

A spatial compositional model for linear unmixing and endmember uncertainty estimation

Yuan Zhou, Anand Rangarajan, *Member, IEEE*, and Paul D. Gader, *Fellow, IEEE*

Abstract

The normal compositional model (NCM) has been extensively used in hyperspectral unmixing. However, previous research has mostly focused on estimation of endmembers and/or their variability, based on the assumption that the pixels are independent random variables. In this paper, we show that this assumption does not hold if all the pixels are generated by a fixed endmember set. This introduces another concept, *endmember uncertainty*, which is related to whether the pixels fit into the endmember simplex. To further develop this idea, we derive the NCM from the ground up without the pixel independence assumption, along with (i) using different noise levels at different wavelengths and (ii) using a spatial and sparsity promoting prior for the abundances. The resulting new formulation is called the *spatial compositional model* (SCM) to better differentiate it from the NCM. The SCM maximum a posteriori (MAP) objective leads to an optimization problem featuring noise weighted least-squares minimization for unmixing. The problem is solved by projected gradient descent, resulting in an algorithm that estimates endmembers, abundances, noise variances, and endmember uncertainty simultaneously. We compared SCM with current state-of-the-art algorithms on synthetic and real images. The results show that SCM can in the main provide more accurate endmembers and abundances. Moreover, the estimated uncertainty can serve as a prediction of endmember error under certain conditions.

Index Terms

endmember extraction, endmember uncertainty, hyperspectral image analysis, linear unmixing, spatial compositional model, normal compositional model

The authors are with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA. E-mail: zhouyuanzxcv@gmail.com, anand.pgader@cise.ufl.edu.

I. INTRODUCTION

A. The linear mixing model

Hyperspectral image unmixing has received wide attention in the remote sensing, signal and image processing communities [1], [2]. The widely researched model in this area is the *linear mixing model* (LMM). Assume we have a hyperspectral image $I(\mathbf{x}) : \mathcal{D} \rightarrow \mathbb{R}^B$ where $\mathcal{D} \subset \mathbb{R}^2$ is the image domain and B the number of wavelengths. LMM assumes that the spectral measurement at each pixel $\mathbf{y}_i \in \mathbb{R}^B$, $i = 1, 2, \dots, N$ (N being the pixel cardinality) is a non-negative linear combination of the spectral signature of some pure materials, called *endmembers*, $\mathbf{m}_j = [m_{j1}, \dots, m_{jB}] \in \mathbb{R}^B$, $j = 1, 2, \dots, M$ (M being the number of endmembers). The governing equation is

$$\begin{aligned} \mathbf{y}_i &= \mathbf{M}^T \boldsymbol{\alpha}_i + \mathbf{n}_i, \text{ with} \\ \alpha_{ij} &\geq 0, \sum_{j=1}^M \alpha_{ij} = 1, \forall i \end{aligned} \quad (1)$$

where $\mathbf{M} := [\mathbf{m}_1, \dots, \mathbf{m}_M]^T \in \mathbb{R}^{M \times B}$. $\boldsymbol{\alpha}_i := [\alpha_{i1}, \dots, \alpha_{iM}]^T \in \mathbb{R}^M$ is the fractional *abundance map* and satisfies the positivity and sum-to-one (simplex) constraints, and $\mathbf{n}_i \in \mathbb{R}^B$ is a small, additive perturbation (noise). As a result, the pixels generated by this model form a simplex in a B -dimensional vector space whose vertices are the endmembers. Combining the above equation for all the pixels, we have the following equation for the LMM:

$$\mathbf{Y} = \mathbf{A}\mathbf{M} + \mathbf{N}, \quad (2)$$

where $\mathbf{Y} := [\mathbf{y}_1, \dots, \mathbf{y}_N]^T \in \mathbb{R}^{N \times B}$, $\mathbf{A} := [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N]^T \in \mathbb{R}^{N \times M}$, and $\mathbf{N} := [\mathbf{n}_1, \dots, \mathbf{n}_N]^T \in \mathbb{R}^{N \times B}$.

The goal of the linear unmixing problem is to retrieve \mathbf{A} and \mathbf{M} given \mathbf{Y} . This is an ill-posed inverse problem as it can have an infinite number of solutions. Fig. 1 shows the difficulties stemming from this underdetermined nature. In the figure, (a), (b) are 2 cases wherein the true endmembers can not be estimated while (c) shows the only case where they can be estimated under the assumption that the endmembers tightly surround the pixels. This has been extensively used in the literature in the form of minimal volume [3], pure pixels [4], or pairwise closeness [5], [6]. In Fig. 1(d), we show a different and interesting problem. It implies that spatial information should be used in the unmixing process: if the pixels from an endmember set lie in one region while those from the other set lie in a different region, the abundances in the intersection can be easily identified by spatial location.

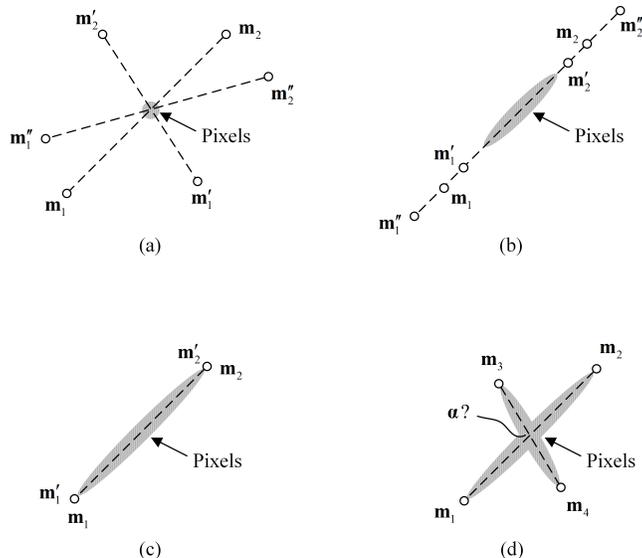


Fig. 1. Difficulties in linear unmixing: \mathbf{m}_1 , \mathbf{m}_2 are the endmembers that generate the pixel data; \mathbf{m}'_1 , \mathbf{m}'_2 and \mathbf{m}''_1 , \mathbf{m}''_2 are the possible endmembers that can be inferred from the data. (a) The pixels (gray area) generated by \mathbf{m}_1 , \mathbf{m}_2 with $\alpha = (0.5, 0.5)^T$. Not only can we infer \mathbf{m}'_1 , \mathbf{m}'_2 or \mathbf{m}''_1 , \mathbf{m}''_2 that generate the same pixels, we can use the entire Euclidean space for the endmembers. (b) The abundances α ranging from $(0.7, 0.3)$ to $(0.2, 0.8)$. Here, we can determine that the endmembers should lie in a line that fits the pixels. However, we still cannot determine the specific position of the endmembers without other information. (c) We have the information that the abundances range from $(1, 0)$ to $(0, 1)$. Now we can find the endmembers as we are not only given the line, but also the relative position of the endmembers to the boundary of the pixels. (d) Suppose we have two endmember sets, \mathbf{m}_1 , \mathbf{m}_2 and \mathbf{m}_3 , \mathbf{m}_4 and we can obtain them. How can we determine the abundances α of the pixels in the intersection? Should it be a linear combination of \mathbf{m}_1 and \mathbf{m}_2 or of \mathbf{m}_3 and \mathbf{m}_4 , or of all of them? However, if we know the spatial location, we may identify them. For example, suppose the endmembers occupy 4 quadrants of the image where \mathbf{m}_1 , \mathbf{m}_2 occupy the top 2 quadrants while \mathbf{m}_3 , \mathbf{m}_4 occupy the bottom 2 quadrants. The mixed pixels appear at the boundary between two materials. Hence all the mixed pixels of \mathbf{m}_1 , \mathbf{m}_2 appear in the top area while the mixed pixels of \mathbf{m}_3 , \mathbf{m}_4 appear in the bottom area. If we know where a pixel is located, we can decide if it belongs to \mathbf{m}_1 , \mathbf{m}_2 or \mathbf{m}_3 , \mathbf{m}_4 .

B. Previous work

The methods developed to solve this problem may be mainly categorized into geometrical, statistical and sparse regression based approaches [2]. Vertex component analysis (VCA) assumes that the endmembers are present in the image pixels and finds them by projection [4]. Iterative constrained endmembers (ICE) minimizes the least-squares error and the pairwise closeness constraint [5]. Minimum volume constrained nonnegative matrix factorization (MVC-NMF) minimizes the same error along with the squared value of the approximation to the volume formed by the endmembers [3]. Graph regularized $L_{1/2}$ nonnegative matrix factorization (GLNMF) [7], [8] replaces the volume constraint by a graph regularizer and a sparsity promoting term. Finally,

piecewise convex multiple-model endmember detection (PCOMMEND) clusters the data into a collection of convex sets and simultaneously estimates endmembers and abundances for each set [6].

Besides these methods that rely only on independent pixels, some recent work introduces spatial information to aid the unmixing process [9], [10], [11], [12], [13]. For example, in [9], two smoothness terms for abundances and endmembers were proposed to utilize the spatial information in terms of wavelength proximity and pixel location. In [10], a Markov Random Field (MRF) Potts-Markov model, was used to model the partitioning of the image which can help the unmixing process. Sampling methods were used to infer the unknown parameters. In [11], minimization of the L_1 norm of the differences between neighboring abundances was proposed to impose spatial correlation. In [12], a similar minimization of the L_2 norm with weights determined by the spectral angular distance was proposed under the rubric of weighted nonnegative matrix factorization (WNMF). In recent work [13], spatial information was used to generate an additional map to guide sparsity.

A different method category is based on modeling the likelihood using Gaussian density functions, also known as the *normal compositional model* (NCM) [14], [15], [16], [17], [18]. The earliest application of NCM to hyperspectral unmixing can be traced back to [14], wherein a maximum likelihood estimation (MLE) approach was presented for NCM endmember extraction. In [15], [16], priors (mainly uniform distributions) were imposed on abundances and endmembers with sampling methods used to maximize the posterior. They assumed the endmember spectra at different wavelengths were independent (i.e. the endmember covariance matrices are diagonal) and estimated one parameter of variability for each endmember. In [17], a more complex NCM was proposed without the assumption of independence of endmember spectra at different wavelengths. Particle swarm optimization based expectation-maximization was used to maximize the log-likelihood. In [18], the image data were partitioned into several convex sets and NCM was applied on each set.

C. Endmember uncertainty

Despite the voluminous previous work on unmixing, there is little previous research on estimating the *model uncertainty* of the endmembers directly from the linear mixing model. That is, given the pixel data and an estimated endmember set, the endmember estimates may have residual uncertainty. For example, Fig. 2 shows 3 possible estimated endmember sets on a

synthetic dataset when $B = 2$. We can expect the endmember set $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3$ to have a small uncertainty since they fit the pixels very well. Allowing them to move around may ruin the fitting. $\mathbf{m}'_1, \mathbf{m}'_2, \mathbf{m}'_3$ are located within the pixels. They should have a large uncertainty because they can move around more freely to better fit the pixels. For $\mathbf{m}''_1, \mathbf{m}''_2, \mathbf{m}''_3$, the uncertainty may be small if we consider that they have already fitted the pixels very well.

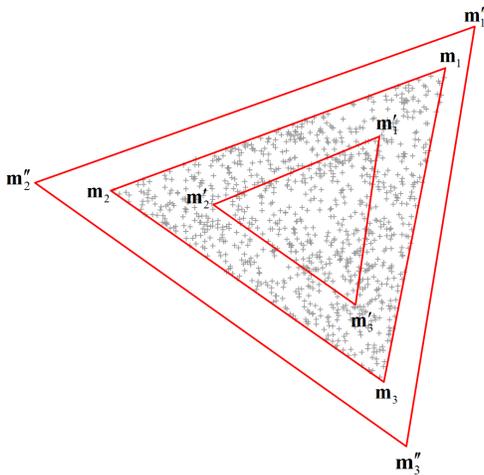


Fig. 2. Model uncertainty of the estimated endmembers at different positions. Intuitively, $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3$ should have a small uncertainty while the uncertainty of $\mathbf{m}'_1, \mathbf{m}'_2, \mathbf{m}'_3$ should be large. The uncertainty of $\mathbf{m}''_1, \mathbf{m}''_2, \mathbf{m}''_3$ could be small if we consider that all the pixels can still fit into them.

The above intuition implies that the uncertainty may reflect the error of endmembers. To show how this intuition formally works in NCM, assume a simple case that an endmember $\mathbf{m} \in \mathbb{R}^2$ follows a Gaussian distribution centered at $\mathbf{r} \in \mathbb{R}^2$ with covariance matrix $\Sigma \in \mathbb{R}^{2 \times 2}$, i.e. $p(\mathbf{m}) = \mathcal{N}(\mathbf{m}|\mathbf{r}, \Sigma)$. Suppose \mathbf{m} is given and \mathbf{r} has been estimated with $\mathbf{m} \neq \mathbf{r}$. We want to find Σ using maximum likelihood estimation (MLE). Maximizing $p(\mathbf{m})$ is equivalent to minimizing

$$-\log p(\mathbf{m}) = \frac{1}{2} \log |\Sigma| + \frac{1}{2} (\mathbf{m} - \mathbf{r})^T \Sigma^{-1} (\mathbf{m} - \mathbf{r}).$$

Let $\Sigma = \mathbf{U} \text{diag}(\sigma_1^2, \sigma_2^2) \mathbf{U}^T$, $\sigma_1 > 0$, $\sigma_2 > 0$ be the eigendecomposition. Then the minimization problem above becomes

$$\log \sigma_1 + \log \sigma_2 + \frac{1}{2} \sigma_1^{-2} z_1^2 + \frac{1}{2} \sigma_2^{-2} z_2^2$$

where $\mathbf{z} := (z_1, z_2)^T := \mathbf{U}^T (\mathbf{m} - \mathbf{r})$. When an eigenvector in \mathbf{U} is not perpendicular to $\mathbf{m} - \mathbf{r}$, i.e. $z_1 \neq 0$, $z_2 \neq 0$, the minimization can be achieved by setting the derivatives with respect to

σ_1 and σ_2 to 0, which leads to $\sigma_1 = |z_1|$, $\sigma_2 = |z_2|$. However, this is not the global minimum because if one eigenvector in \mathbf{U} is perpendicular to $\mathbf{m} - \mathbf{r}$ (the other being parallel), say $z_2 = 0$, $z_1 = \|\mathbf{m} - \mathbf{r}\|$, σ_2 can be arbitrarily close to 0 such that $\log\sigma_2$ goes to negative infinity. Assume $\sigma_i \geq \epsilon$ for a small positive ϵ to make a solution exist, then the global minimum lies at $\sigma_1 = \|\mathbf{m} - \mathbf{r}\|$, $\sigma_2 = \epsilon$. Therefore, we can see that the MLE estimated matrix Σ has the square root of its largest eigenvalue equal to $\|\mathbf{m} - \mathbf{r}\|$ while its eigenvector is parallel to $\mathbf{m} - \mathbf{r}$. For our formulation (2), assume \mathbf{M} follows a Gaussian distribution with centers in \mathbf{R} . We then propose a *fundamental* question:

- Given \mathbf{Y} , can we find the covariance matrices (uncertainty) that measure the difference between the estimated endmembers \mathbf{R} and the ground truth \mathbf{M} ?

If the answer is yes, we have the means via the chosen measure to predict the error without knowing the ground truth. This paper attempts to find such covariance matrices.

The previous NCMs did not solve this problem. The covariance matrices from the previous NCMs represent the *endmember variability* which arises from the assumption that the endmember set used to linearly generate a pixel may vary with location due to illumination conditions, atmospheric, environmental, temporal factors and intrinsic variability in a material [19]. We explain the difference between these two concepts, uncertainty and variability, by first summarizing the previous NCMs. Suppose the j th endmember follows a Gaussian distribution centered at $\mathbf{r}_j \in \mathbb{R}^B$ with covariance matrix Σ_j , i.e. $p(\mathbf{m}_j) = \mathcal{N}(\mathbf{m}_j | \mathbf{r}_j, \Sigma_j)$. Assuming the endmembers to be independent, the random variable transformation (r.v.t.) $\mathbf{y}_i = \mathbf{M}^T \boldsymbol{\alpha}_i$ for each pixel suggests that the probability density function of \mathbf{y}_i can be derived as

$$p(\mathbf{y}_i) = \mathcal{N}\left(\mathbf{y}_i | \mathbf{R}^T \boldsymbol{\alpha}_i, \sum_{j=1}^M \alpha_{ij}^2 \Sigma_j\right) \quad (3)$$

where $\mathbf{R} := [\mathbf{r}_1, \dots, \mathbf{r}_M]^T \in \mathbb{R}^{M \times B}$. Then, NCM assumes the random variables $\{\mathbf{y}_i : i = 1, \dots, N\}$ are independent (henceforth referred to as *pixel independence*). The density function of \mathbf{Y} is then the product of the density functions of each component,

$$\begin{aligned} p(\mathbf{Y}) &= \prod_{i=1}^N p(\mathbf{y}_i) = \prod_{i=1}^N \mathcal{N}\left(\mathbf{y}_i | \mathbf{R}^T \boldsymbol{\alpha}_i, \sum_{j=1}^M \alpha_{ij}^2 \Sigma_j\right) \\ &= \mathcal{N}\left(\text{vec}(\mathbf{Y}^T) | \text{vec}((\mathbf{A}\mathbf{R})^T), \Sigma'_{\mathbf{Y}}\right), \end{aligned} \quad (4)$$

where $\Sigma'_{\mathbf{Y}} \in \mathbb{R}^{NB \times NB}$ is defined as

$$\Sigma'_{\mathbf{Y}} := \begin{bmatrix} \delta_{ij} \sum_{k=1}^M \alpha_{ik}^2 \Sigma_k \end{bmatrix}. \quad (5)$$

We observe that $p(\mathbf{Y})$ is another Gaussian distribution with a *block diagonal* covariance matrix. The estimation of \mathbf{A} , \mathbf{R} and $\{\Sigma_j\}$ are handled differently in different works.

To estimate the uncertainty however, we can not assume the pixels $\{\mathbf{y}_i\}$ to be independent. To see this, suppose $B = 1$. Then we have $\mathbf{M} \in \mathbb{R}^M$, $\mathbf{R} \in \mathbb{R}^M$, $\mathbf{Y} \in \mathbb{R}^N$ which are vectors and the covariance matrix in (4) becomes an N by N diagonal matrix. The independence of endmembers suggests that the density of \mathbf{M} is

$$p(\mathbf{M}) = \mathcal{N}(\mathbf{M}|\mathbf{R}, \Sigma)$$

where Σ is an M by M diagonal matrix with each element being the variance of each endmember. The r.v.t. $\mathbf{Y} = \mathbf{A}\mathbf{M}$ indicates that the density function of \mathbf{Y} *does not even exist*. This is because the domain of $p(\mathbf{M})$ (\mathbb{R}^M) is projected to a subspace of dimension M in \mathbb{R}^N , which has Lebesgue measure 0 (then integrating $p(\mathbf{Y})$ gives value 0). One way to make the density function exist is to add noise, i.e., use equation (2). By assuming the noise to be Gaussian and independent at each pixel, $p(\mathbf{N}) = \mathcal{N}(\mathbf{N}|\mathbf{0}, \mu^2\mathbf{I}_N)$, we can see that the density function of \mathbf{Y} becomes

$$p(\mathbf{Y}) = \mathcal{N}(\mathbf{Y}|\mathbf{A}\mathbf{R}, \mathbf{A}\Sigma\mathbf{A}^T + \mu^2\mathbf{I}_N)$$

where the covariance matrix is not a diagonal matrix, which indicates that the pixels are *not* independent. The general case with $B > 1$ will be derived later.

D. Our contribution

In this paper, we solve the problem of estimating the model uncertainty by proposing a *spatial compositional model* (SCM) based on NCM without assuming pixel independence while utilizing spatial information on the abundances. Hence, the major contribution of this work is that we propose the concept of endmember uncertainty which is related to error prediction, and present a model that calculates the full likelihood to estimate it. Compared to previous NCMs and methods with spatial information, our method also features: (i) a spatial term that integrates smoothness and sparsity of abundances together; (ii) a model considering different noise levels at different wavelengths and finally (iii) the integration of noise estimation, endmember uncertainty and linear unmixing into a common framework. The final minimization problem can be solved by an algorithm that not only provides the endmembers and abundances, but also the noise and the uncertainty. An initial work that shows the benefits of the spatial prior used in the paper was published in [20].

Notation. Throughout the paper, $\text{SPD}(n)$ denotes the set of n by n symmetric positive definite matrices. We use the following notation for operations on a matrix $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n]^T$. We use $\text{Tr}(\mathbf{A})$, $|\mathbf{A}|$, $\|\mathbf{A}\|_F$, $\text{vec}(\mathbf{A})$ to denote the trace, determinant, Frobenius norm and vectorization of \mathbf{A} respectively. The vectorization operator is defined by concatenating its columns, $\text{vec}(\mathbf{A}^T) = [\mathbf{a}_1^T, \dots, \mathbf{a}_n^T]^T$. We use $(\mathbf{A})_{ij}$ to denote the extraction of the element at the i th row and j th column by viewing \mathbf{A} as a matrix, and use $(\mathbf{A})_i$ to denote the extraction of the i th row of \mathbf{A} and transposing it by viewing \mathbf{A} as a vector (hence $(\mathbf{A})_i = \mathbf{a}_i$). We use $[a_{ij}]$ to denote a matrix in which the element at the i th row, j th column is a_{ij} . So the matrix $[\delta_{ij}a_i] = \text{diag}(a_1, \dots, a_n)$ is a diagonal matrix with diagonal elements $\{a_i\}$ by defining $\delta_{ij} = 1$ only when $i = j$ and 0 otherwise. We use $\mathbf{A} \geq 0$ to denote that $a_{ij} \geq 0$ given $\mathbf{A} = [a_{ij}]$. The Kronecker product between two matrices \mathbf{A} and \mathbf{B} is defined by $\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}]$ and the Hadamard product is defined by $\mathbf{A} \circ \mathbf{B} = [a_{ij}b_{ij}]$. We use \mathbf{I}_N for the N by N identity matrix and $\mathbf{1}_N$ as an N by 1 vector consisting of all 1s.

II. THE SPATIAL COMPOSITIONAL MODEL

A. The hyperspectral image likelihood

We are interested in determining the uncertainty of the extracted endmembers. To achieve this, we first model the density function of \mathbf{M} , then use (2) to perform a r.v.t. to get the density function of \mathbf{Y} , and finally maximize the posterior given \mathbf{Y} to find the parameters. Assuming that the endmember \mathbf{m}_j follows a multivariate Gaussian centered at \mathbf{r}_j with covariance matrix Σ_j , i.e.

$$p(\mathbf{m}_j | \mathbf{r}_j, \Sigma_j) = \mathcal{N}(\mathbf{m}_j | \mathbf{r}_j, \Sigma_j),$$

and further assuming that the endmembers are independent, we have the conditional probability density function of the whole endmember set becoming the product of the independent components:

$$p(\mathbf{M} | \mathbf{R}, \Theta) = \mathcal{N}(\text{vec}(\mathbf{M}^T) | \text{vec}(\mathbf{R}^T), [\delta_{ij}\Sigma_j]), \quad (6)$$

where $\mathbf{R} := [\mathbf{r}_1, \dots, \mathbf{r}_M]^T \in \mathbb{R}^{M \times B}$ and $\Theta := \{\Sigma_j : j = 1, \dots, M\}$ is the collection of covariance matrices. Assume that the noise \mathbf{n}_i follows an independent zero mean, μ_k^2 variance Gaussian at the k th wavelength, i.e.,

$$p(\mathbf{n}_i | \{\mu_k\}) = \mathcal{N}(\mathbf{n}_i | \mathbf{0}, \mathbf{D}),$$

where $\mathbf{D} := \text{diag}(\mu_1^2, \mu_2^2, \dots, \mu_B^2)$ is the covariance matrix. Also, since the noise is usually independent at different locations, we have

$$p(\mathbf{N}|\mathbf{D}) = \mathcal{N}(\text{vec}(\mathbf{N}^T)|\mathbf{0}, \mathbf{I}_N \otimes \mathbf{D}). \quad (7)$$

From the probability density functions in (6), (7), and the transformation $\text{vec}((\mathbf{A}\mathbf{M})^T) = (\mathbf{A} \otimes \mathbf{I}_B)\text{vec}(\mathbf{M}^T)$, the r.v.t. (2) indicates that the conditional probability density function of \mathbf{Y} is

$$p(\mathbf{Y}|\mathbf{R}, \boldsymbol{\Theta}, \mathbf{A}, \mathbf{D}) = \mathcal{N}(\text{vec}(\mathbf{Y}^T)|\mathbf{r}_Y, \boldsymbol{\Sigma}_Y), \quad (8)$$

where $\mathbf{r}_Y \in \mathbb{R}^{NB}$, $\boldsymbol{\Sigma}_Y \in \mathbb{R}^{NB \times NB}$ are defined as

$$\begin{aligned} \mathbf{r}_Y &:= (\mathbf{A} \otimes \mathbf{I}_B)\text{vec}(\mathbf{R}^T) \\ &= \text{vec}((\mathbf{A}\mathbf{R})^T), \\ \boldsymbol{\Sigma}_Y &:= (\mathbf{A} \otimes \mathbf{I}_B)[\delta_{ij}\boldsymbol{\Sigma}_j](\mathbf{A} \otimes \mathbf{I}_B)^T + \mathbf{I}_N \otimes \mathbf{D} \\ &= \left[\delta_{ij}\mathbf{D} + \sum_{k=1}^M \alpha_{ik}\alpha_{jk}\boldsymbol{\Sigma}_k \right]. \end{aligned} \quad (9)$$

Note that compared to $\boldsymbol{\Sigma}'_Y$ in (5), the covariance matrix in (9) is not a block diagonal matrix implying that the hyperspectral pixels $\{\mathbf{y}_i : i = 1, \dots, N\}$ are not independent.

B. Modeling the priors

We model the prior probability density of \mathbf{A} by assuming that $\boldsymbol{\alpha}_i$ is a Markov random field (MRF). That is, we treat the image grid as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is the set of graph nodes and \mathcal{E} is the set of edges. The density of the whole grid can be modeled based on a potential function of the neighboring nodes. Considering that neighboring pixels with similar spectra are more likely to be the same mixture of materials, the prior probability density of \mathbf{A} can be assumed to be in favor of smooth assignment of $\boldsymbol{\alpha}_i$ to all these pairs of pixels.

Driven by this intuition, the prior probability density of \mathbf{A} is modeled as

$$\begin{aligned} p(\mathbf{A}) &= \frac{1}{Z} \exp \left\{ -\frac{\beta_1}{4} \sum_{i=1}^N \sum_{j=1}^N w_{ij} \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j\|^2 \right\} \\ &= \frac{1}{Z} \exp \left\{ -\frac{\beta_1}{2} \text{Tr}(\mathbf{A}^T \mathbf{L} \mathbf{A}) \right\}, \end{aligned} \quad (10)$$

where w_{ij} controls the spatial intimacy between node i and node j given by

$$w_{ij} = \begin{cases} e^{-\|\mathbf{y}_i - \mathbf{y}_j\|^2 / 2B\eta^2}, & \text{if } (i, j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases}.$$

$\mathbf{L} := [\delta_{ij} \sum_k w_{ik}] - [w_{ij}] \in \mathbb{R}^{N \times N}$ is the well known symmetric positive semidefinite *graph Laplacian* matrix [21]. $Z = \int \exp \left\{ -\frac{\beta_1}{2} \text{Tr}(\mathbf{A}^T \mathbf{L} \mathbf{A}) \right\} d\mathbf{A}$ is the partition function such that the integration of $p(\mathbf{A})$ is 1. Since the partition function is a constant and does not affect the objective function derived later, we will ignore it henceforth.

In practice, a region may contain only one pure material, which means the abundance maps for many pixels are concentrated on a single component, e.g., $\alpha_{ij} = 1$, $\alpha_{ik} = 0$ for $k \neq j$. This suggests that \mathbf{A} may have a higher prior probability for each α_i being sparse. A common sparsity promoting technique is to minimize the L_1 norm on α_i , which is not applicable here due to the sum-to-one constraint. A previous work uses the $L_{1/2}$ norm $\sum_{i,j} \alpha_{ij}^{1/2}$ to promote sparsity [22]. However, the non-smooth objective requires us to take subgradients which we would prefer to avoid. Here, we introduce a quadratic form $\|\alpha_i\|^2$, which by itself is not sparsity promoting, but does have that effect when maximized subject to the simplex constraint. Fig. 3 shows the sparsity promoting effect if we want to maximize $\|\alpha_i\|^2$ subject to the simplex constraint when $M = 2$. For $M > 2$, a similar result can be achieved. Hence, we can add $\sum_i \|\alpha_i\|^2$ to (10) and have a prior probability defined as

$$\begin{aligned} p(\mathbf{A}) &\propto \exp \left\{ -\frac{\beta_1}{2} \text{Tr}(\mathbf{A}^T \mathbf{L} \mathbf{A}) + \frac{\beta_2}{2} \text{Tr}(\mathbf{A}^T \mathbf{A}) \right\} \\ &= \exp \left\{ -\frac{\beta_1}{2} \text{Tr}(\mathbf{A}^T \mathbf{K} \mathbf{A}) \right\}, \end{aligned} \quad (11)$$

where $\mathbf{K} := \mathbf{L} - \frac{\beta_2}{\beta_1} \mathbf{I}_N$ if $\beta_1 \neq 0$.

The parameters \mathbf{r}_j can also be assumed to be drawn from suitable prior distributions. From the analysis of Fig. 1, to obtain a unique solution, we assume that the endmembers should tightly surround the mixed pixels. To achieve this goal, we introduce a joint prior on \mathbf{R} and \mathbf{D} as

$$\begin{aligned} p(\mathbf{R}, \mathbf{D}) &\propto \exp \left\{ -\frac{\rho}{4} \sum_{i=1}^M \sum_{j=1}^M (\mathbf{r}_i - \mathbf{r}_j)^T \mathbf{D}^{-1} (\mathbf{r}_i - \mathbf{r}_j) \right\} \\ &= \exp \left\{ -\frac{\rho}{2} \text{Tr}(\mathbf{R}^T \mathbf{H} \mathbf{R} \mathbf{D}^{-1}) \right\} \end{aligned} \quad (12)$$

where $\mathbf{H} \in \mathbb{R}^{M \times M}$ is the corresponding Laplacian matrix (\mathbf{H} has -1 everywhere except for diagonal terms taking the value $M - 1$). The prior will be in favor of endmembers that are close to each other. The significance of using the Mahalanobis distance over the traditional distance $\|\mathbf{r}_i - \mathbf{r}_j\|$ is to selectively calculate the pairwise closeness term from the bands not corrupted by the noise. Suppose some bands have larger noise than the others. Then they do not contribute as much to this value as bands with smaller noise. However, the prior will introduce a bias on

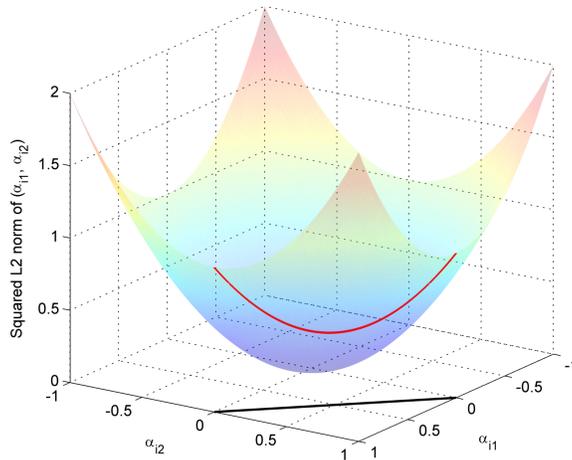


Fig. 3. The sparsity promoting effect of maximizing $\|\alpha_i\|^2$ subject to the simplex constraint when $M = 2$. The black line segment in the plane $z = 0$ is the simplex constraint. The red line on the paraboloid is the projected values of $\|\alpha_i\|^2$ from the simplex. Maximizing $\|\alpha_i\|^2$ will lead to solutions corresponding to the extreme ends of the simplex (sparse solution).

the estimation of the noise \mathbf{D} as it will prefer large noise. Since it is controlled by a parameter ρ , we can assume the bias vanishes when ρ is sufficiently small.

C. Maximizing the posterior

From the prior probability density in (11), (12) and the conditional probability density in (8), we invoke Bayes' theorem to get the posterior probability density used subsequently in posterior maximization, i.e.

$$p(\mathbf{R}, \Theta, \mathbf{A}, \mathbf{D}|\mathbf{Y}) \propto p(\mathbf{A})p(\mathbf{R}, \mathbf{D})p(\mathbf{Y}|\mathbf{R}, \Theta, \mathbf{A}, \mathbf{D})$$

where $p(\Theta)$ and $p(\mathbf{Y})$ are assumed to follow uniform distributions. Maximizing $\log p(\mathbf{R}, \Theta, \mathbf{A}, \mathbf{D}|\mathbf{Y})$ is equivalent to minimizing $\mathcal{E}(\mathbf{R}, \Theta, \mathbf{A}, \mathbf{D})$ as

$$\begin{aligned} \mathcal{E} = & \text{vec}((\mathbf{Y} - \mathbf{AR})^T)^T \Sigma_{\mathbf{Y}}^{-1} \text{vec}((\mathbf{Y} - \mathbf{AR})^T) + \log |\Sigma_{\mathbf{Y}}| \\ & + \beta_1 \text{Tr}(\mathbf{A}^T \mathbf{KA}) + \rho \text{Tr}(\mathbf{R}^T \mathbf{HRD}^{-1}) \end{aligned} \quad (13)$$

where $\Sigma_{\mathbf{Y}}$ is given in (9). Notice that the first term in (13) involves inversion of a large non-sparse NB by NB matrix, which is computationally expensive. We now describe methods to reduce the complexity.

Using the *Woodbury identity* (see Theorem 18.2.8, p. 424 of [23]), $\Sigma_{\mathbf{Y}}^{-1}$ becomes

$$\begin{aligned}
\Sigma_{\mathbf{Y}}^{-1} &= (\mathbf{I}_N \otimes \mathbf{D} + (\mathbf{A} \otimes \mathbf{I}_B)[\delta_{ij}\Sigma_j](\mathbf{A} \otimes \mathbf{I}_B)^T)^{-1} \\
&= (\mathbf{I}_N \otimes \mathbf{D})^{-1} - (\mathbf{I}_N \otimes \mathbf{D})^{-1} (\mathbf{A} \otimes \mathbf{I}_B) \\
&\quad \left\{ [\delta_{ij}\Sigma_j]^{-1} + (\mathbf{A} \otimes \mathbf{I}_B)^T (\mathbf{I}_N \otimes \mathbf{D})^{-1} (\mathbf{A} \otimes \mathbf{I}_B) \right\}^{-1} \\
&\quad (\mathbf{A} \otimes \mathbf{I}_B)^T (\mathbf{I}_N \otimes \mathbf{D})^{-1} \\
&= \mathbf{I}_N \otimes \mathbf{D}^{-1} - \left(\mathbf{A} \otimes \mathbf{D}^{-\frac{1}{2}} \right) \mathbf{Q}^{-1} \left(\mathbf{A} \otimes \mathbf{D}^{-\frac{1}{2}} \right)^T, \tag{14}
\end{aligned}$$

where $\mathbf{Q} \in \mathbb{R}^{MB \times MB}$, $\mathbf{S}_j \in \mathbb{R}^{B \times B}$ are defined as

$$\begin{aligned}
\mathbf{Q} &:= \left[\delta_{ij} \mathbf{D}^{\frac{1}{2}} \Sigma_j^{-1} \mathbf{D}^{\frac{1}{2}} \right] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B \\
&= [\delta_{ij} \mathbf{S}_j] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B, \tag{15}
\end{aligned}$$

and

$$\mathbf{S}_j := \mathbf{D}^{\frac{1}{2}} \Sigma_j^{-1} \mathbf{D}^{\frac{1}{2}}, \quad j = 1, \dots, M. \tag{16}$$

Note that $\mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B$ is a positive semidefinite matrix and therefore $\mathbf{Q} \in \text{SPD}(MB)$ ($\Sigma_j \in \text{SPD}(B)$). Plugging (14) into the first term of the objective function leads to

$$\begin{aligned}
&\text{vec} \left((\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right)^T \Sigma_{\mathbf{Y}}^{-1} \text{vec} \left((\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right) \\
&= \text{vec} \left((\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right)^T (\mathbf{I}_N \otimes \mathbf{D}^{-1}) \text{vec} \left((\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right) \\
&\quad - \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} \\
&= \left\| \mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right\|_F^2 - \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}, \tag{17}
\end{aligned}$$

where $\mathbf{z} \in \mathbb{R}^{MB}$ denotes

$$\begin{aligned}
\mathbf{z} &:= \left(\mathbf{A} \otimes \mathbf{D}^{-\frac{1}{2}} \right)^T \text{vec} \left((\mathbf{Y} - \mathbf{A}\mathbf{R})^T \right) \\
&= \text{vec} \left(\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T \mathbf{A} \right). \tag{18}
\end{aligned}$$

From the *matrix determinant lemma* (see Theorem 18.1.1, p. 416 of [23]), the logarithm term $\log |\Sigma_{\mathbf{Y}}|$ becomes

$$\begin{aligned}
&\log \left| \mathbf{I}_N \otimes \mathbf{D} + (\mathbf{A} \otimes \mathbf{I}_B) [\delta_{ij}\Sigma_j] (\mathbf{A} \otimes \mathbf{I}_B)^T \right| \\
&= \log \left| [\delta_{ij}\Sigma_j]^{-1} + (\mathbf{A} \otimes \mathbf{I}_B)^T (\mathbf{I}_N \otimes \mathbf{D}^{-1}) (\mathbf{A} \otimes \mathbf{I}_B) \right| \\
&\quad |[\delta_{ij}\Sigma_j]| |\mathbf{I}_N \otimes \mathbf{D}| \\
&= \log \left| \mathbf{I}_M \otimes \mathbf{D}^{-\frac{1}{2}} \right|^2 |\mathbf{Q}| |[\delta_{ij}\Sigma_j]| |\mathbf{I}_N \otimes \mathbf{D}| \\
&= \log |\mathbf{Q}| - \log |[\delta_{ij}\mathbf{S}_j]| + N \log |\mathbf{D}|. \tag{19}
\end{aligned}$$

Combining the results in (17) and (19), minimizing (13) becomes equivalent to minimizing $\mathcal{E}_1(\mathbf{R}, \mathbf{A}, \mathbf{D}, \{\mathbf{S}_j\})$ defined as

$$\begin{aligned} \mathcal{E}_1 = & \|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2 - \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} + \log |\mathbf{Q}| \\ & - \log |[\delta_{ij} \mathbf{S}_j]| + N \log |\mathbf{D}| + \beta_1 \text{Tr}(\mathbf{A}^T \mathbf{K} \mathbf{A}) \\ & + \rho \text{Tr}(\mathbf{R}^T \mathbf{H} \mathbf{R} \mathbf{D}^{-1}) \end{aligned} \quad (20)$$

subject to

$$\mathbf{A} \geq 0, \mathbf{A} \mathbf{1}_M = \mathbf{1}_N, \mathbf{R} \geq 0, \mathbf{S}_j \in \text{SPD}(B) \quad (21)$$

where \mathbf{Q} , \mathbf{z} and \mathbf{S}_j are defined in (15), (18) and (16) respectively. Note that letting $\Sigma_j \rightarrow \mathbf{0}$ (i.e. there is little endmember uncertainty) will result in $\mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}$ vanishing as \mathbf{S}_j tends to infinity, and $\log |\mathbf{Q}|$ canceling $\log |[\delta_{ij} \mathbf{S}_j]|$ as $[\delta_{ij} \mathbf{S}_j]$ dominates \mathbf{Q} . For this special case, the entire objective function reduces to the noise weighted least-squares objective.

D. Optimizing the objective function

We show in Appendix A that $\mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}$ and $\log |\mathbf{Q}| - \log |[\delta_{ij} \mathbf{S}_j]|$ are positive and are negligible compared to $\|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2$ when \mathbf{A} , \mathbf{R} , \mathbf{D} are close to the ground truth. This gives us another objective function $\mathcal{E}_2(\mathbf{R}, \mathbf{A}, \mathbf{D})$ that is independent of $\{\mathbf{S}_j\}$:

$$\begin{aligned} \mathcal{E}_2 = & \|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2 + N \log |\mathbf{D}| \\ & + \beta_1 \text{Tr}(\mathbf{A}^T \mathbf{K} \mathbf{A}) + \rho \text{Tr}(\mathbf{R}^T \mathbf{H} \mathbf{R} \mathbf{D}^{-1}), \end{aligned} \quad (22)$$

to be minimized subject to the constraints of \mathbf{A} , \mathbf{R} in (21). We will first minimize (22) with respect to \mathbf{A} , \mathbf{R} , \mathbf{D} and then minimize (20) with respect to \mathbf{D} , $\{\mathbf{S}_j\}$ given the obtained \mathbf{A} , \mathbf{R} . Note that both objectives require optimization over convex sets (\mathbf{A} is restricted to the Cartesian product of simplices, \mathbf{S}_j is restricted to the convex cone of positive definite matrices) and so gradient projection methods can be used to solve these kinds of problems (please see Section 2.3 in [24]).

Possessed with an initial condition, we can alternate between updates of \mathbf{A} , \mathbf{R} and \mathbf{D} to reduce the energy. Taking the derivative of (22) with respect to \mathbf{A} , we have

$$\frac{\partial \mathcal{E}_2}{\partial \mathbf{A}} = 2(-\mathbf{Y} \mathbf{D}^{-1} \mathbf{R}^T + \mathbf{A} \mathbf{R} \mathbf{D}^{-1} \mathbf{R}^T + \beta_1 \mathbf{K} \mathbf{A}). \quad (23)$$

The gradient projection method sets the value of the next iteration, \mathbf{A}^{n+1} , to be the projected value of the steepest descent

$$\mathbf{A}^{n+1} = \phi \left(\mathbf{A}^n - \tau^n \frac{\partial \mathcal{E}_2}{\partial \mathbf{A}} (\mathbf{R}^n, \mathbf{A}^n, \mathbf{D}^n) \right), \quad (24)$$

where

$$\phi : \mathbf{X} \mapsto \arg \min_{\mathbf{Y} \in \mathbb{R}^{N \times M}} \|\mathbf{X} - \mathbf{Y}\|_F^2 \text{ s.t. } \mathbf{Y} \geq 0, \mathbf{Y} \mathbf{1}_M = \mathbf{1}_N$$

projects a matrix to the nearest matrix that satisfies the simplex constraint (e.g. we use the algorithm in Fig. 1 of [25]). $\tau^n > 0$ is the step size and is set by 1D minimization or the familiar Armijo rule (Section 2.2.1 in [24]). It is shown that the sequence generated by (24) is gradient related, i.e. $\langle \frac{\partial \mathcal{E}_2}{\partial \mathbf{A}} (\mathbf{R}^n, \mathbf{A}^n, \mathbf{D}^n), \mathbf{A}^{n+1} - \mathbf{A}^n \rangle < 0$ (Proposition 2.3.1 in [24]), which leads to a stationary point given proper step sizes τ^n such as exact line minimization [26],

$$\tau^n = \arg \min_{\tau \geq 0} \mathcal{E}_2 \left(\mathbf{R}^n, \phi \left(\mathbf{A}^n - \tau \frac{\partial \mathcal{E}_2}{\partial \mathbf{A}} \right), \mathbf{D}^n \right).$$

Numerically, we can use adaptive step sizes that start with a small step and gradually increase it by an order of magnitude until \mathcal{E}_2 starts increasing. Similar gradient descent methods were proposed in [27], [28], and it is shown that such methods have a faster convergence rate than those based on multiplicative update rules [7], [8].

Once we have updated \mathbf{A} , we can update \mathbf{R} by finding a new value that reduces (22). A gradient projection method can also be used for \mathbf{R} because of the positivity constraint. However, the spatial smoothness and sparsity promoting term, $\text{Tr}(\mathbf{A}^T \mathbf{K} \mathbf{A})$, along with the pairwise closeness term actually make \mathbf{R} seldom negative even when just using a closed form solution. Taking the derivative of (22) with respect to \mathbf{R} , we have

$$\frac{\partial \mathcal{E}_2}{\partial \mathbf{R}} = 2 \left(-\mathbf{A}^T \mathbf{Y} \mathbf{D}^{-1} + \mathbf{A}^T \mathbf{A} \mathbf{R} \mathbf{D}^{-1} + \rho \mathbf{H} \mathbf{R} \mathbf{D}^{-1} \right).$$

Letting $\frac{\partial \mathcal{E}_2}{\partial \mathbf{R}} = \mathbf{0}$, we obtain a closed form solution for \mathbf{R} that ignores the positivity constraint,

$$\mathbf{R} = (\mathbf{A}^T \mathbf{A} + \rho \mathbf{H})^{-1} \mathbf{A}^T \mathbf{Y}. \quad (25)$$

Then we need to update \mathbf{D} to further reduce the energy. Considering that $\mathbf{D} := \text{diag}(\mu_1^2, \mu_2^2, \dots, \mu_B^2)$, we take the derivative of (22) with respect to each μ_k and set $\frac{\partial \mathcal{E}_2}{\partial \mu_k} = 0$. The solution turns out to be

$$\mu_k^2 = \frac{1}{N} \left\{ \sum_{i=1}^N (\mathbf{Y} - \mathbf{A} \mathbf{R})_{ik}^2 + \rho (\mathbf{R}^T \mathbf{H} \mathbf{R})_{kk} \right\}, \forall k. \quad (26)$$

Given an initial condition, we can alternate between updates of \mathbf{A} , \mathbf{R} and \mathbf{D} based on (24), (25) and (26) respectively. The details are given in the first two steps of Algorithm 1. Since the energy is lowered at each step, it is assumed that energy minimization leads to convergence.

Note that the estimated \mathbf{D} above is biased. The intuition of $\frac{1}{N} \sum_i (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2$ is straightforward for estimating the noise at the k th band. The additional term $\rho (\mathbf{R}^T \mathbf{H} \mathbf{R})_{kk}$ exists because of our prior in (12), which will prefer larger noise. Hence we should ignore this term by setting $\rho = 0$ for estimating the noise. Also, when it comes to finding \mathbf{D} , $\{\mathbf{S}_j\}$ (hence Σ_j), we should resort to the original objective function (20). Taking the derivative of (20) with respect to \mathbf{D} (μ_k) and setting it to 0 is a little complicated, so we leave the details to Appendix B and merely show the result here. Let $d_k := \mu_k^{-1}$ and $\mathbf{d} := [d_1, \dots, d_B]^T$. The solution for minimizing (20) with respect to \mathbf{d} is a nonlinear system of equations given by

$$\frac{1}{N} (\mathbf{E} - \mathbf{F}^T \mathbf{Q}^{-1} \mathbf{F}) \mathbf{d} = [d_1^{-1}, \dots, d_B^{-1}]^T, \quad (27)$$

where $\mathbf{E} \in \mathbb{R}^{B \times B}$, $\mathbf{F} \in \mathbb{R}^{MB \times B}$ are defined as

$$\mathbf{E} := \text{diag} \left\{ \sum_{i=1}^N ((\mathbf{Y} - \mathbf{A}\mathbf{R}) \circ (\mathbf{Y} - \mathbf{A}\mathbf{R}))_i \right\}, \text{ and} \quad (28)$$

$$\mathbf{F} := [\text{diag} \{ (\mathbf{A}^T (\mathbf{Y} - \mathbf{A}\mathbf{R}))_1 \}, \dots, \text{diag} \{ (\mathbf{A}^T (\mathbf{Y} - \mathbf{A}\mathbf{R}))_M \}]^T. \quad (29)$$

According to Appendix A, when \mathbf{A} , \mathbf{R} are sufficiently accurate, $\mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}$ is small compared to $\|\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2$, hence the term $\mathbf{F}^T \mathbf{Q}^{-1} \mathbf{F}$ can be ignored and $\frac{1}{N} \mathbf{E} \mathbf{d} = [d_1^{-1}, \dots, d_B^{-1}]^T$ is equivalent to (26). However, when \mathbf{A} , \mathbf{R} are not accurate enough, the solution in (26) could be much larger than the true noise, in which case $\mathbf{F}^T \mathbf{Q}^{-1} \mathbf{F}$ slightly compensates for the bias.

Using the chain rule in matrix form to take the derivative of (20) with respect to \mathbf{S}_j , we have

$$\frac{\partial \mathcal{E}_1}{\partial \mathbf{S}_j} = (\mathbf{Q}^{-1} \mathbf{z} \mathbf{z}^T \mathbf{Q}^{-1})_{jj} + (\mathbf{Q}^{-1})_{jj} - \mathbf{S}_j^{-1}, \quad (30)$$

where $(\cdot)_{jj}$ denotes the extraction of the j th diagonal B by B block of the MB by MB matrix. We use alternating updates on \mathbf{D} and \mathbf{S}_j to minimize (20) while keeping \mathbf{A} , \mathbf{R} fixed. To update \mathbf{S}_j , a gradient projection method similar to (24) can be used, where the projection onto the set of positive definite matrices is obtained by truncating the eigenvalues [29]. The details are given in Step 3 of Algorithm 1.

Remark 1. The choice of free parameters should be invariant with respect to the changing magnitudes of each term in (22) for different N , M and B . For example, the first term in (22) has a magnitude of NB . From the banded diagonal nature of $[w_{ij}]$ in (11), $\text{Tr}(\mathbf{A}^T \mathbf{L} \mathbf{A})$

Algorithm 1 The SCM implementation

Input: $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T$, M , β_1 , β_2 , ρ .

- Step 1: Set $\beta_1 \leftarrow \frac{B}{M}\beta_1$, $\beta_2 \leftarrow \frac{B}{M}\beta_2$, $\rho \leftarrow \frac{N}{M^2}\rho$. Construct the matrices \mathbf{K} , \mathbf{H} .¹
- Step 2: Initialize \mathbf{R} to be the centers of M clusters of \mathbf{Y} obtained by fuzzy C-means.² Initialize $\mathbf{A} = \phi\left(\mathbf{Y}\mathbf{R}^T(\mathbf{R}\mathbf{R}^T + 10^{-6}\mathbf{I}_M)^{-1}\right)$, where $\phi: \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{N \times M}$ is the projection function for \mathbf{A} .³ Initialize \mathbf{D} by $\mu_k^2 = \frac{1}{N} \sum_{i=1}^N (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2$. Solve \mathbf{A} , \mathbf{R} by repeating the following three steps until convergence.
 - Update \mathbf{A} by $\phi_A(\tau) := \phi\left(\mathbf{A} - \tau \frac{\partial \mathcal{E}_2}{\partial \mathbf{A}}\right)$, where $\frac{\partial \mathcal{E}_2}{\partial \mathbf{A}}$ is given in (23). If $\mathcal{E}_2(\mathbf{R}, \phi_A(\epsilon), \mathbf{D}) < \mathcal{E}_2(\mathbf{R}, \mathbf{A}, \mathbf{D})$, τ is successively set to $10^i \epsilon$, $i = 0, 1, 2, \dots$ until $\mathcal{E}_2(\mathbf{R}, \phi_A(10^{i+1}\epsilon), \mathbf{D}) \geq \mathcal{E}_2(\mathbf{R}, \phi_A(10^i\epsilon), \mathbf{D})$, otherwise set to zero.
 - Update \mathbf{R} by (25).
 - Update \mathbf{D} by (26).
- Step 3: Initialize $\mu_k^2 = \frac{1}{N} \sum_{i=1}^N (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2$, $\Sigma_j = 0.1^2 \mathbf{I}_B$, $\mathbf{S}_j = \mathbf{D}^{\frac{1}{2}} \Sigma_j^{-1} \mathbf{D}^{\frac{1}{2}}$. Let $\psi: \mathbb{R}^{B \times B} \rightarrow \mathbb{R}^{B \times B}$ be the projection function for \mathbf{S}_j .³ Solve \mathbf{D} , $\{\mathbf{S}_j\}$ by repeating the following two steps until convergence.
 - Update \mathbf{S}_j by $\psi_j(\tau_j) := \psi\left(\mathbf{S}_j - \tau_j \frac{\partial \mathcal{E}_1}{\partial \mathbf{S}_j}\right)$ for $j = 1, \dots, M$, where $\frac{\partial \mathcal{E}_1}{\partial \mathbf{S}_j}$ is given in (30). The step size τ_j is determined in a similar fashion to the \mathbf{A} update in step 2.
 - Update \mathbf{D} by solving (27) numerically with initial condition $d_k = \left(\frac{1}{N} \sum_{i=1}^N (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2\right)^{-\frac{1}{2}}$.

Output: \mathbf{A} , \mathbf{R} , \mathbf{D} , $\Sigma_j = \mathbf{D}^{\frac{1}{2}} \mathbf{S}_j^{-1} \mathbf{D}^{\frac{1}{2}}$.

has a magnitude of NM . So the parameter β_1 should be scaled by $\beta_1 \leftarrow \beta_1 B/M$. Similarly, the parameters β_2 , ρ should be scaled accordingly. When constructing \mathbf{L} , we use $\eta = 0.05$ for calculating $[w_{ij}]$. When constructing $\mathbf{K} = \mathbf{L} - \frac{\beta_2}{\beta_1} \mathbf{I}_N$, we set $\beta_1 \leftarrow \max(\beta_1, 10^{-9})$ to prevent β_1 from becoming zero.

Remark 2. The initial endmembers are important in endmember estimation. Randomly picking pixels as endmembers or deploying fast algorithms such as VCA may provide an initial estimate. We find that fuzzy C-means [30] works well in practical applications. This could be due to the fact that it can cluster the pixels close to the endmembers while being resilient to the highly mixed pixels.

Remark 3. The projection function ϕ for \mathbf{A} is defined as $\phi(\mathbf{A}) = [\max(\alpha_{ij} - \theta_i, 0)]$ where

$\theta_i = \frac{1}{K_i} \left(\sum_{k=1}^{K_i} \alpha'_{ik} - 1 \right)$. $\alpha'_{i1} \geq \dots \geq \alpha'_{iM}$ are sorted versions of $\alpha_{i1}, \dots, \alpha_{iM}$, K_i is the largest k such that $\alpha'_{ik} - \frac{1}{k} \left(\sum_{l=1}^k \alpha'_{il} - 1 \right) > 0$ [25]. The projection function ψ for \mathbf{S}_j is defined as $\psi(\mathbf{X}) = \mathbf{D}^{\frac{1}{2}} \mathbf{U} [\delta_{ij} \min(\max(\lambda_i, \sigma_{\max}^{-2}), \sigma_{\min}^{-2})] \mathbf{U}^T \mathbf{D}^{\frac{1}{2}}$ where $\mathbf{U} [\delta_{ij} \lambda_i] \mathbf{U}^T$ is the eigendecomposition of $\mathbf{D}^{-\frac{1}{2}} \mathbf{X} \mathbf{D}^{-\frac{1}{2}}$ [29]. The eigenvalues of Σ_j are constrained to lie within $[\sigma_{\min}^2, \sigma_{\max}^2]$ with $\sigma_{\min} = 10^{-9}$, $\sigma_{\max} = 1$ so that they cannot become negative and \mathbf{S}_j cannot go to infinity.

III. RESULTS

In the experiments, all the algorithms were implemented in MATLAB[®]. For endmember extraction, we compared the following algorithms:

- (1) SCM is implemented according to Algorithm 1 and publicly available.¹
- (2) An NCM that assumes pixel independence is used in [15], [17], [14] with different optimization approaches where one of the goals was endmember variability. We considered the objective function $-\log p(\mathbf{Y})$ (with $p(\mathbf{Y})$ defined in (4)) with a pairwise closeness constraint on the centers of the Gaussians. We implemented the optimization algorithm using standard projected gradient descent.
- (3) PCOMMEND [6] assumes the endmembers are divided into subsets and hence promotes sparsity. It is shown to work well on the Pavia University dataset.²
- (4) GLNMF treats the pixels as a manifold and constructs a k -neighborhood graph to aid unmixing [7], [8]. It also uses $L_{1/2}$ norm on the abundances to promote sparsity. We implemented it by using the multiplicative update rules in [8].
- (5) WNMF minimizes a combination of the graph based spatial constraint and a least-squares term [12]. The graph is constructed by considering a first order neighborhood with a decreasing function of spectral angular distance as weights. This was re-implemented.

The above implemented algorithms have the same initialization as SCM to ensure a fair comparison (with PCOMMEND directly obtained from the authors). The parameters for all the methods are determined by a greedy algorithm that works as follows. Suppose we have 3 parameters to tune and the step size is (10, 10, 10). With (1, 1, 1) as the starting parameter set, the algorithm searches its 6 neighbors (0.1, 1, 1), (10, 1, 1), (1, 0.1, 1), (1, 10, 1), (1, 1, 0.1), (1, 1, 10) and runs the method to find the parameter set with least error. Then it moves

¹The code is available on GitHub (<https://github.com/zhouyuanxcv/Hyperspectral>).

²The code is available on Alina Zare's PCOMMEND GitHub page (<https://github.com/TigerSense/PCOMMEND/>).

to that point and continues the process until it finds a point that has lower error than all the neighboring points. For the synthetic dataset, we use a step size (10, 10, 10) on a synthetic image with medium noise to determine the optimal parameters for each method. For the two real datasets, we use a step size (2, 2, 2) and round the resulting parameter to the nearest 1, 2, 5 with the same order of magnitude.

Throughout the experiments, we use the mean of absolute difference as the error of abundances or endmembers, i.e.,

$$E_{\mathbf{A}} = \frac{1}{NM} \sum_{i,j} |\alpha_{ij} - \alpha_{ij}^{GT}|, \quad (31)$$

$$E_{\mathbf{m}_j} = \frac{1}{B} \sum_k |m_{jk} - m_{jk}^{GT}|, \quad E_{\mathbf{M}} = \frac{1}{M} \sum_j E_{\mathbf{m}_j}, \quad (32)$$

where α_{ij}^{GT} and m_{jk}^{GT} are the ground truth abundance and endmember values respectively. In the case of real datasets where not all the ground truth abundances are available, $E_{\mathbf{A}}$ is calculated only on the identified pure pixels. The results from algorithms were permuted to match the ground truth for calculating the error.

To measure and visualize the uncertainty from $\{\Sigma_j\}$, recall that the covariance matrix of a Gaussian distribution determines its shape, i.e. the eigenvectors are the directions of the variation patterns while the eigenvalues are the variances of the projected (onto the eigenvectors) 1D data points. The uncertainty can be measured by the largest eigenvalue and its corresponding eigenvector. We use the square root of the largest eigenvalue, σ , as the *uncertainty amount* since it corresponds to the standard deviation. Then, the corresponding eigenvector (normalized), \mathbf{u} , can be viewed as the *uncertainty direction*. Given the estimated endmember \mathbf{r} , the *uncertainty range* can be visualized by two curves $\mathbf{r} \pm 2\sigma\mathbf{u}$.

A. Synthetic images

We first test SCM on synthetic images of size 40 by 40 generated from the true material spectra in the ASTER spectral library [31]. Six rocks—basanite, basalt, quartzite, dacite, rhyolite and limestone are picked as endmembers from the ASTER library in the experiments (with their spectra shown in Fig. 4(a) with wavelengths ranging from $0.4\mu\text{m}$ to $14\mu\text{m}$ and discretized into 200 values). The abundance maps are created randomly in the following way. Suppose basanite is the background material, the other materials are randomly spawned on this background. For

the j th material, 200 blobs are generated and their locations follow a Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{c}_j, \sigma_p^2 \mathbf{I}_2)$ ($\mathbf{c}_j, j = 1, \dots, 4$ are the centers of the 4 quadrants and \mathbf{c}_5 is the center of the image). Each blob imposes an abundance bump with the shape of a Gaussian $\mathcal{N}(\mathbf{a}|\mathbf{x}, s^2 \mathbf{I}_2)$ where the width s is also randomly sampled from a univariate Gaussian distribution $\mathcal{N}(s|1.5, 0.5^2)$ to promote more irregularity. Then the background abundance map is obtained by subtracting the sum of these 5 generated abundances from 1, which usually does not lead to a pure pixel. An example of generated abundance maps is shown in Fig. 4(c). After the signals are created by linearly mixing the endmembers and abundances, we add the noise generated by $\mathcal{N}(\mathbf{n}|\mathbf{0}, \text{diag}(\mu_1^2, \dots, \mu_B^2))$, where each μ_k is again sampled from a uniform distribution on the range $[0, \sigma_Y]$ such that each band has a different noise level. An example of a generated image is shown in Fig. 4(b).

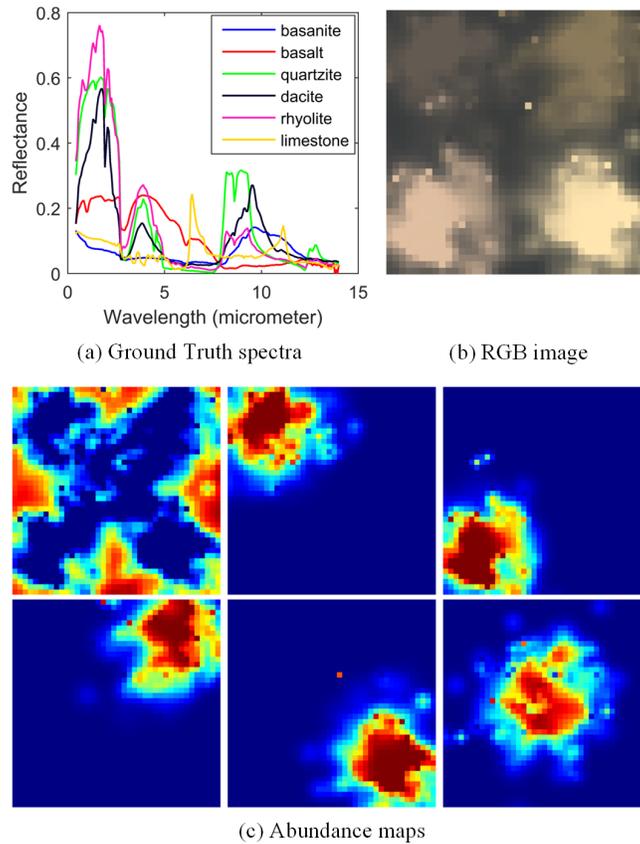


Fig. 4. (a) Spectral signatures of the 6 endmembers used to generate synthetic images. (b) RGB image (using wavelength 488nm for blue, 556nm for green, 693nm for red) from a sample image with $\sigma_Y = 0.0016$. (c) Abundance maps used to generate the image in (b).

For endmember extraction, we compared all the methods based on 10 noise levels, σ_Y ranging

from $\sigma_Y = 0.0001$ to $\sigma_Y = 0.0512$. 20 random images were generated in each case such that the average error (excluding the worst 3 results) can be calculated. The parameters of SCM were $\beta_1 = 10$, $\beta_2 = 10$, $\rho = 0.001$. Fig. 5 shows the errors of all the algorithms. From the plots, we can see that SCM has lower errors than the other methods for all the noise levels, with respect to both endmembers and abundances. The advantage could arise from our noise weighted least-squares term as the synthetic images feature different noise levels at different wavelengths.

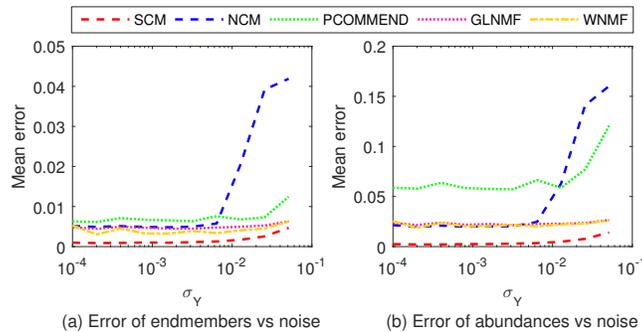


Fig. 5. Error of endmembers and abundances based on the synthetic images. The value at each noise level is the average of errors in (31) and (32) on 20 random images (excluding the worst 3 results).

For uncertainty estimation, we compared the uncertainties of different estimated endmembers for the image in Fig. 4. To achieve this, we changed the value of ρ from large to small gradually. This causes the location of the estimated endmembers to change from being close together inside the pixel cloud to sparsely scattered outside. We average the uncertainty amounts of all endmembers to represent the whole uncertainty. Fig. 6 shows this value along with the error of endmembers versus decreasing ρ . The error of endmembers has its minimum in the middle between 10^{-3} and 10^{-4} . Interestingly, this is also the place where the uncertainty amount starts to decrease to a stable value (nearly 0). This corresponds to the intuition that when the endmembers are outside the pixel cloud (i.e. ρ is small), all the pixels can be well represented by the endmembers thus having a low uncertainty. When the endmembers are inside, the more they are closely packed together, the more the uncertainty as more pixels are beyond their representational capabilities. Recalling our fundamental question about error prediction, the result here implies that it is applicable in terms of amount (when endmembers are close inside).

Fig. 7 shows the uncertainty ranges of close endmembers when $\rho = 0.1$. Fig. 6 implies that the endmembers are actually inside the pixel cloud since ρ is greater than the optimal value. We can see that not only does the uncertainty amount reflect the distance to the ground truth,

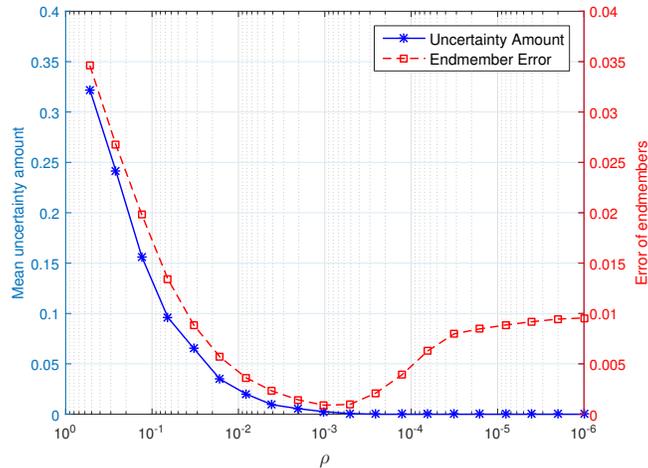


Fig. 6. Effect of pairwise closeness (ρ) on the mean uncertainty amount (blue solid line, scale on the left) and the error of endmembers (red dashed line, scale on the right) for the image in Fig. 4. The minimal error corresponds to the starting point where the uncertainty amount drops to zero.

the uncertainty direction also reflects the distortion of the estimated endmembers. Combining these pieces of information, the uncertainty range is able to cover the ground truth for nearly all of them, except the first material basanite with insufficient uncertainty range to the ground truth, which could be caused by the lack of pure pixels. Therefore the uncertainty estimated can serve as a prediction of the endmember error in this case, given endmembers estimated with a sufficient closeness constraint.

B. Pavia University

The SCM algorithm was applied on the Pavia University dataset, which was recorded by the Reflective Optics System Imaging Spectrometer (ROSIS) during a flight over Pavia, northern Italy. It is a 340 by 610 image with 103 bands with wavelengths ranging from 430nm to 860nm. The real spacing is 1.3 meters/pixel. The image covers both natural and urban areas as shown in Fig. 8(a). There are 9 materials identified as ground truth (shown in Fig. 8(b)). From these pure pixels, average spectra for each material are used as the ground truth endmember signature. Fig. 8(c) shows the ground truth endmembers, from which we find that self-blocking bricks and gravel have very similar spectra, and so do asphalt and bitumen. In this unsupervised unmixing setting, an automated algorithm may distinguish at most 7 endmembers (and indeed in previous research such as [6], only 6 endmembers were estimated).

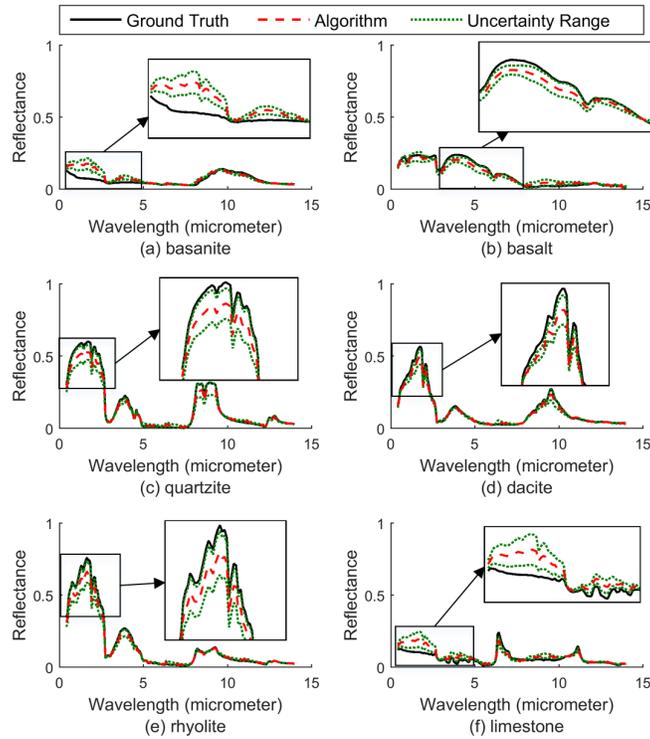


Fig. 7. Uncertainty ranges of endmembers estimated with $\rho = 0.1$ for the image in Fig. 4. The uncertainty ranges nearly cover the ground truth endmembers.

We executed all the methods on this dataset with 7 endmembers (PCOMMEND with 6 endmembers as suggested in [6]). The parameters for SCM are $\beta_1 = 5$, $\beta_2 = 10$, $\rho = 0.05$. Two materials, gravel and bitumen, are excluded in the comparison because they are attributed to self-blocking bricks and asphalt respectively. Fig. 9 shows the abundance maps of all the methods. When compared to the ground truth in Fig. 8(b), we can see that the result of SCM matches the identified pure pixels best. The second best abundance maps are from GLNMF, which could be attributed to its sparsity promoting term. WNMF also has a somewhat nice result despite the fact that some materials, like painted metal sheets, are not obvious. NCM only shows asphalt, trees, painted metal sheets and shadows meaningfully while the others are difficult to identify.

Fig. 10 shows the resulting endmember spectra from all the methods. We also computed the errors for these endmembers and the result is shown in Table I. From these results, we see that SCM performed best overall. The worst estimated endmember of SCM is painted metal sheets and this is due to the complex nature of this dataset wherein objects other than the

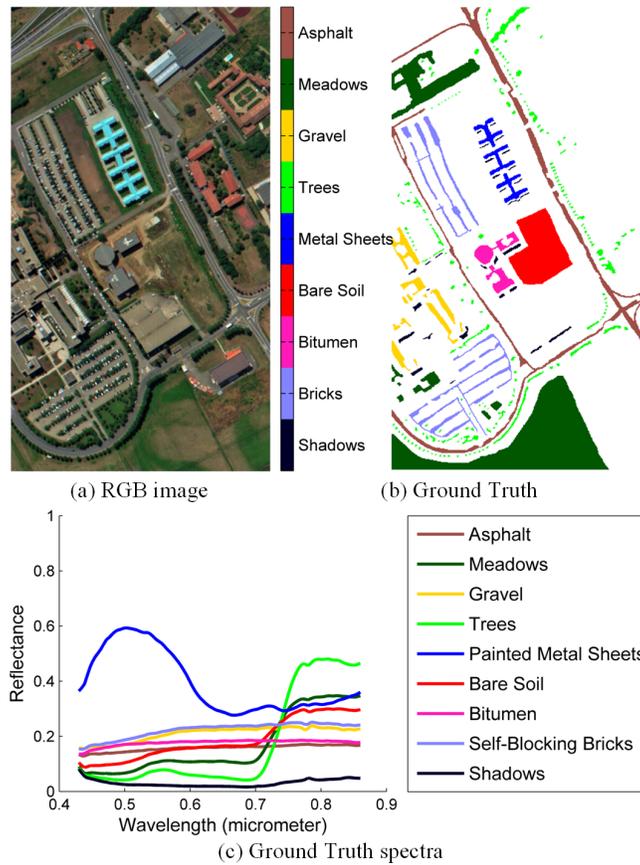


Fig. 8. (a) RGB image for Pavia University and (b) the ground truth pure pixels of 9 materials (each colored region corresponds to a material and the unidentified pixels are left in white). (c) Ground truth spectra obtained by averaging the pure pixels for each endmember. Asphalt and bitumen (gravel and self-blocking bricks) have similar spectral signatures.

identified 9 materials are present. The abundance maps of SCM show that SCM identified painted metal sheets, some sidewalks, and some other types of roof as one endmember, hence mixed their spectra as this endmember. Also, since there are not many pixels of painted metal sheets, the result is strongly biased. Since we are performing unsupervised unmixing, it is difficult to distinguish these additional objects.

The uncertainty ranges of endmembers from SCM are shown in Fig. 11. We see that for the well estimated endmembers, the uncertainties are so small that the endmembers coincide with the uncertainty ranges. For the biased endmember of painted metal sheets, the uncertainty is also large such that it nearly covers the ground truth. For the trees and shadows, the SCM estimated endmembers deviate from the ground truth at the right end and the uncertainty ranges also feature a large gap at the right end.

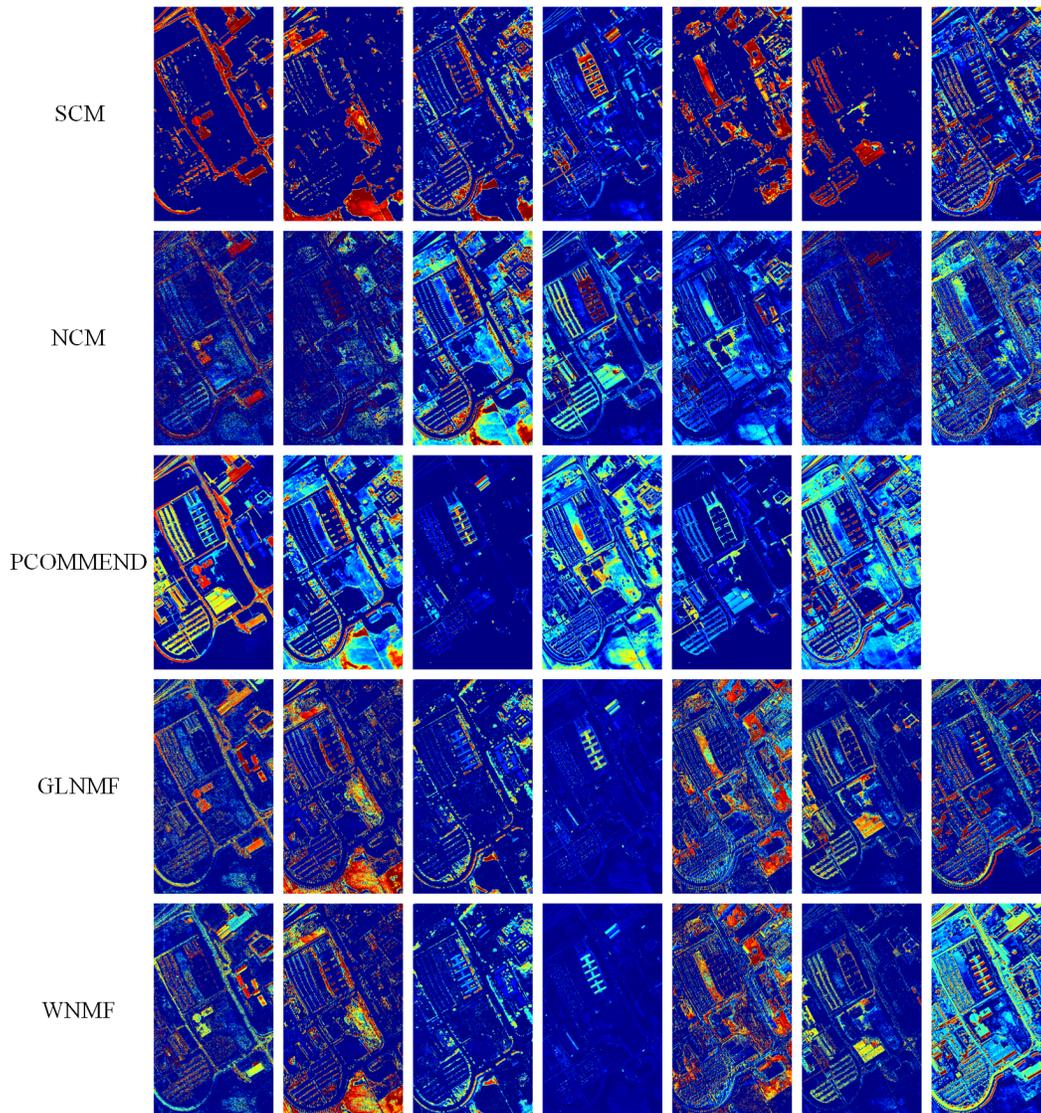


Fig. 9. Abundance maps for Pavia University. The identified materials are asphalt (bitumen), meadows, trees, painted metal sheets, bare soil, self-blocking bricks (gravel), shadows respectively. The abundance errors (calculated by Eq. (31) on those identified pure pixels) for these 5 methods are 0.0864, 0.2257, 0.1533, 0.1219, 0.1499 respectively.

C. Mississippi Gulfport

The dataset was collected over the University of Southern Mississippi's-Gulfpark Campus [32]. It has 72 bands corresponding to wavelengths $0.368\mu m$ to $1.043\mu m$. The spatial resolution is 1 meter/pixel. The scene contains several man-made and natural materials including sidewalks, roads, various types of building roofs, concrete, shrubs, trees, and grasses. We selected the bottom right corner region (185 by 89 pixels) as our dataset, which contains 6 materials. Fig. 12 shows

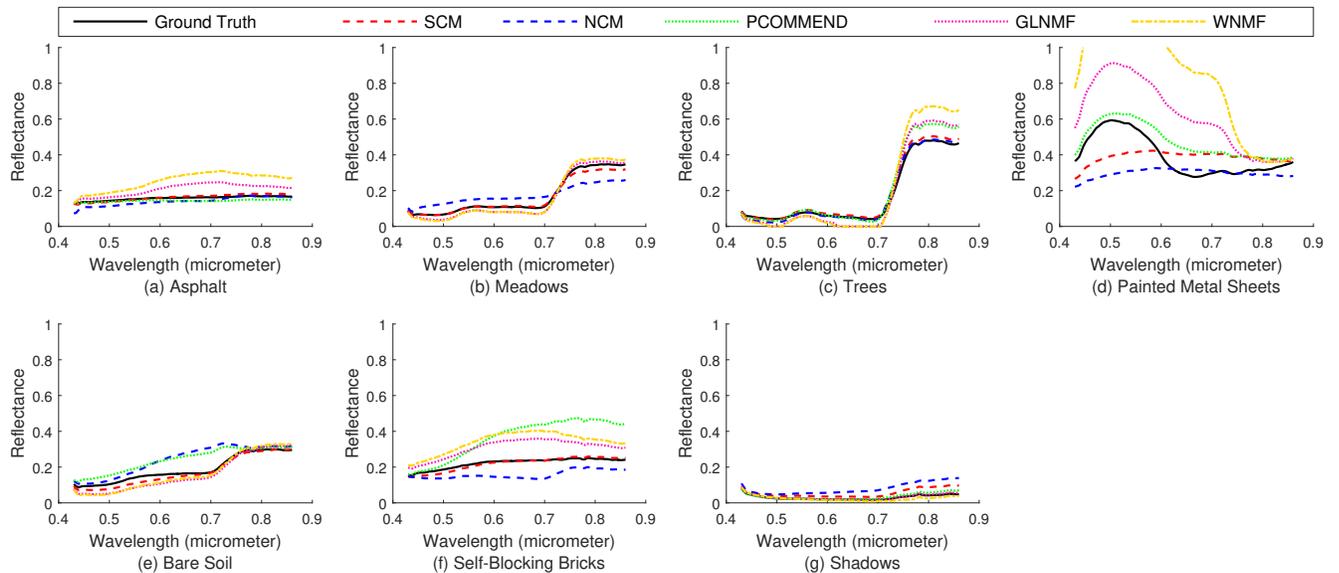


Fig. 10. Qualitative comparison of endmembers for Pavia University.

TABLE I
QUANTITATIVE COMPARISON OF ENDMEMBERS FOR PAVIA UNIVERSITY.

| Error ^a | SCM | NCM | PCM | GLNMF | WNMF |
|--------------------|---------------|---------------|---------------|---------------|--------|
| Asphalt | 0.0078 | 0.0178 | 0.0140 | 0.0494 | 0.0950 |
| Meadows | 0.0092 | 0.0572 | N.A. | 0.0219 | 0.0282 |
| Trees | 0.0142 | 0.0111 | 0.0336 | 0.0523 | 0.0754 |
| Metal Sheets | 0.1024 | 0.1069 | 0.0809 | 0.2274 | 0.4654 |
| Bare Soil | 0.0159 | 0.0555 | 0.0543 | 0.0345 | 0.0337 |
| Bricks | 0.0093 | 0.0647 | 0.1409 | 0.0857 | 0.1191 |
| Shadows | 0.0237 | 0.0486 | 0.0063 | 0.0020 | 0.0082 |
| Average | 0.0261 | 0.0517 | 0.0550 | 0.0676 | 0.1179 |

^a the error was calculated according to Eq. (32) and the minimum value for each row is boldfaced.

its RGB image and ground truth.

The parameters for SCM are $\beta_1 = 10$, $\beta_2 = 50$, $\rho = 0.01$. The abundance maps are shown in Fig. 13. Due to the automatically trained large β_2 , the abundances of SCM show a sharp transition from material to background, which is not usually seen in unmixing scenarios, but is more akin to a segmentation. Note that the dataset has a high spatial resolution (1 meter/pixel) so we could not expect many mixed pixels. Also, the ground truth features many pure pixels, which

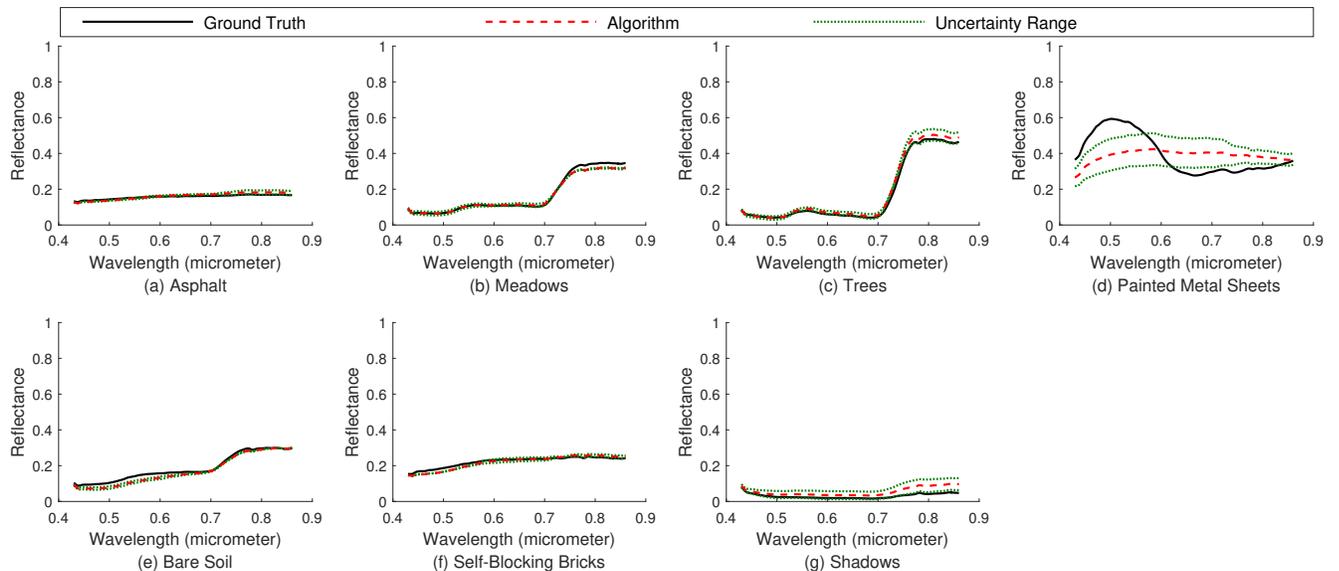


Fig. 11. Uncertainty ranges of endmembers estimated from SCM for Pavia University.

influences the parameters. Compared to the ground truth, SCM still matches it best, followed by GLNMF. The latter also has a sparsity promoting term. The result here indicates that the widely used $L_{1/2}$ sparsity term may not work as well as our integrated smoothness and sparsity prior.

Fig. 14 shows the endmember comparison with quantitative errors in Table II. It can be seen that SCM matches several materials best. The second best is GLNMF with a slightly larger overall endmember error.

TABLE II
QUANTITATIVE COMPARISON OF ENDMEMBERS FOR GULFPORT.

| Error | SCM | NCM | PCM | GLNMF | WNMF |
|------------|---------------|---------------|--------|---------------|--------|
| Dirt | 0.0301 | 0.0174 | 0.0607 | 0.0065 | 0.1227 |
| Asphalt | 0.0057 | 0.0890 | 0.0420 | 0.0093 | 0.0213 |
| Dead Grass | 0.0107 | 0.0672 | 0.0226 | 0.0189 | 0.0305 |
| Grass | 0.0213 | 0.1372 | 0.0515 | 0.0468 | 0.0649 |
| Shadow | 0.0160 | 0.0057 | 0.0064 | 0.0074 | 0.0159 |
| Tree | 0.0029 | 0.1224 | 0.0034 | 0.0084 | 0.0281 |
| Average | 0.0145 | 0.0732 | 0.0311 | 0.0162 | 0.0472 |

The uncertainty ranges of endmembers are shown in Fig. 15. Since all the endmembers are pretty close to the ground truth, all the uncertainties are also small. The only visible uncertainty

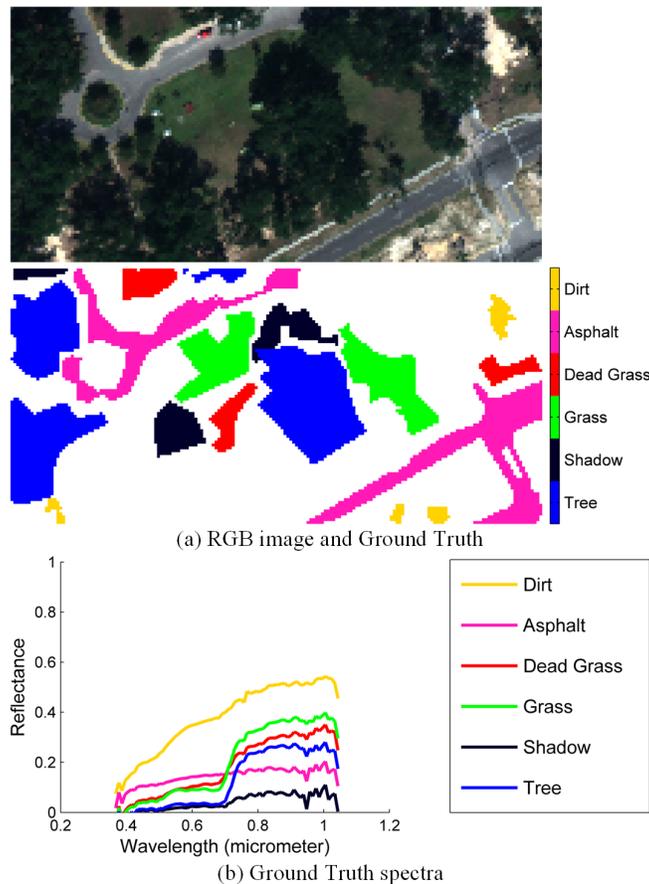


Fig. 12. (a) RGB image and ground truth for the Mississippi Gulfport dataset. (b) Mean spectra of endmembers.

range that is not consistent with the endmember spectra is from dirt, which also features the largest endmember error according to Table II. Note that though SCM also has a relatively large endmember error for grass, the uncertainty is small. This could be due to the fact that the uncertainty captures whether the pixels can be represented by the endmembers. It is possible that the endmembers are far away from the ground truth, but still represent the pixels well (e.g. see Fig. 6: at a small value of ρ , the endmembers are loosely expanded, thus away from the ground truth, but since they still fit the pixels well, the uncertainty is small).

IV. DISCUSSION AND CONCLUSION

In this paper we have presented a spatial compositional model for linear unmixing of hyperspectral images. We use a smoothness and sparsity promoting prior on the abundances, and assume the pixels are not independent to estimate not only the endmembers and abundances,

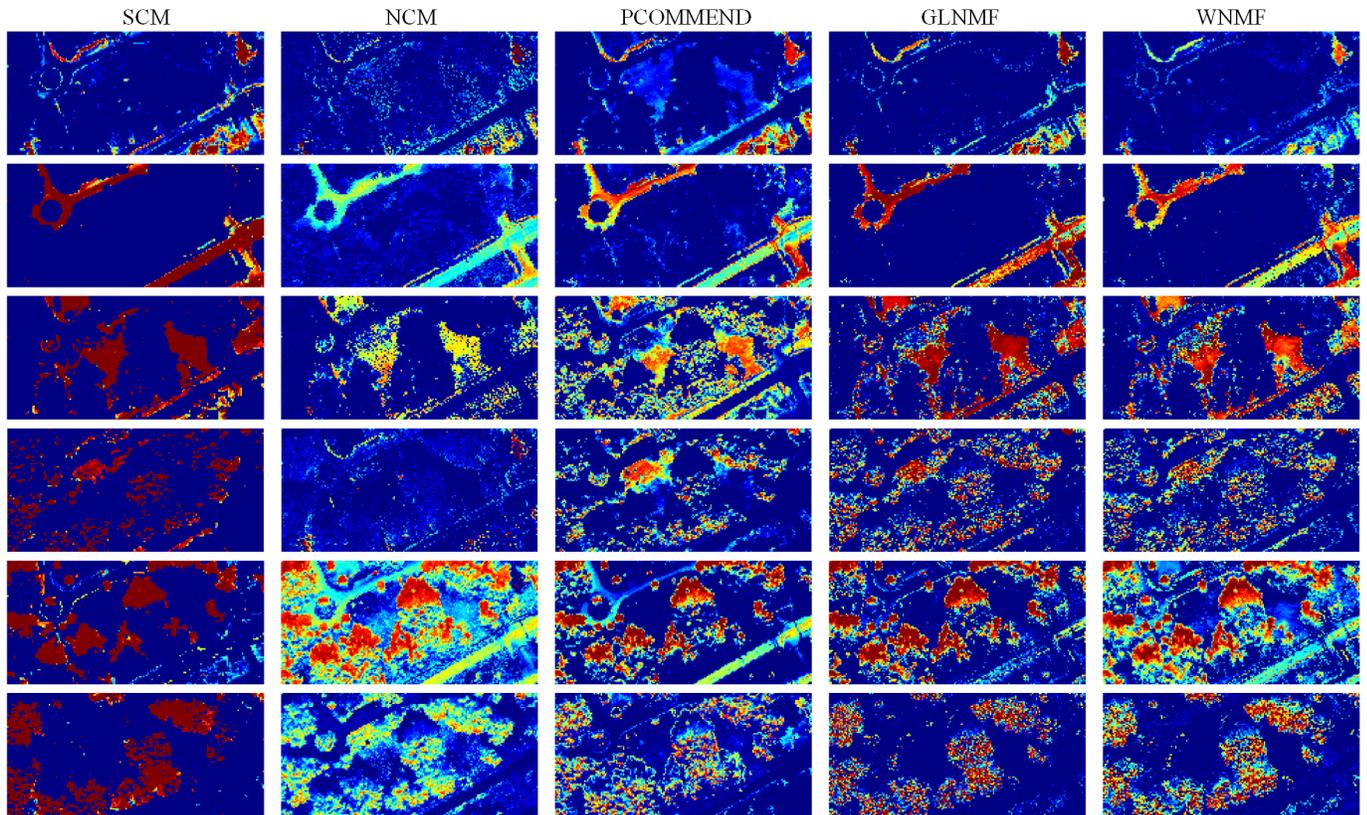


Fig. 13. Abundance maps for the Gulfport dataset. The abundance errors for these 5 methods are 0.1003, 0.1991, 0.1564, 0.1252, 0.1533 respectively.

but also noise variances and endmember uncertainties. The results on synthetic and real datasets show that the estimated endmembers are usually more accurate than the competing methods, and the uncertainty may serve as an error prediction under certain conditions. Finally we discuss some fundamental issues.

1. We model the noise as having different variances at different wavelengths and estimate the noise simultaneously with the abundances and endmembers. However, there is a subtle difference between our estimated noise and the noise estimated in [33], [34], [35]. Due to spectral variability, the LMM in (2) may not hold for real datasets. To be specific, the endmembers that are used to generate the pixels may vary per pixel. This is caused by many factors such as topography, atmosphere, and multiple scattering in canopies (e.g. the painted metal sheets in the Pavia dataset present different reflectances because of the varying angle of placement of these sheets) [19]. Our mathematical development shows that the noise becomes the residual error between the reconstructed signal and the true signal. This is correct if the LMM holds. Since it may not

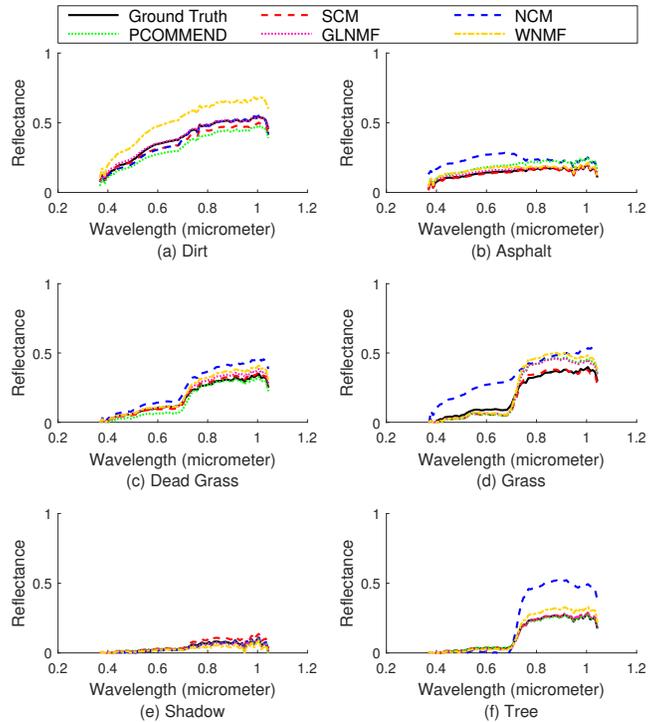


Fig. 14. Qualitative comparison of endmembers for Gulfport.

hold in real scenarios, the noise estimated in this paper is actually not only the noise, but also contains spectral variability.

2. This introduces an issue of system identification. Can we distinguish between endmember variability and noise? If we only know the number of endmembers without other information, the answer is no. Taking the SCM as an example, when the parameters β_1 and β_2 are tuned differently, the solution obtained by minimizing the objective function will be different. Without access to the ground truth, we don't know which solution is better, because one solution may interpret the image as having large endmember variability while another solution may interpret the image as having small endmember variability. And we can not assume that the solution with less endmember variability (i.e. lower estimated noise) is better. Hence, there is no general way to determine these parameters in the absence of application backgrounds (though ρ may be determined by the estimated uncertainty according to Fig. 6), and we recommend training them based on some ground truth information.

3. The uncertainty encoded by the covariance matrix shows that its range can predict the error when the estimated endmembers are inside the pixel cloud. This is determined by the pixel

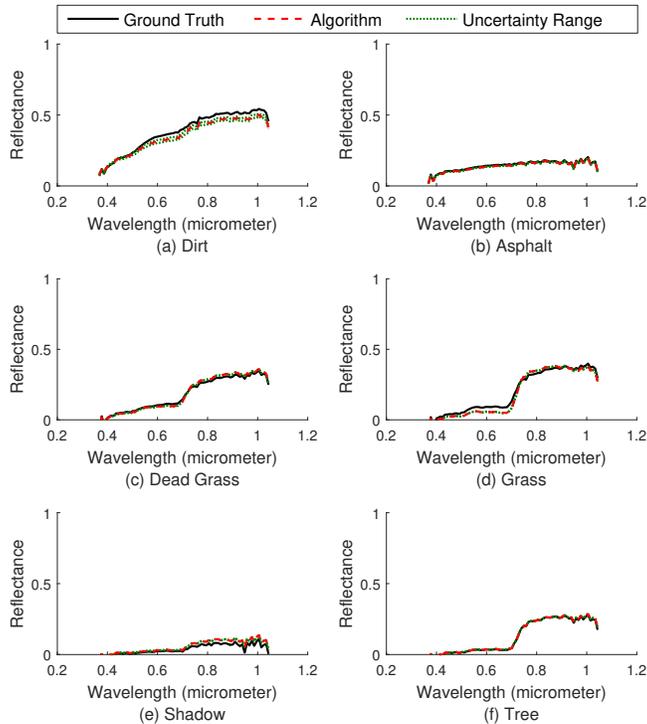


Fig. 15. Uncertainty ranges of endmembers estimated from SCM for Gulfport.

representational capability of the endmembers, i.e. how well do the pixels fit into the simplex formed by the endmembers. Hence, it will fail to predict the error in some cases. One case is when the endmembers are widely expanded such that the simplex is quite large (e.g. the endmembers $\{\mathbf{m}_j''\}$ in Fig. 2 could have a small uncertainty but if $\{\mathbf{m}_j\}$ is the ground truth, the uncertainty does not reflect the error). Another case is if pure pixels do not exist, the uncertainty could at best infer the smallest simplex that represents the pixels, instead of predicting the error (e.g. Fig. 1(b) where the abundances do not span the full range).

In future work, we plan to revisit issues ranging from model selection to endmember variability modeling.

APPENDIX A

RATIONALE FOR THE APPROXIMATION IN (22)

We show that the objective function (20) can be approximated by (22) for minimization with respect to \mathbf{A} , \mathbf{R} in this Appendix, i.e. we show that

$$0 < \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} \ll \|\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{AR})^T\|_F^2,$$

$$0 < \log |\mathbf{Q}| - \log |[\delta_{ij}\mathbf{S}_j]| \ll \|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{AR})^T\|_F^2,$$

assuming \mathbf{A} , \mathbf{R} , \mathbf{D} are close to the ground truth. To be specific, both $\mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}$ and $\log |\mathbf{Q}| - \log |[\delta_{ij}\mathbf{S}_j]|$ are $O(MB)$ while $\|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{AR})^T\|_F^2$ is $O(NB)$.

We first show that $\mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}$ (positive because $\mathbf{Q} \in \text{SPD}(MB)$) is negligible compared to $\|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{AR})^T\|_F^2$. Assume $\mathbf{A}^T \mathbf{A}$ is nonsingular (hence $\mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B \in \text{SPD}(MB)$), from the inequality in Lemma 1 (given at the end of this Appendix), we have

$$\begin{aligned} \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} &= \mathbf{z}^T \left([\delta_{ij}\mathbf{S}_j] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B \right)^{-1} \mathbf{z} \\ &< \mathbf{z}^T \left(\mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B \right)^{-1} \mathbf{z} \\ &= \mathbf{z}^T \left\{ (\mathbf{V}\mathbf{\Lambda} \otimes \mathbf{I}_B) (\mathbf{\Lambda}\mathbf{V}^T \otimes \mathbf{I}_B) \right\}^{-1} \mathbf{z} \\ &= \mathbf{z}^T \left((\mathbf{V}\mathbf{\Lambda})^{-T} \otimes \mathbf{I}_B \right) \left((\mathbf{V}\mathbf{\Lambda})^{-1} \otimes \mathbf{I}_B \right) \mathbf{z} \\ &= \left\| \left((\mathbf{V}\mathbf{\Lambda})^{-1} \otimes \mathbf{I}_B \right) \mathbf{z} \right\|^2 \\ &= \left\| \mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{AR})^T \mathbf{A} (\mathbf{V}\mathbf{\Lambda})^{-T} \right\|_F^2 \\ &= \left\| \mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{AR})^T \mathbf{U} \right\|_F^2 \end{aligned}$$

where $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, $\mathbf{U} \in \mathbb{R}^{N \times M}$, $\mathbf{\Lambda} \in \mathbb{R}^{M \times M}$, $\mathbf{V} \in \mathbb{R}^{M \times M}$ is the compact *singular value decomposition* (SVD) of \mathbf{A} . Since \mathbf{U}^T is part of an orthogonal matrix, $(\mathbf{Y} - \mathbf{AR})^T \mathbf{U}$ can be seen as rotating the columns of $\mathbf{Y} - \mathbf{AR}$ and picking only M elements of the rotated vectors. This is trivial compared to $\mathbf{Y} - \mathbf{AR}$ which has N elements for each column (when \mathbf{A} , \mathbf{R} are sufficiently accurate, $\mathbf{Y} - \mathbf{AR}$ contains only the white noise sampled from an independent Gaussian, which is uncorrelated with \mathbf{U}).

Second, we can show that $\log |\mathbf{Q}| - \log |[\delta_{ij}\mathbf{S}_j]| > 0$ and it is also negligible compared to $\|\mathbf{D}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{AR})^T\|_F^2$. The positivity arises from *Weyl's inequality* (Theorem 4.3.1 in [36]) as the eigenvalues of \mathbf{Q} are greater than those of $[\delta_{ij}\mathbf{S}_j]$. Note that

$$\begin{aligned} \log |\mathbf{Q}| &= \log \left| \left[\delta_{ij} \mathbf{D}^{\frac{1}{2}} \mathbf{\Sigma}_j^{-1} \mathbf{D}^{\frac{1}{2}} \right] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{I}_B \right| \\ &= \log \left| \mathbf{I}_M \otimes \mathbf{D}^{\frac{1}{2}} \right|^2 \left| [\delta_{ij} \mathbf{\Sigma}_j^{-1}] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{D}^{-1} \right| \\ &= M \log |\mathbf{D}| + \log \left| [\delta_{ij} \mathbf{\Sigma}_j^{-1}] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{D}^{-1} \right|, \end{aligned}$$

and

$$\begin{aligned}
\log |[\delta_{ij}\mathbf{S}_j]| &= \log \left| \left[\delta_{ij} \mathbf{D}^{\frac{1}{2}} \boldsymbol{\Sigma}_j^{-1} \mathbf{D}^{\frac{1}{2}} \right] \right| \\
&= \log \left| \left(\mathbf{I}_M \otimes \mathbf{D}^{\frac{1}{2}} \right) \left([\delta_{ij} \boldsymbol{\Sigma}_j^{-1}] \right) \left(\mathbf{I}_M \otimes \mathbf{D}^{\frac{1}{2}} \right) \right| \\
&= M \log |\mathbf{D}| + \log |[\delta_{ij} \boldsymbol{\Sigma}_j^{-1}]|.
\end{aligned}$$

Let $\sigma_{j1}, \dots, \sigma_{jB}$ be the the eigenvalues of $\boldsymbol{\Sigma}_j^{-1}$ in ascending order and $\lambda_1, \dots, \lambda_M$ be the eigenvalues of $\mathbf{A}^T \mathbf{A}$ in ascending order (which implies the eigenvalues of $\mathbf{A}^T \mathbf{A} \otimes \mathbf{D}^{-1}$ are λ_j / μ_k^2 , $j = 1, \dots, M$, $k = 1, \dots, B$). Let $l = \arg \min_k \mu_k$. The sum of the above two expansions leads to

$$\begin{aligned}
&\log |\mathbf{Q}| - \log |[\delta_{ij}\mathbf{S}_j]| \\
&= \log |[\delta_{ij} \boldsymbol{\Sigma}_j^{-1}] + \mathbf{A}^T \mathbf{A} \otimes \mathbf{D}^{-1}| - \log |[\delta_{ij} \boldsymbol{\Sigma}_j^{-1}]| \\
&\leq \log \prod_{j=1}^M \prod_{k=1}^B (\sigma_{jk} + \lambda_M / \mu_l^2) - \log \prod_{j=1}^M \prod_{k=1}^B \sigma_{jk} \\
&= \sum_{j=1}^M \sum_{k=1}^B \log (1 + \lambda_M / (\sigma_{jk} \mu_l^2))
\end{aligned}$$

where Weyl's inequality for eigenvalues is again used. Given that the reflectances of the end-member signatures are bounded from above, the endmember covariance matrix should have σ_{jk} bounded from below. Assuming that μ_l cannot be arbitrarily close to 0, we have $\lambda_M / (\sigma_{jk} \mu_l^2)$ bounded from above. Although the interior of the logarithm could be pretty large, the logarithm makes it small. Compared to $\|\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2 \approx NB$ (by $\mu_k^2 \approx \frac{1}{N} \sum_i (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2$, when \mathbf{A} , \mathbf{R} are sufficiently accurate), $\sum_{j,k} \log (1 + \lambda_M / (\sigma_{jk} \mu_l^2))$ is negligible since $M \ll N$.

Lemma 1. *Let $\mathbf{A} \in \text{SPD}(n)$, $\mathbf{B} \in \text{SPD}(n)$, then for any nonzero $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T (\mathbf{A} + \mathbf{B})^{-1} \mathbf{x} < \mathbf{x}^T \mathbf{A}^{-1} \mathbf{x}$.*

Proof: Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$, $\mathbf{B} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ be the eigendecompositions of \mathbf{A} and \mathbf{B} respectively.

Then

$$\mathbf{x}^T \mathbf{A}^{-1} \mathbf{x} = \mathbf{x}^T (\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T)^{-1} \mathbf{x} = \mathbf{y}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

where $\mathbf{y} := \mathbf{U}^T \mathbf{x}$, while

$$\begin{aligned}
\mathbf{x}^T (\mathbf{A} + \mathbf{B})^{-1} \mathbf{x} &= \mathbf{x}^T (\mathbf{U} (\boldsymbol{\Sigma} + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T) \mathbf{U}^T)^{-1} \mathbf{x} \\
&= \mathbf{y}^T (\boldsymbol{\Sigma} + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T)^{-1} \mathbf{y}
\end{aligned}$$

where $\mathbf{Q} := \mathbf{U}^T \mathbf{V}$. By the Woodbury identity,

$$(\boldsymbol{\Sigma} + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T)^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Q}\mathbf{C}\mathbf{Q}^T\boldsymbol{\Sigma}^{-1}$$

where $\mathbf{C} := (\boldsymbol{\Lambda}^{-1} + \mathbf{Q}^T\boldsymbol{\Sigma}^{-1}\mathbf{Q})^{-1}$. We have

$$\mathbf{y}^T (\boldsymbol{\Sigma} + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T)^{-1} \mathbf{y} = \mathbf{y}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} - \mathbf{z}^T \mathbf{C} \mathbf{z}$$

where $\mathbf{z} := \mathbf{Q}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$. Because $\mathbf{C} \in \text{SPD}(n)$ (since $\boldsymbol{\Lambda}^{-1} \in \text{SPD}(n)$ and $\mathbf{Q}^T \boldsymbol{\Sigma}^{-1} \mathbf{Q} \in \text{SPD}(n)$) and \mathbf{z} is nonzero, $\mathbf{z}^T \mathbf{C} \mathbf{z} > 0$. Then we have $\mathbf{x}^T (\mathbf{A} + \mathbf{B})^{-1} \mathbf{x} < \mathbf{x}^T \mathbf{A}^{-1} \mathbf{x}$. ■

APPENDIX B

CLOSED FORM SOLUTION FOR \mathbf{D} IN (20)

We derive the closed form solution by taking the derivative of (20) with respect to \mathbf{D} and setting it to zero. As in Section II-D, \mathbf{S}_j is treated as a separate variable. Also, we set $\rho = 0$ as estimating the noise does not require the prior of endmembers. So we will only focus on

$$\mathcal{E}'_1 = \|\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2 - \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} + N \log |\mathbf{D}|.$$

To facilitate the process, we denote $d_k := \mu_k^{-1}$, $\mathbf{d} := [d_1, \dots, d_B]^T$ (hence $\mathbf{D} = \text{diag}(\mathbf{d} \circ \mathbf{d})^{-1}$) and calculate the derivative with respect to \mathbf{d} .

We start with the two simple terms. Since

$$\begin{aligned} \mathcal{E}''_1 &= \|\mathbf{D}^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T\|_F^2 + N \log |\mathbf{D}| \\ &= \text{Tr} \left\{ \mathbf{D}^{-1} (\mathbf{Y} - \mathbf{A}\mathbf{R})^T (\mathbf{Y} - \mathbf{A}\mathbf{R}) \right\} + N \log |\mathbf{D}| \\ &= \sum_{k=1}^B d_k^2 \left(\sum_i (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2 \right) - 2N \sum_{k=1}^B \log d_k, \end{aligned}$$

the derivative with respect to d_k can be obtained by

$$\frac{\partial \mathcal{E}''_1}{\partial d_k} = 2 \left(\sum_i (\mathbf{Y} - \mathbf{A}\mathbf{R})_{ik}^2 \right) d_k - 2N d_k^{-1},$$

which can be organized as

$$\frac{\partial \mathcal{E}''_1}{\partial \mathbf{d}} = 2\mathbf{E}\mathbf{d} - 2N [d_1^{-1}, \dots, d_B^{-1}]^T, \quad (33)$$

with $\mathbf{E} \in \mathbb{R}^{B \times B}$ defined in (28).

Then we move to the difficult part—calculating $\frac{\partial \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}}{\partial \mathbf{d}}$ by considering \mathbf{z} defined in (18). Let $\mathbf{Q}^{-1} =: [\mathbf{C}_{ij}]$ with $\mathbf{C}_{ij} \in \mathbb{R}^{B \times B}$, $(\mathbf{Y} - \mathbf{A}\mathbf{R})^T \mathbf{A} =: [\mathbf{x}_1, \dots, \mathbf{x}_M]$ with $\mathbf{x}_j =: [x_{j1}, \dots, x_{jB}]^T$. We have $\mathbf{D}^{-\frac{1}{2}} \mathbf{x}_j = [d_1 x_{j1}, \dots, d_B x_{jB}]^T$ and

$$\begin{aligned} \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z} &= \text{vec} \left(\left[\mathbf{D}^{-\frac{1}{2}} \mathbf{x}_1, \dots, \mathbf{D}^{-\frac{1}{2}} \mathbf{x}_M \right] \right)^T [\mathbf{C}_{ij}] \\ &\quad \text{vec} \left(\left[\mathbf{D}^{-\frac{1}{2}} \mathbf{x}_1, \dots, \mathbf{D}^{-\frac{1}{2}} \mathbf{x}_M \right] \right) \\ &= \sum_{i=1}^M \sum_{j=1}^M \left(\mathbf{D}^{-\frac{1}{2}} \mathbf{x}_i \right)^T \mathbf{C}_{ij} \left(\mathbf{D}^{-\frac{1}{2}} \mathbf{x}_j \right) \\ &= \sum_{i=1}^M \sum_{j=1}^M \left\{ \sum_{k=1}^B \sum_{l=1}^B (\mathbf{C}_{ij})_{kl} d_k x_{ik} d_l x_{jl} \right\} \\ &= \sum_{i=1}^M \sum_{j=1}^M \mathbf{d}^T \text{diag}(\mathbf{x}_i) \mathbf{C}_{ij} \text{diag}(\mathbf{x}_j) \mathbf{d}. \end{aligned}$$

Hence the derivative can be obtained as

$$\begin{aligned} \frac{\partial \mathbf{z}^T \mathbf{Q}^{-1} \mathbf{z}}{\partial \mathbf{d}} &= \sum_{i=1}^M \sum_{j=1}^M \left\{ \text{diag}(\mathbf{x}_i) \mathbf{C}_{ij} \text{diag}(\mathbf{x}_j) \mathbf{d} \right. \\ &\quad \left. + \text{diag}(\mathbf{x}_j) \mathbf{C}_{ij}^T \text{diag}(\mathbf{x}_i) \mathbf{d} \right\} \\ &= \mathbf{F}^T [\mathbf{C}_{ij}] \mathbf{F} \mathbf{d} + \mathbf{F}^T [\mathbf{C}_{ij}]^T \mathbf{F} \mathbf{d} \\ &= 2\mathbf{F}^T \mathbf{Q}^{-1} \mathbf{F} \mathbf{d}, \end{aligned} \tag{34}$$

where $\mathbf{F} \in \mathbb{R}^{MB \times B}$ is defined in (29), and we use the symmetry of \mathbf{Q}^{-1} .

Combining (33) and (34), we have the derivative

$$\frac{\partial \mathcal{E}'_1}{\partial \mathbf{d}} = 2\mathbf{E} \mathbf{d} - 2N [d_1^{-1}, \dots, d_B^{-1}]^T - 2\mathbf{F}^T \mathbf{Q}^{-1} \mathbf{F} \mathbf{d},$$

which leads to the solution in (27) by setting $\frac{\partial \mathcal{E}'_1}{\partial \mathbf{d}} = \mathbf{0}$.

ACKNOWLEDGMENTS

We wish to thank Jose Bioucas-Dias and Alina Zare for useful conversations and the anonymous reviewers for very helpful comments.

REFERENCES

- [1] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.

- [2] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. D. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.
- [3] L. Miao and H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 3, pp. 765–777, 2007.
- [4] J. M. Nascimento and J. M. Bioucas Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.
- [5] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. F. Huntington, "ICE: A statistical approach to identifying endmembers in hyperspectral images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, no. 10, pp. 2085–2095, 2004.
- [6] A. Zare, P. D. Gader, O. Bchir, and H. Frigui, "Piecewise convex multiple-model endmember detection and spectral unmixing," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2853–2862, 2013.
- [7] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.
- [8] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, "Manifold regularized sparse NMF for hyperspectral unmixing," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2815–2826, 2013.
- [9] S. Jia and Y. Qian, "Constrained nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 161–173, 2009.
- [10] O. Eches, N. Dobigeon, and J.-Y. Tourneret, "Enhancing hyperspectral image unmixing with spatial correlations," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4239–4247, 2011.
- [11] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Total variation spatial regularization for sparse hyperspectral unmixing," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4484–4502, 2012.
- [12] J. Liu, J. Zhang, Y. Gao, C. Zhang, and Z. Li, "Enhancing spectral unmixing by local neighborhood weights," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 5, pp. 1545–1552, 2012.
- [13] F. Zhu, Y. Wang, B. Fan, S. Xiang, G. Meng, and C. Pan, "Spectral unmixing via data-guided sparsity," *IEEE Trans. on Image Processing*, vol. 23, no. 12, pp. 5412–5427, 2014.
- [14] D. Stein, "Application of the normal compositional model to the analysis of hyperspectral imagery," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, 2003, pp. 44–51.
- [15] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model: Application to hyperspectral imagery," *IEEE Trans. on Image Processing*, vol. 19, no. 6, pp. 1403–1413, 2010.
- [16] O. Eches, N. Dobigeon, and J.-Y. Tourneret, "Estimating the number of endmembers in hyperspectral images using the normal compositional model and a hierarchical Bayesian algorithm," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 582–591, 2010.
- [17] B. Zhang, L. Zhuang, L. Gao, W. Luo, Q. Ran, and Q. Du, "PSO-EM: A hyperspectral unmixing algorithm based on normal compositional model," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7782–7792, 2014.
- [18] A. Zare, P. Gader, and G. Casella, "Sampling piecewise convex unmixing and endmember extraction," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 51, no. 3, pp. 1655–1665, 2013.
- [19] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 95–104, 2014.
- [20] Y. Zhou, A. Rangarajan, and P. D. Gader, "A spatial compositional model for linear unmixing," in *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2015.

- [21] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [22] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via $L_{1/2}$ sparsity-constrained nonnegative matrix factorization," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4282–4297, 2011.
- [23] D. A. Harville, *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York, 1997.
- [24] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.
- [25] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l_1 -ball for learning in high dimensions," in *Proc. of the 25th International Conference on Machine Learning (ICML)*. ACM, 2008, pp. 272–279.
- [26] W. W. Hager and S. Park, "The gradient projection method with exact line search," *Journal of Global Optimization*, vol. 30, no. 1, pp. 103–118, 2004.
- [27] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. on Image Processing*, vol. 20, no. 7, pp. 2030–2048, 2011.
- [28] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [29] P.-A. Absil and J. Malick, "Projection-like retractions on matrix manifolds," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 135–158, 2012.
- [30] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1981.
- [31] A. Baldridge, S. Hook, C. Grove, and G. Rivera, "The ASTER spectral library version 2.0," *Remote Sensing of Environment*, vol. 113, no. 4, pp. 711–715, 2009.
- [32] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "MUUFL Gulfport hyperspectral and LiDAR airborne data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, 2013.
- [33] L. Gao, Q. Du, B. Zhang, W. Yang, and Y. Wu, "A comparative study on linear regression-based noise estimation for hyperspectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 488–498, 2013.
- [34] R. Roger, "Principal components transform with simple, automatic noise adjustment," *International Journal of Remote Sensing*, vol. 17, no. 14, pp. 2719–2727, 1996.
- [35] R. Roger and J. Arnold, "Reliably estimating the noise in AVIRIS hyperspectral images," *International Journal of Remote Sensing*, vol. 17, no. 10, pp. 1951–1962, 1996.
- [36] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 1985.



Yuan Zhou received the B.E degree in Software Engineering (2008), the M.E. degree in Computer Application Technology (2011), both from Huazhong University of Science and Technology, Wuhan, Hubei, China. Then he worked in Shanghai UIH as a software engineer for two years. Since 2013, he has been a Ph.D. student in the Department of CISE, University of Florida, Gainesville, FL, USA. His research interests include image processing, computer vision and machine learning.



of consciousness.

Anand Rangarajan received the B.Tech degree from the Indian Institute of Technology, Madras, India and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA. He is an Associate Professor in the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA. Prior to this, he was an Assistant Professor in the Departments of Diagnostic Radiology and Electrical Engineering, Yale University, New Haven, CT, USA. His research interests include machine learning, computer vision, medical and hyperspectral imaging, and the science



Paul Gader (M'86–SM'09–F'11) received the Ph.D. degree in mathematics for image-processing-related research from the University of Florida, Gainesville, FL, USA, in 1986. He was a Senior Research Scientist with Honeywell, a Research Engineer and a Manager with the Environmental Research Institute of Michigan, Ann Arbor, MI, USA, and a Faculty Member with the University of Wisconsin, Oshkosh, WI, USA, the University of Missouri, Columbia, MO, USA, and the University of Florida, FL, USA, where he is currently a Professor of Computer and Information Science and Engineering. He performed his first research in image processing in 1984 working on algorithms for the detection of bridges in forward-looking infrared imagery as a Summer Student Fellow at Eglin Air Force Base. He has since worked on a wide variety of theoretical and applied research problems including fast computing with linear algebra, mathematical morphology, fuzzy sets, Bayesian methods, handwriting recognition, automatic target recognition, biomedical image analysis, landmine detection, human geography, and hyperspectral and light detection, and ranging image analysis projects. He has authored/co-authored hundreds of refereed journal and conference papers.