec}(\boldsymbol{S}_r)$ and $\varphi_{q,\varepsilon}(\boldsymbol{x}) = \sum(x_i^2+\varepsilon)^{\frac{q}{2}}$ with $0 < q \leq 1$ and $\varepsilon > 0$. The matrices $\boldsymbol{H}_x$ and $\boldsymbol{H}_y$ are the two "gradient matrices", which are defined as $\boldsymbol{H}_x = \boldsymbol{H}\otimes\boldsymbol{I}$ and $\boldsymbol{H}_y = \boldsymbol{I}\otimes\boldsymbol{H}$, respectively, where $\boldsymbol{I}\in\mathbb{R}^{J_M\times J_M}$ is an identity matrix and

$$\boldsymbol{H} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & -1 \\ -1 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{I_M\times I_M}.$$

Note that when $q < 2$, the ℓ_q function is known to be effective in promoting sparsity [33]. The parameter $\varepsilon > 0$ is employed for making the function smooth when $0 < q \leq 1$. As $q \to 0$ and $\varepsilon \to 0$, $\varphi(\boldsymbol{S}_r) \to \|\boldsymbol{H}_x\boldsymbol{q}_r\|_0 + \|\boldsymbol{H}_y\boldsymbol{q}_r\|_0$. Hence, $\varphi(\boldsymbol{S}_r)$ promotes small 2D TV of $\boldsymbol{S}_r$.

2) Schatten-p Function Based Low-rank Rank Regularization: The regularization $\phi_{p,\tau}(\boldsymbol{S}_r)$ is introduced for promoting a low-rank $\boldsymbol{S}_r$—i.e., as a surrogate for the constraint $\mathrm{rank}(\boldsymbol{S}_r) = L_r$. Here, we propose to employ the smoothed Schatten-p function [34], [35] for this purpose. Specifically, given a matrix $\boldsymbol{X}$, the smoothed Schatten-p function is defined as follows:

$$\phi_{p,\tau}(\boldsymbol{X}) = \sum_{i=1}^M (\sigma_i(\boldsymbol{X})^2+\tau)^{p/2} = \mathrm{tr}((\boldsymbol{X}\boldsymbol{X}^\top+\tau\boldsymbol{I})^{p/2}),$$

where $p > 0$. In the second line we have used the convention $\boldsymbol{A}^p = \boldsymbol{U}\boldsymbol{\Lambda}^p\boldsymbol{U}^\top$ for real symmetric matrix $\boldsymbol{A}$, in which $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ denotes the eigendecomposition of $\boldsymbol{A}$. The Shattern-p function can be understood as an ℓ_p function applied to the singular values of the matrix $\boldsymbol{X}$. The parameter $\tau > 0$ is again for smoothing the function when $0 < p \leq 1$. Also, as pointed out in [2], [36], $\phi_{p,\tau}(\boldsymbol{X})$ can be understood as a nonconvex approximation for the nuclear norm.

The reformulation in (14) admits a continuously differentiable objective function. The reformulation has also avoided directly handling the hard rank constraint via using the Schatten p-function. This allows us to design efficient first-order optimization algorithms for tackling the problem. We should mention that a similar reformulation for (12) can also be readily obtained as follows:

$$\begin{aligned}\min_{\{\boldsymbol{S}_r,\widetilde{\boldsymbol{S}}_r\}_{r=1}^R,\boldsymbol{C}} \quad & \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_H - \sum_{r=1}^R \widetilde{\boldsymbol{S}}_r\circ\boldsymbol{C}(:,r)\right\|_F^2 \\ & + \frac{1}{2}\left\|\underline{\boldsymbol{Y}}_M - \sum_{r=1}^R \boldsymbol{S}_r\circ(\boldsymbol{P}_M\boldsymbol{C}(:,r))\right\|_F^2 \\ & + \sum_{r=1}^R \eta_r\big(\phi_{p,\tau}(\boldsymbol{S}_r)+\phi_{p,\tau}(\widetilde{\boldsymbol{S}}_r)\big) \\ & + \sum_{r=1}^R \theta_r\varphi(\boldsymbol{S}_r) + \frac{\lambda}{2}\|\boldsymbol{C}\|_F^2 \\ \text{s.t. } & \boldsymbol{S}_r\geq\boldsymbol{0}, r=1,\ldots,R, \boldsymbol{C}\geq\boldsymbol{0},\end{aligned} \tag{16}$$

in which we use a low-rank $\widetilde{\boldsymbol{S}}_r$ (where the low-rank property is promoted by $\phi_{p,\tau}(\widetilde{\boldsymbol{S}}_r)$) to replace $\widetilde{\boldsymbol{A}}_r\widetilde{\boldsymbol{B}}_r^\top \in \mathbb{R}^{I_H\times J_H}$ in the formulation in (12).

B. Inexact Alternating Accelerated Projected Gradient

Let us denote the optimization problem in (14) as

$$\begin{aligned}\min_{\boldsymbol{S},\boldsymbol{C}} \ & \mathcal{J}_1(\boldsymbol{S},\boldsymbol{C}) \\ \text{s.t. } & \boldsymbol{S}\geq\boldsymbol{0}, \ \boldsymbol{C}\geq\boldsymbol{0}.\end{aligned} \tag{17}$$

Note that by our reformulation, the function $\mathcal{J}_1(\boldsymbol{S},\boldsymbol{C})$ is continuously differentiable.

1) Basic Algorithmic Structure: To tackle the optimization problem, we propose to employ the alternating projected gradient (APG) strategy. To be specific, the updates can be summarized as follows

$$\begin{aligned}\boldsymbol{C}^{(t+1)} &\leftarrow \max\left\{\boldsymbol{C}^{(t)} - \alpha^{(t)}\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t)}), \boldsymbol{0}\right\} \\ \boldsymbol{S}^{(t+1)} &\leftarrow \max\left\{\boldsymbol{S}^{(t)} - \beta^{(t)}\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)}), \boldsymbol{0}\right\},\end{aligned} \tag{18}$$

where the thresholding operators (i.e., $\max\{\cdot,\boldsymbol{0}\}$) are orthogonal projectors onto the nonnegativity orthant. Note that both partial gradients exist, under our design of the regularization terms. The expressions of $\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t)})$ and $\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)})$ can be found in Appendix C.

The convergence properties of APG algorithm can be shown under the framework in [37]. To be specific, under proper choices of $\alpha^{(t)}$ and $\beta^{(t)}$, one can show that the gradient projection step decreases $\mathcal{J}_1$ by a sufficiently large quantity—which is the key for establishing convergence of the solution sequence. To be specific, we have the following:

Proposition 1: Assume that $\alpha^{(t)} \leq 1/L_{\boldsymbol{C}}^{(t)}$ and $\beta^{(t)} \leq 1/L_{\boldsymbol{S}}^{(t)}$ in all iterations, where

$$\begin{aligned} L_{\boldsymbol{C}}^{(t)} &= \sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}^{(t)}\big) \\ &\quad + \sigma_{\max}(\boldsymbol{P}_M^\top\boldsymbol{P}_M)\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)}\big) + \lambda, \\ L_{\boldsymbol{S}}^{(t)} &= \sigma_{\max}\big((\boldsymbol{C}^{(t+1)})^\top\boldsymbol{C}^{(t+1)}\big)\sigma_{\max}\big((\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top\boldsymbol{P}_2\otimes\boldsymbol{P}_1\big) \\ &\quad + \sigma_{\max}((\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top\boldsymbol{P}_M\boldsymbol{C}^{(t+1)}) + p\max_r \eta_r\sigma_{\max}(\boldsymbol{W}_r^{(t)}) \\ &\quad + q\max_r \theta_r\big(\sigma_{\max}(\boldsymbol{H}_x^\top\boldsymbol{U}_r^{(t)}\boldsymbol{H}_x) + \sigma_{\max}(\boldsymbol{H}_y^\top\boldsymbol{V}_r^{(t)}\boldsymbol{H}_y)\big), \end{aligned} \tag{19}$$

where $\boldsymbol{S}$ is defined as in (2), $\boldsymbol{W}_r^{(t)} = (\boldsymbol{S}_r^{(t)}(\boldsymbol{S}_r^{(t)})^\top + \tau\boldsymbol{I})^{\frac{p-2}{2}}$, $\boldsymbol{U}_r^{(t)}$ and $\boldsymbol{V}_r^{(t)}$ are diagonal matrices with $[\boldsymbol{U}_r^{(t)}]_{i,i} = ([\boldsymbol{H}_x\boldsymbol{q}_r^{(t)}]_i^2 + \varepsilon)^{\frac{q-2}{2}}$, and $[\boldsymbol{V}_r^{(t)}]_{i,i} = ([\boldsymbol{H}_y\boldsymbol{q}_r^{(t)}]_i^2 + \varepsilon)^{\frac{q-2}{2}}$, $r = 1,\ldots,R$. In addition, assume that $L_{\boldsymbol{C}}^{(t)} < \infty$ and $L_{\boldsymbol{S}}^{(t)} < \infty$ for all t. Then, every limit point of the solution sequence produced by the algorithm in (18) is a stationary point of Problem (14).

The proof of Proposition 1 is relegated to Appendix D in the supplementary materials.

2) Per-iteration Complexity: To implement the algorithm, one needs to compute the gradients w.r.t. $\boldsymbol{C}$ and $\boldsymbol{S}$, respectively. The major computation burden lies in constructing $\boldsymbol{W}_r^{(t)}$ since it involves full SVD of $\boldsymbol{S}_r^{(t)}$. Consequently, the per-iteration complexity is dominated by this step, which costs $\mathcal{O}(RI_M^2\max(I_M, J_M, K_H))$ flops (which is approximately $\mathcal{O}(I_M^3)$ if $I_M \approx J_M \geq K_H$); see the detailed complexity analysis in Appendix C.

Another part that may incur many computational flops is to compute $\alpha^{(t)}$ and $\beta^{(t)}$ in each iteration. To reduce the computational burden in each iteration, we first pre-compute a number of terms, i.e., $\sigma_{\max}(\boldsymbol{P}_M^\top\boldsymbol{P}_M)$ and $\sigma_{\max}\big((\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top\boldsymbol{P}_2\otimes\boldsymbol{P}_1\big)$, since they do not change over the iterations. Second, instead of directly computing $L_{\boldsymbol{C}}^{(t)}$ and $L_{\boldsymbol{S}}^{(t)}$, which need the exact values of the terms $\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}^{(t)}\big)$, $\sigma_{\max}(\boldsymbol{H}_x^\top\boldsymbol{U}_r^{(t)}\boldsymbol{H}_x)$ and $\sigma_{\max}(\boldsymbol{H}_y^\top\boldsymbol{V}_r^{(t)}\boldsymbol{H}_y)$, we compute their upper bounds—since we only need the inequalities $\alpha^{(t)} \leq 1/L_{\boldsymbol{C}}^{(t)}$ and $\beta^{(t)} \leq 1/L_{\boldsymbol{S}}^{(t)}$ to hold. Take the first term as an example. Its upper bound can be obtained via the following:

$$\begin{aligned} &\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}^{(t)}\big) \\ &\qquad \leq \sigma_{\max}^2(\boldsymbol{S}^{(t)})\sigma_{\max}\big((\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top\boldsymbol{P}_2\otimes\boldsymbol{P}_1\big). \end{aligned}$$

Note that the right hand side only costs $\mathcal{O}(NR^2)$ flops. Similar, we compute the upper bound $\sigma_{\max}(\boldsymbol{H}_y^\top\boldsymbol{V}_r^{(t)}\boldsymbol{H}_y) \leq \sigma_{\max}(\boldsymbol{H}_y^\top)\sigma_{\max}(\boldsymbol{V}_r^{(t)})\sigma_{\max}(\boldsymbol{H}_y)$, where $\sigma_{\max}(\boldsymbol{V}_r^{(t)})$ is the largest diagonal entry of $\boldsymbol{V}_r^{(t)}$ and the other two terms are pre-computed. Using the above approach, the step size computation's complexity is almost negligible relative to that of the gradient computation.

In summary, the per-iteration complexity is in the order of $\mathcal{O}(I_M^3)$ (assuming $I_M \approx J_M \geq K_H$). For HSR tasks, I_M often lies in the range of several hundreds, which is affordable in most cases. One way to reduce complexity is to consider patch-by-patch execution of the coupled decomposition task.

3) Extrapolation: The APG algorithm is conceptually simple to implement. However, as a first-order optimization algorithm, the iteration complexity is normally not low—i.e., empirically, it often takes many iterations for the algorithm to converge to a reasonably "good" solution. One way to improve this situation without increasing per-iteration complexity is to employ the so-called extrapolation strategy. This strategy was first proposed by Nesterov [38] for convex optimization, and then was extended to handle multiblock nonconvex optimization in [37]. To be specific, instead of computing the partial gradients w.r.t. $\boldsymbol{C}$ and $\boldsymbol{S}$ at the current iterates, one can compute them at some extrapolated points $\check{\boldsymbol{C}}^{(t+1)}$ and $\check{\boldsymbol{S}}^{(t+1)}$. In a nutshell, the extrapolated point, e.g., $\check{\boldsymbol{C}}^{(t+1)}$, is a linear combination of the current $\boldsymbol{C}^{(t+1)}$ and the previous $\boldsymbol{C}^{(t)}$ using specially designed combination coefficients—and we use the combination coefficients designed by Nesterov [38] (cf. line 9 in Algorithm 1). The variables are updated via gradient projection at those extrapolated points (cf. Algorithm 1).

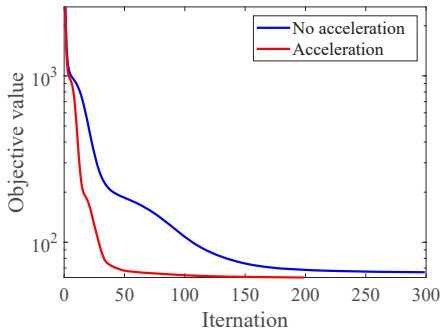


Fig. 6. Comparison of the objective values under SC-LL1 with acceleration or not. (The setting of this figure is same as the experiment shown in Table II in Section V-B2.)

Algorithm 1: SC-LL1 for solving (14).

1 **Input:** HSI $\underline{\boldsymbol{Y}}_H$, MSI $\underline{\boldsymbol{Y}}_M$, starting points $\boldsymbol{C}^{(0)}$, $\boldsymbol{S}^{(0)}$, degraded matrices $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_M$.
2 **Parameters:** λ, $\{\theta_r\}_{r=1}^R$, $\{\eta_r\}_{r=1}^R$, $\gamma_1^{(0)} = \gamma_2^{(0)} = 1$, p, q, ε, τ.
3 $\check{\boldsymbol{C}}^{(0)} = \boldsymbol{C}^{(0)}$, $\check{\boldsymbol{S}}^{(0)} = \boldsymbol{S}^{(0)}$.
4 $t = 0$.
5 **repeat**
6 %% **update** $\boldsymbol{C}$ %%
7 $\boldsymbol{C}^{(t+1)} \leftarrow \max\big\{\check{\boldsymbol{C}}^{(t)} - \alpha^{(t)}\nabla_{\check{\boldsymbol{C}}}\mathcal{J}_1\big(\boldsymbol{S}^{(t)}, \check{\boldsymbol{C}}^{(t)}\big), \boldsymbol{0}\big\}$;
8 $\gamma_1^{(t+1)} = \frac{1+\sqrt{1+4\left(\gamma_1^{(t)}\right)^2}}{2}$;
9 $\check{\boldsymbol{C}}^{(t+1)} = \boldsymbol{C}^{(t+1)} + \left(\frac{\gamma_1^{(t)}-1}{\gamma_1^{(t+1)}}\right)(\boldsymbol{C}^{(t+1)} - \boldsymbol{C}^{(t)})$;
10 %% **update** $\boldsymbol{S}$ %%
11 $\boldsymbol{S}^{(t+1)} \leftarrow \max\big\{\check{\boldsymbol{S}}^{(t)} - \beta^{(t)}\nabla_{\check{\boldsymbol{S}}}\mathcal{J}_1\big(\check{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)}\big), \boldsymbol{0}\big\}$;
12 $\gamma_2^{(t+1)} = \frac{1+\sqrt{1+4\left(\gamma_2^{(t)}\right)^2}}{2}$;
13 $\check{\boldsymbol{S}}^{(t+1)} = \boldsymbol{S}^{(t+1)} + \left(\frac{\gamma_2^{(t)}-1}{\gamma_2^{(t+1)}}\right)(\boldsymbol{S}^{(t+1)} - \boldsymbol{S}^{(t)})$;
14 $t = t + 1$;
15 **until** some stopping criterion are satisfied.
16 **Output:** $\widehat{\boldsymbol{C}} = \boldsymbol{C}^{(t)}$ and $\widehat{\boldsymbol{S}} = \boldsymbol{S}^{(t)}$. Reconstruct $\underline{\boldsymbol{Y}}_S$ using $\underline{\widehat{\boldsymbol{Y}}}_S(i,j,k) = \sum_{i=1}^R \widehat{\boldsymbol{S}}_r(i,j)\widehat{\boldsymbol{C}}(k,r)$.

Fig. 6 shows the convergence curves of the original algorithm in (18) and the accelerated version. The curves are averaged from 20 trials with Gaussian noise (the signal-to-noise ratio is 30dB). The initialization and noise at each trial

are generated randomly. Here, the goal is to fuse an HSI and an MSI whose sizes are $64 \times 64 \times 103$ and $256 \times 256 \times 4$, respectively; more details can be found in Section V. One can see that, under this simulation setting, the accelerated version uses about 50 iterations to reach a fairly low objective value level, while the original algorithm takes more than 200 iterations to reach the same level. Because of such effectiveness, in the experiments, we will always use the accelerated version.

We summarize the proposed accelerated algorithm for solving (14) in Algorithm 1, which will be referred to as the *structured coupled LL1 decomposition* (SC-LL1) algorithm in the sequel.

Remark 3: In terms of convergence, the SC-LL1 algorithm has similar properties as those of the original version in (18) [37]. The subtlety is that the accelerated version does not always produce a nonincreasing sequence of the cost function value, which may pose difficulties in convergence analysis. The work in [37] offered a fix via checking the cost value in each iteration. Nevertheless, we observe that such a fix is mostly for the theoretical proof purpose, while does not make practical differences. Checking the objective value in every iteration may increase the computational complexity. Hence, we do not include it in our algorithm.

Algorithm 2: BSC-LL1 for solving (16).

Input: HSI $\underline{\boldsymbol{Y}}_H$, MSI $\underline{\boldsymbol{Y}}_M$, starting points $\boldsymbol{C}^{(0)}$, $\boldsymbol{S}^{(0)}$, $\widetilde{\boldsymbol{S}}^{(0)}$, degraded matrix $\boldsymbol{P}_M$.
Parameters: λ, $\{\theta_r\}_{r=1}^R$, $\{\eta_r\}_{r=1}^R$, $\gamma_1^{(0)} = \gamma_2^{(0)} = \gamma_3^{(0)} = 1$, p, q, ε, τ.
$\check{\boldsymbol{C}}^{(0)} = \boldsymbol{C}^{(0)}$, $\bar{\boldsymbol{S}}^{(0)} = \widetilde{\boldsymbol{S}}^{(0)}$, $\check{\boldsymbol{S}}^{(0)} = \boldsymbol{S}^{(0)}$.
$t = 0$.
repeat
%% **update** $\boldsymbol{C}$ %%
$\boldsymbol{C}^{(t+1)} \leftarrow \max\left\{\check{\boldsymbol{C}}^{(t)} - \alpha^{(t)} \nabla_{\check{\boldsymbol{C}}} \mathcal{J}_2\left(\boldsymbol{S}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \check{\boldsymbol{C}}^{(t)}\right), \boldsymbol{0}\right\}$;
$\gamma_1^{(t+1)} = \frac{1+\sqrt{1+4\left(\gamma_1^{(t)}\right)^2}}{2}$;
$\check{\boldsymbol{C}}^{(t+1)} = \boldsymbol{C}^{(t+1)} + \left(\frac{\gamma_1^{(t)}-1}{\gamma_1^{(t+1)}}\right)\left(\boldsymbol{C}^{(t+1)} - \boldsymbol{C}^{(t)}\right)$;
%% **update** $\boldsymbol{S}$ %%
$\boldsymbol{S}^{(t+1)} \leftarrow \max\left\{\check{\boldsymbol{S}}^{(t)} - \beta^{(t)} \nabla_{\check{\boldsymbol{S}}} \mathcal{J}_2\left(\check{\boldsymbol{S}}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)}\right), \boldsymbol{0}\right\}$;
$\gamma_2^{(t+1)} = \frac{1+\sqrt{1+4\left(\gamma_2^{(t)}\right)^2}}{2}$;
$\check{\boldsymbol{S}}^{(t+1)} = \boldsymbol{S}^{(t+1)} + \left(\frac{\gamma_2^{(t)}-1}{\gamma_2^{(t+1)}}\right)\left(\boldsymbol{S}^{(t+1)} - \boldsymbol{S}^{(t)}\right)$;
%% **update** $\widetilde{\boldsymbol{S}}$ %%
$\widetilde{\boldsymbol{S}}^{(t+1)} \leftarrow \bar{\boldsymbol{S}}^{(t)} - \zeta^{(t)} \nabla_{\bar{\boldsymbol{S}}} \mathcal{J}_2\left(\boldsymbol{S}^{(t+1)}, \bar{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)}\right)$;
$\gamma_3^{(t+1)} = \frac{1+\sqrt{1+4\left(\gamma_3^{(t)}\right)^2}}{2}$;
$\bar{\boldsymbol{S}}^{(t+1)} = \widetilde{\boldsymbol{S}}^{(t+1)} + \left(\frac{\gamma_3^{(t)}-1}{\gamma_3^{(t+1)}}\right)\left(\widetilde{\boldsymbol{S}}^{(t+1)} - \widetilde{\boldsymbol{S}}^{(t)}\right)$;
$t = t + 1$;
until some stopping criterion are satisfied.
Output: $\widehat{\boldsymbol{C}} = \boldsymbol{C}^{(t)}$ and $\widehat{\boldsymbol{S}} = \boldsymbol{S}^{(t)}$. Reconstruct $\underline{\boldsymbol{Y}}_S$ using $\underline{\widehat{\boldsymbol{Y}}}_S(i,j,k) = \sum_{i=1}^R \widehat{\boldsymbol{S}}_r(i,j)\widehat{\boldsymbol{C}}(k,r)$.

C. Algorithm for Semi-blind Cases

Consider the optimization problem (16) for handling HSR when the spatial degradation operator is unknown. Similarly, we express the problem in (16) as follows:

$$\min_{\boldsymbol{S},\widetilde{\boldsymbol{S}},\boldsymbol{C}} \ \mathcal{J}_2(\boldsymbol{S}, \widetilde{\boldsymbol{S}}, \boldsymbol{C}) \quad \text{s.t. } \boldsymbol{S} \geq \boldsymbol{0},\ \boldsymbol{C} \geq \boldsymbol{0}. \tag{20}$$

We again tackle the problem using APG with the acceleration strategy, and refer the corresponding algorithm as *blind structured coupled LL1 decomposition* (BSC-LL1) algorithm. The detailed steps are summarized in Algorithm 2. The expressions of $\nabla_{\boldsymbol{C}} \mathcal{J}_2(\boldsymbol{S}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t)})$, $\nabla_{\boldsymbol{S}} \mathcal{J}_2(\boldsymbol{S}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)})$, and $\nabla_{\widetilde{\boldsymbol{S}}} \mathcal{J}_2(\boldsymbol{S}^{(t+1)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)})$, and the step sizes $\alpha^{(t)}$, $\beta^{(t)}$ and $\zeta^{(t)}$ (and the corresponding Lipchitz constants $L_{\boldsymbol{C}}^{(t)}$, $L_{\boldsymbol{S}}^{(t)}$, and $L_{\widetilde{\boldsymbol{S}}}^{(t)}$) can be found in Appendix E.

V. Experiments

In this section, we present various experiments on semi-real data and real data to demonstrate the effectiveness of the proposed HSR framework.

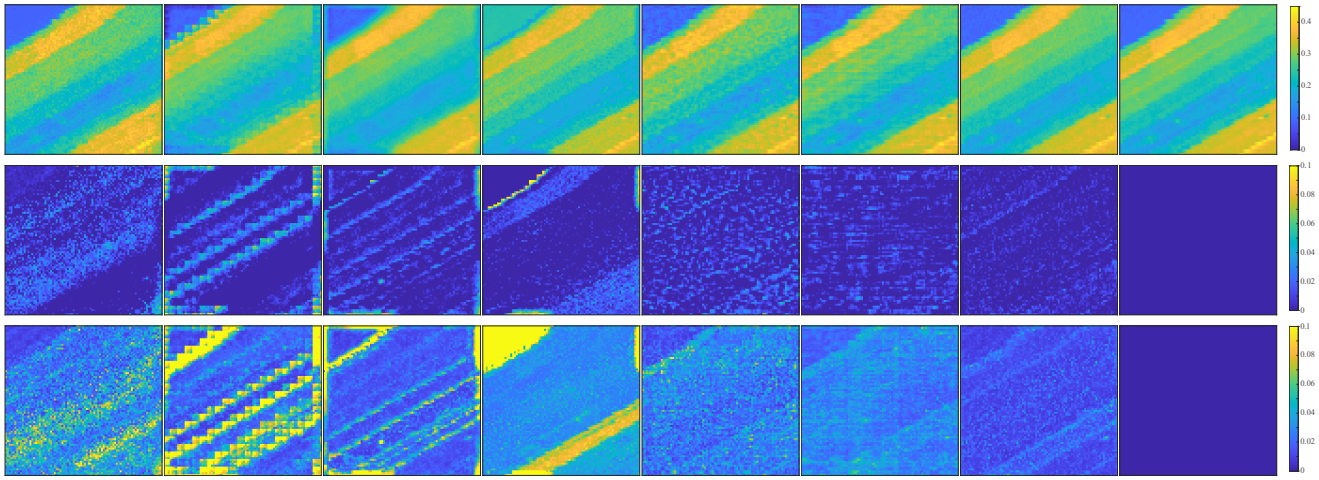
A. Experiment Setup

We benchmark our algorithm using CNMF [3], HySure [4], FUSE [5], MFpoly [7], SCOTT [10], STEREO [9]. In particular, SCOTT and STEREO are the Tucker and CPD model based HSR approaches, respectively. All simulations are coded using MATLAB 2019b and the experiments are run on a desktop with 3.4 GHz i7 CPU and 16 GB RAM. The Matlab code of our implementation can be downloaded from https://github.com/MengDing56/Code_SCLL1_HSR.

1) Degradation Model: For the semi-real data experiments, we follow the convention using a real hyperspectral image to act as the "ground-truth" SRI—so that the recovery performance can be measured [3]–[5], [9], [10]. The pair of simulated HSI and MSI images are generated following the Wald's protocol [39]. The degradation from the SRI to the HSI is as follows: the SRI is first blurred by a 9×9 Gaussian kernel and then downsampled every 4 pixels along each spatial dimension. For the degradation from the SRI to the MSI, we follow the setting in [9]. Specifically, we form the degradation matrix $\boldsymbol{P}_M$ according to the specifications of two multispectral sensors, namely, LANDSAT[2] and QuickBird[3]. The spectral degradation matrix $\boldsymbol{P}_M$ is a band selection and aggregating matrix; see details in [9]. In addition, zero-mean white Gaussian noise is added to HSI and MSI with signal-to-noise ratio (SNR) being 30 dB, if not otherwise specified. All the experiments are averaged from 20 random trials with different noise terms.

2) Metrics: In the semi-real data simulations, to evaluate the quality of the recovered SRIs, we employ a number of widely used metrics from the literature [1], [40]. In particular, we employ the reconstruction signal-to-noise ratio (R-SNR), structural similarity index (SSIM), cross correlation (CC), universal image quality index (UIQI), root mean square error (RMSE), relative dimensional global error (ERGAS), and spectral angle mapper (SAM). The detailed definitions can

[2]https://landsat.gsfc.nasa.gov/
[3]https://www.satimagingcorp.com/satellite-sensors/quickbird/



(a) CNMF (b) HySure (c) FUSE (d) MFpoly (e) SCOTT (f) STEREO (g) SC-LL1 (h) SRI

Fig. 7. The recovered SRIs, the corresponding residual images (the 34-th band), and the SAM maps of Salinas. First row: the recovered SRIs of the 34-th band; Second row: the corresponding residual images of the 34-th band; Third row: the SAM maps.

be found in [40], [41]. Higher R-SNR, SSIM, CC, and UIQI values and lower RMSE, ERGAS, and SAM values indicate a more preferred reconstruction performance.

3) Parameter Selection: In the proposed algorithms, we set $\theta_1 = \cdots = \theta_R = \theta$ and $\eta_1 = \cdots = \eta_R = \eta$. The regularization parameters θ, η and λ can be tuned by comparing the reconstructed HSI and/or MSI (using the estimated SRI and the degradation matrices) and the observed HSI and/or MSI [e.g., via observing metrics such as SAM and R-SNR computed using $\widehat{\underline{\boldsymbol{Y}}}_H$ (resp. $\widehat{\underline{\boldsymbol{Y}}}_M$) and $\underline{\boldsymbol{Y}}_H$ (resp. $\underline{\boldsymbol{Y}}_M$)] . Note that this is a heuristic that may be prone to "overfitting" to the observed data, but seems to work reasonably well in practice; also see Sec. V-B6 for numerical results showing the effect of changing these parameters. In addition, we fix the parameters $p = 0.5$, $\tau = 1$ in the Schatten-p function and $q = 0.5$, $\varepsilon = 10^{-3}$ in the TV regularization (15). The entries of the initialization terms $\boldsymbol{S}^0$, $\widetilde{\boldsymbol{S}}^0$, and $\boldsymbol{C}^0$ are drawn uniformly at random from 0 to 1. We terminate the proposed algorithm when the relative error of the objective value between two consecutive iterations is below 10^{-4} or when the number of iterations exceeds 300 and 600 for SC-LL1 and BSC-LL1, respectively. We tune the parameters of the baseline algorithms following the respective papers' instructions.

B. Semi-real Experiments with Known $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$

We first test the proposed method in cases where all $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, and $\boldsymbol{P}_M$ are known.

1) Salinas Dataset: The first experiment uses a subscene of the Salinas dataset with a size of $80 \times 84 \times 204$ (after removing 20 bands corrupted by water absorption), which is collected by AVIRIS sensor [42] over Salinas Valley [43]. Applying the described spatial degradation and the LANDSAT spectral degradation, we generate $\underline{\boldsymbol{Y}}_H \in \mathbb{R}^{20\times 21\times 204}$ and $\underline{\boldsymbol{Y}}_M \in \mathbb{R}^{80\times 84\times 6}$. We set $R = 6$, which is according to the number of materials that was reported in the literature [43]. Fig. 7 presents the 34-th band of the estimated SRIs (a band that is not contained in the MSI), the corresponding residual images (i.e., $\underline{\boldsymbol{Y}}_S(:,:,k) - \widehat{\underline{\boldsymbol{Y}}}_S(:,:,k)$ for $k = 34$), and the SAM maps. From Fig. 7, one can see that SC-LL1 has small residues across all pixels, while other algorithms' residual maps are less smooth. The proposed method also outputs an SAM map that is closer to the ideal one (cf. the rightmost column). Both results indicate that the proposed method produces an estimated SRI that captures the details of the SRI well.

TABLE I
PERFORMANCE FOR SALINAS WITH THE DEGRADATION KNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	SCOTT	STEREO	SC-LL1
R-SNR (∞)	24.30	21.16	25.24	19.55	29.89	30.25	**33.94**
SSIM (1)	0.9688	0.9588	0.9773	0.9579	0.9757	0.9751	**0.9894**
CC (1)	0.9337	0.9032	0.9204	0.7298	0.9607	0.9148	**0.9762**
UIQI (1)	0.6715	0.6484	0.6876	0.6486	0.7312	0.6484	**0.8263**
RMSE (0)	0.0122	0.0175	0.0109	0.0210	0.0064	0.0061	**0.0040**
ERGAS (0)	0.4768	0.4671	0.5396	1.1637	0.2895	0.8533	**0.2128**
SAM (0)	0.0348	0.0502	0.0315	0.0560	0.0283	0.0288	**0.0154**
Time (0)	0.70	1.47	**0.21**	2.33	3.09	0.74	23.93

We also test all the algorithms methods under different SNRs that range from 20 dB to 50 dB. Fig. 8 shows the averaged evaluation results under different noise levels. One can see that the proposed SC-LL1 method consistently outperforms the baselines under all metrics, showing promising performance. More detailed numerical comparison under SNR=30dB can be found in Table I. One can see that our method stands out under all accuracy metrics. As a price to pay, the proposed algorithm is computationally more intensive and costs more runtime.

TABLE II
PERFORMANCE FOR PAVIA UNIVERSITY WITH THE DEGRADATION KNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	SCOTT	STEREO	SC-LL1
R-SNR (∞)	20.31	16.00	20.20	15.60	21.38	24.26	**26.48**
SSIM (1)	0.9420	0.8929	0.9344	0.9262	0.9178	0.9460	**0.9730**
CC (1)	0.9807	0.9449	0.9769	0.9522	0.9816	0.9898	**0.9937**
UIQI (1)	0.9100	0.8541	0.9009	0.8884	0.8767	0.9064	**0.9467**
RMSE (0)	0.0218	0.0357	0.0220	0.0374	0.0192	0.0138	**0.0107**
ERGAS (0)	0.5501	0.8582	0.5382	0.9407	0.4414	0.3290	**0.2564**
SAM (0)	0.0814	0.1116	0.0827	0.1109	0.0776	0.0709	**0.0519**
Time (0)	4.61	8.55	0.54	20.97	**0.34**	8.54	84.44

2) Pavia University Dataset: For the second experiment, we use a subimage of the Pavia University. This dataset is captured by the ROSIS sensor [44]. The sizes of the SRI and the HSI are $256 \times 256 \times 103$ and $64 \times 64 \times 103$, respectively.

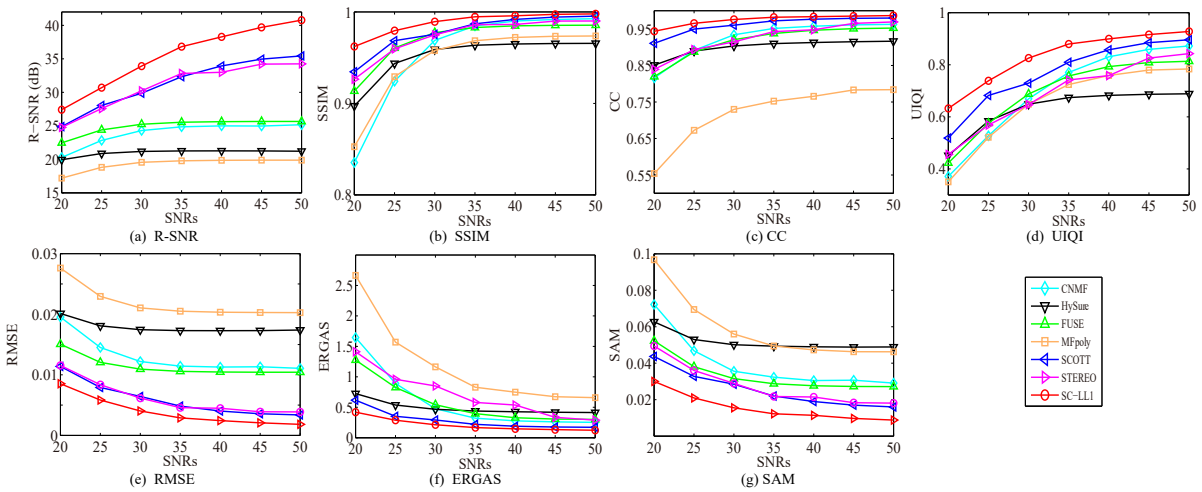


Fig. 8. Reconstruction metrics for Salinas under different noises.

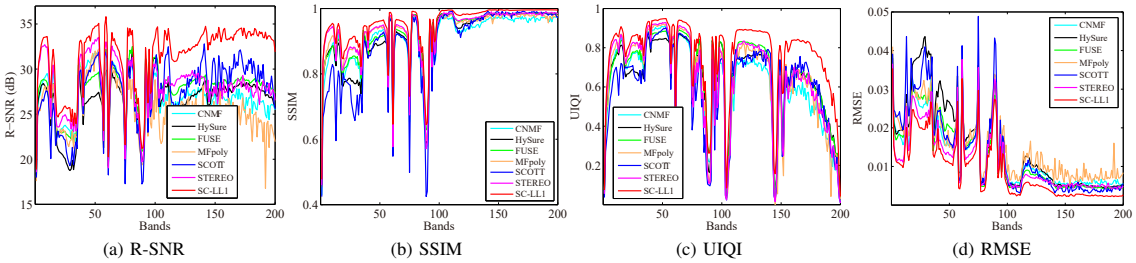


Fig. 9. R-SNR, SSIM, UIQI, and RMSE values of each band of the Indian Pines image.

We generate the MSI with a size of $256 \times 256 \times 4$ through the QuickBird spectral degradation pattern. We set $R = 4$ as the number of endmembers in this simulation.

Table II shows the reconstruction performance of algorithms under SNR=30dB. Similar to the previous experiment, one can see that SC-LL1 consistently evaluates the best over different metrics. In particular, the R-SNR output by the proposed algorithm is at least 2dB higher than that of the best baseline, which is considered a notable margin (improvement by 58%).

3) Indian Pines Dataset: The third dataset that we use is a subscene of the Indian Pines data, which is again acquired by the AVIRIS sensor. This subscene consists of $R = 16$ different prominent materials as reported in [45]. After removing water-absorption contaminated bands, we have an SRI $\underline{\boldsymbol{Y}}_S$ that has a size of $144 \times 144 \times 200$. Then, the HSI $\underline{\boldsymbol{Y}}_H \in \mathbb{R}^{36\times36\times200}$ is generated using the aforementioned spatial degradation and the MSI $\underline{\boldsymbol{Y}}_S \in \mathbb{R}^{144\times144\times6}$ is generated using the LANDSAT spectral degradation specification.

TABLE III
PERFORMANCE FOR INDIAN PINES WITH THE DEGRADATION KNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	SCOTT	STEREO	SC-LL1
R-SNR (∞)	26.45	24.70	26.49	25.54	25.04	27.38	**28.77**
SSIM (1)	0.9113	0.9150	0.9269	0.9278	0.8932	0.9339	**0.9514**
CC (1)	0.8454	0.8494	0.8654	0.8123	0.8364	0.8505	**0.9158**
UIQI (1)	0.6190	0.6434	0.6704	0.6438	0.6111	0.6534	**0.7685**
RMSE (0)	0.0153	0.0188	0.0153	0.0170	0.0180	0.0138	**0.0117**
ERGAS (0)	0.2005	0.2226	0.1844	0.2423	0.2120	0.1790	**0.1372**
SAM (0)	0.0435	0.0510	0.0419	0.0491	0.0516	0.0406	**0.0339**
Time (0)	9.29	19.05	0.48	22.53	**0.21**	2.72	129.50

From Table III, one can see that the proposed method again exhibits the most promising performance over all evaluation metrics. In addition, Fig. 9 shows the R-SNR, SSIM, UIQI, and RMSE curves against the spectral bands. Again, the propose approach has a more favorable performance over different frequency bands.

4) Jasper Ridge Dataset: The last dataset that we employ under the settings where $\boldsymbol{P}_1, \boldsymbol{P}_2$ are known is the Jasper Ridge data with a size of $100 \times 100 \times 198$ (after removing bands corrupted by dense water vapor and atmospheric effects). The HSI and MSI are with sizes of $25\times25\times198$ and $100\times100\times6$, respectively. We use $R = 4$ following [46]. Table IV shows the performance of all methods. Similar as before, the proposed SC-LL1 exhibits promising performance over this dataset.

TABLE IV
PERFORMANCE FOR JASPER RIDGE WITH THE DEGRADATION KNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	SCOTT	STEREO	SC-LL1
R-SNR (∞)	25.10	20.07	22.63	16.34	26.47	25.43	**27.16**
SSIM (1)	0.9524	0.9050	0.9320	0.8457	0.9555	0.9415	**0.9731**
CC (1)	0.9880	0.9701	0.9815	0.9512	0.9902	0.9840	**0.9921**
UIQI (1)	0.8651	0.7959	0.8449	0.7266	0.8589	0.8396	**0.9022**
RMSE (0)	0.0162	0.0288	0.0214	0.0445	0.0138	0.0155	**0.0127**
ERGAS (0)	0.3845	0.6308	0.4872	0.9681	0.3489	0.5205	**0.3283**
SAM (0)	0.0801	0.1169	0.0937	0.1658	0.0838	0.0935	**0.0675**
Time (0)	0.84	1.53	**0.27**	8.79	0.33	2.04	31.18

5) Effect of Regularization Terms: Table V shows the recovered results with different combinations of priors (low-rank (LR), TV, and combined as in SC-LL1) under known $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ under the proposed algorithmic structure. One can see that, only using the LR regularization (i.e., the LL1 model) the performance is already much improved compared to the no prior case—showing the difference between LL1 based tensor modeling and the matrix factorization models. With LR and TV combined, the performance is further improved. This suggests the recoverability-guaranteed LL1 modeling together

with plausible prior information may offer the most promising performance.

TABLE V
PERFORMANCE FOR INDIAN PINES WITH THE DEGRADATION KNOWN UNDER DIFFERENT REGULARIZATION TERMS.

Method (ideal)	Salinas				Indian Pines			
	no Priors	only LR	only TV	SC-LL1	no Priors	only LR	only TV	SC-LL1
R-SNR (∞)	29.44	32.44	33.50	**34.07**	26.09	27.71	28.56	**28.80**
SSIM (1)	0.9659	0.9851	0.9880	**0.9896**	0.8814	0.9406	0.9503	**0.9525**
CC (1)	0.9411	0.9688	0.9739	**0.9775**	0.7893	0.8907	0.9182	**0.9236**
UIQI (1)	0.6521	0.7758	0.8102	**0.8274**	0.5429	0.7164	0.7769	**0.7867**
RMSE (0)	0.0067	0.0048	0.0042	**0.0040**	0.0160	0.0133	0.0120	**0.0117**
ERGAS (0)	0.3714	0.2425	0.2120	**0.2014**	0.2545	0.1576	0.1387	**0.1347**
SAM (0)	0.0307	0.0199	0.0166	**0.0152**	0.0471	0.0387	0.0347	**0.0339**

6) Effect of Parameter Tuning: In Fig. 10, we show the sensitivity of the proposed method against the three key parameters, i.e., θ, η and λ, under different SNRs. From the figure, one can see that there is a relatively wide range of values these three parameters could take, without affecting the HSR performance heavily.

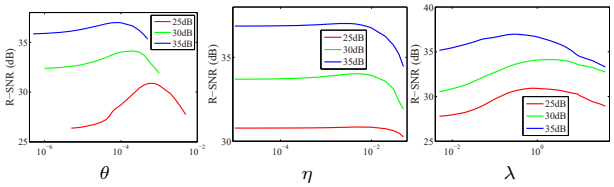


Fig. 10. R-SNR (dB) under different SNR, θ, η and λ on the Salinas data. When changing one parameter, the others are fixed to the "optimal values" under the corresponding SNRs as revealed in the three figures.

C. Semi-real Experiments with Unknown $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$

In this subsection, we test the proposed BSC-LL1 under the case where the spatial degradation matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are unknown. Note that the coupled CPD and coupled Tucker methods can also work under such a scenario. Hence, we employ their "semi-blind" versions, namely, the BSTEREO [9] and BSCOTT [10] algorithms, as our baselines in the pertinent experiments. We use the same Pavia University and Jasper Ridge datasets as before. All the settings remain the same, except that the spatial degradation operators are assumed to be unknown.

Fig. 11 shows the recovered SRIs, residual images, and the SAM maps obtained on the Pavia Unversity dataset. One can see that the proposed BSC-LL1 method keeps the edges of the SRI better compared to the baselines. This again shows our method's ability for striking a good balance between spatial smoothness and detail sharpness.

TABLE VI
PERFORMANCE FOR PAVIA UNIVERSITY WITH THE SPATIAL DEGRADATION UNKNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	BSCOTT	BSTEREO	BSC-LL1
R-SNR (∞)	20.08	16.01	17.18	15.62	24.82	23.90	**26.05**
SSIM (1)	0.9406	0.8930	0.9038	0.9304	0.9551	0.9439	**0.9710**
CC (1)	0.9797	0.9450	0.9541	0.9540	0.9907	0.9890	**0.9932**
UIQI (1)	0.9088	0.8542	0.8582	0.8936	0.9212	0.9033	**0.9422**
RMSE (0)	0.0224	0.0356	0.0311	0.0372	0.0129	0.0144	**0.0112**
ERGAS (0)	0.5650	0.8576	0.7228	0.9373	0.3185	0.3447	**0.2694**
SAM (0)	0.0816	0.1115	0.0923	0.1084	0.0628	0.0736	**0.0544**
Time (0)	4.72	8.51	0.83	4.79	**0.21**	9.75	137.10

Tables VI and VII show the corresponding evaluation results on the Pavia Unversity dataset and the Jasper Ridge dataset, respectively. Again, BSC-LL1 performs better than other methods under all metrics. For both datasets, the R-SNR performance of the proposed method largely exceeds that of the baselines.

TABLE VII
PERFORMANCE FOR JASPER RIDGE WITH THE SPATIAL DEGRADATION UNKNOWN.

Method (ideal)	CNMF	HySure	FUSE	MFpoly	BSCOTT	BSTEREO	BSC-LL1
R-SNR (∞)	25.15	20.08	16.37	15.66	25.11	24.25	**27.02**
SSIM (1)	0.9532	0.9051	0.8299	0.8524	0.9538	0.9302	**0.9721**
CC (1)	0.9883	0.9701	0.9523	0.9477	0.9847	0.9805	**0.9919**
UIQI (1)	0.8665	0.7962	0.6904	0.7263	0.8574	0.8238	**0.8984**
RMSE (0)	0.0161	0.0288	0.0441	0.0478	0.0161	0.0178	**0.0129**
ERGAS (0)	0.3818	0.6302	0.8283	1.0365	0.4333	0.5842	**0.3339**
SAM (0)	0.0807	0.1168	0.1386	0.1709	0.0883	0.1040	**0.0696**
Time (0)	0.86	1.47	0.37	1.44	**0.16**	2.01	45.88

D. Real-data Experiment

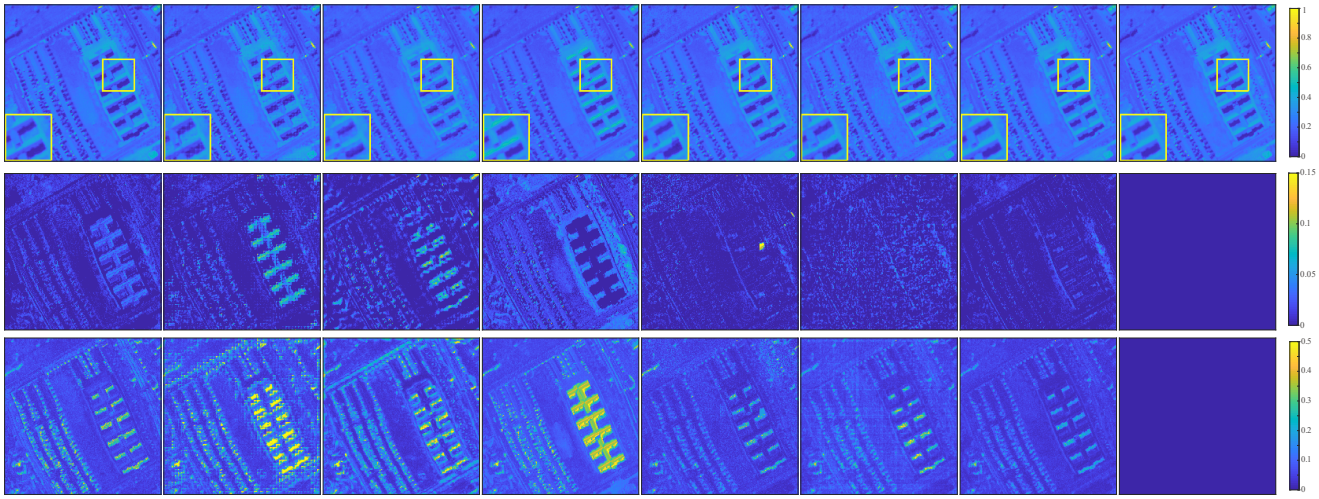
We also test the algorithms on real co-registered hyperspectral and multispectral images, where are offered in a recent work [19]. The HSI data with a size of $120 \times 120 \times 89$ is acquired by the Hyperison sensor. The 89 spectral bands are obtained after removing heavily corrupted bands. The MSI data with a size of $360 \times 360 \times 4$ is acquired by the Sentinel-2A satellite, which provides MSIs with 13 bands. We follow the demosntration method in [2] and select 4 bands whose pixels are 10m$\times$10m areas on the ground. The central wavelengths of these 4 bands are 490 nm, 560 nm, 665 nm, and 842 nm, respectively. Note that the Sentinel-2A image has more than 4 bands, but the other bands admit different spectral resolutions and thus are excluded in this experiment.

Note that both the spatial and spectral degradation operators are unknown in the real experiment. Hence, we apply the proposed BSC-LL1 for this dataset. As for the tensor-based baselines, we employ the semi-blind versions, i.e., BSCOTT [10] and BSTEREO [9]. For the matrix-based baselines, i.e., CNMF [3], HySure [4], FUSE [5], MFpoly [7], we use their own modules for estimating $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_M$ from the input HSI and MSI. The spectral degradation matrix $\boldsymbol{P}_M$ for the tensor-based methods is estimated from the available HSI and MSI images using the algorithm in [4]. For the number of materials, we set $R = 4$, which is by visual inspection.

Fig. 12 shows the recovered performance of real data at two selected bands, namely, the 23-th and 50-th bands—which are not included in the MSI. One can see that all the algorithms work to a certain extent, outputting higher spatial resolution images compared to the HSI (see column (g)). Nevertheless, one can observe that the FUSE, HySure, and MFpoly [7] algorithms both have undesired strip noise in the recovered SRI. The CNMF obtains sharp edges, which indicates satisfactory performance. However, the recovered SRIs by CNMF loses some objects and details in the spatial domain. The tensor based baselines, i.e., BSCOTT and BSTEREO, both work reasonably well. However, the proposed BSC-LL1 seems to perform better in terms of keeping edges and smoothness of the recovered SRI; see the highlighted regions.

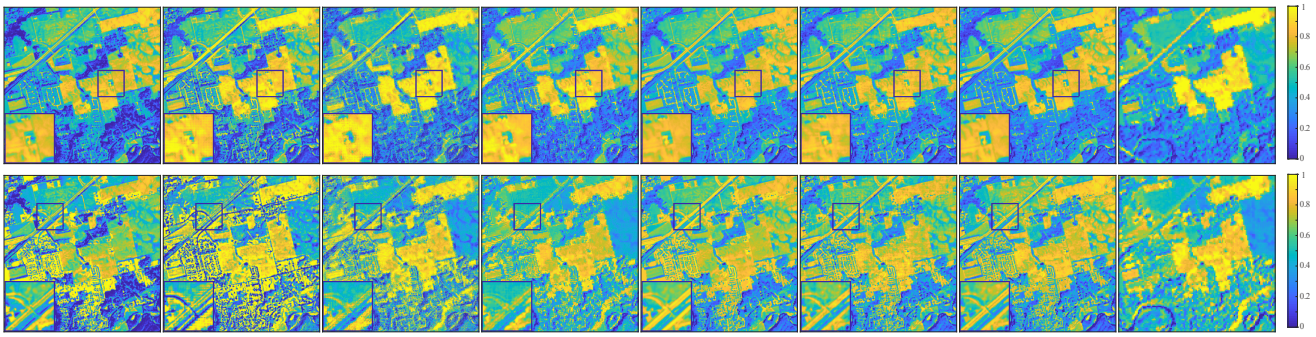
VI. CONCLUSION

In this paper, we proposed a coupled LL1 block-term tensor decomposition framework for HSR. We showed that



(a) CNMF (b) HySure (c) FUSE (d) MFpoly (e) BSCOTT (f) BSTEREO (g) BSC-LL1 (h) SRI

Fig. 11. The results on the Pavia University dataset under semi-blind settings. First row: the recovered SRIs of the 73-th band; Second row: the corresponding residual images of the 73-th band; Third row: the SAM maps.



(a) CNMF (b) HySure (c) FUSE (d) MFpoly (e) BSCOTT (f) BSTEREO (g) BSC-LL1 (h) HSI

Fig. 12. The recovered SRIs of real-data experiment. First row: the 23-th band of the recovered SRIs; Second row: the 50-th band of the recovered SRIs.

the SRI recoveriability is guaranteed under mild conditions. Hence, using the advocated LL1 modeling instead of the CPD/Tucker modeling for spectral images does not lose theoretical guarantees. A salient feature of the proposed framework is that the latent factors under the LL1 model admit physical meaning. Therefore, structural constraints and regularization terms that reflect prior information about the latent factors can be flexibly incorporated in our HSR formulations—thereby enhancing performance under noisy scenarios. We recast the proposed HSR criteria into forms that take into consideration of the physical characterizations of the endmembers' spectral signatures and abundance maps (e.g., nonnegativity and spatial smoothness), and proposed a convergence-guaranteed accelerated alternating gradient projection algorithmic framework to tackle these problems. We tested the proposed algorithms over extensive semi-real and real experiments, which both showed consistently promising performance on HSR tasks. A future direction is to study scalable algorithms that can reduce the runtime while keeping a comparable HSR performance.

Appendix A
Proof of Theorem 1

To proceed, let us consider the following important lemmas.

Lemma 2: [9] Let $\widetilde{\boldsymbol{A}} = \boldsymbol{P}\boldsymbol{A}$, where $\boldsymbol{A}$ is drawn from any absolutely continuous joint distribution in $\mathbb{R}^{I\times L}$ and $\boldsymbol{P} \in \mathbb{R}^{I'\times I}$ is full row rank. Then, $\widetilde{\boldsymbol{A}}$ follows a joint absolutely continuous joint distribution in $\mathbb{R}^{I'\times L}$.

Let $(\{\boldsymbol{A}_r, \boldsymbol{B}_r^\top\}_{r=1}^R, \boldsymbol{C})$ denote the ground-truth factors of the SRI tensor. Assume that $(\{\boldsymbol{A}_r^\star, (\boldsymbol{B}_r^\star)^\top\}_{r=1}^R, \boldsymbol{C}^\star)$ represents an optimal solution of the formulation (11). Then, we have

$$\begin{aligned}\underline{\boldsymbol{Y}}_M &= \sum_{r=1}^R \left(\boldsymbol{A}_r^\star(\boldsymbol{B}_r^\star)^\top\right) \circ \boldsymbol{P}_M\boldsymbol{C}^\star(:,r)\\ &= \sum_{r=1}^R \left(\boldsymbol{A}_r\boldsymbol{B}_r^\top\right) \circ \boldsymbol{P}_M\boldsymbol{C}(:,r), \qquad (21)\\ \underline{\boldsymbol{Y}}_H &\sum_{r=1}^R \left(\boldsymbol{P}_1\boldsymbol{A}_r^\star(\boldsymbol{P}_2\boldsymbol{B}_r^\star)^\top\right) \circ \boldsymbol{C}^\star(:,r)\\ &= \sum_{r=1}^R \left(\boldsymbol{P}_1\boldsymbol{A}_r(\boldsymbol{P}_2\boldsymbol{B}_r)^\top\right) \circ \boldsymbol{C}(:,r). \qquad (22)\end{aligned}$$

To proceed, note that $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_M$ all have full row rank. By Lemma 2, $\boldsymbol{P}_1\boldsymbol{A}$, $\boldsymbol{P}_2\boldsymbol{B}$, and $\boldsymbol{P}_M\boldsymbol{C}$ all follow certain joint absolutely continuous distributions. Therefore, under the conditions $I_M J_M \geq L^2 R$ and $\min\left(\lfloor I_M/L\rfloor, R\right) + \min\left(\lfloor J_M/L\rfloor, R\right) + \min(K_M, R) \geq 2R+2$, and by Lemma 1,

the LL1 decomposition of $\underline{\boldsymbol{Y}}_M$ is essentially unique—which means that from (21) one can conclude the following:

$$\boldsymbol{S}^\star = \boldsymbol{S}\boldsymbol{\Pi}\boldsymbol{\Lambda},\ \boldsymbol{P}_M\boldsymbol{C}^\star = \boldsymbol{P}_M\boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\Lambda}^{-1}, \tag{23}$$

where $\boldsymbol{S}^\star = [\text{vec}(\boldsymbol{S}_1^\star), \ldots, \text{vec}(\boldsymbol{S}_R^\star)]$, $\boldsymbol{S}_r^\star = \boldsymbol{A}_r^\star (\boldsymbol{B}_r^\star)^\top$ and $\boldsymbol{S} = [\text{vec}(\boldsymbol{S}_1), \ldots, \text{vec}(\boldsymbol{S}_R)]$, $\boldsymbol{S}_r = \boldsymbol{A}_r\boldsymbol{B}_r^\top$. Define $\widetilde{\boldsymbol{S}} = [(\boldsymbol{P}_2\boldsymbol{B}_1 \odot_c \boldsymbol{P}_1\boldsymbol{A}_1)\boldsymbol{1}_L \ldots, (\boldsymbol{P}_2\boldsymbol{B}_R \odot_c \boldsymbol{P}_1\boldsymbol{A}_R)\boldsymbol{1}_L]$. One can easily see that

$$\widetilde{\boldsymbol{S}}^\star = \widetilde{\boldsymbol{S}}\boldsymbol{\Pi}\boldsymbol{\Lambda}, \tag{24}$$

where $\widetilde{\boldsymbol{S}}^\star = [(\boldsymbol{P}_2\boldsymbol{B}_1^\star \odot_c \boldsymbol{P}_1\boldsymbol{A}_1^\star)\boldsymbol{1}_L \ldots, (\boldsymbol{P}_2\boldsymbol{B}_R^\star \odot_c \boldsymbol{P}_1\boldsymbol{A}_R^\star)\boldsymbol{1}_L]$.

Next, we show that $\widetilde{\boldsymbol{S}}$ has full column rank almost surely. To see this, note that the matrix $\boldsymbol{P}_2\boldsymbol{B} \odot \boldsymbol{P}_1\boldsymbol{A} = [\boldsymbol{P}_2\boldsymbol{B}_1 \otimes \boldsymbol{P}_1\boldsymbol{A}_1, \ldots, \boldsymbol{P}_2\boldsymbol{B}_R \otimes \boldsymbol{P}_1\boldsymbol{A}_R]$ admits full column rank almost surely if all $\boldsymbol{A}_r$ and $\boldsymbol{B}_r$ are drawn from any joint absolutely continuously distribution and if $I_H J_H \geq LR$ [47, Lemma 3.3], where we have used the notation $\boldsymbol{B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_R]$ and $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_R]$. Since the matrix $\widetilde{\boldsymbol{P}} = [\boldsymbol{P}_2\boldsymbol{B}_1 \odot_c \boldsymbol{P}_1\boldsymbol{A}_1, \ldots, \boldsymbol{P}_2\boldsymbol{B}_R \odot_c \boldsymbol{P}_1\boldsymbol{A}_R]$ is a submatrix of $\boldsymbol{P}_2\boldsymbol{B} \odot \boldsymbol{P}_1\boldsymbol{A}$, it also has full column rank. Note that $\widetilde{\boldsymbol{S}} = [(\boldsymbol{P}_2\boldsymbol{B}_1 \odot_c \boldsymbol{P}_1\boldsymbol{A}_1)\boldsymbol{1}_L \ldots, (\boldsymbol{P}_2\boldsymbol{B}_R \odot_c \boldsymbol{P}_1\boldsymbol{A}_R)\boldsymbol{1}_L]$. Our claim is that $\widetilde{\boldsymbol{S}}$ has to be full column rank. This can be seen by contradiction. Suppose that $\widetilde{\boldsymbol{S}}$ is rank deficient. Then, there exists $(\alpha_1, \ldots, \alpha_R) \neq \boldsymbol{0}$ such that

$$(\boldsymbol{P}_2\boldsymbol{B}_1 \odot_c \boldsymbol{P}_1\boldsymbol{A}_1)\boldsymbol{1}_L\alpha_1 + \ldots + (\boldsymbol{P}_2\boldsymbol{B}_R \odot_c \boldsymbol{P}_1\boldsymbol{A}_R)\boldsymbol{1}_L\alpha_R = \boldsymbol{0}. \tag{25}$$

However, (25) also means that the columns of $\widetilde{\boldsymbol{P}}$ (and thus those of $\boldsymbol{P}_2\boldsymbol{B} \odot \boldsymbol{P}_1\boldsymbol{A}$) are linearly dependent, which is a contradiction.

To continue, we rearrange the tensor using the following rule:

$$\boldsymbol{Y}_H^{(3)} = [\underline{\boldsymbol{Y}}_H(1,1,:), \underline{\boldsymbol{Y}}_H(2,1,:), \ldots, \underline{\boldsymbol{Y}}_H(I_H, J_H, :)]^\top.$$

The above matrix can also be represented as [17]: $\boldsymbol{Y}_H^{(3)} = \widetilde{\boldsymbol{S}}\boldsymbol{C}^\top$. By our identification criterion and (22), we also have $\boldsymbol{Y}_H^{(3)} = \widetilde{\boldsymbol{S}}^\star(\boldsymbol{C}^\star)^\top$. Combining the above two equalities and using (24), we have

$$\widetilde{\boldsymbol{S}}\boldsymbol{C}^\top = \widetilde{\boldsymbol{S}}\boldsymbol{\Pi}\boldsymbol{\Lambda}(\boldsymbol{C}^\star)^\top \Longrightarrow \boldsymbol{C}^\star = \boldsymbol{C}\boldsymbol{\Pi}\boldsymbol{\Lambda}^{-1}, \tag{26}$$

where we have used the fact that $\widetilde{\boldsymbol{S}}$ has full column rank.

Combining (23) and (26), one can see that the matrix unfolding of the SRI (following the same unfolding rule as in $\boldsymbol{Y}_H^{(3)}$), i.e., $\boldsymbol{Y}_S^{(3)}$, can be recovered by $\boldsymbol{Y}_S^{(3)} = \boldsymbol{S}^\star(\boldsymbol{C}^\star)^\top$. This completes the proof.

Appendix B
Proof of Theorem 2

The proof of Theorem 2 is by applying Lemma 1 to the HSI and MSI individually. Then, the coupled tensor modeling is utilized for "aligning" the permutation and scaling ambiguities automatically.

Assume that $(\{\widetilde{\boldsymbol{A}}_r^\star, \widetilde{\boldsymbol{B}}_r^\star\}_{r=1}^R, \{\boldsymbol{A}_r^\star, \boldsymbol{B}_r^\star\}_{r=1}^R, \boldsymbol{C}^\star)$ is an optimal solution of the formulation (12). Using Lemma 1 and the assumptions, the following expressions hold

$$\boldsymbol{S}^\star = \boldsymbol{S}\boldsymbol{\Pi}_1\boldsymbol{\Lambda}_1,\ \boldsymbol{P}_M\boldsymbol{C}^\star = \boldsymbol{P}_M\boldsymbol{C}\boldsymbol{\Pi}_1\boldsymbol{\Lambda}_1^{-1}, \tag{27}$$

where $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Lambda}_1$ are a permutation matrix and a nonsingular scaling matrix associated with the MSI decomposition, i.e., the decomposition model in (12b).

Moreover, applying Lemma 1 to the HSI and the equality constraint in (12a), the following expressions also hold

$$\widetilde{\boldsymbol{S}}^\star = \widetilde{\boldsymbol{S}}\boldsymbol{\Pi}_2\boldsymbol{\Lambda}_2,\ \boldsymbol{C}^\star = \boldsymbol{C}\boldsymbol{\Pi}_2\boldsymbol{\Lambda}_2^{-1}, \tag{28}$$

where $\widetilde{\boldsymbol{S}}^\star$ and $\widetilde{\boldsymbol{S}}$ are denoted as in Appendix A, and $\boldsymbol{\Pi}_2$ and $\boldsymbol{\Lambda}_2$ are a permutation matrix and a nonsingular scaling matrix associated with the decomposition model in (12a).

Plugging $\boldsymbol{C}^\star$ into (27), we have

$$\boldsymbol{P}_M\boldsymbol{C}\boldsymbol{\Pi}_2\boldsymbol{\Lambda}_2^{-1} = \boldsymbol{P}_M\boldsymbol{C}\boldsymbol{\Pi}_1\boldsymbol{\Lambda}_1^{-1}. \tag{29}$$

By Lemma 2, $\boldsymbol{P}_M\boldsymbol{C}$ is drawn from an absolutely continuous joint distribution. Hence, the Kruskal rank of $\boldsymbol{P}_M\boldsymbol{C}$ is $\min(K_M, R)$ almost surely.

First notice that if $R = 1$, then $\boldsymbol{\Pi}_1 = \boldsymbol{\Pi}_2$ and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ hold trivially. Second, consider the case where $R \geq 2$. Under such cases and by assuming that $K_M \geq 2$, we have $\min(K_M, R) \geq 2$. Hence, any two columns of $\boldsymbol{P}_M\boldsymbol{C}$ are linearly independent by the fact that $\boldsymbol{P}_M\boldsymbol{C}$ is drawn from a joint absolutely continuous distribution. Our claim is that $\boldsymbol{\Pi}_1 = \boldsymbol{\Pi}_2$ and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ still hold under such circumstances. To see this, let us re-write (29) as follows:

$$\boldsymbol{P}_M\boldsymbol{C}\boldsymbol{Z}_1(:,r) = \boldsymbol{P}_M\boldsymbol{C}\boldsymbol{Z}_2(:,r),\ r = 1, \ldots, R, \tag{30}$$

where $\boldsymbol{Z}_i = \boldsymbol{\Pi}_i\boldsymbol{\Lambda}_i^{-1}$ for $i = 1, 2$. Consequently, we have

$$\boldsymbol{P}_M\boldsymbol{C}\left(\boldsymbol{Z}_1(:,r) - \boldsymbol{Z}_2(:,r)\right) = \boldsymbol{0},\ r = 1, \ldots, R. \tag{31}$$

Note that the vector $(\boldsymbol{Z}_1(:,r) - \boldsymbol{Z}_2(:,r))$ has at most two nonzero elements. However, any two columns of $\boldsymbol{P}_M\boldsymbol{C}$ are linearly independent. This means that (31) holds if and only if $\boldsymbol{Z}_1(:,r) = \boldsymbol{Z}_2(:,r)$ for $r = 1, \ldots, R$, which leads to $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2$ and $\boldsymbol{\Pi}_1 = \boldsymbol{\Pi}_2$. As a result, we have the following equality: $\boldsymbol{C}^\star = \boldsymbol{C}\boldsymbol{\Pi}_1\boldsymbol{\Lambda}_1^{-1}$. Then, the matrix unfolding of the SRI, i.e., $\boldsymbol{Y}_S^{(3)}$, can be recovered by $\boldsymbol{Y}_S^{(3)} = \boldsymbol{S}^\star(\boldsymbol{C}^\star)^\top$. This completes the proof.

Appendix C
Gradients in (18)

The $\boldsymbol{C}$-subproblem is a quadratic program. The gradient can be readily derived:

$$\begin{aligned}\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)}, \boldsymbol{C}^{(t)}) &= \boldsymbol{C}^{(t)}(\boldsymbol{S}^{(t)})^\top(\boldsymbol{P}_2 \otimes \boldsymbol{P}_1)^\top(\boldsymbol{P}_2 \otimes \boldsymbol{P}_1)\boldsymbol{S}^{(t)} \\ &+\boldsymbol{P}_M^\top\boldsymbol{P}_M\boldsymbol{C}^{(t)}(\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)} + \lambda\boldsymbol{C}^{(t)} \\ &-\left(\boldsymbol{Y}_H^{(3)}\right)^\top(\boldsymbol{P}_2 \otimes \boldsymbol{P}_1)\boldsymbol{S}^{(t)} + \boldsymbol{P}_M^\top\left(\boldsymbol{Y}_M^{(3)}\right)^\top\boldsymbol{S}^{(t)},\end{aligned} \tag{32}$$

where $\boldsymbol{Y}_M^{(3)} = [\underline{\boldsymbol{Y}}_M(1,1,:), \underline{\boldsymbol{Y}}_M(2,1,:), \ldots, \underline{\boldsymbol{Y}}_M(I_M, J_M, :)]^\top$.

The detailed computational complexity analysis for instantiating $\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)}, \boldsymbol{C}^{(t)})$ can be found in Table VIII. The computational complexity may be reduced by exploiting the structure of $\boldsymbol{P}_H = \boldsymbol{P}_2 \otimes \boldsymbol{P}_1$. Note that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are often sparse [5], [9]. For example, $d \times d$ Gaussian blurring kernels and downsampling operators are widely used to model $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, and such modeling makes these matrices sparse—the number of nonzero element $\text{nnz}(\boldsymbol{P}_1) = dI_H$ and $\text{nnz}(\boldsymbol{P}_2) = dJ_H$; see, e.g., [9], for details.

TABLE VIII
COMPLEXITY OF COMPUTING $\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t)})$.

Term	Complexity
$\boldsymbol{C}^{(t)}(\boldsymbol{S}^{(t)})^\top\boldsymbol{P}_H^\top\boldsymbol{P}_H\boldsymbol{S}^{(t)}$	$\mathcal{O}(R((d^2+R)I_HJ_H+RK_H))$
$\boldsymbol{P}_M^\top\boldsymbol{P}_M\boldsymbol{C}^{(t)}(\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)}$	$\mathcal{O}(RI_MJ_MK_H)$
$\left(\boldsymbol{Y}_H^{(3)}\right)^\top\boldsymbol{P}_H\boldsymbol{S}^{(t)}$	$\mathcal{O}(K_H(RI_MJ_M+d^2I_HJ_H))$
$\boldsymbol{P}_M^\top\left(\boldsymbol{Y}_M^{(3)}\right)^\top\boldsymbol{S}^{(t)}$	$\mathcal{O}(RI_MJ_MK_H)$

TABLE IX
COMPLEXITY OF COMPUTING $\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)})$.

Term	Complexity
$\boldsymbol{P}_H^\top(\boldsymbol{P}_H\boldsymbol{S}^{(t)}(\boldsymbol{C}^{(t+1)})^\top-\boldsymbol{Y}_H^{(3)})\boldsymbol{C}^{(t+1)}$	$\mathcal{O}(I_HJ_H(d^2R+RK_H+d^2K_H))$
$(\boldsymbol{S}^{(t)}(\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top-\boldsymbol{Y}_M^{(3)})\boldsymbol{P}_M\boldsymbol{C}^{(t+1)}$	$\mathcal{O}(RI_MJ_MK_H)$
$\text{vec}(\boldsymbol{W}_r^{(t)}\boldsymbol{S}_r^{(t)})$	$\mathcal{O}(RI_M^2(I_M+J_M))$
$\boldsymbol{H}_x^\top\boldsymbol{U}_r^{(t)}\boldsymbol{H}_x\boldsymbol{q}_r^{(t)}$, $\boldsymbol{H}_y^\top\boldsymbol{V}_r^{(t)}\boldsymbol{H}_y\boldsymbol{q}_r^{(t)}$	$\mathcal{O}(RI_MJ_M)$

To see the analytical form of $\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S},\boldsymbol{C}^{(t+1)})$, we construct a tight upper bounded function $\mathcal{F}(\boldsymbol{S},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})$ at tth iteration such that $\mathcal{F}(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})\geq\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)})$, and $\nabla_{\boldsymbol{S}}\mathcal{F}(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})=\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)})$. Then, we will compute $\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S},\boldsymbol{C}^{(t+1)})$ through computing $\nabla_{\boldsymbol{S}}\mathcal{F}(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})$.

It was shown in [2], [35] that, for $0<p\leq 1$, $\phi_{p,\tau}(\boldsymbol{Z})$ admits a quadratic tight upper bound (i.e., a *majorizer*) at $\boldsymbol{Z}^{(t)}$:

$$\begin{aligned}&\widetilde{\phi}(\boldsymbol{Z},\boldsymbol{W}^{(t)})\\&=\frac{p}{2}\text{tr}\big(\boldsymbol{W}^{(t)}(\boldsymbol{Z}\boldsymbol{Z}^\top+\tau\boldsymbol{I})\big)+\frac{2-p}{p}\text{tr}\big((\boldsymbol{W}^{(t)})^{\frac{p}{p-2}}\big),\end{aligned}\tag{33}$$

where $\boldsymbol{W}^{(t)}=\big(\boldsymbol{Z}^{(t)}(\boldsymbol{Z}^{(t)})^\top+\tau\boldsymbol{I}\big)^{\frac{p-2}{2}}$. In addition, it is shown in [29] that a quadratic majorizer $\widetilde{\varphi}(\boldsymbol{z},\boldsymbol{z}^{(t)})$ of $\varphi_{q,\varepsilon}(\boldsymbol{z})$ $(0<q\leq 1)$ is as follows

$$\begin{aligned}\widetilde{\varphi}(\boldsymbol{z},\boldsymbol{z}^{(t)})&=\sum_i[\boldsymbol{w}^{(t)}]_i[\boldsymbol{z}]_i^2+\frac{2-q}{2}\Big(\frac{2}{q}[\boldsymbol{w}^{(t)}]_i\Big)^{\frac{q}{q-2}}+\varepsilon[\boldsymbol{w}^{(t)}]_i\\&=\frac{q}{2}\boldsymbol{z}^\top\boldsymbol{U}^{(t)}\boldsymbol{z}+\text{const},\end{aligned}\tag{34}$$

where $[\boldsymbol{w}^{(t)}]_i=\frac{q}{2}\big(([\boldsymbol{z}^{(t)}]_i)^2+\varepsilon\big)^{\frac{q-2}{2}}$, $\boldsymbol{U}^{(t)}$ is a diagonal matrix with $[\boldsymbol{U}^{(t)}]_{i,i}=[\boldsymbol{w}^{(t)}]_i$ and const is a constant. Combining (33) and (34), we obtain the quadratic majorizer function $\mathcal{F}(\boldsymbol{S},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})$ as follows:

$$\begin{aligned}&\mathcal{F}(\boldsymbol{S},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})=\frac{1}{2}\|\boldsymbol{Y}_H^{(3)}-(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}(\boldsymbol{C}^{(t+1)})^\top\|_F^2\\&+\frac{1}{2}\|\boldsymbol{Y}_M^{(3)}-\boldsymbol{S}(\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top\|_F^2+\sum_{r=1}^R\widetilde{\phi}(\boldsymbol{S}_r,\boldsymbol{W}_r^{(t)})\\&+\sum_{r=1}^R\big(\widetilde{\varphi}(\boldsymbol{H}_x\boldsymbol{q}_r,\boldsymbol{H}_x\boldsymbol{q}_r^{(t)})+\widetilde{\varphi}(\boldsymbol{H}_y\boldsymbol{q}_r,\boldsymbol{H}_y\boldsymbol{q}_r^{(t)})\big).\end{aligned}\tag{35}$$

The gradient of $\mathcal{F}(\boldsymbol{S},\boldsymbol{C}^{(t+1)};\boldsymbol{S}^{(t)})$ w.r.t. $\boldsymbol{S}$ can be expressed as follows:

$$\begin{aligned}&\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t+1)})\\&=(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top\big((\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}^{(t)}(\boldsymbol{C}^{(t+1)})^\top-\boldsymbol{Y}_H^{(3)}\big)\boldsymbol{C}^{(t+1)}\\&+\big(\boldsymbol{S}^{(t)}(\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top-\boldsymbol{Y}_M^{(3)}\big)\boldsymbol{P}_M\boldsymbol{C}^{(t+1)}\\&+p[\eta_1\text{vec}(\boldsymbol{W}_1^{(t)}\boldsymbol{S}_1^{(t)}),\ldots,\eta_R\text{vec}(\boldsymbol{W}_R^{(t)}\boldsymbol{S}_R^{(t)})]\\&+q[\theta_1\boldsymbol{H}_x^\top\boldsymbol{U}_1^{(t)}\boldsymbol{H}_x\boldsymbol{q}_1^{(t)},\ldots,\theta_R\boldsymbol{H}_x^\top\boldsymbol{U}_R^{(t)}\boldsymbol{H}_x\boldsymbol{q}_R^{(t)}]\\&+q[\theta_1\boldsymbol{H}_y^\top\boldsymbol{V}_1^{(t)}\boldsymbol{H}_y\boldsymbol{q}_1^{(t)},\ldots,\theta_R\boldsymbol{H}_y^\top\boldsymbol{V}_R^{(t)}\boldsymbol{H}_y\boldsymbol{q}_R^{(t)}],\end{aligned}$$

where $\boldsymbol{W}_r^{(t)}=(\boldsymbol{S}_r^{(t)}(\boldsymbol{S}_r^{(t)})^\top+\tau\boldsymbol{I})^{\frac{p-2}{2}}$, $[\boldsymbol{U}_r^{(t)}]_{i,i}=\big([\boldsymbol{H}_x\boldsymbol{q}_r^{(t)}]_i^2+\varepsilon\big)^{\frac{q-2}{2}}$, and $[\boldsymbol{V}_r^{(t)}]_{i,i}=\big([\boldsymbol{H}_y\boldsymbol{q}_r^{(t)}]_i^2+\varepsilon\big)^{\frac{q-2}{2}}$, $r=1,\ldots,R$. The complexity of constructing $\nabla_{\boldsymbol{C}}\mathcal{J}_1(\boldsymbol{S}^{(t)},\boldsymbol{C}^{(t)})$ can be seen in Table IX.

REFERENCES

[1] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geosc. Remote Sens. Mag.*, vol. 5, no. 2, pp. 29–56, 2017.

[2] R. Wu, W.-K. Ma, X. Fu, and Q. Li, "Hyperspectral super-resolution via global-local low-rank matrix estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7125–7140, 2020.

[3] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, 2012.

[4] M. Simões, J. M. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, 2015.

[5] Q. Wei, N. Dobigeon, and J. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, 2015.

[6] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[7] Q. Li, W.-K. Ma, and Q. Wu, "Hyperspectral super-resolution: Exact recovery in polynomial time," in *IEEE Workshop Stat. Signal Process.*, 2018, pp. 378–382.

[8] H. Liu, R. Wu, and W.-K. Ma, "Is there any recovery guarantee with coupled structured matrix factorization for hyperspectral super-resolution?" in *Proc. IEEE Int. Workshop Comput. Adv. Mult.-Sens. Adapt. Process.*, 2019.

[9] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6503–6517, 2018.

[10] C. Prévost, K. Usevich, P. Comon, and D. Brie, "Hyperspectral super-resolution with coupled Tucker approximation: Recoverability and SVD-based algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 931–946, 2020.

[11] R. Borsoi, C. Prévost, K. Usevich, D. Brie, J. Bermudez, and C. Richard, "Coupled tensor decomposition for hyperspectral and multispectral image fusion with variability," *arXiv:2006.16968v1*, 2020.

[12] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3862–3871.

[13] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, 2018.

[14] E. Wycoff, T.-H. Chan, K. Jia, W.-K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 1409–1413.

[15] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3586–3594.

[16] Q. Wei, J. M. Bioucas-Dias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, 2015.

[17] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms—Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[18] Y. Qian, F. Xiong, S. Zeng, J. Zhou, and Y.-Y. Tang, "Matrix-vector nonnegative tensor factorization for blind unmixing of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1776–1792, 2017.

[19] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, pp. 1–23, 2018.

[20] G. Zhang, X. Fu, K. Huang, and J. Wang, "Hyperspectral super-resolution: A coupled nonnegative block-term tensor decomposition approach," in *Proc. IEEE Int. Workshop Comput. Adv. Mult.-Sens. Adapt. Process.*, 2019, pp. 470–474.

[21] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *J. Math. Phys.*, vol. 6, no. 1-4, pp. 164–189, 1927.

[22] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[23] F. Xiong, Y. Qian, J. Zhou, and Y.-Y. Tang, "Hyperspectral unmixing via total variation regularized nonnegative tensor factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2341–2357, 2019.

[24] G. Zhang, X. Fu, J. Wang, X.-L. Zhao, and M. Hong, "Spectrum cartography via coupled block-term tensor decomposition," *IEEE Trans. Signal Process.*, vol. 68, pp. 3660–3675, 2020.

[25] M. Sørensen and L. De Lathauwer, "Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms—Part I: Uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 36, no. 2, p. 496–522, 2015.

[26] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and M. Akcakaya, "Tensor completion from regular sub-Nyquist samples," *IEEE Trans. Signal Process.*, vol. 68, pp. 1–16, 2020.

[27] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1200–1210, 2012.

[28] M. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4484–4502, 2012.

[29] X. Fu, K. Huang, W.-K. Ma, N. D. Sidiropoulos, and R. Bro, "Joint tensor factorization and outlying slab suppression with applications," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6315–6328, 2015.

[30] A. Zymnis, S. Kim, J. Skaf, M. Parente, and S. Boyd, "Hyperspectral image unmixing via alternating projected subgradients," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput*, 2007, pp. 1164–1168.

[31] H. K. Aggarwal and A. Majumdar, "Hyperspectral unmixing in the presence of mixed noise using joint-sparsity and total variation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4257–4266, 2016.

[32] S. A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. Optimiz.*, vol. 20, no. 3, pp. 1364–1377, 2010.

[33] X. Fu, W.-K. Ma, J. M. Bioucas-Dias, and T.-H. Chan, "Semiblind hyperspectral unmixing in the presence of spectral library mismatches," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5171–5184, 2016.

[34] R. Bouldin, "Best approximation of a normal operator in the Schatten p-norm," in *Proc. Amer. Math. Soc.*, vol. 80, no. 2, 1980, pp. 277–282.

[35] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res.*, vol. 13, pp. 3441–3473, 2012.

[36] M.-J. Lai, Y. Xu, and W. Yin, "Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, 2013.

[37] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imaging Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.

[38] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Dokl. Akad. Nauk SSSR*, vol. 269, pp. 543–547, 1983.

[39] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[40] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simões, J. Tourneret, M. A. Veganzones, G. Vivone, Q. Wei, and N. Yokoya, "Hyperspectral pansharpening: A review," *IEEE Geosc. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, 2015.

[41] Q. Wei, J. M. Bioucas-Dias, N. Dobigeon, J. Tourneret, M. Chen, and S. Godsill, "Multiband image fusion based on spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7236–7249, 2016.

[42] R. O. Green and M. L. Eastwood, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, 1998.

[43] J. A. Gualtieri, S. R. Chettri, R. F. Cromp, and L. F. Johnson, "Support vector machine classifiers as applied to AVIRIS data," in *Proc. JPL Airborne Geosci. Workshop*, 1999, pp. 217–227.

[44] B. Kunkel, F. Blechinger, R. Lutz, R.Doerffer, H. van der Piepen, and M. Schroder, "ROSIS (reflective optics system imaging spectrometer)—a candidate instrument for polar platform missions," in *Proc. SPIE*, vol. 868, 1988, p. 134–142.

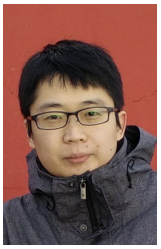
[45] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band AVIRIS hyperspectral image data set: June 12, 1992 Indian Pine test site 3," in *Purdue University Research Repository*, 2015.

[46] F. Zhu, Y. Wang, B. Fan, S. Xiang, G. Meng, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5412–5427, 2014.

[47] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms—Part I: Lemmas for partitioned matrices," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1022–1032, 2008.

[48] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, 2014.

Meng Ding received the B.S. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Mathematical Sciences, UESTC. From 2019 to 2020, he was an exchange Ph.D. Student at the Oregon State University, Corvallis, Oregon, United States, supported by the China Scholarship Council. His current research interests include image processing, tensor analysis, and machine learning.



Xiao Fu is an Assistant Professor in the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, United States. He received his Ph.D. degree in Electronic Engineering from The Chinese University of Hong Kong (CUHK), Hong Kong, in 2014. He was a Postdoctoral Associate in the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, United States, from 2014 to 2017. His research interests include the broad area of signal processing and machine learning.

Ting-Zhu Huang received the B.S., M.S., and Ph.D. degrees in computational mathematics from the Department of Mathematics, Xi'an Jiaotong University, Xi'an, China. He is currently a Professor with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include scientific computation and applications, numerical algorithms for image processing, numerical linear algebra, preconditioning technologies, and matrix analysis with applications.

Jun Wang received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in communication and information systems from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1997, 2000, and 2009, respectively. Since 2000, he has been with the National Key Laboratory of Science and Technology on Communications, UESTC, where he is currently a Professor. His research interests include signal processing for wireless communications, cognitive radio, machine learning for wireless communications.

Xi-Le Zhao received the M.S. and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2009 and 2012, respectively. He is currently a Professor with the School of Mathematical Sciences, UESTC. His current research interests include image processing, computer vision, and machine learning.

Supplementary Materials for "Hyperspectral Super-Resolution via Interpretable Block-Term Tensor Modeling"

M. Ding, X. Fu, T.-Z. Huang, J. Wang, and X.-L. Zhao

APPENDIX D
PROOF OF PROPOSITION 1

With the derivations in Appendix C, it is readily seen that the largest eigenvalue of the Hessian of $\mathcal{J}_1(\boldsymbol{S}^{(t)}, \boldsymbol{C}^{(t)})$ w.r.t. $\boldsymbol{C}^{(t)}$ is upper boounded by the following:

$$\big(\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)^\top(\boldsymbol{P}_2\otimes\boldsymbol{P}_1)\boldsymbol{S}^{(t)}\big) + \sigma_{\max}(\boldsymbol{P}_M^\top\boldsymbol{P}_M)\times\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)}\big)+\lambda\big),$$

since the subproblem is a quadratic program. Hence, the above is a legitimate Lipschitz constant $L_{\boldsymbol{C}}^{(t)}$ of the corresponding gradient.

Similarly, a Lipschitz constant $L_{\boldsymbol{S}}^{(t)}$ of $\nabla_{\boldsymbol{S}}\mathcal{J}_1(\boldsymbol{S}^{(t)}, \boldsymbol{C}^{(t+1)})$ can be derived—which is as specified in Proposition 1.

With these two constants being bounded $L_{\boldsymbol{C}}^{(t)} < \infty$ and $L_{\boldsymbol{S}}^{(t)} < \infty$, it can be seen that the alternating gradient projection algorithm in Algorithm 1 falls into the category of inexact block coordinate descent with *sufficient decrease* guarantees in each iteration [37]—if the step sizes are chosen following the rules stated in the statement of Proposition 1. Then, invoking the convergence theory (i.e., Theorem 2.8 [37]) there, one can easily show that every limit point of solution sequence produced by the proposed algorithm is a stationary point of the optimization problem of interest.

APPENDIX E
GRADIENTS AND LIPSCHITZ CONSTANTS IN ALGORITHM 2

When the spatial degradation operators are unknown, similar gradient calculations as in Appendices C-D can be applied. One can show the following:

$$\begin{aligned}
&\nabla_{\boldsymbol{C}}\mathcal{J}_2(\boldsymbol{S}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t)})\\
&\quad= \boldsymbol{C}^{(t)}(\widetilde{\boldsymbol{S}}^{(t)})^\top\widetilde{\boldsymbol{S}}^{(t)}\\
&\quad+ \boldsymbol{P}_M^\top\boldsymbol{P}_M\boldsymbol{C}^{(t)}(\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)} + \lambda\boldsymbol{C}^{(t)}\\
&\quad- (\boldsymbol{Y}_H^{(3)})^\top\widetilde{\boldsymbol{S}}^{(t)} + \boldsymbol{P}_M^\top(\boldsymbol{Y}_M^{(3)})^\top\boldsymbol{S}^{(t)},\\
&\nabla_{\boldsymbol{S}}\mathcal{J}_2(\boldsymbol{S}^{(t)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)})\\
&\quad= \big(\boldsymbol{S}^{(t)}(\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top - \boldsymbol{Y}_M^{(3)}\big)\boldsymbol{P}_M\boldsymbol{C}^{(t+1)}\\
&\quad+ p[\eta_1\text{vec}(\boldsymbol{W}_1^{(t)}\boldsymbol{S}_1^{(t)}), \ldots, \eta_R\text{vec}(\boldsymbol{W}_R^{(t)}\boldsymbol{S}_R^{(t)})]\\
&\quad+ q[\theta_1\boldsymbol{H}_x^\top\boldsymbol{U}_1^{(t)}\boldsymbol{H}_x\boldsymbol{q}_1^{(t)}, \ldots, \theta_R\boldsymbol{H}_x^\top\boldsymbol{U}_R^{(t)}\boldsymbol{H}_x\boldsymbol{q}_R^{(t)}]\\
&\quad+ q[\theta_1\boldsymbol{H}_y^\top\boldsymbol{V}_1^{(t)}\boldsymbol{H}_y\boldsymbol{q}_1^{(t)}, \ldots, \theta_R\boldsymbol{H}_y^\top\boldsymbol{V}_R^{(t)}\boldsymbol{H}_y\boldsymbol{q}_R^{(t)}],\\
&\nabla_{\widetilde{\boldsymbol{S}}_r}\mathcal{J}_2(\boldsymbol{S}^{(t+1)}, \widetilde{\boldsymbol{S}}^{(t)}, \boldsymbol{C}^{(t+1)})\\
&\quad= \big(\widetilde{\boldsymbol{S}}^{(t)}(\boldsymbol{C}^{(t+1)})^\top - \boldsymbol{Y}_H^{(3)}\big)\boldsymbol{C}^{(t+1)}\\
&\quad+ p[\eta_1\text{vec}(\widetilde{\boldsymbol{W}}_1^{(t)}\widetilde{\boldsymbol{S}}_1^{(t)}), \ldots, \eta_R\text{vec}(\widetilde{\boldsymbol{W}}_R^{(t)}\widetilde{\boldsymbol{S}}_R^{(t)})].
\end{aligned}$$

In addition, we have

$$\begin{aligned}
L_{\boldsymbol{C}}^{(t)} &= \sigma_{\max}(\boldsymbol{P}_M^\top\boldsymbol{P}_M)\times\sigma_{\max}\big((\boldsymbol{S}^{(t)})^\top\boldsymbol{S}^{(t)}\big)\\
&\quad+ \sigma_{\max}\big((\widetilde{\boldsymbol{S}}^{(t)})^\top\widetilde{\boldsymbol{S}}^{(t)}\big) + \lambda,\\
L_{\boldsymbol{S}}^{(t)} &= \sigma_{\max}((\boldsymbol{C}^{(t+1)})^\top\boldsymbol{P}_M^\top\boldsymbol{P}_M\boldsymbol{C}^{(t+1)})\\
&\quad+ p\max_{r=1,\ldots,R}\eta_r\sigma_{\max}(\boldsymbol{W}_r^{(t)})\\
&\quad+ q\max_{r=1,\ldots,R}\theta_r\sigma_{\max}(\boldsymbol{H}_x^\top\boldsymbol{U}_r^{(t)}\boldsymbol{H}_x)\\
&\quad+ q\max_{r=1,\ldots,R}\theta_r\sigma_{\max}(\boldsymbol{H}_y^\top\boldsymbol{V}_r^{(t)}\boldsymbol{H}_y),\\
L_{\widetilde{\boldsymbol{S}}}^{(t)} &= \sigma_{\max}((\boldsymbol{C}^{(t+1)})^\top\boldsymbol{C}^{(t+1)})\\
&\quad+ p\max_{r=1,\ldots,R}\eta_r\sigma_{\max}(\widetilde{\boldsymbol{W}}_r^{(t)}),
\end{aligned}$$

where $\widetilde{\boldsymbol{W}}_r^{(t)} = (\widetilde{\boldsymbol{S}}_r^{(t)}(\widetilde{\boldsymbol{S}}_r^{(t)})^\top + \tau\boldsymbol{I})^{\frac{p-2}{2}}$, $r = 1, \ldots, R$.

APPENDIX F
BRIEF DESCRIPTION OF THE BASELINE ALGORITHMS

We briefly describe the baseline algorithms including CNMF [3], HySure [4], FUSE [5], MFpoly [7], SCOTT [10], and STEREO [9].

The methods in [3]–[5], [7], i.e., CNMF, HySure, FUSE, and MFpoly, all perform low-rank matrix estimation techniques using the matricized HSI and MSI in order to estimate the SRI. In particular, MFpoly [7] uses a two-stage matrix estimation-based approach that admits recoverability guarantees and a polynomial time algorithm for the HSR task under some stringent conditions, e.g., when every pixel only contains a small number of endmembers and and the so-called "pure pixels" (see definition in [48]) exist. The STEREO algorithm [9] models the spectral images as tensors with low-rank CPD representations. Utilizing the algebraic properties of CPD, recoverability of the SRI is shown to hold under mild conditions. Later on, the SCOTT algorithm [10] adopts a similar idea but used the Tucker tensor model, which facilitates a fast algorithm leveraging on the higher-order SVD.