Estimation of the Number of Endmembers Using Robust Outlier Detection Method

Charoula Andreou, *Student Member, IEEE*, and Vassilia Karathanassi

***Abstract*—This paper introduces a novel approach for estimating the numbers of endmembers in hyperspectral imagery. It exploits the geometrical properties of the noise hypersphere and considers the signal as outlier of the noise hypersphere. The proposed method, called *outlier detection method (ODM)*, is automatic and non-parametric. In a principal component space, noise is spherically symmetric in all directions and lies on the surface of a hypersphere with a constant radius. Reversely, signal radiuses are much larger that noise radius and vary in all directions, thus signal lies in a hyperellipsoid. The proposed method involves three steps: 1) noise estimation; 2) minimum noise fraction transformation; and 3) outlier detection using inter quartile range. Estimation of the number of endmembers is accomplished by the estimation of the number of noise hypersphere outliers using a robust outlier detection method. The ODM was evaluated using simulated and real hyperspectral data, and it was also compared with well-known methods for estimating the number of endmembers. Evaluation of the method showed that the method produces robust and satisfactory results, and outperforms in relation to its competitors.**

***Index Terms*—Hyperspectral imagery, outlier detection method (ODM), signal processing, signal subspace.**

I. Introduction

ESTIMATION of the number of signals is a fundamental problem in signal processing. In the scientific field of hyperspectral imagery, signals are related to the unique constituent deterministic spectral signatures, called endmembers [1]. A predetermined number of endmembers is required by the majority of the existing endmember extraction methods in order to detect the optimal set of endmembers. Estimation of the correct number of endmembers has significant impact on the performance of the endmember extraction algorithms and consequently on the accuracy of the spectral unmixing process. According to [2], the accuracy of spectral unmixing will be the highest when the exact number of endmembers that are required to account for the spectral variability is utilized in the model. Using fewer endmembers than the actual number would lead to the increase of the root mean square error between the original and the reconstructed image, while too many endmembers would make the model sensitive to instrumental noise, atmospheric influences, and natural variability in spectra, resulting in abundance estimation error. Furthermore, the number of endmembers is associated with the intrinsic or, in a more wide sense, with the virtual dimensionality of a hyperspectral dataset [3], [4], as it determines the optimal number of dimensions to be retained after dimensionality reduction in order to represent the dataset. Hence, an accurate determination of the number of the endmembers significantly contributes to the accuracy of the spectral unmixing processing and enables low-dimensional representation of spectral vectors, yielding gains in computational time and complexity, data storage and signal-to-noise ratio (SNR) [5].

In recent years, many algorithms have been developed which contribute to the estimation of the number of endmembers. The available methods can be classified into separate categories. The first category comprises eigen-based energy methods [6], [7]. These methods involve a dimensionality reduction method and estimate the minimum number of the transformed components for which the total variance of the data is equal to a specified percentage of energy. However, the cut-off threshold should be manually chosen, which is very difficult to determine since the eigenvalues corresponding to signals and noise are sometimes very similar [8]. In the second category, information criteria based on likelihood functions [9], [10] are included. Two well-known information criteria for model order selection are Akaike information criterion (AIC) [9] and minimum description length (MDL) [10]. Since the criteria require the prior knowledge of the mixture model or likelihood function, the estimation may suffer from model mismatch errors resulting from incorrect prior information. Moreover, it has been shown in [4] that the results of AIC and MDL when applied to hyperspectral data are seriously overestimated due to the invalid Gaussian distribution assumption made on the abundances [8]. The third category consists of eigenvalue-based methods [4], [8]. Harsanyi-Farrand-Chang (HFC) and noise-whitened HFC (NWHFC) [4] methods estimate the virtual dimensionality (VD) based on the fact that the eigenvalues of the correlation matrix and of the covariance matrix will be equal if noise exists. Thus, eigenvalues of both data correlation and covariance matrices are calculated and if their difference is positive—according to a determined probability false alarm parameter—then a signal source is present. VD methods might overestimate the number of the endmembers because they estimate the spectrally distinct signal sources which could comprise known and unknown image endmembers, background signatures, interferences and anomalies [4]. HFC and NWHFC methods impose limitations to automation since they result in different estimates for different false alarm parameters. Recently, a new empirical method for estimating the

Manuscript received January 18, 2013; revised April 13, 2013; accepted April 23, 2013. This work was carried out within the framework of the project "Automatic Oil-Spill Recognition and Geopositioning integrated in a Marine Monitoring Network (ARGOMARINE)" 2009–2012 and supported by CEC under Contract FP7-SST-2008-RTD-1.

The authors are with Laboratory of Remote Sensing, School of Rural and Surveying Engineering, National Technical University of Athens, 15780 Athens, Greece (e-mail: candreou@central.ntua.gr; karathan@survey.ntua.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2013.2260135

number of endmembers presented in [8] modifies the VD concept. The method is called eigenvalue likelihood maximization (ELM), and it is based on the fact that the eigenvalues which correspond to the noise are identical in the covariance and the correlation matrices, while eigenvalues corresponding to the signal are larger in the correlation matrix than in the covariance matrix. The eigenvalue-based methods are based only on the eigenvectors of the observed data correlation or covariance matrix. Since signal subspace dimension is unknown in most real applications, it must be inferred from data leading to a model-order problem which may lead to poor results [11]. Authors in [11] presented hyperspectral signal subspace identification by minimum error (HySime) method which selects the subset of eigenvectors that best represents the signal subspace in the minimum mean square error sense.

All of the aforementioned methods arguably consider the existence of two different distributions, the one related to noise and the other related to signal, or in geometrical approach they consider two different subspaces one of noise and one of signal. However, in hyperspectral space, signal vectors are very few in order to estimate their population distribution properly or to statistical analyze them.

In this paper, a new automatic nonparametric method for estimating the number of endmembers is introduced. Its novelty lies in the fact that it considers only the existence of noise and treats signals as outliers of noise. No estimation of statistical distributions is required. The new method, called outlier detection method (ODM), explores the geometrical properties of the noise hypersphere. It searches for the signals whose radius is by far larger than the one of the noise introducing for the first time in virtual dimension theory a robust outlier detection method. In particular, the ODM implements noise estimation and whitening process. Afterwards, observed data are transformed into a new principal component space, where noise is expected to lie in a hypersphere of constant radius. Estimation of the number of noise hypersphere outliers using a robust inter quartile range based outlier detection method [12] results in the estimation of the number of endmembers. In [13] an empirical method for estimating the number of endmembers is presented which implies the approach adopted by ODM.

The remainder of this paper is organized as follows. Section II formulates the estimation of the signal subspace dimension problem. Section III describes the theoretical fundamentals which substantiate the proposed approach. Section IV presents two well-known noise estimation methods and Section V provides elaborate description of the proposed method. Evaluation of the proposed method compared with state-of-the-art relevant methods using simulated and real data is given in Section VI. Finally, Section VII provides concluding remarks.

II. Data Model and Problem Formulation

Consider that, if L is the total number of bands, each observed spectral vector y, $y \in \Re^L$, consists of a signal vector x, $x \in \Re^L$, and an error term n, $n \in \Re^L$, for additive noise which includes sensor noise, endmember variability, and other model inadequacies [1], [8], [11], [14]:

$$y = x + n. \tag{1}$$

Furthermore, a signal vector lies in an unknown p-dimensional subspace of the band space, where $p < L$, and it is described by

$$x = \sum_{k=1}^{p} a_k s_k = S\alpha. \tag{2}$$

Under the subspace model scenario, the $L \times 1$ signal vectors s_k are linearly independent (or otherwise S is a full rank $L \times p$ matrix), serving as a basis for the spectral subspace [1] and α is considered a $p \times 1$ vector containing coefficients a_k. Under the linear spectral mixing concept [14], matrix $S = [s_1, \ldots, s_p]$ comprises the endmember spectra and $\alpha = [a_1, \ldots, a_p]$ their corresponding abundances. The latter should obey to sum-to-one and positivity constraints in order to be physically meaningful. In this paper, we study the subspace model which specifies the linear vector subspace region of the spectral space in which spectral vectors are allowed to reside regardless the adopted spectral mixing model, linear or nonlinear [1], [11].

According to [15], in the case of independent and identically distributed (i.i.d.) zero mean noise with variance σ_n^2I, signal subspace can be estimated, even if signal vectors are unknown, by the orthogonal decomposition of the covariance matrix of the observed vectors y, R_y. The estimate of the signal subspace is the span of the eigenvectors of R_y, $\langle M \rangle = \langle [e_1, e_2, \ldots, e_p] \rangle$, whose respective eigenvalues $l_1 > l_2 > \ldots > l_p$ are larger than σ_n^2 of noise. Of course, in most real applications, the dimension p of the signal subspace is unknown and noise is not i.i.d.. Therefore, in many cases, noise estimation is a prerequisite for the denoising or whitening process which is discussed in Section IV. A plethora of signal subspace estimation methods [4], [6]–[10] are based on the eigenvalues of the covariance or correlation matrix of the observed spectral vectors y. The drawbacks of using only the eigenvalues are presented in [11]. In this paper, the estimation of the signal subspace dimension is based on the transformation of the observed vectors y using the eigenvectors of R_y. The new transformed space is then statistically analyzed based on information theory concepts which are presented in the following section.

III. Definition of Noise Hypersphere

A. Multivariate Normal Distribution

Let $X = [X_1, X_2, \ldots, X_L]^T$ be a $L \times 1$ random vector. Its mean value is given by $m = \mathrm{E}(X)$, $E(\cdot)$ stands for expected value, and its covariance matrix by $R_X = \mathrm{E}(X-m)(X-m)^T$.

Assuming that random X is multivariate normal and R_X is a nonsingular matrix, the following quadratic form:

$$r^2 = (x-m)^T R_X^{-1} (x-m) \tag{3}$$

is a weighted norm which is called the Mahalanobis distance from x to m. The locus of points x for which r^2 is constant is also a locus of points for which the density $f(x)$ is constant. In case that the locus is a hypersphere, its radius is equal to r [15].

B. Noise Hypersphere

Based on information theory [16], [17] the zero mean white Gaussian noise vector $n \sim N(0, \sigma_n^2 I)$ has constant noise spectral density N_0. It is spherically symmetric in all directions in the spectral space and lies on the surface

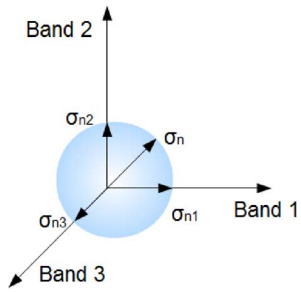

Fig. 1. Illustration of noise hypersphere in three dimensions ($\mathrm{n} \sim N(0, \sigma^2 \boldsymbol{I})$).

of a hypersphere with radius equal to σ_n. More specifically, in an L-dimensional hypersphere, the distance of point $n = [n_1, n_2, \ldots n_L]^T$ from its origin (zero) point is according to (3) $r_n = \sqrt{n_1^2 + n_2^2 + \ldots + n_L^2}$ and the distance of the normalized noise vector is $r_n/\sqrt{L}$, which is the σ_n. The advantages of considering the normalized version of noise vector as well as further details related to the above can be found in [16]–[18].

Thereupon, $\sigma_{n1} = \sigma_{n2} = \ldots = \sigma_{nL}$ are the standard deviations of the normalized noise vector in each dimension of the hypersphere and are equal to its radius as it shown in Fig. 1. The signal vector x has evidently $\sigma_x > \sigma_n$ and, since σ_x varies in all directions, it lies in a hyperellipsoid. Further analysis of the signal and noise locus is provided for a given dataset in Section V-A. In order to utilize the aforementioned properties of the noise hypersphere, it is requisite that the noise is zero mean i.i.d. or that noise is known, and therefore it can be transformed to zero mean i.i.d.. Both requisites do not stand in real applications. However, many approaches have been developed for noise estimation. Two of them are presented in Section IV.

IV. Noise Estimation

Noise estimation is of great importance not only for hyperspectral imagery but generally for signal processing. Here, nearest neighbor difference (NND) [19] and multiple regression theory [20] based methods are analyzed since these are widely used by signal subspace estimation algorithms [10], [11]. Both of these noise estimation methods are evaluated using simulated data in Section VI.

A. Nearest Neighbor Difference

The nearest neighbour difference (NND) method [19], also called shift difference method, is considered to be the easiest method for noise estimation. The procedure exploits the fact that signal exhibits strong spatial correlation among nearby pixels in an image, while the spatial correlation for noise is very weak. Therefore, it is assumed that noise samples are independent and have the same statistics [11]. The shift difference method should be applied on a homogeneous area. More precisely, it is performed on the data by differencing the two adjacent pixels to the right and above each pixel and averaging the results to obtain the noise value to assign to the pixel being processed. The idea can be illustrated using two adjacent observed vectors, y_1 and y_2, with essentially the same target. Subtracting them yields

$$y_1 - y_2 = (x_1 + n_1) - (x_2 + n_2) \approx n_1 - n_2 \tag{4}$$

where x_1, x_2 are the signal vectors and n_1, n_2 are noise vectors. Depending on the image, the noise estimation may be performed in a homogeneous subset of pixels, assuming that noise is the same throughout the whole image. Therefore the covariance matrix of noise R_n can be estimated, instead of noise value per observed spectral vector. The drawback of the NND method is that due to its assumption that adjacent pixels have the same signal information, the method is not proper for all the datasets, because the amount of pixels belonging to homogeneous areas may not be adequate for an accurate calculation of noise statistics.

B. Multiple Regression Theory-Based Method

The multiple regression theory-based approach [11], [20] is amenable to hyperspectral data since it can accommodate many explanatory variables which may be correlated, such as data in adjacent spectral bands. In particular, let Y be a $L \times N$ data matrix, where N are the $L \times 1$ observed spectral vectors, and L be the spectral bands. Define $Z = Y^T$, a $N \times 1$ vector $z_i = [Z]_{:,i}$ containing the values of all of the pixels in band i and $Z_{\partial i} = [z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_L]$ is a $N \times (L-1)$ matrix containing the pixel values of all the bands except for band i. Assuming that vector z_i can be expressed as a linear combination of the remaining data of $L-1$ bands, the following equation can be written:

$$z_i = Z_{\partial i} b_i + n_i \tag{5}$$

where $Z_{\partial i}$ is the $N \times (L-1)$ explanatory data matrix, b_i is the $(L-1) \times 1$ regression vector, and n_i is the residual error of size $N \times 1$. The linear regression coefficients are determined by

$$\hat{b}_i = \left(Z_{\partial i}^T Z_{\partial i}\right)^{-1} Z_{\partial i}^T z_i. \tag{6}$$

Noise estimation of band i is accomplished by the following equation:

$$\hat{n}_i = z_i - Z_{\partial i} \hat{b}_i. \tag{7}$$

V. Outlier Detection Method (ODM)

Here, the proposed method for estimating the number of endmembers, ODM, is introduced and described analytically. The method is fully automatic and nonparametric. It comprises three steps: 1) noise estimation; 2) MNF transformation; and 3) outlier detection. The main key points of the proposed method are summarized here.

- There is a big effort in hyperspectral community to define a threshold between signal and noise [8]. The ODM introduces a new concept which considers only the existence of noise and treats signal as outlier. Consequently, no threshold is needed.
- Contrary to the existing relevant algorithms which focus on signal subspace, the ODM exploits the properties of

noise subspace. It relies on the mathematical description of the noise hypersphere radius which is given by information theory.
- A new modified version of MNF is introduced which initially performs multiple regression theory based method for noise estimation, instead of NND. Results showed that this modification optimize the MNF method.
- For the first time in virtual dimensionality theory, a robust outlier detection method is used, called inter quartile range (IQR)-based method. Its benefit lies in the fact that it can be used when data distribution is unknown and thus, no statistical parameter estimation is needed. The risk of estimating erroneously the signal distribution due to its small population is omitted.
- The proposed method is characterized by its simplicity.

The first step of the proposed method is noise estimation. Experiments with simulated and real data (Section VI) show that the performance of the proposed method is better when multiple regression based method is applied in comparison to NND method.

The second step includes noise whitening and transformation into a new principal component space. More analytically, the noise covariance matrix R_n is estimated. The orthogonal decomposition of R_n results in the matrix $D_n = [d_{n1}, d_{n2}, \ldots, d_{nL}]$ of size $L \times L$ which consists of noise eigenvectors d_n, each one of size $L \times 1$. Suppose that the observed $L \times N$ data matrix Y, where N are the $L \times 1$ observed spectral vectors and L the spectral bands, is transformed using the noise eigenvectors. The transformed data F is given by

$$F = D_n^T Y \tag{8}$$

The matrix F of size $L \times N$ consists of N transformed spectral vectors f of size $L \times 1$. Define $W = F^T$, a $N \times 1$ vector $w_i = [W]_{:,i}$ contains the values of all the transformed pixels in band i. Dividing each data of band i with the standard deviation of noise $\hat{\sigma}_{ni}$ (symbol $\wedge$ stands for the estimated value) of the corresponding band i

$$W' = \left[\frac{w_1}{\hat{\sigma}_{n1}}, \frac{w_2}{\hat{\sigma}_{n2}}, \ldots, \frac{w_L}{\hat{\sigma}_{nL}}\right] \tag{9}$$

results in the $N \times L$ matrix W', which is the transformed data with equal noise variance $\hat{\sigma}_n^2$ in each band, which means that noise is whitened in the transformed space. The next step is the orthogonal decomposition of the covariance matrix of W'^T, which results in the $L \times L$ matrix $D_{W'^T}$ containing the $d_{w'^T}$ eigenvectors of size $L \times 1$. The transformation of W'^T using the eigenvectors of $D_{W'^T}$

$$Y' = D_{W'^T}^T W'^T \tag{10}$$

defines a new principal component space in which transformed data of $L \times N$ matrix Y' consists of uncorrelated noise which increases with the component rank. Thus, the well-known MNF [19] is modified by applying different noise estimation method.

Assuming that noise is white, rotation of a signal structure [i.e., in (10)] does not change the noise distribution [18]. Consequently, noise remains spherically distributed about the mean value and lies in a hypersphere of radius $\hat{\sigma}_n$. It is reasonable that noise estimation comprises an error, which is justified in term of fluctuations. Therefore, it cannot be expected that standard deviations of noise components are exactly equal to σ_n but it should be expected to be close to zero mean value as the minimum standard deviation of a component corresponds to the maximum noise fraction [19]. Conversely, standard deviation $\hat{\sigma}_{xi}$ of signal x_i, where $i = 1, \ldots, p$, is larger than $\hat{\sigma}_n$ and decreases as component's rank increases.

The third step of ODM includes outlier detection using quartile range. Outlier detection is widely used to detect and/or remove anomalous observations from the data. It is a primary step in many data-mining applications [12]. There are many definitions given for outliers. The one that fits on the particular approach is given by Hawkins [21] who defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

The sample mean and the sample variance give good estimation for data location and data shape, but they are affected by outliers. IQR-based method [12], [22] is one of the most common methods for outlier detection as IQR is a robust statistic compared to total range and standard deviation. The method can be used when data distribution is unknown. Assume that observed values are placed in ascending order. The lower quartile Q_1 is the observation at the 25th percentile, the second quartile Q_2 is defined the observation at the 50th percentile, and the third quartile Q_3 is the observation at the 75th percentile. The quantity $\mathrm{Q}_3 - \mathrm{Q}_1$ is called the inter quartile range (IQR) and it provides a means to indicate the boundary beyond which the data will be labelled as outliers. More precisely, if an observation is below $\mathrm{Q}_1 - 1.5 \times \mathrm{IQR}$ or above $\mathrm{Q}_3 + 1.5 \times \mathrm{IQR}$, it is viewed as being too far from the central values to be reasonable.

In most real applications, signal subspace, and consequently signal vectors are unknown and even if they are known they are very few in order to be statistically analyzed. Noise subspace consists always of some hundreds of components which are much more than the signal components in the transformed hyperspectral space. Assuming that standard deviations of all the principal components correspond to noise, it is expected that the whole data lies in a hypersphere of radius σ_n. Thus, signal components can be considered as outliers of noise hypersphere.

As was mentioned in the previous section, the radius of noise hypersphere is much smaller than the radius of the signal hyperellipsoid and since search is focused on detecting noise hypersphere outliers, only the upper bound is of interest in this particular procedure. Let us assume that $\Sigma = [\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_L]$ is a $L \times 1$ vector which consists of the standard deviations $\hat{\sigma}_i$ of each i^{th} transformed component. The transformed components are ranked according to the SNR, which implies that $\hat{\sigma}_i$ are in descending order and thus, the first p values of Σ correspond to signal vectors. Taking the $\Sigma^R = [\hat{\sigma}_L, \hat{\sigma}_{L-1}, \ldots, \hat{\sigma}_1]$ as the $L \times 1$ vector which consists of the standard deviations $\hat{\sigma}_i$ in reverse order, meaning in ascending order, the first $L - p$ values of Σ^R correspond to noise vectors. As p is unknown, we suppose that all of the values of Σ^R correspond to noise. Euclidean distance (ED) of adjacent values of Σ^R, $ED = [ED(\hat{\sigma}_L, \hat{\sigma}_{L-1}), ED(\hat{\sigma}_{L-1}, \hat{\sigma}_{L-2}), \ldots, ED(\hat{\sigma}_2, \hat{\sigma}_1)]$ reflects possible divergences which are considered reasonable

when $\hat{\sigma}_i$ corresponds to noise, but outliers when $\hat{\sigma}_i$ corresponds to signal.

Proposed Algorithm (ODM)

Data: The L × N matrix Y, where N are the $L \times 1$ observed spectral vectors and L the spectral bands
Result: Estimation of the number of endmembers p
Step 1: *Noise estimation*
$\mathrm{Z} = \mathrm{Y}^{\mathrm{T}}$
for $i = 1$ to L, $z_i = [Z]_{:,i}$ all the pixels of band i

$Z_{\partial i} = [z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_L]$ the pixels of all the bands except i
$z_i = Z_{\partial i} b_i + n_i$ express band i as linear combination of L-1 bands
$\hat{b}_i = (Z_{\partial i}^T Z_{\partial i})^{-1} Z_{\partial i}^T z_i$ calculation of coefficient b
$\hat{n}_i = z_i - Z_{\partial i} \hat{b}_i$ noise estimation for band i

end for
OUTPUT $\hat{n}$
Step 2: *White noise data transformation*
Estimation of the noise covariance matrix R_n
Orthogonal decomposition of R_n: $D_n = [d_{n1}, d_{n2}, \ldots, d_{nL}]$

$F = D_n^T Y$
$W = F^T$,
$w_i = [W]_{:,i}$ are the transformed pixels in band i

for $i = 1$ to L, estimation of $\hat{\sigma}_{ni}$ (standard deviation of noise)

$w_i' = w_i / \hat{\sigma}_{ni}$

end for
Estimation of the covariance matrix of W'^T
Orthogonal decomposition of W'^T: $D_{W'^T}$
Transformation of whitened data $Y' = D_{W'^T}^T W'^T$
OUTPUT: Transformed whitened data Y'
Step 3: *Outlier detection*
Estimation of standard deviation of Y' : $\Sigma = [\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_L]$
Normalization of Σ
Put $\hat{\sigma}_i$ in descending order $\Sigma^R = [\hat{\sigma}_L, \hat{\sigma}_{L-1}, \ldots, \hat{\sigma}_1]$
Calculation of Euclidean distance for $\hat{\sigma}_i$ of adjacent bands

$$ED = [ED_{L,L-1}(\hat{\sigma}_L, \hat{\sigma}_{L-1}), ED_{L-1,L-2}(\hat{\sigma}_{L-1}, \hat{\sigma}_{L-2}), \ldots, ED_{2,1}(\hat{\sigma}_2, \hat{\sigma}_1)]$$

Retrieval of quartiles from EDs:

for i = 1 to 4

$\mathrm{k} = \mathrm{i} * (25\%) * (\mathrm{L} - 1)$
$\mathrm{Q}_i = ED_{\mathrm{k},\mathrm{k}-1}$

end for

$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$

Definition of the number p of endmembers

$\mathrm{p} = 0$

for i = 2 to L

if $ED_{i,i-1}$ is greater than Q3 + 1.5 * IQR
p++

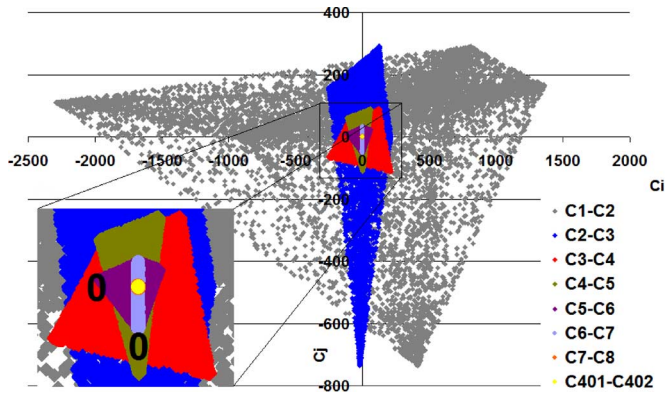

Fig. 2. Overlapping scattergrams of pairs of adjacent principal components. Only two axes are used (i, j). Each component (B) is kept on the same axis for the pairs in which is encountered.

end for
OUTPUT: Number of endmembers p

A. Geometrical Concept

For illustration purposes, the following experiment on simulated hyperspectral data is implemented. The simulated data generated according to the linear mixing scattering mechanism using seven random spectral signatures from the U.S. Geological Survey (USGS) digital spectral library and consist of 10^4 pixels and 423 spectral bands. The abundance fractions follow a Dirichlet distribution according to [11] enforcing positivity and full additivity constraints. Gaussian colored noise was added to the data resulting in an SNR of 20 dB. The procedure described previously containing noise estimation using multiple regression theory, noise whitening and transformation into a new principal component space and outlier detection is implemented.

Fig. 2 shows the distribution of the transformed data through overlapping scattergrams of pairs of adjacent components. Only two axes are used (i, j). Each component is kept on the same axis for the two pairs in which is encountered. The extent of each scattergram in i and j directions implies the magnitude of the standard deviation of Ci and Cj component, respectively. As it is observed, standard deviations of the first six principal components are relatively high and as band rank increases, standard deviations increase. More precisely, suppose $\hat{\sigma}_i$ denotes the standard deviation of band i. As shown in the overlapping scattergram, the following relation exists: $\hat{\sigma}_1 > \hat{\sigma}_2 > \ldots > \hat{\sigma}_6 \gg \hat{\sigma}_7 \approx \hat{\sigma}_8 \approx \ldots \approx \hat{\sigma}_{402} \approx \hat{\sigma}_L$. Furthermore, it is remarkable that noise circle (in this case, it is not hypersphere since scattergrams are shown in two dimensions) can be detected from the C7–C8 pair (orange circle) and after. This means that, in the hyperspectral space, the radius of the noise hypersphere is associated with the standard deviation of the seventh component which is right after the $p - 1$ component. It should be noted that, since simulated data are generated according to a linear mixing model, the dimension of the signal subspace is $p - 1$. The scattergram of the 401st and 402nd components was randomly selected to testify the equality of the noise standard deviation σ_n in all of the directions (Fig. 2). The difference between the radius of the orange circle (Components 7–8) compared with the radius of the

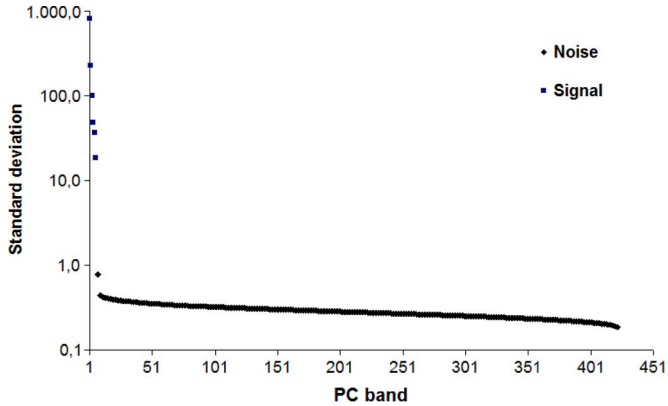

Fig. 3. Diagram of the standard deviations of each principal component ($p = 7$). A logarithmic scale is used on the y-axis.

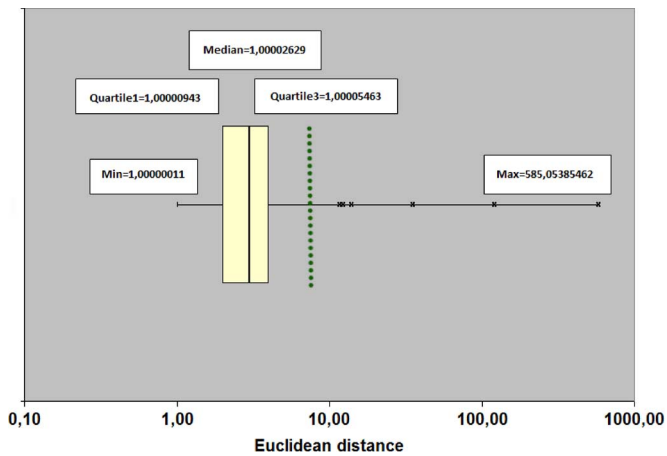

Fig. 4. Box plot indicating the existence of outliers (on the right side of the green line). A logarithmic scale is used on y-axis.

yellow circle (Components 401–402) can be considered without loss of generality as a result of fluctuations.

Another approach for studying the standard deviations of the principal components is by plotting them on a diagram. Fig. 3 shows the standard deviations of each principal component resulted from the above experiment. By observing the diagram, it is easy to perceive that standard deviations of the noise (in black) are almost similar while standard deviations of signal (in blue) differ greatly from each other. The optimum threshold by which signals are discerned from noise is estimated by using the IQR-based method.

A graphical display on which outliers can be indicated is a Box plot [12]. Fig. 4 shows the Box plot created by using the EDs between the standard deviation values. The majority of the EDs are close to one and reside on the left side of the green line, which indicates the upper bound. Black points represent the outliers. As can be observed, the differences in EDs between the three quartiles are negligible compared to the values of the six outliers.

VI. Experiments

A. Simulated Data Experiments

The ODM algorithm was applied on simulated data and compared with the state-of-the-art signal subspace methods, the HySime method and the NWHFC eigen-based Neyman-Pearson detector. The simulated data were generated by a random set of 15 spectral signatures with 423 spectral bands from the U.S. Geological Survey (USGS) digital spectral library. The abundance fractions follow a Dirichlet distribution according to [11] enforcing positivity and full additivity constraints. Experiments were conducted with respect to: 1) the size of the image N; 2) the number of endmembers p; 3) the SNR values; 4) the type of noise (white noise and Gaussian shaped noise)[1]; and 5) the existence of outliers.

The reason that different image sizes are introduced is twofold. First, due to sampling error, estimation of the noise covariance matrix R_n and estimation of standard deviation $\hat{\sigma}$ are both affected by the sample size, and they should be examined and evaluated using a smaller image size, as well. Second, recent developed endmember extraction methods tend to integrate spatial information into the endmember extraction process [23]. Towards this direction, these methods search for local endmembers in subsets of image data. Therefore, effectiveness of the ODM is examined for such a scenario. Thus, two sets of simulated hyperspectral images were created which differ in size, containing 2500 and 10^4 pixels, respectively. Furthermore, evaluation of the proposed method regarding various numbers of endmembers should also be tested. According to [14], the number of endmembers that may be practically identified typically ranges from three to seven, depending on the number of bands and the spectral variability of the scene components. In the case of high spectral resolution, the hyperspectral datasets may comprise even more, i.e., AVIRIS Cuprite image consists of at least 18 distinct spectral signatures according to the USGS. Therefore, the number of endmembers p was determined to be 3, 7, and 15. Two different types of noise were added in the simulated images; white noise and Gaussian shaped noise with variance σ_n^2 equal to 0.02, leading to SNR values of 50, 30, 20, and 10 dB. A noise estimation step is required in order to transform noise to zero mean i.i.d.. In the case of simulated images with white noise, the last is spherically symmetric in all directions and lies on the surface of a hypersphere with a constant radius Therefore, noise estimation can be omitted.

Figs. 5 and 6 show the standard deviation values for each transformed component of images with $N = 10^4$, with p equal to 3, 7 and 15 and white and colored noise respectively. It is observed that standard deviation values minimize and stabilize when the number of the transformed components is equal to the number of the endmembers. For clarity purposes, it was chosen to present a subset of the transformed components of all the simulated images in a stacked plot, and therefore the scale of the values in Fig. 5 and Fig. 6 has changed. Table I shows the results of the applied methods for images with white noise. As it is concluded from the results, regarding the images of 2500 pixels, the ODM yielded quite satisfactory results outperforming the HySime and NWHFC algorithms when SNR values were very low (30 dB-10 dB). For the images of 10^4 pixels, all the applied methods yielded the same high performance for the images which contain 3 endmembers, regardless the amount of noise.

[1]The algorithm which was used for the generation of the simulated data is available at http://www.lx.it.pt/~bioucas/code.htm.

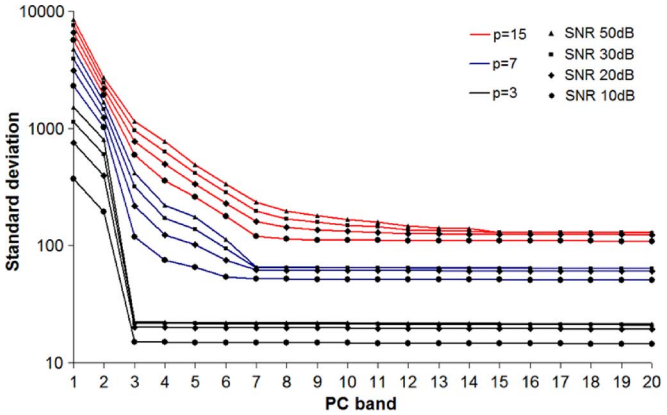

Fig. 5. Stacked plots of standard deviation values for each PC band for the images with $N = 10^4$ and white noise.

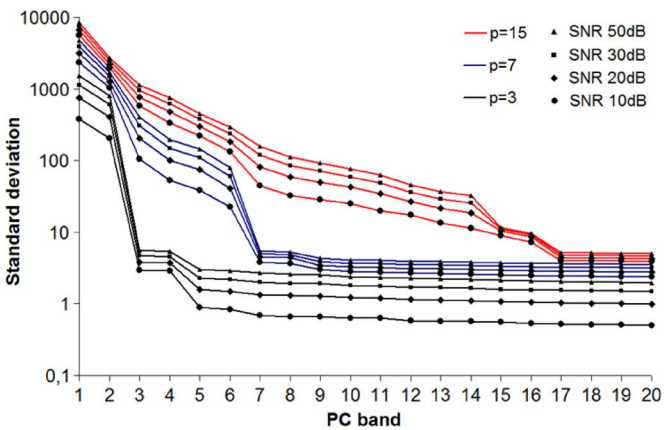

Fig. 6. Stacked plots of standard deviation values for each PC band for the images with $N = 10^4$ and Gaussian shaped noise.

When the space dimension increased, with $p = 7$, the proposed method outperformed the NWHFC method and it had the same high performance with HySime, except for the image with SNR of 10 dB for which the proposed method performed better. For the images with 15 endmembers, the proposed method yielded systematically better results than both HySime and NWHFC.

Table II shows the results of the applied methods for images with Gaussian shaped noise. Two different methods for noise estimation were implemented. As it was expected, NND noise estimation led to the worst results. This is reasonable because NND needs to calculate the shift difference in homogeneous area while pixels in simulated data were created randomly without homogeneous areas. The most satisfactory results were given by ODM for both image sizes when multiple regression theory based method was used for noise estimation. Especially in case of low SNR, results are much more satisfactory compared to the results from HySime, while both methods presented similar results for high SNR. The NWHFC method, as implemented in [24], [25], presented the worst results.

In order to test the method's resistance to outliers, simulated images containing 7 endmembers and Gaussian shaped noise with SNR values of 30 dB and 50 dB were used. Outliers were added to the images by randomly sampling three outlying points from a uniform distribution, according to [26]. Table III reports the results. Estimations of the proposed method are satisfactory and testify its resistance to outliers.

B. Real Data Experiments

The proposed algorithm was applied on two real hyperspectral remote sensing images in order to be evaluated in case of unequally distributed noise. The first image was acquired in June, 1992 by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over an agricultural area of north-western Indiana (Indian Pines) (Fig. 7). It consists of 145 $\times$ 145 pixels with 220 spectral bands covering a spectral range from 400 to 2500 nm. The number of bands was reduced to 186 after removing 34 bands due to water absorption and low SNR. According to the associate ground based observations,[2] 16 land cover classes exist in the image; alfalfa, corn-notill, corn-mintill, corn, grass-pasture, grass-trees, grass-pasture-mowed, hay-windrowed, oats, soybean-notill, soybean-mintill, soybean-clean, wheat, woods, buildings-grass-trees-drives and stone-steel-towers. It should be noted that the aforementioned classes do not represent the entire scene and some of them are not associated with pure materials. Consequently, the number of the endmembers is expected to be higher than 16. Fig. 8 shows the standard deviation values for each transformed component of the AVIRIS image and Table IV shows the estimated number of endmembers from the applied methods.

As it is listed in Table IV, the ODM using NDD and multiple regression theory based method for noise estimation and the NWHFC resulted in a reasonable number of the distinct classes while the HySime underestimated it. The fact that NWHFC estimates were much higher than its competitors is reasonable since the method searches for signal sources which may include not only endmembers but also unknown interferences, such as clutters, background signatures and anomalies [5].

The second real hyperspectral dataset which has been used for evaluation was collected in 1997 by the AVIRIS sensor over a well-known mining region of Cuprite in Nevada. The image scene is well understood mineralogically and the ground truth spectral signatures are available in the USGS digital library. According to the associated ground based observations and the mineral map produced in 1995 by USGS,[3] 18 minerals can be identified in the image. Besides minerals, there should be other distinct classes depicted in the image, whose amount is unknown. Thus, the number of endmembers is expected to be higher than 18. The original image has 220 spectral bands covering a spectral range from 0.4 to 2.5 μm. The number of bands was reduced to 188 after removing bad bands due to water absorption and low SNR. Fig. 9 shows the subimage scene of 351 $\times$ 350 pixels with reflectance values which was selected for the experiments. Fig. 10 shows the standard deviation values for each transformed component of the AVIRIS image. Table V shows the estimated number of endmembers from the applied methods.

[2][Online]. Available: https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.

[3][Online]. Available: http://speclab.cr.usgs.gov/cuprite95.tgif.2.2um_map.gif.

TABLE I
ESTIMATED NUMBER OF ENDMEMBERS FROM IMAGE WITH WHITE NOISE AS FUNCTION OF SNR, p AND N

SNR	Method	White Noise					
		2500 pixels			10^4 pixels		
		p=3	p=7	p=15	p=3	p=7	p=15
50 dB	ODM	3	7	15	3	7	15
	HySime	3	7	15	3	7	15
	NWHFC $P_f=10^{-3}$	3	4	5	3	5	9
	$P_f=10^{-4}$	3	4	4	3	5	7
	$P_f=10^{-5}$	3	4	4	3	5	7
30 dB	ODM	3	7	15	3	7	15
	HySime	21	23	24	3	7	14
	NWHFC $P_f=10^{-3}$	3	5	4	3	6	6
	$P_f=10^{-4}$	3	4	3	3	5	6
	$P_f=10^{-5}$	3	4	3	3	5	5
20 dB	ODM	3	7	13	3	7	13
	HySime	23	26	25	3	7	11
	NWHFC $P_f=10^{-3}$	3	4	1	3	6	5
	$P_f=10^{-4}$	3	4	1	3	5	4
	$P_f=10^{-5}$	3	4	1	3	5	3
10 dB	ODM	3	7	10	3	7	11
	HySime	20	27	24	3	6	7
	NWHFC $P_f=10^{-3}$	3	3	1	3	5	3
	$P_f=10^{-4}$	3	3	1	3	5	3
	$P_f=10^{-5}$	3	3	1	3	4	3

TABLE II
ESTIMATED NUMBER OF ENDMEMBERS FROM IMAGE WITH GAUSSIAN SHAPED NOISE AS FUNCTION OF SNR, p AND N [1] STANDS FOR MULTIPLE REGRESSION, [2] STANDS FOR NND

SNR	Method	Gaussian shaped noise (σ^2=0.02)					
		2500 pixels			10^4 pixels		
		p=3	p =7	p =15	p =3	p =7	p =15
50 dB	ODM [1]	3	7	15	4	7	15
	ODM [2]	49	39	48	32	38	33
	HySime	3	7	15	3	7	15
	NWHFC $P_f=10^{-3}$	62	13	8	47	23	13
	$P_f=10^{-4}$	53	12	7	41	22	11
	$P_f=10^{-5}$	52	12	6	37	20	11
30 dB	ODM [1]	3	8	16	4	7	15
	ODM [2]	39	40	47	34	31	35
	HySime	3	7	13	3	7	14
	NWHFC $P_f=10^{-3}$	45	58	37	66	48	13
	$P_f=10^{-4}$	41	54	31	59	43	13
	$P_f=10^{-5}$	38	47	29	56	36	12
20 dB	ODM [1]	4	8	17	4	8	16
	ODM [2]	48	48	45	30	25	42
	HySime	3	6	6	3	6	8
	NWHFC $P_f=10^{-3}$	30	23	72	23	56	45
	$P_f=10^{-4}$	24	20	60	18	48	38
	$P_f=10^{-5}$	23	17	49	17	43	32
10 dB	ODM [1]	5	9	17	5	8	17
	ODM [2]	47	38	40	37	34	32
	HySime	3	4	5	3	5	5
	NWHFC $P_f=10^{-3}$	12	16	5	56	67	45
	$P_f=10^{-4}$	8	14	5	48	60	36
	$P_f=10^{-5}$	5	14	4	45	50	32

As it is shown in Table V, the ODM using NDD and multiple regression theory based method for noise estimation and the NWHFC resulted in number of the distinct classes higher than 18 while the HySime underestimated it. Particularly, ODM using NDD significantly overestimates the number of endmembers since the Cuprite image does not include adequate number of pixels belonging to homogeneous areas.

TABLE III
ESTIMATED NUMBER OF ENDMEMBERS FROM IMAGE WITH THREE OUTLIERS

Method	p=7	
	50dB	30dB
ODM[2]	7	8
HySime	7	6
NWHFC $P_f=10^{-3}$	7	6
$P_f=10^{-4}$	6	5
$P_f=10^{-5}$	6	5

Fig. 7. AVIRIS Indian pines hyperspectral dataset.

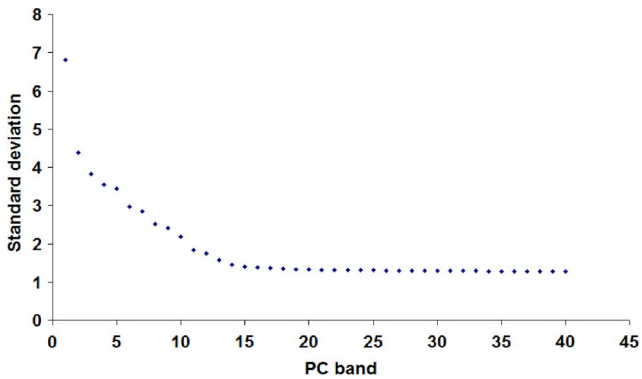

Fig. 8. Standard deviation values of the transformed bands for AVIRIS Indian pine image.

TABLE IV
ESTIMATED NUMBER OF ENDMEMBERS FOR THE AVIRIS INDIAN PINES IMAGE [1] STANDS FOR MULTIPLE REGRESSION, [2] STANDS FOR NND

Method	Estimated number of endmembers (reference number: higher than 16)
ODM [1]	17
ODM [2]	24
HySime	14
NWHFC $P_f=10^{-3}$	27
$P_f=10^{-4}$	23
$P_f=10^{-5}$	22

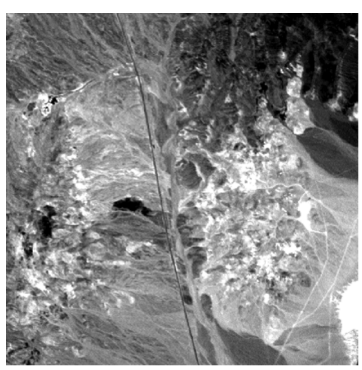

Fig. 9. AVIRIS Cuprite hyperspectral data.

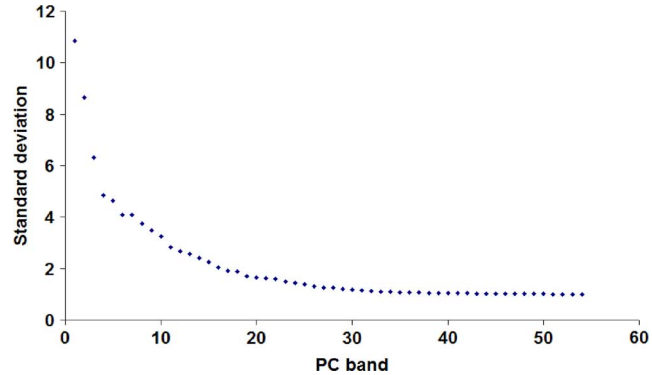

Fig. 10. Standard deviation values of the transformed bands for AVIRIS Cuprite image.

TABLE V
ESTIMATED NUMBER OF ENDMEMBERS FOR THE AVIRIS CUPRITE IMAGE [1] STANDS FOR MULTIPLE REGRESSION, [2] STANDS FOR NND

Method	Estimated number of endmembers (reference number: higher than 18)
ODM [1]	20
ODM [2]	29
HySime	15
NWHFC $P_f=10^{-3}$	22
$P_f=10^{-4}$	21
$P_f=10^{-5}$	19

VII. CONCLUSION

In this paper, a new automatic and nonparametric method for the estimation of the number of the endmembers in hyperspectral imagery was introduced. The proposed method, called the outlier detection method (ODM) develops a novel approach considering signal as an outlier of the noise hypersphere. In particular, after noise estimation and whitening process, the transformed data reside in a principal component space where noise presents spherically symmetry towards all the directions, having a constant radius. Conversely, signal radius varies in all of the directions, and it is much larger than the noise radius in the components which include it. Estimation of the number of noise hypersphere outliers using a robust IQR-based outlier detection method results in the estimation of the number of endmembers. The proposed method is characterized by its simplicity and its significant benefit to refrain from estimation of statistical distributions.

Experiments using simulated data proved the efficiency of the ODM which outperformed compared with its competitors. The performance of the proposed method is quite satisfactory in real data, as well. Through this particular work it is concluded that a successful estimation of the number of endmembers strongly depends on how well signal and noise are discerned. Outlier detection theory could be efficiently used for this goal. Additional experiments could contribute to a superior performance of the proposed method. Future research should also focus on combining the proposed method with endmember extraction methods which integrate spatial information, taking advantage of its successful estimation of the number of endmembers in small-sized images.

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Nascimento and Dr. P. Kolokoussis for their valuable advice and comments as well as the associate editor and the anonymous reviewers whose constructive suggestions contributed to the improvement of this manuscript.

REFERENCES

[1] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *MIT Lincoln Lab. J.*, vol. 14, no. 1, pp. 9–116, 2003.

[2] D. E. Sabol, J. B. Adams, and M. O. Smith, "Quantitative sub-pixel spectral detection of targets in multispectral images," *J. Geophys. Res.*, vol. 97, no. E2, pp. 2659–2672, 1992.

[3] K. Fukunaga, "Intrinsic dimensionality extraction," in *Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, vol. 2, pp. 347–360.

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

[6] J. A. Richards, *Remote Sensing Digital Image Analysis, An Introduction*, 2nd ed. New York, NY, USA: Springer-Verlag, 1993.

[7] S. Moussaoui, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J. Benediktsson, "On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation," *Neurocomput. Vis. Res.; Adv. Blind Signal Process.*, vol. 71, no. 10–12, pp. 2194–2208, Jun. 2008.

[8] B. Luo, J. Chanussot, S. Doute, and L. Zhang, "Empirical automatic estimation of the number of endmembers in hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 24–28, Jan. 2013.

[9] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[10] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[11] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 8, pp. 2435–2445, Aug. 2008.

[12] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley, 1977.

[13] C. Andreou and V. Karathanassi, "New automated method for estimating the number of endmembers in hyperspectral images," in *Proc. 4th Wkshp Hyperspectral Image and Signal Process.: Evolution in Remote Sens.*, pp. 1–4.

[14] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 44–57, Jan. 2002.

[15] L. L. Scharf, *Statistical Signal Processing, Detection Estimation and Time Series Analysis*. Reading, MA, USA: Addison-Wesley, 1991.

[16] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[17] J. R. Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise*, 2nd ed. New York, NY, USA: Dover.

[18] F. M. J. Willems, *Information and Communication Theory: Communication Theory*. Eindhoven, The Netherlands: Eindhoven Univ. Technol., 2010.

[19] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of images quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, Jan. 1988.

[20] R. E. Roger and J. F. Arnold, "Reliably estimating the noise in AVIRIS hyperspectral imagers," *Int. J. Remote Sens.*, vol. 17, no. 10, pp. 1951–1962, 1996.

[21] D. Hawkins, *Identification of Outliers*. London, U.K.: Chapman and Hall, 1980.

[22] I. Ben-Gal, "Outlier detection," in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, O. Maimon and L. Rockach, Eds. Amsterdam, The Netherlands: Kluwer Academic, 2005.

[23] A. Plaza, G. Martın, J. Plaza, M. Zortea, and S. Sanchez, "Recent developments in endmember extraction and spectral unmixing," in *Optical Remote Sensing: Advances in Signal Processing and Exploitation Techniques*, S. Prasad, L. Bruce, and J. Chanussot, Eds. Berlin, Germany: Springer, 2011, pp. 235–267.

[24] Open Source MATLAB Hyperspectral Toolbox 2012, ver. 0.06 [Online]. Available: http://matlabhyperspec.sourceforge.net/

[25] Endmember Induction Algorithms (EIAs) Toolbox Grupo de Inteligencia Computacional, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Spain [Online]. Available: http://www.ehu.es/computationalintelligence/index.php/Endmember_Induction_Algorithms

[26] A. Zare and P. Gader, "L1-endmembers: A robust endmember detection and spectral unmixing algorithm," in *Proc. SPIE Defense, Security and Sensing*, Orlando, FL, USA, 2010, vol. 7695.

Charoula Andreou (S'11) received the Dipl.Ing. degree in rural and surveying engineering from National Technical University of Athens, Athens, Greece, in 2008, where she is currently working toward the Ph.D. degree in hyperspectral remote sensing.

Since 2009, she is a Research Associate with the Laboratory of Remote Sensing, National Technical University of Athens, Athens, Greece, participating in several European and national research projects. She is currently a Visiting Researcher with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Munich, Germany. Her main research interests are focused on hyperspectral remote sensing, spectral unmixing, feature extraction, and target detection.

Vassilia Karathanassi received the Dip.Ing. degree from the National Technical University of Athens (NTUA), Athens, Greece, in 1984, the D.E.A. degree in "aménagement urbanisme geographie" from the University of Paris IV, Paris, France, in 1985, the D.E.A. degree in "methodes physiques en teledetection" from the University of Paris VII, Paris, in 1986, and the Ph.D. degree from the School of Rural and Surveying Engineering (SRSE), NTUA, in 1990.

She is currently an Assistant Professor with the School of Rural and Surveying Engineering, National Technical University of Athens, Athens, Greece. She has taught courses in photointerpretation, hyperspectral remote sensing, remote sensing of the environment, microwave remote sensing, and synthetic aperture radar (SAR) interferometry and polarimetry. She has participated in research projects as a Principal Investigator (eight projects) and Project Leader (eight projects). She has authored and coauthored 23 papers in referred journals and more than 50 papers in proceedings of national and international congresses. She has been a reviewer for the *Journal of Photogrammetry and Remote Sensing*, *Photogrammetric Engineering and Remote Sensing*, and the *International Journal of Remote Sensing*. Her main research areas are remote sensing, image processing, feature extraction, texture analysis, neural networks, SAR applications, and hyperspectral data processing.

Prof. Karathanassi is a reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.