Subspace-Based Preprocessing Module for Fast Hyperspectral Endmember Selection

Xiangfei Shen, Wenxing Bao, and Kewen Qu

Abstract—Endmember extraction algorithms (EEAs) play a crucial role in hyperspectral image (HSI) perception, and yet they normally suffer from three flaws: 1) High computational burden, 2) weak noise robustness, and 3) high outlier sensitivity. To solve these problems, this article proposes a fast subspace-based preprocessing module, called fast subspace-based preprocessing module (FSPM), to select a high-quality data subset for subsequent endmember extraction. Specifically, FSPM first transforms an HSI into a low-dimensional data subspace using singular value decomposition. For each component pair, FSPM then detects its convex hull vertices and proposes a local outlier score measurement to remove potential outliers. FSPM finally transforms determined data points into noise-reduced data space for the sake of identifying endmembers. The proposed FSPM sheds new light on the current preprocessing field, which can fast reduce noises and outliers as well as remove redundant data points. Based on various validation metrics, experiments conducted on both synthetic and real HSIs indicate that the proposed FSPM is superior to current state-of-the-art preprocessing techniques.

Index Terms—Convex hull, endmember extraction, fast preprocessing, hyperspectral image (HSI), outlier detection.

I. INTRODUCTION

In hyperspectral images (HSIs), mixed pixels are a mixture of several distinct substances known as endmembers, weighted by their corresponding fractional abundances, which is primarily due to the limited spatial resolution of a sensor or mixed homogeneous area of distinct materials [1], [2]. To exploit mixed pixels, the process of decomposing a mixed pixel into a collection of endmembers and a set of abundances is called hyperspectral unmixing (HU), involving endmember extraction and abundance estimation procedures [3], [4].

However, when the light involves multiple scattering phenomena, endmembers are mixed based on a nonlinear mixture model (NLMM), otherwise, follow a linear mixture model (LMM) which provides clear physical meanings. Based on the LMM assumption, HU algorithms nowadays are explored mainly from three perspectives: 1) Statistical-based unmixing strategy, 2) spectral library-based sparse unmixing, and 3) fully constrained least-square-based endmember extraction.

Statistical-based algorithms recently received plenty of attention partly because those algorithms can simultaneously estimate endmembers and abundances, but principally because of their Bayesian paradigm which can formulate HU issues using explicitly mathematical and physical meanings. Classic statistical frameworks concentrate on the nonnegative matrix factorization (NMF) [5] which blindly decomposes an HSI as the product of endmember and abundance matrices by alternatively optimizing two matrices. By imposing various constraints on the basic NMF, lots of NMF-based variants can capture desirable unmixing performances such as minimum simplex volume constraint [2], ℓ1/2 sparsity regularizer [6], graph constraint [7], collaborative sparsity (ℓ1,2) regularizer [8], etc.

Comparing with statistical frameworks, sparse unmixing algorithms hinge on an existing spectral library instead of exploiting endmembers, meaning that they focus solely on estimating an abundance matrix. Several intrinsic abundance properties are generally explored including total variation [9], collaborative sparsity [10], nonlocal information [11], spatial discontinuity information [12], etc. Nevertheless, if we apply a sparse unmixing algorithm to obtain abundances, a spectral library is demanded, which is not consistently assured.

For the fully constrained least-square-based endmember extraction, it concerns two independent parts that first extract endmembers and then apply a fully constrained least-square algorithm [13] to estimate abundances. Furthermore, endmember extraction algorithms (EEAs) normally identify desired endmembers from whole hyperspectral data space based on four research lines. First, extreme subspace projection-based EEAs regard pixels with extreme projections as endmembers. These include pixel purity index (PPI) [14], orthogonal subspace projection (OSP) [15], and vertex component analysis (VCA) [16]. Second, maximum simplex volume-based EEAs identify endmembers by specifying a maximum volume among all simplexes. These include N-FINDR [17], simplex growing algorithm (SGA) [18], and alternative volume maximization (AVMAX) [19]. Third, minimum simplex volume-based EEAs fit a minimum simplex to enclose entire data points using a set of generated endmembers. These include minimum volume enclosing simplex (MVES) [20] and minimum volume simplex analysis (MVSA) [21]. Fourth, spatial-spectral information-based EEAs investigate a process of utilizing spatial-spectral attributes to explore endmembers. These include automatic morphological endmember extraction (AMEE) [22], spatial-spectral-based...
endmember extraction (SSEE) [23], spatially weighted simplex strategy (SWSS) [24], and spatial energy-constrained maximum simplex volume (SENMAV) [25].

Despite the EEAs that determine endmembers from the entire data in an unsupervised manner, recent attempts equally highlight the field of preprocessing algorithm (PPA) that normally searches a combination of spatial-spectral information to select a data subset for subsequent endmember extraction task. Representative modules include spatial preprocessing (SPP) [26], spatial-spectral preprocessing (SSPP) [27], spatial-spectral preprocessing module (SPMM) [28], regional clustering-based spatial preprocessing (RCSSP) [29], etc. More specifically, those PPA follow similar paradigms that contain two strategies: 1) Clustering- or superpixel-based techniques, including k-means [28], iterative self-organizing data analysis (ISO- DATA) [27], fuzzy C-means [30], and simple linear iteration clustering (SLIC) [29], are used to exploit spatial information by specifying spatially local homogeneity of each target pixel owing to an underlying fact that the potential endmembers generally originate from spatially homogeneous areas, and 2) a set of projection bases normally generated by principal component analysis (PCA) are considered to identify pixels’ spectral purity by identifying their projections onto the bases [27]–[29]. Those pixels who maintain extreme projections are more likely to possess high spectral purity than other counterparts.

These strategies play an important role in the existing PPAs which have captured beneficial performances to some extent. But in the meantime, the differences bound to the spatial-spectral combination, to spatial homogeneity exploitation, to computational cost, to outliers, and noises. First, they rely grossly on the process of exploiting spatial and spectral information from two independent steps, in other words, they separately extract valuable information from rich spectral bands. Then, they deeply link the determination of spatial information with homogeneous areas. Next, they suffer heavily from high execution time, indicating that the joint action of processing the image employing the SPP or the SSPP algorithms and then extracting the endmembers from the resulting image takes more time than directly applying the VCA algorithm to the original nonpreprocessed image [31]. To reduce the computational requirements of SPP and SSPP, two parallel computing-based extensions [32], [33] were recently proposed which to some extent show a potential research line. More importantly, some of them neglect the impact of outliers prior to endmember extraction. Normally, outliers are special patterns in data without a well-defined notion of normal behavior, and their signatures are spectrally distinct from their surroundings or background representation [25], [34], [35]. In this regard, we regard the generation process of outliers as two types: Geographically anomalous type and functionally anomalous type. The first is generally seen as the rare materials that maintain negligible proportions in the observed HSI. The second is more likely to be generated during detector failure, data transfer, and improper data correction [36]. It is worth noting that once there exist potential outliers, they are more likely to be selected as endmembers by most spectral-based EEAs because they deeply deviate from the background data, and they may force a vertex of the simplex to reside at a point beyond the nominal position of the endmember in order to enclose every point [1]. Compared to outliers, noisy pixels are normal spectral signatures corrupted by Gaussian noise or even sparse noise, which can be recovered from noises based on a specific denoising model. One more thing to mention, the existing PPAs emphasize providing endmember candidates but barely considering noises that extensively affect endmember extraction accuracy in poor noise scenarios. For SPP and SSPP, pixel reconstruction and multiscale Gaussian filtering processes are, respectively, adopted to alleviate noises, but the noise reduction is not sufficient.

Moreover, most PPAs that utilize spatial information depend heavily on two facts. The first is that HSIs intrinsically contain spatial and spectral attributes, where spatial contextual can characterize pixels’ spatial correlations. Those pixels who are spatially close to each other have higher correlations than those distant pixels. Under this fact, SPP [26] holds a belief that endmembers are more likely to be found in homogeneous areas rather than in transition or heterogeneous areas, which guides most subsequent PPAs’ algorithmic structures. The second is that spatial information provides a crucial tool for PPAs to detect and remove outliers that inevitably exist in the dataset, principally owing to that outliers generally deviate from normal data behaviors. By applying clustering methods, homogeneous areas are filled with consistent categories, while outliers’ labels distinctly differ from those of their neighborhoods. Compared with the process of specifying spectral information, however, the process of spatial exploitation is time-consuming since it generally involves traversing entire pixels as well as their nearby pixels. Therefore, we investigate potential research lines involving how to elegantly extract spatial-spectral information and supply a noise and outlier removal module. The former one highly affects PPAs’ computational cost, and the latter one can effectively promote the quality of endmember candidates.

We set out to test the hypothesis that traditional spatial-spectral-based data preprocessing is reflected in subspace-based data exploitation processes that have less computational requirements and weak noise/outlier interference. The aim of the proposed fast subspace-based preprocessing module (FSPM) pays closer attention to the following two facts: 1) Homogeneous pixels equally locate in the simplex, but outliers deviate from the simplex. Therefore, the proposed FSPM avoids traversing pixels from the HSI but only needs to remove potential outliers beyond the simplex, 2) the methods that regard projections as pixels’ spectral purity can be effectively replaced by finding vertices or boundary points from the simplex, which may require less computational burden and produce more beneficial pixels. In this regard, FSPM does not investigate spatial-spectral information of the HSI but rather reduces the HSI into a low-dimensional subspace by performing singular value decomposition (SVD). The exploitation of spatial and spectral information in FSPM can be, respectively, regarded as the process of masking outliers which goes beyond the convex hull and specifying convex hull vertices. By iteratively identifying desired data points from a subspace, FSPM then transforms determined data points into noise-reduced data space. Fig. 1 illustrates the preprocessing
procedures of FSPM. Based on different validation metrics, experiments conducted on synthetic and real HSIs indicate that FSPM not only requires less computational requirements but also provides a more preferable endmember candidate set than other PPAs.

We make four chief contributions.

1) This article introduces a preprocessing module for fast selecting a data subset. Instead of using traditional strategies that narrowly search for a combination of spatial-spectral information, the proposed FSPM treats preprocessing issues as a problem of subspace-based data subset selection.

2) This article unifies two core tasks of data preprocessing, including fast spatial-spectral information exploitation and noise reduction. The former one highly depends on iteratively specifying convex hull vertices from affine space and removing outliers far away from the convex hull, and the latter one reduces noises using SVD. Under these conditions, the proposed FSPM can simultaneously obtain a desired endmember subset with negligible processing time.

3) To remove outliers, FSPM develops a new outlier removal method by considering the neighbor sparsity of convex hull vertices especially vertically and horizontally extreme points.

4) Besides the current metrics such as spectral angle distance (SAD) and root-mean-square error (RMSE), this article proposes a new metric to validate PPAs’ performance called preprocessing performance measure (PPM), involving evaluating endmember accuracy and computational cost of the PPA which coupled with EEAs.

This article is divided into five sections, which can be given as follows. Section II reviews several important issues regarding EEAs and PPAs. The FSPM is introduced in Section III in detail. Section IV displays experimental results with regard to the FSPM and other comparison algorithms. Section V concludes this article.

II. BACKGROUND

A. LMM

LMM considers that light follows a linear combination of different materials (or endmembers). Suppose \( Y = [y_1, y_2, \ldots, y_n]_{L \times N} \) is an HSI with \( L \) bands and \( N \) total pixels. It can be formulated by considering endmembers, abundance, and noise matrices

\[
Y = MA + W
\]

where \( M \in \mathbb{R}^{L \times p} \), \( A \in \mathbb{R}^{p \times N} \), and \( W \in \mathbb{R}^{L \times N} \) are endmember, abundance, and additive noise matrices, respectively, and \( p \) is the number of endmembers, which can be estimated by virtual dimensionality (VD) [37]. Besides, abundance nonnegative and sum-to-one constraints are adopted, such that \( A \succeq 0 \) and \( 1^T_p A = 1^T_N \), where \( 1 \) denotes the vector of all ones.

B. EEA

1) Simplex Projection-Based EEA: Simplex projection-based EEAs such as PPI [14], OSP [15], and VCA [16] normally investigate the extreme projections on determined projection bases. It is worth mentioning that there are some differences among such EEAs. PPI demands to project entire pixel signatures onto loads of randomly generated vectors, a.k.a. skewers. OSP considers projecting hyperspectral data onto an orthogonal subspace spanned by already determined endmembers. VCA performs an affine transformation and finds potential endmembers by projecting entire subspace data onto skewers. Although the endmembers extracted by VCA are inconsistent if we perform multiple independent runs, they can avoid noise interference.

2) Simplex Volume-Based EEA: Simplex volume-based EEAs can be further divided into two categories. The first one inflates a maximum matrix determinant by successively replacing
pixels so as to reach the largest inner simplex volume. Representative methods include NFINDR [17], SGA [19], AVMAX [19], and fast gram determinant-based algorithm (FGDA) [38]. The second one shrinks a minimum exterior simplex volume by iteratively updating the objection function. Classic methods include simplex identification via split augmented Lagrangian (SISAL) [39], MVES [20], and MVSA [21]. The biggest difference behind these two strategies is that the former one selects endmembers from the entire data space and it can obtain the desired endmember extraction performance especially for those HSIs who contain pure pixels, while the latter one has the ability on handling both pure and nonpure assumption-based scenarios.

3) NMF: NMF-based HU algorithms involve jointly estimating endmembers and abundances from the HSI. Based on the previous notations, NMF can be expressed as

\[
Y \approx MA
\]

where \( M \succeq 0 \) and \( A \succ 0 \). In order to measure the difference between the original signal \( Y \) and reconstructed signal \( MA \), its cost function is considered as

\[
\min_{M, A} \frac{1}{2} \| Y - MA \|_F^2
\]

where \( \| \cdot \|_F \) denotes the Frobenius norm, and the cost function also subjects to \( I_p^T A = I_n \). By considering endmember-oriented constraints, NMF-based variants can capture better unmixing performance even though there exists noises or nonpure pixels, given as

\[
\min_{M, A} \frac{1}{2} \| Y - MA \|_F^2 + \phi(M)
\]

where \( \phi(M) \) denotes an endmember regularizer. Representative regularizer include minimum distance [40], minimum simplex volume [2], [8], endmember dissimilarity [41], and minimum dispersion [42].

Fig. 2 provides visual descriptions regarding four endmember extraction strategies. Extreme projection and volume-based EEAs regard vertices of the simplex as endmembers. Minimum volume-based EEA shrinks a minimum exterior volume which generates optimized endmembers. NMF not only can produce endmembers but also has the ability on estimating abundances. Each arrow equally represents that the current EEA can be applied to HSI for endmember extraction purposes.

C. PPA

1) Pixel Reconstruction-Based PPA: Pixel reconstruction-based PPAs highlight the process of assessing spatial weights between target pixels and their surroundings and reconstructing target pixels using assessed weights. The basic formula can be simply expressed as

\[
y = w^T N \text{ s.t. } w \succeq 0
\]

where \( w \) and \( N \) denote a weight vector and neighbor pixels, respectively. In SPP [26], it measures weights between each pixel and its adjacent pixels based on SAD calculations. In [43], it uses the local linear embedding (LLE) to generate spatially correlated weights for SPP. Additionally, Zhang et al. [44] introduces a local sparse representation model to exploit the sparsity of the weight vector. Nevertheless, the abovementioned PPAs demand to process entire pixels without discarding redundant pixels, which barely reduce the computational cost of subsequent EEAs as well as themselves. Another inevitable shortcoming is that those PPAs can improve endmember accuracy, especially for poor noise scenarios, owing to pixel’s reconstruction process. When there are relatively light noises such as 40 dB, however, they may, in turn, affect endmember accuracy primarily because those pixels who have high spectral purity are reconstructed by their 5 x 5 or even 7 x 7 neighborhoods, so as to degrade their spectral purity.

2) Pixel Removal-Based PPA: Compared to pixel reconstruction-based PPAs, pixel removal-based PPAs adopt a strategy that removes plenty of redundant pixels that normally situate in heterogeneous areas or maintain relatively low spectral purity. By setting a specific retaining ratio, such as 20% of total pixels, the rest of the pixels are removed, meaning that the subsequent endmember extraction process will be faster. Those PPAs who follow this strategy are the dominant way to preprocess hyperspectral data in the past decade, including SSPP [27], SSPM [28], RCSPP [29], spatial edges and spectral extremes preprocessing (SE²PP) [31], region-based SPP [45], geodesic and Euclidean distance-based preprocessing (GEPP) [46], and spatial energy and spectral purity-based preprocessing (SESPP) [47]. However, such PPAs are concerned with traditional spatial exploitation processes such as k-means and SLIC, which are time-consuming. They also simultaneously resort to projecting entire pixels onto a set of bases to specify pixels’ spectral purity. By fusing spatial and spectral information, they design a specific joint type to select desired endmember candidates. More importantly, in order to produce desired preprocessing performances, they involve several cut-off parameters such as the size of the sliding kernel...
In order to avoid time-consuming steps, FSPM uses SVD to reduce hyperspectral data into $p$-dimensional subspace. Specifically, it first captures a sample correlation matrix by calculating $YY^T/N$ and extracts the left eigenvector matrix $U$ from that matrix. Then, it projects the original data $Y$ into $U$ to generate a low-dimensional data subspace $X = U^TY$, which measures intrinsically principal information of hyperspectral data. In other to avoid displacements between the original data and projected ones [21], we utilize the projective projection by introducing a scaling factor $u = \frac{1}{p}\sum_{i=1}^{p} U_{i,i}$, such that $x^T u > 0$ (see [3], [16], and [21] for more details). This step provides two obvious advantages: 1) It converts $L$-dimensional data into $p$-dimensional projected data which promotes computational efficiency, and 2) it masks a large percentage of noises so that the convex hull vertices are more likely to be accurately found.

It is worth mentioning that there are plenty of methods to generate the subspace such as randomly generated orthonormal bases [48], principal component analysis [16], or NMF. There are two chief reasons why we adopt SVD to generate the subspace instead of other alternatives. The first one is that SVD
For any positive integer $p$, are the $D$-s with radius fixed in two and three dimensions, with $p \neq 1$. By applying the above description shows that the possible combinations, where $k$ points that have the $\delta$-log($D$) matrix to be low-rank.

2) Detection of Convex Hull Vertices: By applying the abovementioned data transformation, FSPM emphasizes the process of finding convex hull vertices because endmembers are normally specified from vertices if there exist pure pixels or formed by the boundary pixels if the pure-pixel assumption cannot hold. With the increase of data dimensions, however, the computational cost of convex hull algorithms such as Graham scan [49] and Jarvis march [50] go up exponentially. Even though $p$-dimensional data has achieved a low-dimensional space with each dimension representing a primary component, it is still impossible to detect convex hull vertices using entire components. In [48], Heylen and Scheuners restrict to the values $D \in \{1, 2, 3\}$, where $D$ denotes the data dimension, because it is acceptable for endmember extraction not only for its light time complexity but also due to less memory requirements.

In this article, we restrict the detection process of convex hull vertices to 2D and select 2D subspace combinations by $F = \{0 < i \neq j \leq p|X_{[i,j,:]}\}$. It is worth mentioning that there exist $C(p, 2)$ possible combinations, where $C(p, 2) = \frac{p!}{2!(p-2)!} = \frac{p^2}{2}(p-2)$. Besides, we adopt quick hull (Qhull) [51] to find vertices of each 2D subspace owing to its reasonable complexity $O(N \log(h))$ in two and three dimensions, with $h$ as the number of vertices spanning the convex hull. Fig. 5 visually displays the detection process of convex hull vertices. Qhull first starts by identifying the leftmost and rightmost points and then finds the convex hulls of the points on two sides. By using the recursive strategy, all vertices can be quickly obtained.

3) Outliers Removal: The above description shows that the subspace-based view is meaningful and adequate to quickly select a high-quality data subset for the sake of identifying endmembers, but not satisfactory for the general case when potential outliers exist. Traditionally, PPAs such as SPP and SSPM concern a spatial exploitation process that can remove outliers to mitigate error propagation of subsequent endmember extraction. These outliers govern abnormal spectral behavior which tends to be regarded as endmembers in the spectral domain but is liable to be excluded by considering their spatial neighbors in the spatial domain. However, the process of spatial information exploration is time-consuming. In this regard, FSPM discards this process but takes an outlier detector to remove outliers from the data simplex. The outlier detector relies on the fact that homogeneous pixels normally exist in the simplex, while endmembers and outliers potentially situate at its vertices. Fig. 4 visually describes the fact that homogeneous pixels are interior data located in the simplex and outliers distinctly differ from normal data representation.

The process of outlier detection hinges grossly on three points.

1) Outliers are seen as abnormal points beyond the simplex, indicating that they normally vertically or horizontally contain extreme values.

2) The neighborhood of outliers is extremely sparser than that of normal data.

3) The local density of normal data is significantly close to that of their neighborhood, while outliers are quite on the contrary.

It is worth mentioning that, for point 2), the neighborhood sparsity relies on the fact that a normal neighborhood set contains spectrally similar signatures and spatially consistent patterns while an abnormal neighborhood set maintains some spectrally distinct signatures. In this regard, when a dataset that contains outliers is projected to a subspace, the data points in the subspace will show the fact that the neighborhood of outliers is sparse while the neighborhood of normal data points is still correlated to each other. In other words, outliers that existed in high-dimensional data can be clearly observed in low-dimensional intrinsic signal space in which the neighborhood of outliers is sparse. Therefore, FSPM pays more attention to four vertically and horizontally extreme points, because endmembers and outliers are more likely to have extreme coordinate values in the convex hull. For the four extreme points, their k-nearest neighbors (KNN) are specified to calculate local density. To reduce the computational cost of neighborhood searching, we set a radius-fixed searching region to define points’ KNN.

Definition 1: For any positive integer $k$, the k-nearest neighbors of a point $x$ within a searching region $D$ with radius fixed at $\delta$, such that:

1) if there are $s(s \leq k)$ data points within $D$, the number of nearest neighbors of $x$ is $s$, otherwise;

2) the nearest neighbors of $x$ are the $k$ points that have the closest Euclidean distances to $x$. 

Fig. 5. Process of finding convex hull vertices. Qhull first identifies leftmost and rightmost points (a) and then finds the vertex on two parts (c). By recursively determining vertices, Qhull can extract entire vertices of the convex hull. Red and blue points denote the detected vertices and interior data.
For each extreme point, FSPM calculates its local density concerning Euclidean distances between its neighbors and itself. For each component, i.e., 1-D subspace $X_i$, however, its data mean and standard deviation explicitly differs from other counterparts, implying that directly specifying neighborhoods by Euclidean distance on the original component pair tends to be affected by different data variance. In this regard, each component is standardized by fixing the data mean and standard deviation as 0 and 1, respectively, given by

$$X_{i,:} \leftarrow \frac{X_{i,:} - \mu}{\sigma}$$

where $\mu$ and $\sigma$, respectively, represent the data mean and standard deviation. It is worth noting that when there are only $s (s \leq k)$ neighbors in a fixed searching region $D$, we consider the relation between $k$ and $s$ and use it to weight target pixels’ local density. Specifically, if $k = s$, the target point’s local density is weighted by $k/s = 1$, otherwise its weight is less than 1, which can weaken outlier’s local density.

**Definition 2**: Given a KNN set $N = \{z_i\}_{i=1}^s$ of $x$, its local density (LoD) is defined as:

$$\text{LoD}(x) = \frac{s}{\sum_{i=1}^{s} k \|x - z_i\|_2^2}. \quad (7)$$

We are also aware that the outlier’s LoD is different from that of its neighbors, triggering that its neighbors’ LoD is equally required to specify. However, when entire neighbors’ LoD are demanded to calculate, it is a waste of execution cost because there are lots of shared neighbors. Considering this fact, FSPM highlights the farthest neighbor lying in $D$ and calculates its LoD.

**Definition 3**: Given a KNN set $N = \{z_i\}_{i=1}^s$ of $x$, its farthest neighbor $\hat{z} \in N$ is defined as

$$\hat{z} = \arg \max_{z \in N} \|x - z\|_2. \quad (8)$$

By calculating the LoD of the target point and its farthest neighbor point, we can obtain the target point’s local outlier scores (LOS). The most crucial property behind LOS is that LOS explicitly closes to 1 or even larger than 1 when it involves high reconstruction errors during the data transformation process, which can quickly identify outliers. In this regard, a cut-off threshold can be considered to distinguish normal points from outliers.

**Definition 4**: The local outlier score (LOS) of $x$ is defined as

$$\text{LOS}(x) = \frac{\text{LoD}(x)}{\text{LoD}(\hat{z})}. \quad (9)$$

We are also aware that there exist a lot of outlier detectors that have captured desired detection performances in the field of data mining including statistical/probabilistic-based, linear model-based, density-based, and so on [52]. Statistical/probabilistic-based detectors such as the 3σ rule [53] and percentile-based box theory are the fast strategy to identify outliers, and yet they assume that the data points follow a Gaussian distribution, which is not always ensured. Linear model-based detectors such as principal component analysis assume that outliers can generate higher reconstruction errors during the data transformation process than those of normal data points. Density-based detectors such as local outlier factor (LOF) [54] and DBSCAN [55] identify outliers by considering the local density of entire data points. Our proposed outlier removal method, i.e., LOS, shows several clear advantages.

1) LOS only takes into account convex hull vertices especially vertically and horizontally extreme points rather than entire data points, which can quickly identify outliers.
2) LOS calculates the target point’s weighted local density from its specified s-NN.
3) LOS generates outlier scores using the target point and farthest neighbor’s LoD instead of calculating the LoD of each data point within the neighborhood.

**Algorithm 1: Pseudo-Code of FSPM.**

1: **Input:** HSI: $Y = \{y_1, y_2, \ldots, y_n\}_{1 \times N}$; the number of endmembers: $p$; pixel selection ratio: $\lambda$; the radius of searching region: $\delta$; the number of neighborhoods: $k$; and the separation threshold of outliers and normal data: $\eta$.
2: **Output:** Hyperspectral endmember subset: $R$.
3: Memory pre-allocation: $v = [], X = \text{zeros}(p, N), K = 0$.
4: $U = \text{svd}(YY^T/N)$.
5: $X = U^TY$.
6: $u = \frac{1}{p} \sum_{i=1}^{p} U_{i,:}$.
7: $\lambda_{i,j} = X_{i,:}/X_{i,:}^T u$; % Projective projection.
8: $F = \{0 < i \neq j \leq p | \lambda_{i,j}; \}$; % Determine all subspace combinations.
9: while $K < \lfloor \lambda \times N \rfloor$ do
10: for $i \rightarrow \frac{1}{2}p(p - 1)$ do
11: $Q = \text{standardization}(F_i)$;
12: $\phi = \text{convexhull}(Q)$;
13: $E = Q(:,\phi)$; % Extract vertically and horizontally extreme points %
14: for $j \rightarrow 4$ do
15: Specify $D$ using fixed radius $\delta$;
16: Specify $\text{LoD}(E(:,j))$ by considering its KNN in $D$;
17: Specify $\text{LoD}(\hat{z})$ by considering its KNN in corresponding $D$;
18: if $\text{LOS}(E(:,j)) < \eta$ then
19: Remove $E(:,j)$ and update $\phi$;
20: end if
21: end for
22: Update $v$ by $v = [v, \phi]$;
23: end for
24: if length($v$) > $\lfloor \lambda \times N \rfloor$ then
25: break;
26: else
27: Update $F$ by $F(:,\phi) = []$;
28: Update $K$ by $K = \text{length}(v)$;
29: end if
30: end while
31: Output data subset $R = UX(:, v)$;
et al. at 0.1, which can be defined as

\[ \delta + (10) \]

and MVC-NMF [2], their corresponding Max - k and Max - 1 = m were integrated to verify exist in created outliers by considering the idea of avoiding repetitive selection which not only can avoid a repetitive selection process but also gradually shrink the data scale [see Line 25]. For k, it determines the number of nearest neighbors and it is fixed at 5 in our subsequent experiments. In terms of \( \eta \), it is a cut-off threshold used to separate outliers and normal points. We are aware that this threshold is difficult to tune partly because of various anomalous types and mainly because of the potential sparsity of different datasets.

In [34], Xu et al. created outliers by considering the idea that the spectral reflectance of several consecutive bands is significantly higher than that of other bands. In [3] and [25], the authors created outliers based on LMM, defined as

\[ a = m_p \gamma + M_{p-1} \zeta_{p-1} + \omega. \tag{10} \]

In this formula, we allow that the abundance proportions of an endmember signature \( m_p \) is \( \gamma \in [1, 1, 2] \). For the rest of endmember signatures \( M_{p-1} = \{ m_i \}_{i=1}^{p-1} \), their corresponding abundance vector is \( \zeta_{p-1} \) subjecting to an important ASC constraint \( \gamma + 1^T \cdot \zeta_{p-1} = 1 \). In this outlier assumption, we use it to generate outliers for Experiment 2 and fix \( \eta \) at 0.1, which can accurately exclude outliers.

In Algorithm 1, we use a while-do loop to iteratively select convex hull vertices until we reach the desired number \( |\lambda \times N| \). It is worth mentioning that we remove already determined vertices from \( F \) which not only can avoid a repetitive selection process but also gradually shrink the data scale [see Line 25].

IV. RESULTS

This section details several experiments, which were conducted on synthetic and real HSIs. All the algorithms ran on a PC with an Intel Core i7-2600 K (at 3.4 GHz) and 16 GB RAM. Experimental results obtained from all the algorithms were averaged on 10 independent runs.

A. Experimental Preparations

1) Benchmark Methods: Four state-of-the-art PPAs, i.e., SPP [26], SSP [27], SSPP [28] (implemented by the authors), and RCSPP [29] (implemented by the authors), were compared to our proposed FSPM algorithm, and five representative EEs, i.e., OSP [15], N-FINDR [17], VCA [16], MVSA [21], and MVC-NMF [2] were integrated to verify experimental performances. A spatial-spectral-based EEA called SENMAV [25] is considered to supply experimental result comparisons. Besides, the parameters of four benchmark PPAs are carefully set according to their original publications [26]–[29].

In order to visually describe the intrinsic idea of LOS, Fig. 6 provides four key parts in detail. Noting that the process of removing potential outliers (i.e., extreme points in the vertical and horizontal axes) can be parallel to the extraction process of convex hull vertices, which is an underlying solution to further accelerate the preprocessing procedures of FSPM. Besides, Algorithm 1 provides detailed algorithmic structures to elaborate proposed FSPM.

C. Implementation Issues

This section provides several crucial implementation details regarding FSPM.

There are four parameters, including the ratio of pixel selection \( \lambda \), the radius of searching region \( \delta \), the number of neighborhoods \( k \), and the cut-off threshold \( \eta \). \( \lambda \) determines how many pixels will be extracted from the original hyperspectral data. According to the descriptions of pseudo-code of FSPM [see Algorithm 1], line 11 is the most time-consuming step that requires at least \( \frac{1}{2} p(p - 1) \) runs with each complexity \( O(N \log(h)) \). In this regard, if \( \lambda \) is too large, it obviously affects FSPM’s time performance. Besides, since FSPM pays more attention to convex hull vertices that are seen as potential endmembers, \( \lambda \) can be fixed at a relatively small ratio, such as 1% of total pixels. \( \delta \) specifies a neighborhood searching region that avoids finding KNN from the entire data. Normally, \( \delta \) can determine a circle searching region, but it involves a relatively complex searching procedure. In this article, we fix \( \delta \) according to vertically and horizontally extreme points because those points can determine a proximate data range, and they are also easy to be found according to points’ coordinates. Assume Max_v and Max_h, respectively, denote vertical and horizontal maxima, and Min_v and Min_h, respectively, denote vertical and horizontal minima. Here, we define \( \delta \) from two dimensions, i.e., \( \delta_v = (\text{Max}_v - \text{Min}_v)/10 \) and \( \delta_h = (\text{Max}_h - \text{Min}_h)/10 \). By applying \( \delta_v \) and \( \delta_h \), each target point can quickly define its rectangle searching region. Besides, if we applied a circle searching region, \( \delta \) would be fixed at \( \sqrt{\delta^2_v + \delta^2_h} \). For \( k \), it determines the number of nearest neighbors and it is fixed at 5 in our subsequent experiments.

In terms of \( \eta \), it is a cut-off threshold used to separate outliers and normal points. We are aware that this threshold is difficult to tune partly because of various anomalous types and mainly because of the potential sparsity of different datasets.

In [34], Xu et al. created outliers by considering the idea that the spectral reflectance of several consecutive bands is significantly higher than that of other bands. In [3] and [25], the authors created outliers based on LMM, defined as

\[ a = m_p \gamma + M_{p-1} \zeta_{p-1} + \omega. \tag{10} \]

In this formula, we allow that the abundance proportions of an endmember signature \( m_p \) is \( \gamma \in [1, 1, 2] \). For the rest of endmember signatures \( M_{p-1} = \{ m_i \}_{i=1}^{p-1} \), their corresponding abundance vector is \( \zeta_{p-1} \) subjecting to an important ASC constraint \( \gamma + 1^T \cdot \zeta_{p-1} = 1 \). In this outlier assumption, we use it to generate outliers for Experiment 2 and fix \( \eta \) at 0.1, which can accurately exclude outliers.

In Algorithm 1, we use a while-do loop to iteratively select convex hull vertices until we reach the desired number \( |\lambda \times N| \). It is worth mentioning that we remove already determined vertices from \( F \) which not only can avoid a repetitive selection process but also gradually shrink the data scale [see Line 25].
2) Evaluation Metrics:
1) SAD: SAD assesses the spectral similarity between two spectra, given by

\[ SAD(\alpha, \beta) = \arccos \left( \frac{\alpha^T \beta}{\| \alpha \| \| \beta \|} \right) \]  \hspace{1cm} (11)

where \( \alpha \) and \( \beta \) denote extracted endmember spectra and library, respectively. The higher the spectral similarity between \( \alpha \) and \( \beta \), the smaller the SAD.

2) RMSE: RMSE evaluates the reconstruction error between two images, and it is given by

\[ \text{RMSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sqrt{\frac{1}{L \times N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \]  \hspace{1cm} (12)

where \( \mathbf{Y} \) and \( \hat{\mathbf{Y}} \) represent original and estimated image, respectively. Both them have \( L \) bands and \( N \) total pixels. The lower the RMSE values, the better the reconstruction performance.

3) Speedup: Traditionally, researchers consider recording the computational time of PPAs and EEAs for the purpose of evaluating the performance of PPAs. Although most PPAs have relatively low computational requirements, the existing PPAs suffer from the fact that the joint action of processing the image employing PPA and then extracting the endmembers from the resulting image takes more than directly applying EEA. In other words, they barely produce desired acceleration performances for subsequent EEAs. In this context, we consider the speedup metric to verify the acceleration performance of PPA. Speedup measures computational cost ratio between EEAs without coupling PPAs and on itself with coupling, and it is defined as follows:

\[ \text{Speedup} = \frac{T_{\text{eea}}}{T_{\text{ppa}} + T_{\text{ppa-eea}}} \]  \hspace{1cm} (13)

where \( T_{\text{eea}} \), \( T_{\text{ppa}} \), and \( T_{\text{ppa-eea}} \) stand for EEA’s execution time on original hyperspectral data, PPA’s preprocessing cost, and EEA’s execution time on preprocessed data using PPA, respectively. If speedup results are greater than 1, it infers that PPA achieves the goal of accelerating endmember extraction of EEA.

4) PPM: In the previous publications with regard to PPAs, the authors separately record the endmember accuracy of PPA-EEAs and EEAs but fail to consider whether PPAs have the ability on improving the endmember accuracy. In this regard, we consider evaluating the endmember accuracy between PPA-EEAs and EEAs and propose a new preprocessing performance measurement (PPM) to evaluate the effectiveness of PPAs from time and endmember accuracy perspectives. The main motivation behind this metric is that the existing metrics such as SAD and RMSE in the field of hyperspectral unmixing barely provide the difference measure of the endmember extraction accuracy obtained from the preprocessed image and original non-preprocessed image. We use a log operation to measure the difference between SAD results obtained from EEA and SAD results obtained from the PPA-EEA combination.

We also unify the speedup results and SAD differences, defined as

\[ \text{PPM} = \text{Speedup} \times \log \frac{\text{SAD}_{\text{eea}}}{\text{SAD}_{\text{ppa-eea}}} \]  \hspace{1cm} (14)

The chief benefit of PPM is that it can directly provide performance comparisons among PPAs. Once the endmember extraction accuracy obtained from the PPA-EEA combination is higher than that of EEA, PPM will be positive and negative otherwise. An optimal or ideal PPA should have the ability to generate PPM values tending to positive infinity. It is worth mentioning that we can also use RMSE to measure PPM when SAD is not available [see Experiment 7].

3) Hyperspectral Datasets:

a) Synthetic dataset (DC1): DC1 is a representative synthetic image that can generate images according to different parameter settings [2]. Assume there is a desired image that contains \( n \times n \) pixels. In order to generate mixing pixels, the image is divided into plenty of equal blocks with each of them filling with one of the \( p \) spectral signatures and performing a \( w \times w \) spatial low-pass filter. For the purpose of validating preprocessing performances, the parameters were set and reported according to specific experiment requirements [see Fig. 7(a)]. Besides, endmembers are selected from the U.S. Geological Survey (USGS).\(^6\)

b) Indian Pines: Indian Pines was gathered by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in Northwest Indiana. This image contains 145 × 145 pixels, 224 bands, and 16 ground-truth classes [see Fig. 7(b)]. After removing the noise and water absorption bands (the excluded bands were 104–108, 150–163, 220), 200 bands were retained.

c) Cuprite: Cuprite is a real hyperspectral dataset, with data captured by the AVIRIS in Las Vegas, NV, USA. After removing the noise and water absorption bands (the excluded bands were 1–6, 105–115, 150–170, and 221–224) and capturing a desired region of interest, we use a 250×190×182 hyperspectral cube [see Fig. 7(c)]. In [16], 14 types of endmembers are identified for endmember extraction, which can be estimated via VD [37]. In [6], the authors believe that there exist 10 endmembers. In [7],

\(^6\)[Online]. Available: http://speclab.cr.usgs.gov/spectral-lib.htm
TABLE I

Experimental Settings Used in This Article Regarding Image Size, Low-Pass Filter Window, Endmember Number, Noise Level, and Algorithmic Parameter

| Experiments | Image size | Low-pass filter window | Endmember number | Noise level | Algorithmic parameter |
|-------------|------------|------------------------|------------------|-------------|-----------------------|
| Experiment 1 | 257×257×257 | 20                     | 3                | 30dB        | 1% 0.1                |
| Experiment 2 | 247×247×224 | 10                     | 4                | —           | 1% 0.1                |
| Experiment 3 | 247×247×224 | 10                     | 3                | 30dB        | 1% 0.1                |
| Experiment 4 #1 | 247×247×224 | 10                     | 3                | 30dB        | 1% 0.1                |
| Experiment 4 #2 | 257×257×224 | 20                     | 3                | 30dB        | 1% 0.1                |
| Experiment 5 | 257×257×224 | 20                     | 3                | 30dB        | 1% 0.1                |
| Experiment 6 | 145×145×200 | /                      | 16               | /           | 3% 0.1                |
| Experiment 8 | 250×190×182 | /                      | 12               | /           | 3% 0.1                |

a “—” denotes that the experimental setting is an experimental variable.
b “/” denotes that the experimental setting is not available.

Lu et al. hold that there is a subtle difference among some spectra of the same mineral with different chemical compositions, the estimated number of endmembers is reduced to 12 for unmixing. In this manuscript, 12 endmembers are adopted for subsequent unmixing tasks including Alunite, Andradite, Buddingtonite, Dumortierite, Kaolinite#1, Kaolinite#2, Muscovite, Montmorillonite, Nontronite, Pyrope, Sphene, and Chalcedony. The Cuprite and Indian Pines datasets were obtained from the publicly available website of the Computational Intelligence Group from the University of the Basque Country (UPV/EHU).

B. Experimental Results

1) Experiment 1 (Parameter Sensitivity Analysis): To provide visual descriptions of impacts of different $\lambda$ on experimental performances, this experiment was conducted on DC1 with 257×257 pixels, 224 bands, three endmembers, and 30-dB noises. The detailed parameter settings are tabulated in Table I. As can be seen from Fig. 8(a), when combined with FSPM, N-FINDR can provide robust and accurate endmember extraction performances but the endmember accuracy has weak relations to $\lambda$. When $\lambda$ goes up, the time consumption of FSPM also reaches a growth, but it is still lower than that of N-FINDR [see Fig. 8(b)]. Besides, the speedup results have a declining trend when $\lambda$ increases, mainly because small $\lambda$ can provide a few pixels for subsequent endmember identification so that N-FINDR can assure a very fast unmixing task [see Fig. 8(c)].

Fig. 9 visually displays the preprocessed data with different $\lambda$ from 0.1% to 2.5%. Compared to Fig. 9 (b), FSPM can provide desired high-quality preprocessed pixels that normally lie in the vertices and boundaries of the simplex.

2) Experiment 2 (Outlier Sensitivity Analysis): To assess the experimental performance of PPAs on outliers, this experiment considered two types of DC1. The first one contained three endmembers and three outliers while the second one did not maintain outliers. In order to generate outliers, we first randomly selected three normal pixels from DC1 and then applied (10) to recreate outliers according to the endmembers and true abundance fractions of three normal pixels. The three normal pixels were finally replaced by three outliers. In this context, we can generate a synthetic dataset containing outliers for this experiment.

Fig. 10(a) and (b) provide a dataset used in this experiment and outlier spectra, respectively. Table II displays SAD results obtained from N-FINDR and other PPA-EEA combinations. It can be seen from this table that when the dataset meets outliers, N-FINDR and RCSPP-NFINDR show different SAD results on three endmembers compared to those without outliers. Four PPA-EEA combinations including SPP-NFINDR, SSP-PNFINDR, SSPM-NFINDR, and FSPM-NFINDR, are robust to outliers, especially for FSPM which not only has the ability to remove outliers but rather provides the lowest and most consistent SAD results.

![Graph](image-url)
3) Experiment 3 (Visual Comparisons of Spectral Signatures Obtained From EEA and PPA-EEA Combinations): This experiment visually verified the spectral signatures extracted from different algorithms, where the low-pass filter window was fixed at 10 and noise levels varied from 10 to 40 dB and four endmembers were considered. As can be found in Fig. 11, compared to the endmember signatures obtained from four PPA-EEA combinations and NFINDR, FSPM-NFINDR has the ability on extracting endmember signatures that can accurately match USGS spectra. Also, under entire scenarios, FSPM-NFINDR shows minimum SAD values compared to other algorithms.

4) Experiment 4 (Visualization of Preprocessed Data): This experiment was conducted on DC1 that contains 237 × 237 pixels with three endmembers and 224 spectral bands. The noise level was fixed at 30 dB. To simulate real-world scenarios, we used two spatial low-pass filters to generate mixed pixels. The first one used 20 × 20 low-pass filter window, so that there are no pure pixels [see Fig. 12(a)] and the second used 10 × 10 low-pass filter window to retain pure pixels [see Fig. 13(a)]. As can be seen from Fig. 12(a)–(g), FSPM (green points) not only mitigates noises but also can provide a few high-quality data points that lie in the vertices and boundaries of the simplex. Compared to FSPM, SPP (orange points) and SSPP (pink points) to some extent can alleviate noises, but they barely produce desired high-quality data points. Likewise, in Fig. 13(a)–(g), FSPM equally finds a set of endmember candidates from noisy data. Although SPP removes a large percentage of noise owing to the process of pixel reconstruction, it also affects the vertices of a simplex which may have an impact on subsequent endmember accuracy. For SSPP, it can identify the position of pure pixels, but the preprocessed data are still affected by noise. In terms of RCSPP, it applies a superpixel algorithm to segment this dataset into a set of superpixels and retains high-quality pixels from
Fig. 11. Visual comparisons of four endmember signatures extracted from FSPM-NFINDR and other benchmark methods. SAD results obtained from different methods are displayed in the parentheses. (a) Endmember#1(10dB). (b) Endmember#2(10dB). (c) Endmember#3(10dB). (d) Endmember#4(10dB). (e) Endmember#1(20dB). (f) Endmember#2(20dB). (g) Endmember#3(20dB). (h) Endmember#4(20dB). (i) Endmember#1(30dB). (j) Endmember#2(30dB). (k) Endmember#3(30dB). (l) Endmember#4(30dB). (m) Endmember#1(40dB). (n) Endmember#2(40dB). (o) Endmember#3(40dB). (p) Endmember#4(40dB).

5) Experiment 5 (Impact of Image Size on PPAs): This experiment evaluated the impact of image size on PPAs regarding endmember accuracy, speedup, and time-consuming. Five DC1 with different image size were considered including $13 \times 13$, $45 \times 45$, $109 \times 109$, $237 \times 237$, $493 \times 493$. The noise level in each image was fixed at 30 dB, and three endmembers were randomly selected from the USGS library.

Fig. 14(a) displays mean SAD (mSAD) trends averaged on three endmembers. Compared to N-FINDR and other PPA-EEA combinations, when coupled with FSPM, N-FINDR produces higher endmember accuracy. Fig. 14(b) illustrates speedup results on different datasets. Because FSPM ensures a few endmember candidates and negligible time burden itself, it provides relatively larger acceleration performance than that of other PPAs on all the datasets when combined with N-FINDR. In terms of time-consuming, Fig. 14(c) tabulates time details regarding PPAs and N-FINDR. Compared with N-FINDR and other PPAs, FSPM demands little time requirements, while the computational burden of SPP exponentially goes up as the image size increases, followed by RCSPP which needs the second largest time consumption. Besides, after applying preprocessing stages, N-FINDR displays various time performances, where FSPM-NFINDR has the fewest computational time [see Fig. 14(d)].

6) Experiment 6 (Comparison Between EEAs and PPA-EEA Combinations Under Different Noise on DC1): This experiment provides overall results obtained from EEAs and PPA-EEA combinations on different noisy datasets. Four evaluation metrics were used to jointly verify PPAs’ preprocessing performances. As can be seen from Table III that when coupled with FSPM, N-FINDR and OSP can generate the best experimental results including mSAD, RMSE, Speedup, and PPM under all noise scenarios. Although VCA achieves a very fast endmember extraction task, when it combines with FSPM, it still improves its endmember extraction speed with speedup results over 3 or even 7. For MVSA, it involves generating endmembers by specifying a minimum external simplex volume, which depends highly on each superpixel. In this regard, it hardly provides a desirable data subset than other PPAs [see Figs. 12 and 13(f)].
the simplex boundary. When there are worse noises, MVSA may find it difficult to determine the desired simplex. In this regard, SPP can reconstruct each pixel using its neighborhoods, which alleviates noises to recover data simplex to some degree. Besides, FSPM can improve MVSA’s endmember accuracy according to PPM results, and it also provides good acceleration performance for MVSA. In terms of MVC-NMF, it normally requires a lot of computational time owing to its nonconvex regularizer that hinders its optimization speed. When combined with FSPM, MVC-NMF has a large acceleration performance over 130 times while it simultaneously has improvements on endmember accuracy. SENMAV provides a comparison from a spatial-spectral-based endmember extraction viewpoint. Compared with SENMAV, some PPAs can promote EEA’s endmember extraction performance from a spatial-spectral perspective.

7) Experiment 7 (Comparison Between EEAs and PPA-EEA Combinations on Indian Pines Dataset): This experiment was conducted on Indian Pines that contains 145 × 145 pixels, 200 pruned bands, and 16 classes. Since there are no available reference spectra for 16 classes, mSAD was not provided for this dataset, but 16 classes were regarded as the number of endmembers.
TABLE III

OVERALL RESULTS OBTAINED FROM DIFFERENT ALGORITHMS ON DC1 WITH DIFFERENT NOISE LEVELS

| Algorithms | Method 1 | Method 2 | Method 3 | Method 4 | Method 5 |
|------------|----------|----------|----------|----------|----------|
|            | RMSE     | Speedup  | PP | RMSE     | Speedup  | PP | RMSE     | Speedup  | PP | RMSE     | Speedup  | PP |
| NPNDR      | 0.099    | 2.004    | 6.253 | 0.094    | 2.004    | 6.253 |           |          |    |           |          |    |
| SPP+NPVDR  | 0.102    | 2.004    | 6.253 | 0.094    | 2.004    | 6.253 |           |          |    |           |          |    |
| SSP+NPVDR  | 0.207    | 2.004    | 6.253 | 0.100    | 2.004    | 6.253 |           |          |    |           |          |    |
| RCPSP+NPVDR| 0.320    | 2.004    | 6.253 | 0.118    | 2.004    | 6.253 |           |          |    |           |          |    |
| FSP+NPVDR  | 0.099    | 2.004    | 6.253 | 0.094    | 2.004    | 6.253 |           |          |    |           |          |    |

a “—” denotes that the evaluation results are not available.

b "‡" represents that the algorithm fails to extract endmembers.

The best results are in bold.

Table IV details the experimental results obtained from all algorithms on Indian Pines. After coupling with FSPM, N-FINDR, OSP, VCA, and MVSA can produce relatively better experimental performances regarding reconstruction errors and time details compared with other PPAs. In terms of MVC-NMF, all PPAs cannot promote lower RMSE results for MVC-NMF but SSP, SSPM, RCSPP, and FSPM can highly improve its unmixing speed. According to RMSE results provided in Table IV, reconstruction error maps obtained from all algorithms are displayed in Fig. 15.

8) Experiment 8 (Comparison Between EEAs and PPA-EEA Combinations on Cuprite Dataset): This experiment evaluated endmember extraction performances associated with five EEAs and five EEA-PPAs on the Cuprite dataset. It is worth noting that when the extracted endmembers are matched with USGS library by considering SAD, this matching procedure is suboptimal because it may strongly depend on the order in which endmembers

Fig. 14. Impact of (a) mSAD, (b) speedup, and (c) and (d) time details on different image size.

Table IV

OVERALL RESULTS (RMSE, SPEEDUP, AND EXECUTION TIME (IN SECONDS)) OF ALGORITHMS FOR EXPERIMENTS CONDUCTED ON INDIAN PINES DATASET

| Algorithm     | RMSE | Speedup | Execution Time (s) |
|---------------|------|---------|--------------------|
| PSPP          | 0.120| 2.004   | 6.253              |
| SPP+PSPP      | 0.207| 2.004   | 6.253              |
| SSP+PSPP      | 0.320| 2.004   | 6.253              |
| RCPSP+PSPP    | 0.118| 2.004   | 6.253              |
| FSP+PSPP      | 0.099| 2.004   | 6.253              |

a “—” denotes that the evaluation results are not available.

b "‡" represents that the algorithm fails to extract endmembers.

The best results are in bold.
are matched. In this case, the endmember that is matched in the first place may indeed affect subsequent matching of the remaining endmembers [26]. In order to avoid this problem, we select six representative minerals to calculate mSAD because they can be accurately matched with most of the relevant minerals in the USGS library under different endmember extraction scenarios. Six representative minerals include alunite, buddingtonite, dumortierite, kaolinite, muscovite, and montmorillonite.

Table V tabulates the overall experimental results, including mSAD, RMSE, Speedup, PPM, and time details for PPA, EEA, and the PPA-EEA combination (in seconds) of all the algorithms on Cuprite hyperspectral datasets. As can be seen from Table V, when combined with FSPM, N-FINDR, and VCA can generate the best experimental results regarding all evaluation metrics and also provide the lowest time consumption. For MVSA, SSPM provides the best acceleration performance but barely promotes its endmember accuracy. Compared to other PPAs, FSPM provides lower SAD values as well as PPM for MVSA. In terms of MVC-NMF, SPP can produce better endmember accuracy, but FSPM has the lowest reconstruction errors and the highest acceleration performance. For OSP, although FSPM provides the best speedup results, SSPM can provide the lowest SAD, RMSE, and PPM results.

The reconstruction error maps obtained from all algorithms are displayed in Fig. 16. We are aware that the number of estimated endmembers was fixed at 12 potentially resulting in higher RMSE results in some regions than those in 14 types of endmembers. However, algorithmic structures of different endmember extraction methods also to some extent lead to insufficient endmember extraction performance. For instance, compared to pure pixel assumption-based EEAs, NMF-based algorithms such as MVC-NMF can produce lower RMSE results under the same setting of endmember numbers.

V. CONCLUSION

This article investigates a new fast preprocessing algorithm based on subspace exploitation, called FSPM. Specifically, FSPM does not consider the traditional process of spatial-spectral exploration of the HSI but rather first reduces the HSI into a low-dimensional subspace by performing SVD. FSPM then, respectively, treats spatial and spectral information exploitation as the process of removing outliers beyond the simplex and specifying convex hull vertices. After iteratively extracting desired data points from the subspace, FSPM finally transforms determined data points into noise-reduced data space. Compared
to existing representative PPAs, FSPM shows its advantages on the quality of endmember candidates and time requirements. Based on different validation metrics, experiments conducted on synthetic and real hyperspectral images indicate that FSPM not only requires less computational requirements but also provides a more preferable data subset than other counterparts.

We are also aware that there are underlying flaws, which somewhat demands future research works. The first one is that the proposed FSPM requires several parameters to separate the outliers and normal data. Although we find that there are several parameters, including the ratio of pixel selection $\lambda$, the radius of searching region $\delta$ and the number of neighborhoods $k$, can be easily tuned, the most important parameter, i.e., the cut-off threshold $\eta$, requires a more careful tuning process according to specific dataset and outlier scenarios. The second one is that the computational burden of FSPM is affected by the number of endmembers.

ACKNOWLEDGMENT

The authors would like to thank the Image and Intelligence Information Processing Innovation Team of the State Ethnic Affairs Commission of China for their support, and also thank the Associate Editor who handled this manuscript and the two anonymous Reviewers for their outstanding comments and suggestions, which greatly helped to improve the technical quality and presentation of this article.

REFERENCES

[1] N. Keshava and J. F. Mustard, “Spectral unmixing,” IEEE Signal Process. Mag., vol. 19, no. 1, pp. 44–57, Jan. 2002.
[2] L. Miao and H. Qi, “Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization,” IEEE Trans. Geosci. Remote Sens., vol. 45, no. 3, pp. 765–777, Mar. 2007.
[3] J. M. Bioucas-Dias et al., “Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches,” IEEE J. Sel. Topics Appl. Earth Obs. Remote. Sens., vol. 5, no. 2, pp. 354–379, Apr. 2012.
[4] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, N. Nasrabad, and J. Chanussot, “Hyperspectral remote sensing data analysis and future challenges,” IEEE Geosci. Remote Sens. Mag., vol. 1, no. 2, pp. 6–36, Jun. 2013.
[5] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[6] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, “Hyperspectral unmixing via $l_1 / l_2$ sparsity-constrained nonnegative matrix factorization,” IEEE Trans. Geosci. Remote Sens., vol. 49, no. 11, pp. 4282–4297, Nov. 2011.
[7] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, “Manifold normalized sparse NMF for hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 5-1, pp. 2815–2826, May 2013.
[8] J. Li, J. M. Bioucas-Dias, A. J. Plaza, and L. Liu, “Robust collaborative nonnegative matrix factorization for hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 6076–6090, Oct. 2016.
[9] 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, Nov. 2012.
[10] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, “Collaborative sparse regression for hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 1, pp. 341–354, Jan. 2014.
[11] Y. Zhong, R. Feng, and L. Zhang, “Non-local sparse unmixing for hyperspectral remote sensing imagery,” IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 7, no. 6, pp. 1889–1909, Jun. 2014.
[12] S. Zhang, J. Li, Z. Wu, and A. Plaza, “Spatial discontinuity-weighted sparse unmixing of hyperspectral images,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 10, pp. 5767–5779, Oct. 2018.

[13] D. C. Heinz and C. Chang, “Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 39, no. 3, pp. 529–545, Mar. 2001.
[14] J. W. Boardman, F. A. Kruse, and R. O. Green, “Mapping target signatures via partial unmixing of AVIRIS data, in summaries,” in Proc. 5th IPI, Albers, Airborne Earth Science Workshop, 1995, pp. 3–6.
[15] J. Harsanyi and C. Chang, “Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach,” IEEE Trans. Geosci. Remote Sens., vol. 32, no. 4, pp. 779–785, Jul. 1994.
[16] J. M. P. Nascimento and J. M. B. Dias, “Vertex component analysis: A fast algorithm to unmix hyperspectral data,” IEEE Trans. Geosci. Remote Sens., vol. 43, no. 4, pp. 898–910, Apr. 2005.
[17] M. E. Winter, “N-FINDR: An algorithm for fast autonomous spectral endmember determination in hyperspectral data,” in Proc. SPIE, vol. 3753, pp. 266–276, 1999.
[18] C. Chang, C. Wu, W. Liu, and Y. C. Ouyang, “A new growing method for simplex-based endmember extraction algorithm,” IEEE Trans. Geosci. Remote Sens., vol. 44, no. 10, pp. 2804–2819, Oct. 2006.
[19] T. Chan, W. Ma, A. Ambikapathi, and C. Chi, “A simplex volume maximization framework for hyperspectral endmember extraction,” IEEE Trans. Geosci. Remote Sens., vol. 49, no. 11, pp. 4177–4193, Nov. 2011.
[20] T. Chan, C. Chi, Y. Huang, and W. Ma, “A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing,” IEEE Trans. Signal Process., vol. 57, no. 11, pp. 4418–4432, Nov. 2009.
[21] J. Li, A. Agathos, D. Zaharie, J. M. Bioucas-Dias, A. J. Plaza, and X. Li, “Minimum volume simplex analysis: A fast algorithm for linear hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 9, pp. 5067–5082, Sep. 2015.
[22] A. Plaza, P. Martínez, R. M. Pérez, and J. Plaza, “Spatial/spectral endmember extraction by multidimensional morphological operations,” IEEE Trans. Geosci. Remote Sens., vol. 40, no. 9, pp. 2025–2041, Sep. 2002.
[23] D. M. Rogge, B. Rivard, J. Zhang, A. Sanchez, I. Harrits, and J. Feng, “Integration of spatial-spectral information for the improved extraction of endmembers,” Remote Sens. Environ., vol. 110, no. 3, pp. 287–303, 2007.
[24] X. Shen and W. Bao, “Hyperspectral endmember extraction using spatially weighted simplex strategy,” Remote Sens., vol. 11, no. 18, 2019, Art. no. 2147.
[25] X. Shen, W. Bao, and K. Qu, “Spatial-spectral hyperspectral endmember extraction using a spatial energy prior constrained maximum simplex volume approach,” IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens., vol. 13, pp. 1347–1361, 2020. [Online]. Available: https://doi.org/10.1109/JSTARS.2020.2981402
[26] M. Zortea and A. Plaza, “Spatial preprocessing for endmember extraction,” IEEE Trans. Geosci. Remote Sens., vol. 47, no. 8, pp. 2679–2693, Aug. 2009.
[27] G. Martín and A. Plaza, “Spatial-spectral preprocessing prior to endmember identification and unmixing of remotely sensed hyperspectral data,” IEEE J. Sel. Topics Appl. Earth Obs. Remote. Sens., vol. 5, no. 2, pp. 380–395, Apr. 2012.
[28] F. Kowkabi, H. Ghasssemian, and A. Keshavarz, “A fast spatial-spectral preprocessing module for hyperspectral endmember extraction,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 6, pp. 782–786, Jun. 2016.
[29] X. Xu, J. Li, C. Wu, and A. Plaza, “Regional clustering-based spatial preprocessing for hyperspectral unmixing,” Remote Sens. Environ., vol. 204, pp. 333–346, 2018.
[30] F. Kowkabi, H. Ghasssemian, and A. Keshavarz, “Hybrid preprocessing algorithm for endmember extraction using clustering, over-segmentation, and local entropy criterion,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 10, no. 6, pp. 2940–2949, Jun. 2017.
[31] S. López, J. F. Moure, A. Plaza, G. M. Callicó, J. F. López, and R. Sarmiento, “A new preprocessing technique for fast hyperspectral endmember extraction,” IEEE Geosci. Remote Sens. Lett., vol. 10, no. 5, pp. 1070–1074, Sep. 2013.
[32] J. Delgado, G. Martín, J. Plaza, L. Jiménez, and A. Plaza, “Fast spatial preprocessing for spectral unmixing of hyperspectral data on graphics processing units,” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 9, no. 2, pp. 952–961, Feb. 2016.
[33] L. Jiménez, G. Martín, S. Sánchez, J. Plaza, and A. Plaza, “Fast spatial-spectral preprocessing for endmember extraction and spectral unmixing using graphic processing units,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., Beijing, China, 2016, pp. 7038–7041.
[34] M. Xu, L. Zhang, and B. Du, “An image-based endmember bundle extraction algorithm using both spatial and spectral information,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 6, pp. 2607–2617, Jun. 2015.

[35] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys (CSUR), vol. 41, no. 3, 2009, Art. no. 15.

[36] T. Han, D. G. Goodenough, A. Dyk, and J. Love, “Detection and correction of abnormal pixels in hyperion images,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2002, vol. 3, pp. 1327–1330.

[37] C. Chang and Q. Du, “Estimation of number of spectrally distinct signal sources in hyperspectral imagery,” IEEE Trans. Geosci. Remote. Sens., vol. 42, no. 3, pp. 608–619, Mar. 2004.

[38] K. Sun, X. Geng, P. Wang, and Y. Zhao, “A fast endmember extraction algorithm based on Gram determinant,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 6, pp. 1124–1128, Jun. 2014.

[39] J. M. Bioucas-Dias, “A variable splitting augmented lagrangian approach to linear spectral unmixing,” in Proc. First Workshop Hyperspectral Image Signal Processing: Evol. Remote Sens., Grenoble, France, 2009, pp. 1–4.

[40] Y. Yu and W. Sun, “Minimum distance constrained non-negative matrix factorization for the endmember extraction of hyperspectral images,” in Proc. SPIE, vol. 6790, 2007, Art. no. 679015.

[41] N. Wang, B. Du, and L. Zhang, “An endmember dissimilarity constrained non-negative matrix factorization method for hyperspectral unmixing,” IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 6, no. 2, pp. 554–569, Apr. 2013.

[42] A. Huck, M. Guillaume, and J. Blanc-Talon, “Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data,” IEEE Trans. Geosci. Remote. Sens., vol. 48, no. 6, pp. 2590–2602, Jun. 2010.

[43] S. Mei, Q. Du, M. He, and Y. Wang, “Spatial preprocessing for spectral endmember extraction by local linear embedding,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., Milan, Italy, 2015, pp. 5027–5030.

[44] G. Zhang, S. Mei, J. Tian, Y. Feng, and Q. Du, “Local sparse representation based spatial preprocessing for endmember extraction,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., Yokohama, Japan, 2019, pp. 278–281.

[45] G. Martin and A. Plaza, “Region-based spatial preprocessing for endmember extraction and spectral unmixing,” IEEE Geosci. Remote. Sens. Lett., vol. 8, no. 4, pp. 745–749, Jul. 2011.

[46] F. Kowkabi and A. Keshavarz, “Using spectral geodesic and spatial euclidean weights of neighbourhood pixels for hyperspectral endmember extraction preprocessing,” ISPRS J. of Photogrammetry Remote Sens., vol. 158, pp. 201–218, 2019.

[47] X. Shen and W. Bao, “A spatial energy and spectral purity based preprocessing algorithm for fast hyperspectral endmember extraction,” in Proc. 10th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens., Amsterdam, Netherlands, 2019, pp. 1–3.

[48] R. Heylen and P. Scheunders, “Multidimensional pixel purity index for convex hull estimation and endmember extraction,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 7/1, pp. 4059–4069, Jul. 2013.

[49] R. L. Graham, “An efficient algorithm for determining the convex hull of a finite planar set,” Inf. Process. Lett., vol. 1, no. 4, pp. 132–133, 1972.

[50] R. A. Jarvis, “On the identification of the convex hull of a finite set of points in the plane,” Inf. Process. Lett., vol. 2, no. 1, pp. 18–21, 1973.

[51] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, “The quickhull algorithm for convex hulls,” ACM Trans. Math. Softw., vol. 22, no. 4, pp. 469–483, 1996.

[52] C. C. Aggarwal, Outlier Analysis. New York, NY, USA: Springer, 2013.

[53] Y. Su, A. Marinoni, J. Li, J. Plaza, and P. Gamba, “Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing,” IEEE Geosci. Remote. Sens. Lett., vol. 15, no. 9, pp. 1427–1431, Sep. 2018.

[54] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, Dallas, Texas, USA, 2000, pp. 93–104.

[55] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. Second Int. Conf. Knowl. Discov. Data Mining, 1996, pp. 226–231.

Xiangfei Shen received the B.S. degree in computer science and technology from North Minzu University, Yinchuan, China, 2017. He is currently working toward the M.S. degree at the School of Computer Science and Engineering, North Minzu University, Yinchuan, China.

His research interests include hyperspectral image processing, pattern recognition, and machine learning.

Wenxing Bao received the B.Eng. degree in industrial automation from Xidian University, Xi’an, China, in 1993, the M.Sc. degree in electrical engineering from Xi’an Jiaotong University, Xi’an, China, in 2001, and the Ph.D. degree in electronic science and technology from Xi’an Jiaotong University, Xi’an, China, in 2006.

He is currently a Professor and Vice President with North Minzu University, Yinchuan, China. His research interests include digital image processing, remote sensing image classification, and fusing.

Kewen Qu received the B.Eng. degree in computer science from Henan University, Kaifeng, China, in 2008, and the M.Sc. degree in computer science from North Minzu University, Yinchuan, China, in 2012, and the Ph.D. degree from the School of Computer and Information, Hefei University of Technology, Hefei, China, in 2020.

His current research interests include hyperspectral image processing, machine learning, and computer vision.