CONVEX GEOMETRY BASED ESTIMATION OF NUMBER OF ENDMEMBERS IN HYPERSPECTRAL IMAGES

ArulMurugan Ambikapathi, Tsung-Han Chan, and Chong-Yung Chi

Inst. Commun. Eng., National Tsing Hua Univ., Hsinchu, Taiwan

E-mail: aareul@ieee.org,(tsunghan@mx,cychi@ee).nthu.edu.tw

ABSTRACT

Hyperspectral unmixing is a process of decomposing the hyperspectral data cube into endmember signatures and their corresponding abundance maps. For the unmixing results to be completely interpretable, the number of materials (or endmembers) present in that area should be known *a priori*, which however is unknown in practice. In this work, we use hyperspectral data geometry and successive endmember estimation strategy of an endmember extraction algorithm (EEA) to develop two novel algorithms for estimating the number of endmembers, namely geometry based estimation of number of endmembers - convex hull (GENE-CH) algorithm and affine hull (GENE-AH) algorithm. The proposed GENE algorithms estimate the number of endmembers by using Neyman-Pearson hypothesis testing over the endmembers sequentially estimated by an EEA until the estimate of the number of endmembers is obtained. Monte-Carlo simulations demonstrate the efficacy of the proposed GENE algorithms, compared to some existing benchmark methods for estimating number of endmembers.

***Index Terms*—** Hyperspectral unmixing, Successive endmember extraction, Estimation of number of endmembers, Neyman-Pearson hypothesis testing

1. INTRODUCTION

Hyperspectral unmixing (HU) is a process of decomposing the hyperspectral observations over multiple bands into a collection of endmember signatures and their corresponding proportions or abundances, under the assumption that the number of substances (or endmembers) present in that geographical area of interest is given *a priori*. Existing methods for estimating the number of endmembers can be broadly classified into two categories: information theoretic criteria based methods and eigenvalue thresholding methods. The information theoretic criteria based algorithms include Akaike's information criterion (AIC) [1], minimum description length (MDL) [2], and Bayesian information criterion (BIC) [3], to name a few. The estimation results of these algorithms may suffer from model mismatch errors resulting from incorrect prior information [4]. The eigenvalue thresholding based algorithms include Neyman-Pearson detection theory based method [5] (also referred to as virtual dimensionality (VD) in [4]), and hyperspectral signal subspace identification by minimum error (HySiMe) [6], to name a few.

In this work, we propose two hyperspectral data geometry based algorithms for estimating the number of endmembers, namely geometry based estimation of number of endmembers - convex hull (GENE-CH) algorithm and affine hull (GENE-AH) algorithm. The proposed algorithms exploit successive estimation property of a pure-pixel based EEA, and aim to decide when the EEA should stop estimating the next endmember. The GENE-CH and GENE-AH algorithms are devised based on the data geometry fact that all the observed pixel vectors should lie in the convex hull (CH) and affine hull (AH) of the endmember signatures, respectively. In the noisy scenario, the decision of whether the current endmember estimate is in the CH/AH of the previously found endmembers can be formulated as a binary hypothesis testing problem, which can be dealt using Neyman-Pearson detection theory. The performances of the proposed GENE algorithms are demonstrated through Monte-Carlo simulations for various scenarios.

The notations used throughout this paper are standard. $\mathbb{R}^M$ and $\mathbb{R}^{M\times N}$ represent a set of real $M \times 1$ vectors and $M \times N$ matrices, respectively, $\mathbf{1}_N$ represents an $N \times 1$ all-one vector and $\mathbf{0}$ is an all-zero vector of proper dimension. The symbol $\succeq$ denotes the componentwise inequality, $\|\cdot\|_2$ represents the Euclidean norm, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2. PROBLEM STATEMENT AND ASSUMPTIONS

Owing to low spatial resolution, each observed pixel vector represents a mixture of multiple distinct substances and each pixel vector of the hyperspectral images measured over M spectral bands can then be represented by the following $M \times N$ linear mixing model:

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{w}[n], \tag{1}$$

$$\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n] = \sum_{i=1}^{N} s_i[n]\mathbf{a}_i, \ \forall n = 1, \ldots, L. \tag{2}$$

In (1), $\mathbf{y}[n] = [\ y_1[n], \ldots, y_M[n]\ ]^T$ denotes the nth observed pixel vector comprising M spectral bands, $\mathbf{x}[n] = [\ x_1[n], \ldots, x_M[n]\ ]^T$ is its noise-free counterpart, and $\mathbf{w}[n] = [\ w_1[n], \ldots, w_M[n]\ ]^T$ is the noise vector. In (2), $\mathbf{A} = [\ \mathbf{a}_1, \ldots, \mathbf{a}_N\ ] \in \mathbb{R}^{M\times N}$ is the endmember signature matrix with the ith column vector $\mathbf{a}_i$ being the ith endmember signature, $\mathbf{s}[n] = [\ s_1[n], \ldots, s_N[n]\ ]^T \in \mathbb{R}^N$ is the nth abundance vector comprising N fractional abundances, and L is the total number of observed pixels. The noise vector $\mathbf{w}[n]$ is independent and identically distributed (i.i.d.) zero-mean Gaussian with covariance matrix $\mathbf{D} = E\{\mathbf{w}[n]\mathbf{w}[n]^T\} = \text{diag}(\sigma_1^2, \ldots, \sigma_M^2)$, an $M \times M$ diagonal matrix with the ith diagonal entry σ_i^2 denoting the noise variance in the ith spectral band.

Estimation of the number of endmembers is to estimate N from the given hyperspectral data $\mathbf{y}[1], ..., \mathbf{y}[L]$. Generally, hyperspectral image analysis has the following standard non-statistical assumptions: (A1) $s_i[n] \geq 0 \ \ \forall i, n$, (A2) $\sum_{i=1}^{N} s_i[n] = 1 \ \ \forall n$, (A3) $\min\{L, M\} \geq N$ and $\mathbf{A}$ is of full column rank, (A4) (Pure pixel assumption) there exists at least an index set $\{l_1, \ldots, l_N\}$ such that $\mathbf{x}[l_i] = \mathbf{a}_i$, for $i = 1, \ldots, N$.

Two important convex geometry concepts, namely affine hull and convex hull [7], which will play a significant role in the en-

This work was supported by the National Science Council (R.O.C.) under Grant NSC 99-2221-E-007-003-MY3.

suing development, are briefly introduced here. The *affine hull* of $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} \subset \mathbb{R}^M$ is defined as

$$\text{aff}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} = \left\{ \mathbf{x} = \sum_{i=1}^{N} \theta_i \mathbf{a}_i \middle| \mathbf{1}_N^T \boldsymbol{\theta} = 1, \boldsymbol{\theta} \in \mathbb{R}^N \right\}, \quad (3)$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]^T$. Its affine dimension P is no larger than $N-1$. If $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ is affinely independent (i.e., the vectors $\mathbf{a}_1 - \mathbf{a}_N, \ldots, \mathbf{a}_{N-1} - \mathbf{a}_N$ are linearly independent), then $P = N-1$.

The *convex hull* of $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} \subset \mathbb{R}^M$ is defined as

$$\text{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} = \left\{ \mathbf{x} = \sum_{i=1}^{N} \theta_i \mathbf{a}_i \middle| \mathbf{1}_N^T \boldsymbol{\theta} = 1, \boldsymbol{\theta} \succeq \mathbf{0} \right\}. \quad (4)$$

The convex hull $\text{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ is called an $N-1$ dimensional simplex in $\mathbb{R}^M$ if $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\} \subset \mathbb{R}^M$ is affinely independent. Such a simplex $\text{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ has only N extreme points (vertices), exactly being $\mathbf{a}_1, \ldots, \mathbf{a}_N$.

3. DIMENSION REDUCTION AND DATA GEOMETRY

The proposed GENE algorithms (to be presented in Section 4) estimates N using the endmember estimates provided by a successive EEA, and hence the specified dimension reduction with N unknown for any reliable successive EEA is presented.

Assume that we only have prior knowledge on the maximum bound of the number of endmembers $N_{\max}$ where $N \leq N_{\max} \leq M$. In the noisy scenario, similar to the dimension reduction procedure in [8], the dimension-reduced pixel vectors $\tilde{\mathbf{y}}[n]$ can be obtained by the following affine transformation of $\mathbf{y}[n]$:

$$\tilde{\mathbf{y}}[n] = \boldsymbol{\mathcal{C}}^T(\mathbf{y}[n] - \mathbf{d}) \in \mathbb{R}^{N_{\max}-1}, \quad (5)$$

where

$$\mathbf{d} = \frac{1}{L}\sum_{n=1}^{L} \mathbf{y}[n] = \frac{1}{L}\sum_{n=1}^{L} \mathbf{x}[n] + \frac{1}{L}\sum_{n=1}^{L} \mathbf{w}[n], \quad (6)$$

$$\boldsymbol{\mathcal{C}} = [\ \boldsymbol{q}_1(\mathbf{U}_y\mathbf{U}_y^T - L\mathbf{D}), \ldots, \boldsymbol{q}_{N_{\max}-1}(\mathbf{U}_y\mathbf{U}_y^T - L\mathbf{D})\ ], \quad (7)$$

in which $\mathbf{U}_y = [\ \mathbf{y}[1] - \mathbf{d}, \ldots, \mathbf{y}[L] - \mathbf{d}\ ] \in \mathbb{R}^{M \times L}$, and $\boldsymbol{q}_i(\mathbf{R})$ denotes the unit-norm eigenvector of $\mathbf{R}$ associated with the ith principal eigenvalue. In practical situations, the multiple regression analysis based noise covariance estimation method reported in HySiMe [6] can be used to estimate $\mathbf{D}$. Further, due to (1), (2), and (A2), we have

$$\tilde{\mathbf{y}}[n] = \tilde{\mathbf{x}}[n] + \tilde{\mathbf{w}}[n],\ n = 1, \ldots, L, \quad (8)$$

where

$$\tilde{\mathbf{x}}[n] = \sum_{i=1}^{N} s_i[n]\boldsymbol{\alpha}_i,\ n = 1, ..., L, \quad (9)$$

in which $\boldsymbol{\alpha}_i = \boldsymbol{\mathcal{C}}^T(\mathbf{a}_i - \mathbf{d}) \in \mathbb{R}^{N_{\max}-1}$ is the ith dimension-reduced endmember, and $\tilde{\mathbf{w}}[n] \triangleq \boldsymbol{\mathcal{C}}^T\mathbf{w}[n] \in \mathbb{R}^{N_{\max}-1}$ is i.i.d. Gaussian noise with zero mean and covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \boldsymbol{\mathcal{C}}^T\mathbf{D}\boldsymbol{\mathcal{C}} \in \mathbb{R}^{(N_{\max}-1)\times(N_{\max}-1)}. \quad (10)$$

Some convex geometries of the noise-free dimension-reduced data $\tilde{\mathbf{x}}[n]$ given by (9) which will lay a solid platform for the ensuing algorithmic developments, are as follows:

(F1) By (A1)-(A4), any dimension-reduced pixel vectors $\tilde{\mathbf{x}}[n]$ should lie in $\text{conv}\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$ and

$$\text{conv}\{\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[L]\} = \text{conv}\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}, \quad (11)$$

where $\text{conv}\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$ is a simplex with N extreme points being $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$.

(F2) Considering only (A2) and (A3), any dimension-reduced pixel vectors $\tilde{\mathbf{x}}[n]$ should lie in $\text{aff}\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}$ and

$$\text{aff}\{\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[L]\} = \text{aff}\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N\}, \quad (12)$$

where its affine dimension is equal to $N-1$.

4. GEOMETRY BASED ESTIMATION OF NUMBER OF ENDMEMBERS (GENE) ALGORITHMS

In the first subsection, suppose that (A1) – (A4) hold true. Then, the corresponding noisy pixels are given by

$$\tilde{\mathbf{y}}[l_i] = \boldsymbol{\alpha}_i + \tilde{\mathbf{w}}[l_i],\ \forall i = 1, \ldots, N. \quad (13)$$

We propose the GENE-CH algorithm based on the convex hull geometry (F1). For data with (A4) violated, the GENE-AH algorithm is proposed in the subsequent subsection.

4.1. GENE-Convex Hull (GENE-CH) Algorithm

Suppose that a reliable, successive EEA has found the pixel indices $l_1, \ldots, l_N, l_{N+1}, \ldots, l_{k-1}, l_k$, in which $l_1, \ldots, l_N$ are pure pixel indices and $k \leq N_{\max}$. Then by (8),(9), and (13),

$$\tilde{\mathbf{y}}[l_i] = \tilde{\mathbf{x}}[l_i] + \tilde{\mathbf{w}}[l_i],\ i = 1, \ldots, k, \quad (14)$$

where

$$\tilde{\mathbf{x}}[l_i] = \begin{cases} \boldsymbol{\alpha}_i, & i = 1, \ldots, N, \\ \sum_{j=1}^{N} s_j[l_i]\boldsymbol{\alpha}_j, & i = N+1, \ldots, k. \end{cases} \quad (15)$$

As has been depicted in (F1), the total number of extreme points in $\text{conv}\{\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[L]\}$ is N in the absence of noise. That is to say, if $\tilde{\mathbf{x}}[l_k] \in \text{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then it can be inferred by (15) that all the endmembers are already found, i.e., $k \geq N+1$. However, in a real scenario, since only noisy $\tilde{\mathbf{y}}[l_1], \ldots, \tilde{\mathbf{y}}[l_k]$ are available (rather than $\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_k]$), we propose a Neyman-Pearson hypothesis [9] testing based method to determine whether $\tilde{\mathbf{x}}[l_k] \in \text{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, or not, based on noisy $\tilde{\mathbf{y}}[l_1], \ldots, \tilde{\mathbf{y}}[l_k]$. The idea is to find the smallest k for which $\tilde{\mathbf{y}}[l_k]$ is closest to $\text{conv}\{\tilde{\mathbf{y}}[l_1], \ldots, \tilde{\mathbf{y}}[l_{k-1}]\}$ in some optimal sense. To do so, let us consider the following constrained least squares problem:

$$\boldsymbol{\theta}^\star = \arg \min_{\boldsymbol{\theta} \succeq \mathbf{0}, \mathbf{1}_{k-1}^T\boldsymbol{\theta} = 1} \|\tilde{\mathbf{y}}[l_k] - \mathbf{A}_{k-1}\boldsymbol{\theta}\|_2^2, \quad (16)$$

where

$$\mathbf{A}_{k-1} = [\tilde{\mathbf{y}}[l_1], \ldots, \tilde{\mathbf{y}}[l_{k-1}]] \in \mathbb{R}^{(N_{\max}-1)\times(k-1)}. \quad (17)$$

The optimization problem in (16) is convex and can be solved by using available convex optimization solvers such as `SeDuMi` [10]. Then, the optimal fitting error vector $\boldsymbol{e}$ is given by

$$\boldsymbol{e} = \tilde{\mathbf{y}}[l_k] - \mathbf{A}_{k-1}\boldsymbol{\theta}^\star \quad (18)$$

$$= \boldsymbol{\mu}_k + \left(\tilde{\mathbf{w}}[l_k] - \sum_{i=1}^{k-1} \theta_i^\star \tilde{\mathbf{w}}[l_i] \right) \in \mathbb{R}^{N_{\max}-1}, \quad (19)$$

where the second equality is due to (14), and

$$\boldsymbol{\mu}_k = \tilde{\mathbf{x}}[l_k] - \sum_{i=1}^{k-1} \theta_i^\star \tilde{\mathbf{x}}[l_i]. \quad (20)$$

Then the following can be observed from (19):

- If $\tilde{\mathbf{x}}[l_k] \in \operatorname{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then it implies that $\tilde{\mathbf{x}}[l_k] - \sum_{i=1}^{k-1} \theta_i' \tilde{\mathbf{x}}[l_i] = \mathbf{0}$, for some $\boldsymbol{\theta}' = [\theta_1', \ldots, \theta_{k-1}']^T \succeq \mathbf{0}$, $\mathbf{1}_{k-1}^T \boldsymbol{\theta}' = 1$. In the noise-free case (i.e., $\tilde{\mathbf{y}}[l_i] = \tilde{\mathbf{x}}[l_i]$, for all i), $\boldsymbol{\theta}^\star = \boldsymbol{\theta}'$ and $\boldsymbol{e} = \mathbf{0}$, while in the presence of noise $\boldsymbol{\theta}^\star \simeq \boldsymbol{\theta}'$, implying that $\boldsymbol{e}$ can be approximated as a random vector with distribution $\mathcal{N}(\mathbf{0}, \xi^\star \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ has been defined in (10) and

$$\xi^\star = 1 + \theta_1^{\star 2} + \theta_2^{\star 2} + \cdots + \theta_{k-1}^{\star 2} \tag{21}$$

  since $\tilde{\mathbf{w}}[n]$ is i.i.d.

- If $\tilde{\mathbf{x}}[l_k] \notin \operatorname{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{\mu}_k, \xi^\star \boldsymbol{\Sigma})$.

Now let us define

$$r = \boldsymbol{e}^T (\xi^\star \boldsymbol{\Sigma})^{-1} \boldsymbol{e}. \tag{22}$$

When $\tilde{\mathbf{x}}[l_k] \in \operatorname{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, it is easy to see that r can be approximated as a random variable following central Chi-square distribution $\chi^2(Z)$, and otherwise r follows non-central Chi-square distribution $N\chi^2(Z, \boldsymbol{\mu}_k)$ [11] where $Z = N_{\max} - 1$ denotes the degrees of freedom. Hence, we consider the following two hypotheses:

$$\begin{aligned} &H_0 \ (\tilde{\mathbf{x}}[l_k] \in \operatorname{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}) : \ r \sim \chi^2(Z), \\ &H_1 \ (\tilde{\mathbf{x}}[l_k] \notin \operatorname{conv}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}) : \ r \sim N\chi^2(Z, \boldsymbol{\mu}_k). \end{aligned}$$

Since $\boldsymbol{\mu}_k$ is unknown so is $N\chi^2(Z, \boldsymbol{\mu}_k)$, we use Neyman-Pearson classifier rule for the above hypothesis testing problem:

$$\text{Decide } H_0 \ \text{ if } \ r < \kappa \tag{23a}$$

$$\text{Decide } H_1 \ \text{ if } \ r > \kappa, \tag{23b}$$

Denoting the probability density function (pdf) of the central Chi-square distribution by $f_{\chi^2}(x, Z)$, we define

$$\psi(r) \triangleq \int_r^\infty f_{\chi^2}(x, Z) dx = 1 - \frac{\gamma(r/2, Z/2)}{\Gamma(Z/2)}, \tag{24}$$

where $\gamma(x/2, Z/2)$ is the lower incomplete Gamma function [12]. Then, by Neyman-Pearson lemma [9], the optimal threshold κ for problem (23) satisfies

$$\psi(\kappa) = P_{\text{FA}}, \tag{25}$$

where P_{FA} is the preassigned acceptable false alarm rate. There is no closed-form expression for the inverse function of $\psi(\cdot)$, and hence we formulate the decision rule in (23) to

$$\text{Decide } H_0 \ \text{ if } \ \psi(r) > P_{\text{FA}}, \tag{26a}$$

$$\text{Decide } H_1 \ \text{ if } \ \psi(r) < P_{\text{FA}}. \tag{26b}$$

The integral $\psi(r)$ defined in (24) can be easily computed by using available packages such as MATLAB. Once $\psi(r)$ is evaluated, one of the two hypotheses is decided, based on (26). The pseudo-code of the proposed GENE-CH algorithm is given in Table 1.

4.2. GENE-Affine Hull (GENE-AH) Algorithm

When (A4) is violated, the dimension-reduced endmembers estimated by an EEA can also be expressed as in (14), where

$$\tilde{\mathbf{x}}[l_i] = \sum_{j=1}^{N} s_j[l_i] \boldsymbol{\alpha}_j, \ \forall i = 1, \ldots, k. \tag{27}$$

Table 1. Pseudo-codes of GENE-CH and GENE-AH algorithms.

Given	noisy hyperspectral data $\mathbf{y}[n]$, maximum number of endmembers $N \leq N_{\max} \leq M$, false alarm probability P_{FA}, and estimate of noise covariance matrix $\mathbf{D}$; a chosen successive EEA.
Step 1.	Compute $(\boldsymbol{\mathcal{C}}, \mathbf{d})$ given by (6) and (7).
Step 2.	Obtain the first pixel index l_1 by the successive EEA and compute $\tilde{\mathbf{y}}[l_1]$ by (5). Set $k = 2$.
Step 3.	Obtain the kth pixel index l_k by the successive EEA and compute $\tilde{\mathbf{y}}[l_k]$ by (5), and form $\mathbf{A}_{k-1}$ by (17).
Step 4.	Use Sedumi [10] to solve problem (16) for GENE-CH problem (28) for GENE-AH for the optimal $\boldsymbol{\theta}^\star$ and $\boldsymbol{e} = \tilde{\mathbf{y}}[l_k] - \mathbf{A}_{k-1}\boldsymbol{\theta}^\star$.
Step 5.	Compute $r = \boldsymbol{e}^T (\xi^\star \boldsymbol{\Sigma})^{-1} \boldsymbol{e}$, where $\xi^\star = 1 + \boldsymbol{\theta}^{\star T} \boldsymbol{\theta}^\star$ and $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{C}}^T \mathbf{D} \boldsymbol{\mathcal{C}}$, and $\psi(r)$ by (24).
Step 6.	If $\psi(r) > P_{\text{FA}}$, then output $k - 1$ as the estimate of N, else $k := k + 1$ and if $k \leq N_{\max}$ go to **Step 3**.

For such hyperspectral data, it can be shown that GENE-CH may result in an overestimation of the number of endmembers. Hence we next propose the GENE-AH algorithm which does not require the pure pixel assumption (A4). The GENE-AH algorithm uses the fact (F2), which states that in the noise-free case, the affine dimension of $\operatorname{aff}\{\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[L]\}$ is $N - 1$. This implies that in the noise-free case, if $\tilde{\mathbf{x}}[l_k] \in \operatorname{aff}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then $k \geq N+1$. Here again we use Neyman-Pearson hypothesis [9] testing to find the smallest k such that the hypothesis $\tilde{\mathbf{x}}[l_k] \in \operatorname{aff}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$ is true with the given P_{FA}, based on noisy $\tilde{\mathbf{y}}[l_1], \ldots, \tilde{\mathbf{y}}[l_k]$. As in (16), we consider solving the following constrained least squares problem:

$$\boldsymbol{\theta}^\star = \arg \min_{\mathbf{1}_{k-1}^T \boldsymbol{\theta} = 1} \ \|\tilde{\mathbf{y}}[l_k] - \mathbf{A}_{k-1}\boldsymbol{\theta}\|_2^2, \tag{28}$$

where $\mathbf{A}_{k-1}$ has been defined in (17). By defining the optimal fitting error vector $\boldsymbol{e}$ as in (18), we have the following inferences:

- if $\tilde{\mathbf{x}}[l_k] \in \operatorname{aff}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then it can be approximated that $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \xi^\star \boldsymbol{\Sigma})$.
- if $\tilde{\mathbf{x}}[l_k] \notin \operatorname{aff}\{\tilde{\mathbf{x}}[l_1], \ldots, \tilde{\mathbf{x}}[l_{k-1}]\}$, then $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{\mu}_k, \xi^\star \boldsymbol{\Sigma})$,

where $\boldsymbol{\mu}_k$, $\xi^\star$, and $\boldsymbol{\Sigma}$ have been defined in (20), (21), and (10), respectively. Defining the random variable r as in (22), a similar Neyman-Pearson hypothesis testing procedure can be devised for GENE-AH to estimate the number of endmembers present in the data. The procedure for GENE-AH is also summarized in Table 1.

Most existing benchmark algorithms are directly or indirectly developed based on that the range space of the endmembers is the same as that of the hyperspectral data, i.e., they are based only on (A3) (or subspace geometry). However, the GENE algorithms not only make use of (A3), but also (A2) (or affine geometry) for GENE-AH algorithm, and (A1) and (A2) (or convex geometry) for GENE-CH algorithm. The advantages of considering the assumptions (A1) and (A2) on abundances will be more evident in the simulations. It should be noted that the performance of both GENE-CH and GENE-AH algorithms depends on the performance of the successive EEA used. Hence, the successive EEA algorithm employed by both GENE-CH and GENE-AH algorithms need to be reproducible (without any initialization) and can sequentially provide endmember estimates without repetition. For instance, one successive EEA recently proposed in [13], called p-norm based pure pixel identification (TRI-P) algorithm, not only possesses the above performance

Table 2. Mean±standard deviation of the estimated number of endmembers for various algorithms over 100 independent runs for various $P_{\rm FA}$ (whenever applicable), number of endmembers N, and purity levels ρ, as well as SNR=30 dB, $N_{\max} = 25$, $L = 5000$, and $M = 224$.

Methods	$P_{\rm FA}$	$\rho = 1$, SNR=30 dB				$N = 8$, SNR=30 dB			
		Number of Endmembers N				Purity Level ρ			
		8	12	16	20	0.8	0.85	0.9	0.95
GENE-CH (TRI-P, $p = 2$)	10^{-4}	8.02±0.17	12.04±0.19	**15.77±0.46**	**19.78±0.50**	12.82±2.94	10.97±2.38	9.28±1.28	8.33±0.58
	10^{-5}	**8.00±0**	12.03±0.17	**15.77±0.46**	**19.74±0.52**	12.13±2.70	10.23±1.92	8.85±0.98	8.17±0.45
	10^{-6}	**8.00±0**	12.02±0.14	**15.72±0.47**	**19.70±0.52**	11.65±2.69	9.85±1.71	8.61±0.92	8.11±0.38
GENE-AH (TRI-P, $p = 2$)	10^{-4}	**8.00±0**	**12.00±0**	14.76±0.42	17.75±0.50	8.02±0.14	8.01±0.09	**8.00±0**	8.01±0.09
	10^{-5}	**8.00±0**	**12.00±0**	14.57±0.49	17.51±0.54	**8.00±0**	8.01±0.09	**8.00±0**	**8.00±0**
	10^{-6}	**8.00±0**	**12.00±0**	14.32±0.46	17.17±0.66	**8.00±0**	**8.00±0**	**8.00±0**	**8.00±0**
HYSIME [6]	–	**8.00±0**	**12.00±0**	14.00±0	16.15±0.35	**8.00±0**	**8.00±0**	**8.00±0**	**8.00±0**
HFC [4]	10^{-4}	5.00±0	7.14±0.68	8.66±0.27	4.19±0.63	5.00±0	5.00±0	5.00±0	5.00±0
	10^{-5}	5.00±0	6.44±0.53	7.93±0.25	3.67±0.60	5.00±0	5.00±0	5.00±0	5.00±0
	10^{-6}	5.00±0	6.10±0.46	7.76±0.47	3.23±0.52	5.00±0	5.00±0	5.00±0	5.00±0
NW-HFC [4]	10^{-4}	5.00±0	7.18±0.70	9.15±0.35	6.23±0.69	5.00±0	5.00±0	5.00±0	5.00±0
	10^{-5}	5.00±0	6.46±0.62	8.97±0.30	5.46±0.77	5.00±0	5.00±0	5.00±0	5.00±0
	10^{-6}	5.00±0	5.96±0.58	8.80±0.42	4.78±0.70	5.00±0	5.00±0	5.00±0	5.00±0

merits, but also is reliable with theoretical support for endmember identifiability for the noise-free case. Therefore, TRI-P algorithm is a good EEA candidate for the proposed GENE algorithms.

5. SIMULATIONS AND CONCLUSION

One hundred Monte-Carlo runs are performed to study the effectiveness of the proposed GENE-CH and GENE-AH algorithms that employ TRI-P with p=2 (i.e., 2-norm) [13] to acquire endmember estimates for various scenarios. Algorithms considered for comparison are HySiMe [6], HFC [4], and NW-HFC [4]. The GENE algorithms, HFC, and NW-HFC are evaluated under the following false alarm probabilities: 10^{-4}, 10^{-5} and 10^{-6}, and for GENE and NW-HFC algorithms, the true noise covariance matrix is supplied for each simulated realization. The endmembers are chosen from the USGS library [14] with $M = 224$. The maximum bound of the number of endmembers used in GENE algorithms is $N_{\max} = 25$. The abundance vectors $\mathbf{s}[n],\ n = 1, \ldots, L$ are generated by following the Dirichlet distribution [8] for generating synthetic data with different purity levels ρ (defined as $\rho = \max\{\|\mathbf{s}[n]\|_2, n = 1, \ldots, L\}$ [8]).

Table 2 displays mean±standard deviation of the number of endmembers estimated by the algorithms under test for two scenarios, where all the hyperspectral data with $L = 5000$ are corrupted by white Gaussian noise [8] with $\mathrm{SNR} = 30$ dB. The estimated number of endmembers closest to the true number of endmembers are highlighted by bold-faced numbers. In the first scenario, the number of endmember N is allowed to vary as $8, 12, 16$ and 20, while maintaining the purity level $\rho = 1$ (indicating the existence of pure pixels in the data). It can be observed that for higher number of endmembers $N = 16, 20$ GENE-CH yields the best performance followed by GENE-AH. For $N = 8, 12$ both GENE-AH and HySiMe yield best performance. In the second scenario where the purity level ρ of the hyperspectral data varies from 0.8 to 0.95 (indicating no pure pixels in the data) while maintaining $N = 8$, it can be readily seen that when purity level is smaller, GENE-CH overestimates the number of endmembers. On the other hand, GENE-AH with $P_{\rm FA} = 10^{-6}$ and HySiMe correctly estimate the number of endmembers.

In conclusion, we have presented two convex geometry based algorithms for estimating the number of endmembers, namely GENE-CH and GENE-AH algorithms, based on (F1) and (F2), respectively. The GENE algorithms employ a Neyman-Pearson hypothesis testing strategy and they must operate in conjunction with a successive EEA in a synchronization fashion. Simulation results confirm the superior efficacy of the proposed GENE-CH and GENE-AH algorithms over some existing benchmark methods, because of more focused geometry (convex and affine sets rather than range space) considered for the hyperspectral data.

6. REFERENCES

[1] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Control*, vol. AC-19, no. 6, pp. 716-723, Dec. 1974.

[2] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465-471, Sep. 1978.

[3] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, Mar. 1978.

[4] C.-I. 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.

[5] J. Harsanyi, W. Farrand, and C.-I Chang, “Determining the number and identity of spectral endmembers: An integrated approach using Neyman-Pearson eigenthresholding and iterative constrained RMS error minimization,” in *Proc. 9th Thematic Conf. Geologic Remote Sensing*, Feb. 1993.

[6] J. M. B. Dias and J. M. P. Nascimento, “Hyperspectral subspace identification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435-2445, Aug. 2008.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, UK: Cambridge Univ. Press, 2004.

[8] A. Ambikapathi, T.-H. Chan, W.-K. Ma, and C.-Y. Chi, “Chance-constrained robust minimum-volume enclosing simplex algorithm for hyperspectral unmixing,” *IEEE Trans. Geosci. Remote Sens. - Special Issue on Spectral Unmixing of Remotely Sensed Data*, vol. 49, no. 11, pp. 4194-4209, Nov. 2011.

[9] L. C. Ludeman, *Random Processes Filtering, Estimation, and Detection*, Wiley-Interscience Publication, 2003.

[10] J. F. Sturm, “Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones”, *Optimization Methods and Software* vol. 11-12, pp. 625-653, 1999.

[11] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*, 2nd ed., Wiley-Infterscience Publication, 1993.

[12] G. Arfken and H. Weber, *Mathematical Methods for Physicists*, Harcourt Academic Press, 2000.

[13] A. Ambikapathi, T.-H. Chan, C.-Y. Chi and K. Keizer, “Two effective and computationally efficient pure-pixel based algorithms for hyperspectral endmember extraction,” in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 22-27, 2011, pp. 1369-1372.

[14] R. N. Clark, G. A. Swayze, A. Gallagher, T.V. King, and W. M. Calvin, “The U.S. geological survey digital spectral library: Version 1: 0.2 to 3.0,” *U. S. Geol. Surv., Open File Rep.* pp. 93-592, 1993.