Biomedical Signal Processing and Control 95 (2024) 106386

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Unsupervised method for estimating the number of endmembers in hyperspectral images☆

Karina Baños [a,*], Julio Esclarín [b], Juan Ortega [a]

[a] Institute for Technological Development and Innovation in Communications, University of Las Palmas de Gran Canaria, Canary Islands, 35001, Spain
[b] Institute of Cybernetics, Companies and Society, University of Las Palmas de Gran Canaria, Canary Islands, 35001, Spain

ARTICLE INFO

Keywords:
Hyperspectral imaging
Endmember
Unmixing
Solution
Dimensional reduction
Explained variance

ABSTRACT

Accurately determining the number of pure elements, or endmembers, in a mixture is crucial for unmixing applications in hyperspectral image processing. This work introduces a new unsupervised method, called 'Number of Endmembers by Energy Criteria' (NEEC), for estimating the number of endmembers in homogeneous solutions of organic compounds in the liquid state, such as esters, hydrocarbons, and alcohols. The NEEC method utilizes eigenvalue analysis and incorporates an energy concept based on the eigenvalues of the sample correlation matrix. Experiments were conducted on both real and synthetic samples to assess the effectiveness of the proposed algorithm. Synthetic mixtures were created using a non-linear method. The results demonstrate that the NEEC method is highly effective, achieving 86.6% accuracy in estimating the number of endmembers. This highlights its potential for analyzing non-linear samples. This research contributes to the advancement of hyperspectral image processing techniques for unmixing applications.

1. Introduction

Hyperspectral imaging (HSI) is a non-invasive technique that captures a spectrum for each pixel in an image. It facilitates object detection, material identification, and process monitoring by providing both spatial information and high spectral resolution. HSI utilizes hundreds or thousands of wavelength channels, depending on the measuring instrument.

HSI has been widely applied in various scientific disciplines, such as pharmaceutical research [1,2], environmental contamination [3], and medicine [4]. In an HSI scene, pixels are considered mixtures of multiple pure components, known as endmembers. A fundamental challenge in hyperspectral image processing is to separate these endmembers, estimate the optimal number of endmembers in a pixel, extract the spectral signature of the endmembers, and determine the abundance of each endmember in the pixel.

Spectral unmixing is a process that varies depending on the nature of mixtures. Linear mixing occurs when the incident light from the pure components is sufficiently separated to allow mixing within the measuring instrument.

Non-linear mixing, on the other hand, occurs when light scatters due to interaction with various materials in the scene before reaching the instrument or when materials are homogeneously mixed and molecules interact, as in a soluble liquid solution [4].

In this context, unmixing techniques are necessary to separate and characterize the individual contributions of each source due to the superposition of signals from different sources within a single image pixel. The Linear Unmixing Model (LUM) is an essential tool for interpreting and exploiting hyperspectral data.

LUM is a mathematical approach that assumes a hyperspectral observation can be represented as a linear combination of the spectral signatures present in the scene. This model allows estimation of the fraction of each source in each pixel of the image, providing valuable information about the composition and distribution of the endmembers.

LUM is widely accepted due to its ease of implementation compared to non-linear algorithms. The linear approach provides a good initial approximation for non-linear applications. Hapke proposed a macroscopic model by linearizing the non-linear intimate model. This model assumes that the observed pixel spectrum is a weighted linear combination of the endmember spectra.

Advancements in this field are crucial for unlocking its full potential in various applications, such as precision agriculture, geological exploration, crop health monitoring, and water pollution detection. Unmixing techniques are crucial for monitoring water quality as they allow for the identification and quantification of various pollutants such as oils, chemicals, suspended solids, hydrocarbons, and other compounds in aquatic environments. This capability highlights the

☆ This work has been supported by The Ministry of Science and Innovation of the Government of Spain.
* Corresponding author.
E-mail addresses: karina.banos@ulpgc.es (K. Baños), julio.esclarin@ulpgc.es (J. Esclarín), juan.ortega@ulpgc.es (J. Ortega).

https://doi.org/10.1016/j.bspc.2024.106386
Received 20 November 2023; Received in revised form 12 March 2024; Accepted 23 April 2024
Available online 22 May 2024

importance of HSI in water resource management. It enables early detection of pollution and comprehensive assessments of its scope and impact [5–8].

This research focuses on the initial step of the unmixing process, which involves accurately selecting the number of endmembers. Determining the number of components or sources in multivariate data analysis using a linear mixture model is a critical step that significantly impacts the accuracy and utility of unmixing results. In chemometrics, this often involves identifying the chemical rank. Accurately executing self-modeling curve resolution techniques requires precise estimation of the number of chemical components. This precision ensures the correctness of the curve resolution process, as noted in [9].

Three methodologies have been developed to estimate the number of endmembers: information theory-based algorithms, eigenvalue thresholding techniques, and geometric characterization methods. Information-theoretic criteria are used in the first category of algorithms, including methods based on the minimum description length [10], the Akaike information criterion [11], and the Bayesian information criterion [12], among others.

The second category is focused on the eigenvalue thresholding techniques [13–15]. These techniques involve setting a threshold on the eigenvalues obtained from the eigendecomposition of data matrices. This process is foundational to subspace analysis methods such as Principal Component Analysis (PCA) [16], hyperspectral signal subspace estimation through minimum error (Hysime) [17], and the Harsanyi–Farrand–Chang (HFC) analysis [18]. The HFC method evaluates the number of endmembers by comparing the eigenvalues of the correlation matrix to those of the covariance matrix.

This approach has led to the emergence of various methods, such as Das et al. [19] who proposed a technique to determine the optimal eigenvalue cutoff for signal components by applying principles from random matrix theory. Similarly, Zhu et al. [20] analyzed the discrepancies between the eigenvalues of the correlation and covariance matrices, providing another perspective on discerning the dimensional structure of the data. Das et al. [21] also contributed by creating the GAP index, a metric designed to precisely estimate the number of endmembers.

The third category includes methods that use geometric analysis to estimate the number of endmembers, including algorithms such as convex hull (GENE-CH) and affine hull (GENE-AH) [22]. The algorithms assume that the data samples are enclosed within a geometric shape, specifically either a convex hull (CH) or an affine hull (AH), with the vertices of these shapes representing the endmember signatures. This approach uses the spatial relationships and distributions of data points in a multidimensional space to identify the smallest set of points (endmembers) that can represent all other points within the dataset.

Furthermore, recent methodologies have gained recognition, particularly those that incorporate saliency analysis [23], Matrix Theory [24], and the use of simplified fuzzy sets [25]. Graña et al. [26] proposes the theory of lattice-based autoassociative memories to develop new methods for autonomous endmember determination [27, 28].

Additionally, there has been a significant increase in the application of machine learning, especially neural networks, to address the challenges of HSI unmixing [29,30]. These learning-based strategies, which include both supervised and unsupervised learning algorithms, have introduced a new dimension to the field [31].

This paper presents the NEEC method, which uses eigenvalue analysis for the estimation of the number of endmembers in non-linear homogeneous liquid mixtures. These mixtures are created by combining pure liquids from categories such as alcohols, hydrocarbons, and esters in various proportions. Notably, the NEEC method does not require any parameters and involves an eigenvalue transformation of the correlation matrix from an HSI sample to compute a functional from these ordered eigenvalues.

This transformation allows the calculation of a sequence that represents the ratio of consecutive values of the previously obtained functional. The index of the maximum value in this sequence indicates the estimated number of endmembers.

Both real and simulated images from our own collections and external databases were used in the experiments of this study. Simulated spectra were generated using the Linear Quadratic Mixing (LQM) model to reflect the non-linear aspects of the experiments. An extensive comparison with other methods was also performed to evaluate the efficiency and robustness of the proposed method.

2. Related works

2.1. HySIME

This method is based on the principle of minimizing the mean square error (MSE) to infer the subspace of the signal. HySime [17] stands out as an unsupervised method that does not require tuning parameters. Its algorithm estimates the signal and noise correlation matrices and represents the signal subspace by a subset of the eigenvectors of the signal correlation matrix. This subspace is inferred by minimizing the sum of the projection error power with the noise power.

From the correlation matrix of the sample $R_{l,l} = \frac{1}{n}\sum_{h=1}^{n} x_h x_h^T$ and considering the noise, ϵ_i, the following equation follows $\hat{R}_{l,l} = \frac{1}{n}\sum_{h=1}^{n}(y_h-\epsilon_i)(y_h-\epsilon_i)^T$ can be written as the sample correlation matrix of the signal.

Furthermore, performing eigenvalue decomposition of the signal sample correlation matrix $\hat{R}_{l,l}$, which is: $\hat{R}_{l,l} = E\Lambda E^T$ where $E = \{\xi_1, \xi_2, \ldots, \xi_l\}$ is a matrix with the columns being the eigenvectors of $\hat{R}_{l,l}$ and Λ is the eigenvalue matrix of $\hat{R}_{l,l}$. The signal subspace dimension is obtained by selecting a subset of eigenvectors $E_K = \{\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_K}\}$ where $\{i_1, \ldots., i_K\}$ is a subset of $\{1, 2, \ldots., l\}$. The optimal signal subspace is obtained by minimizing the MSE between the signal x_i and $\hat{x}_i = E_K E_K^T y_i$ where y_i is the projection of its observed sample.

2.2. Incremental lattice source induction algorithm (ILSIA)

ILSIA [27] starts with an initial selection of pixels that are considered potential endmembers. It then engages in an iterative process to add and refine the identified endmembers. During each iteration, the coherence between new pixels and existing endmembers is examined. Pixels with sufficient spectral difference, as measured by a spectral angle mapper (SAM) or Euclidean distance, are added. This can be evaluated quantitatively as follows:

$$D(pixel, endmember) = \arccos \frac{\langle pixel, endmember\rangle}{\|pixel\|\|endmember\|}$$

where D is the spectral difference, and the goal is to select pixels with D exceeding a certain threshold. As new endmembers are added, the algorithm adjusts the corresponding spectral abundances to reflect each contribution of endmembers to the image pixels. The abundance estimation for pixel i with respect to endmember j can be represented by:

$$A_{ij} = argmin_A \|I_i - \sum_j A_{ij} E_j\|^2$$

where A_{ij} is the abundance of endmember j in pixel i, I_i is the intensity of pixel i, and E_j is the spectral signature of endmember j. This optimization is subject to the constraints that all abundances are non-negative and their sum for each pixel is one. This process of endmember and abundance estimation continues until a predefined convergence criterion, such as a minimal change in the spectral information divergence (SID) between iterations, is met:

$$\text{SID} = \sum_i (A_{i,new} \log \frac{A_{i,new}}{A_{i,old}} + A_{i,old} \log \frac{A_{i,old}}{A_{i,new}}) < \epsilon$$

where $A_{i,new}$ and $A_{i,old}$ represent the abundance vectors of the current and previous iterations, respectively, and ϵ is a small positive number.

2.3. Uniform Manifold Approximation and Projection (UMAP)

UMAP [25] is an algorithm for dimensionality reduction that applies principles from topology and graph theory.

Its goal is to simplify multidimensional data while retaining essential structures, based on the theories of variety and Riemannian geometry. The fundamental idea is to model high-dimensional data as a smooth surface, suggesting that data points are evenly spread across an undefined shape. The method uses the Mahalanobis distance to calculate proximity among data points within a linear framework and transforms these distances into likelihoods of connection.

The main steps of the method are described below:

- For the set $X = x_i$, where $i = 1...N$, of data points in multidimensional space $\Re^M$, sets μ_i are found. Each set consists of k neighboring points for each data point x_i.
- For each data point i, the nearest neighbor and distance are found: $\rho_i = min(d(x_i, x_j)|x_j \in \mu_i, d(x_i, x_j) > 0)$ and also value σ_i such that

$$\sum_{x_j \in \mu_j} exp(\frac{-max(0, d(x_i, x_j)) - \rho_j}{\sigma_i}) = \log_2 k$$

- A UMAP graph G is constructed as an undirected weighted graph with an adjacency matrix $B = A + A^t - A \cdot A^t$ where the elements of A are given by the weights in the corresponding directed graph:

$$w(x_i, x_j) = exp(\frac{-max(0, d(x_i, x_j)) - \rho_j}{\sigma_i})$$

To organize the data in a more understandable way, the coordinates $y_i, i = 1...N$, of the data points in low-dimensional space are determined by force-directed placement of the graph using forces of attraction F^a and repulsion F^r between vertices i and j [32].

2.4. Saliency-based autonomous endmember detection (SAED)

The SAED [33] algorithm is based on the principle that certain endmembers, which were not initially detected, can be prominently identified within the abundance anomaly (AA) subspace as significant points. This technique enhances the Saliency-based Endmember Detection (SED) [23] approach by incorporating superpixels for a better spatial understanding of endmembers. SAED uses superpixels to identify the optimal number of endmembers by detecting when there are no additional prominent objects in the AA subspace.

This approach creates a saliency map from the hyperspectral image, evaluating each pixel based on its saliency rather than traditional intensity or spectral metrics. The saliency value for pixel i can be calculated using the following formula: $S(i) = \sum_{j \in N(i)} w_{ij} \cdot |I(i) - I(j)|$ where $N(i)$ represents the neighboring pixels around pixel i, w_{ij} is the weight that indicates the spatial distance between pixels i and j, and $|I(i) - I(j)|$ measures the difference in intensity or spectral information between pixels i and j.

The algorithm selects the most significant pixels as endmember candidates based on their unique spectral features. The selection process relies on an adjustable saliency threshold, T_s, to refine detection.

$$\text{Endmember candidates} = \{i|S(i) > T_s\}$$

After the selection process, a refinement stage is carried out to ensure that the chosen endmembers are both effective and distinct. This may involve optimizing the spatial distribution or spectral distinctiveness of the selected endmembers.

2.5. Linear and non-linear unmixing models

The mixing model can be linear or non-linear depending on how the light reaches the sensor. In the LUM, mixing occurs inside the sensor as a linear combination of the endmembers that compose it. The coefficients of these linear combinations represent the abundances of each endmember.

The observation of a hyperspectral pixel vector $x_h \in R^P$ at pixel h is denoted by $y_h \in R^L$. Here, L is the number of wavelength bands and P is the number of pixels. The LUM is defined as follows:

$$y_h = x_h + \epsilon_h = \sum_{i=1}^{K} e_i s_i + \epsilon_h = ES + \epsilon, i = 1, \dots, n \tag{1}$$

where $E = [e_1, \dots, e_K] \in R^K$ is the endmember matrix within each column, e_j stands for an endmember, and s_i is the abundance vector, where K is the number of endmembers present in the scene. It is commonly assumed that ϵ is the vector of the additional uncorrelated Gaussian noise.

As noted in previous work [34,35], abundances must satisfy positive and sum-to-one constraints:

$$s_i \geq 0, \qquad \forall i \in \{1, ..., P\} \qquad \text{and} \qquad \sum_{i=1}^{K} s_i = 1.$$

The Bilinear method is one of the most commonly used methods for non-linear unmixing. This method considers the presence of multiple photon interactions between the final members i and j (for $i, j = 1, \dots, p$ and $i \neq j$) so that the observed mixed pixel can be written as:

$$y = SE + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \beta_{ij} e_i \odot e_j + \epsilon \tag{2}$$

where E and S are defined in LUM, and $\odot$ is the Hadamard (term-by-term) product operation: $e_i \odot e_j = (e_{1,i}, e_{2,i}, \dots .e_{p,i})^T \odot (e_{1,j}, e_{2,j}, \dots .e_{p,j})^T = (e_{1,i} * e_{1,j}, e_{2,i} * e_{2,j}, \dots .e_{p,i} * e_{p,j})^T$.

Regarding the β_{ij} parameters and their constraints, some of the proposed approaches [36] can be seen in the Table 1.

Table 1
Parameters and constraints of bilinear models, where $s_i \geq 0$.

Name	Parameters	Constraints
Nascimento	$\forall i \geq j : \beta_{ij} = 0$ $\forall i < j : \beta_{ij} \geq 0$	$\sum_i s_i + \sum_{ij} \beta_{ij} = 1$
Fan	$\forall i \geq j : \beta_{ij} = 0$ $\forall i < j : \beta_{ij} = s_i * s_j$	$\sum_i s_i = 1$
GBM	$\forall i \geq j : \gamma_{ij} = 0$ $\forall i < j : 0 \leq \gamma_{ij} \leq 1$	$\sum_i s_i = 1$

The General Bilinear Model (GBM) can be derived by rewriting the Bilinear Model defined in Eq. (2):

$$y = SE + \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \gamma_{ij} s_i s_j e_i \odot e_j + \epsilon \tag{3}$$

The coefficient γ_{ij} controls the interactions between the endmembers i, j in the pixel. If $\gamma_{ij} = 0$, the bilinear model becomes the LUM. If $\gamma_{ij} = 1$ the bilinear model becomes the Fan model [37].

All of the bilinear models exclusively account for interactions between components $m_i \odot m_j$, with $i \neq j$, neglecting any interactions within the components themselves $m_i \odot m_i$.

In [38] the authors introduced a non-linear unmixing model by employing a Radiative Transfer model and employing successive approximations and simplifying assumptions, the resulting is the LQM model.

$$y = SE + \sum_{i=1}^{K} \sum_{j=1}^{K} \gamma_{ij} e_i \odot e_j + \epsilon \tag{4}$$

with $s_i > 0 \ \forall i$, $\sum_{i=1}^{M} s_i = 1$ and $\gamma_{ij} \in (0, 1)$.

3. Proposed method

This study presents the NEEC method, which is designed to determine the number of endmembers in chemical solutions without requiring parameter inputs. The method relies on eigenvalue analysis.

The sample images contain a combination of signal and noise, which is dependent on the conditions of data acquisition and processing, including thermal noise, quantization noise, and spatial noise. The images in this study were acquired under laboratory conditions, resulting in minimal noise. To further reduce noise, the HFC noise reduction method was employed, as described in [18].

In HSI, each pixel is characterized by its spectral information distributed across numerous bands. A methodological approach to signal analysis in such images involves the application of the correlation matrix:

$$R_{L,L} = \frac{1}{P}\sum_{j=1}^{P} y_j * y_j^T \tag{5}$$

In this context, the variable y represents the data matrix with dimensions $L \times P$, where L is the number of spectral channels and P is the number of pixels in the sample. This matrix quantifies the degree of linear dependence between the spectral bands. Examining the signal within the eigenvalues of this matrix reveals the effective dimension of the image. This dimension represents the minimum number of bands required to preserve most of the spectral information.

Therefore, the process begins by establishing the correlation matrix $R_{L,L}$ for the sample. Subsequently, the eigenvalues λ_i $(i = 1, \ldots, L)$ are computed and arranged in descending order, with the initial elements representing the bands with the greatest explained variance. In essence, the explained variance can be understood as the quantity of energy present in the signal.

The first K eigenvalues accurately represent the signal, while the remaining ones are considered noise. Therefore, the principal K eigenvectors of the correlation matrix constitute the signal subspace, as proposed in [34,39,40].

The aim of this research is to distinguish the signal subspace from the noise subspace by identifying the optimal value of K, which represents the number of endmembers in the sample.

Usually, the first eigenvalues are significantly larger than the others, which can create a dominance effect. To solve this problem, a non-linear function:

$$g(x) = \frac{x}{x+1} \tag{6}$$

is used to transform the original eigenvalues and limit them to the range $[0, 1)$. This transformation simultaneously augments the ratio between nonzero transformed eigenvalues. Then, these ratios are used to define the criterion described below.

Zhu et al. [41] demonstrated that it is impossible to choose a function that achieves both the best compression and augmentation effect simultaneously.

However, this transformation has several advantages, such as reducing the skewness of the data and achieving a uniform distribution. Additionally, it reduces the impact of the highest values by placing them near 1.

The transformation of the eigenvalues performed by the function in Eq. (6) satisfies the following conditions:

- All eigenvalues must be less than 1.
- For non-zero eigenvalues λ_i and λ_{i+1}, the inequality $\frac{g(\lambda_{i+1})}{g(\lambda_i)} > \frac{\lambda_{i+1}}{\lambda_i}$ (or equivalently $\frac{g(\lambda_i)}{g(\lambda_{i+1})} < \frac{\lambda_i}{\lambda_{i+1}}$) holds true.

In this context, all eigenvalues are limited to values less than 1. The derivative of the function is $g'(x) = \frac{1}{(1+x)^2}$, which decreases as x increases in the positive domain. Therefore, the growth rate of $g(x)$ for positive values of x is less steep compared to that of the identity function $id(x) = x$, which has a constant derivative of 1.

Thus, the transformation applied to the eigenvalues, λ_i, yields: $\mu_j = \frac{\lambda_j}{\lambda_j+1}$, where each μ_j falls within the interval $[0, 1)$ for $j = 1, \ldots, L$. Subsequently, a functional was designed to establish a criterion for computing K, expressed as:

$$I[h] = \int_0^1 h^2(x)dx \tag{7}$$

This process requires defining:

- A partition $\mathcal{P}$ of the interval $[0, 1]$ where

$$\mathcal{P} = \begin{cases} t_0 = 0 \\ t_k = \frac{\mu_k + \mu_{k+1}}{2} \forall k \in \mathbb{Z} : 1 \le k \le L-1 \\ t_L = 1 \end{cases}$$

- A family function $f_k : [0, 1) \longrightarrow [0, 1)$ with

$$f_k(x) = \begin{cases} \mu_k & if\ x \in [t_{k-1}, t_k) \\ 0 & otherwise \end{cases}$$

After applying the functional described in Eq. (7) to the set of functions f_k, a series of values γ_k are derived, as shown in Eq. (8). This process considers both the values of μ_k and the distance between μ_{k-1} and μ_{k+1}.

$$\gamma_k = I[f_k(x)] = \int_0^1 f_k^2(x)dx \tag{8}$$

The γ_k values are then arranged in descending order. This results in a scenario where the functional values corresponding to eigenvalues representative of noise approach zero. In contrast, the functional values for the remaining eigenvalues are significantly greater than zero. This distinction helps to effectively separate noise from meaningful signal components within the data.

An analysis is conducted on the sequence defined by the quotients of successive γ_k values, represented as:

$$\delta_j = \frac{\gamma_j}{\gamma_{j+1}} \tag{9}$$

The maximum ratio, described in Eq. (9), is obtained by dividing the last signal value by the first noise value. This means that the position j, which represents the last value of the signal, signifies the dimension of the signal subspace K.

To avoid a zero denominator and to smooth out abrupt fluctuations within the noise segment, a constant C_P is incorporated into both the denominator and numerator. This inclusion ensures a more stable and reliable determination of K.

The constant C_P, also known as the crest factor in statistical literature [42], is defined in this study as $C_P = \frac{log(P)}{\sqrt{P}}$. It depends solely on the size of the image. However, Zhu et al. [20] determined that C_P not only depends on the size of the image but also on the value of its pixels. Additionally, determining C_P requires an external threshold parameter, indicating a more complex dependency.

The NEEC algorithm, introduced in this study, has a key advantage in its autonomy from external parameters. The computation of C_P is exclusively dependent on the intrinsic properties of the sample image, specifically the total number of pixels and the value of each pixel. This self-sufficiency enhances the applicability and reliability of the algorithm, as it leverages the fundamental aspects of the image data itself for analysis without the need for additional external inputs.

Since the number of bands (L) in the HSI is considerably high, a subset consisting of the second decile of the γ_j values is chosen for determining the constant C_P, as the number of endmembers is less than $\frac{L-1}{10}$. Consequently, C_P is calculated as follows:

$$C_P = \frac{10}{L}\sum_{m=\frac{L}{10}+1}^{\frac{2*L}{10}} \gamma_m \tag{10}$$

The ratio sequence, δ_j, in Eq. (9) is modified by the value of C_P as follows:

$$\phi_j = \frac{\gamma_j + C_P}{\gamma_{j+1} + C_P} \tag{11}$$

The sequence $\{\phi_j\}_{j=1}^{L-1}$ satisfies Theorem 1:

Theorem 1. *Let L be the number of eigenvalues of the correlation matrix $(R_{L,L})$ and let λ_q be the last signal value. Then, the sequence $\{\phi_j\}_{j=1}^{L-1}$ satisfies:*

$$\phi_j = \begin{cases} \frac{\gamma_j}{\gamma_{j+1}} \in \Re^+, & for\ 1 < j < q \\ \max(\phi_j), & for\ j = q \\ 1, & for\ q < j < L-1 \end{cases}$$

To demonstrate Theorem 1, consider the effect of C_P on ϕ_j:

- If $1 < j < q, C_P \ll \{\gamma_j\}_{j=1}^{q} \Rightarrow \phi_j = \frac{\gamma_j + C_P}{\gamma_{j+1} + C_P} \approx \frac{\gamma_j}{\gamma_{j+1}} \in \Re^+$
- If $q < j < L-1, \{\gamma_j\}_{j=q+1}^{L-1} \ll C_P \Rightarrow \phi_j = \frac{\gamma_j + C_P}{\gamma_{j+1} + C_P} \approx \frac{C_P}{C_P} = 1$
- If $j = q$, then ϕ_q represents the quotient between the last value of the signal γ_q and the first value of the noise γ_{q+1}. Since the noise is zero-mean, this quotient is the maximum value of ϕ_j

The algorithm of the proposed method is presented below:

Algorithm 1: Estimating the number of endmembers

Data: matrix (samples x wavelengths)
Result: number of endmembers
1. Calculate the sample correlation matrix $R_{L,L}$ Eq. (5)
2. Calculate and sort λ_i of $R_{L,L}$ in descending order.
3. Transform the eigenvalues λ_i in μ_i into [0 1) Eq. (6)
4. Calculate and sort values γ_k in descending order by Eq. (8)
5. Calculate $\delta_j = \frac{\gamma_j}{\gamma_{j+1}}$ Eq. (9)
6. Calculate C_P Eq. (10)
7. Calculate $\phi_j = \frac{\gamma_j + C_P}{\gamma_{j+1} + C_P}$ Eq. (11)
8. Number of endmembers = j, index of Max(ϕ_h) in ϕ_h $(h = 1, ..., L-1)$

4. Experiments

The experimental methodology consisted of four stages. In the first stage, various mixtures at different ratios and pure liquids were prepared under strictly controlled laboratory conditions. The second stage involved acquiring HSI for the prepared samples. In the third stage, the obtained images underwent a pre-processing procedure to enhance their quality, ensuring they were optimally prepared for analysis. The final stage involved assessing the NEEC method, which was applied to both a real liquid database and a real mineral database.

Additionally, the NEEC method was evaluated using a synthetic liquid database and a synthetic database of different materials. The latter was generated by the Synthesis tools package.

Finally, a comparative evaluation was conducted among other methods, including ILSIA,[1] HySime,[2] SAED[3] and UMAP,[4] which were accessed through the websites recommended by their respective authors. Fig. 1 illustrates the different processes carried out in the study.

[1] https://www.ehu.eus/ccwintco/index.php?title=Endmember_Induction_Algorithms
[2] http://www.lx.it.pt/~bioucas/code.htm
[3] https://github.com/Xinyu-Wang/SGSNMF_TGRS
[4] https://es.mathworks.com/matlabcentral/fileexchange/71902-uniform-manifold-approximation-and-projection-umap

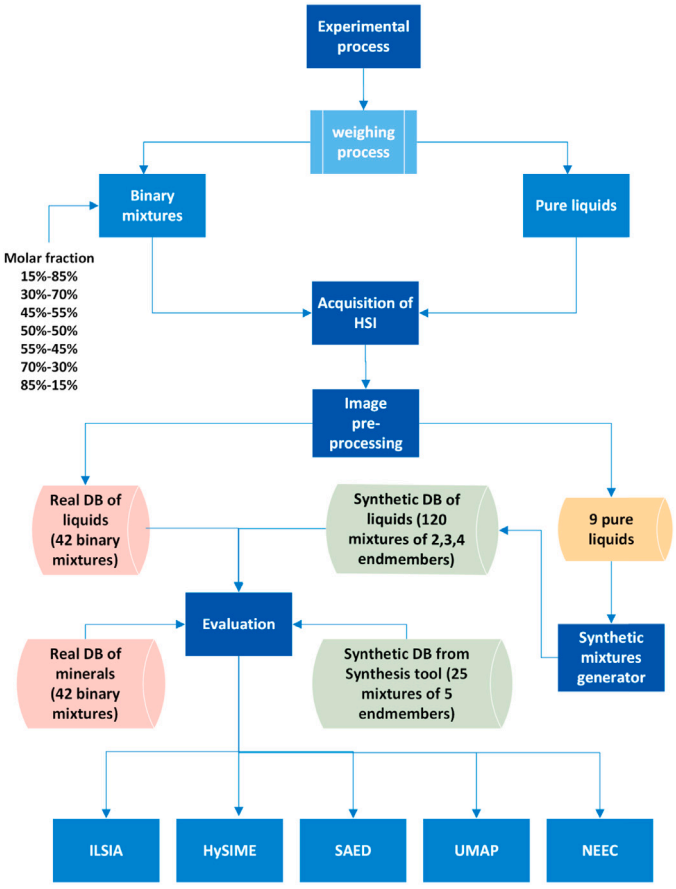

Fig. 1. Workflow of the experimental process.

4.1. Real database of liquids

4.1.1. Weighing process

Four organic compounds with different chemical properties were used to create the mixtures. The selected compounds include a saturated hydrocarbon (hexane), two normal alcohols (2-propanol and 2-butanol), and an ester (ethyl acetate).

The laboratory preparation of the mixtures was carried out under controlled conditions at a room temperature of (21 ±1) °C.

Table 2 provides information on the density and molecular weight of the pure liquids used in the mixture preparation process.

Table 2
Properties of pure liquids.

Liquid	Density (gr/ml)	Molecular weight (gr/mol)
2-Propanol	0.786	60.09
2-Butanol	0.808	74.12
Ethyl acetate	0.902	88.11
Hexane	0.654	86.18

Forty-two homogeneous mixtures were weighed to obtain a total volume of 20 ml each. The preparation took into account the molar composition of the organic compounds as well as their density and molecular mass. The six binary combinations considered in the study are presented in Fig. 2.

Seven samples with different mole fractions were prepared for each of these binary combinations: 15–85, 30–70, 45–55, 50–50, 55–45, 70–30 and 85–15, which were selected for the preparation of the two-component mixtures.

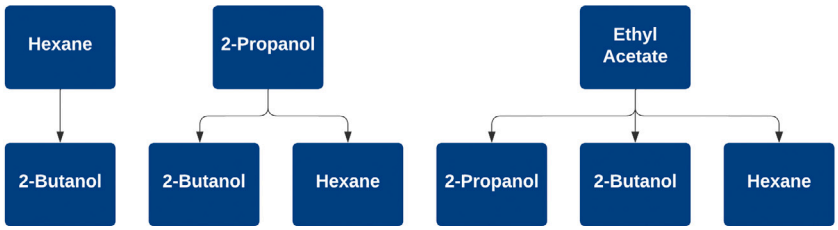

Fig. 2. 6 homogeneous solutions of 2 components.

4.1.2. Acquisition of HSI

The image acquisition system consisted of two hyperspectral cameras, the illumination system, the displacement system and a glass container, as shown in Fig. 3.

This setup allowed detailed spectral information to be captured through precise and controlled imaging of the samples.

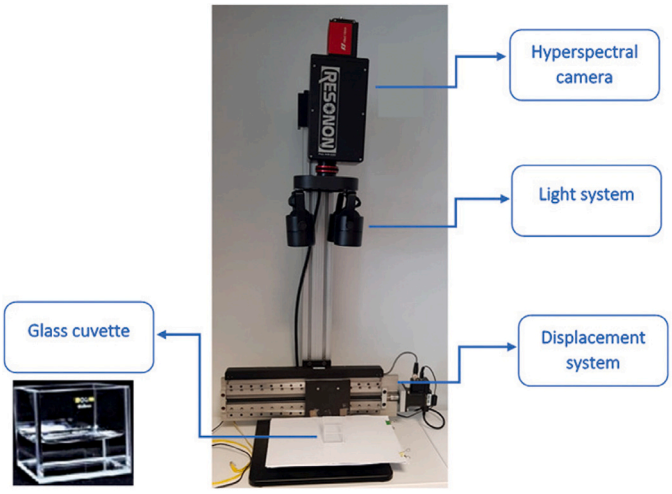

Fig. 3. Hyperspectral acquisition system.

The HSI capture was performed using two different wavelength ranges. The Resonon Pika L camera was used for the visible and near-infrared (VNIR) spectral range from 400 nm to 1000 nm, while the PIKA NIR (near-infrared) camera was used for the range from 900 nm to 1700 nm.

Table 3 gives an overview of the main characteristics of these cameras.

Table 3
Specifications hyperspectral cameras.

Specifications	PIKA L	PIKA NIR
Wavelength range (nm)	400–1000	900–1700
Spectral channels	281	164
Spatial channels	900	320
Spectral resolution FWHM (nm)	3.7	9.7
Sampling resolution	2.1	4.9
Max Frame rate (fps)	249	520

In addition, a TechniQuip Model 21 DC Fibre Optic Illuminator was used to provide stable illumination during the experiments. The lighting system was positioned above the sample, allowing the experiments to be carried out in reflection mode.

On the other hand, 40 × 40 × 40 mm optical glass cells supplied by Hellma GmbH were used. These cells offer a transmission rate of more than 80% over a wide spectral range from 360 nm to 2500 nm.

The use of these high quality cuvettes ensured efficient light transmission and accurate measurements throughout the specified wavelength range.

Image acquisition was performed in a laboratory with standardized conditions, maintaining a constant temperature and uniform lighting conditions using consistent reflectance techniques.

As a result, the signal-to-noise ratio (SNR) remained stable. The SNR values observed in the study ranged between 35 and 40 dB, indicating optimal signal quality with minimal noise interference.

4.1.3. Pre-processed

Each HSI acquired underwent a calibration process using Eq. (12), which effectively converted the raw intensity values into reflectance values.

The calibration process used three key images: the dark image (I_{dark}), obtained by covering the camera shutter; the white-reference image (I_{white}), taken with an empty cell; and the image of the sample under investigation (I_{raw}), taken with the solution in the cell and a white background.

$$I_{ref} = \frac{I_{raw} - I_{dark}}{I_{white} - I_{dark}} \tag{12}$$

A region of interest (ROI) was then manually selected on the image, as shown in Fig. 4(b).

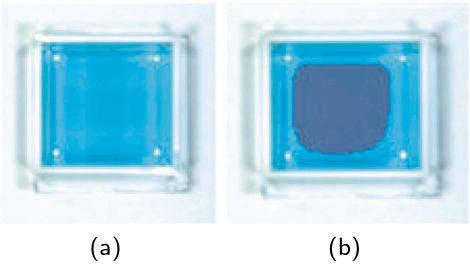

Fig. 4. (a) RGB image captured by the hyperspectral acquisition system showing a mixture of Ethyl acetate and 2-Butanol in a 50%–50% proportion in the NIR range. (b) Selection of the region of interest (ROI).

The contour of the glass cuvette was deliberately excluded from the ROI selection, as these areas tend to cause glare when exposed to light. For each solution and proportion, an image size of 825 pixels (33 × 25) was obtained.

Finally, a normalization process was performed to account for variations in light intensity throughout the image to ensure uniformity across all samples. This normalization step played a critical role in standardizing the light intensity values for further analysis and comparison.

Fig. 5(a–l) display the spectral database obtained after completing the pre-processing stage for the hyperspectral cubes. These spectra offer valuable information and establish the basis for analyzing and interpreting the data.

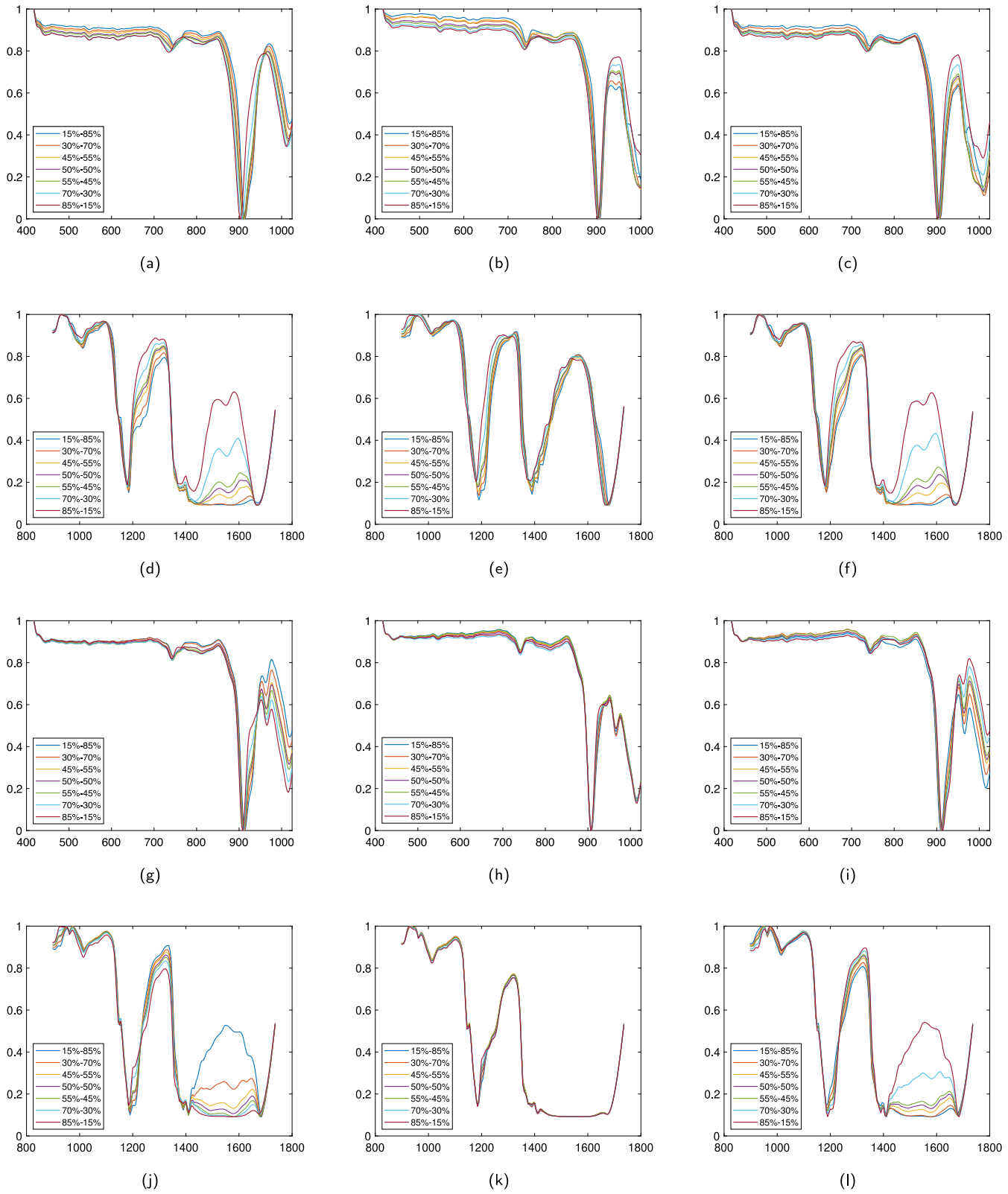

Fig. 5. The *y*-axis represents the normalized reflectance, while the *x*-axis represents the wavelengths. The spectra obtained after pre-processing for six solutions are plotted in the VNIR and NIR ranges, respectively. (a) (d) Ethyl acetate × 2-Propanol (b) (e) Ethyl acetate × Hexane (c) (f) Ethyl acetate × 2-Butanol (g) (j) 2-Propanol × Hexane (h) (k) 2-Propanol × 2-Butanol (i) (l) Hexane × 2-Butanol.

4.2. Real database of minerals

Synthetic planetary materials, or "simulants", designed to replicate one or more properties of a reference sample have been developed by several groups.

The Mars Global Simulant (MGS-1) is an open standard designed as a high-fidelity mineralogical analog to the global basaltic regolith of Mars, represented by the Rocknest Aeolian deposit in Gale Crater [43].

A collection of real-world hyperspectral images from the Harvard Dataverse[5] was used in the study, covering a spectral range of 0.9 to 2.6 μm. The images had a spectral range of 8.98 nm and a spatial resolution of 0.34 mm per pixel [44].

Binary mixtures were prepared in sample trays with dimensions of 2.5 cm × 2.5 cm × 1 cm for analysis as shown in Fig. 6.

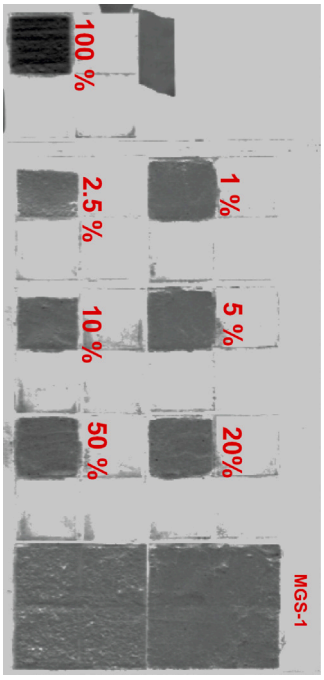

Fig. 6. This is an example Martian tray layout showing the distribution of mixtures containing different proportions of minerals and MGS-1.

Gypsum, calcite, montmorillonite, nontronite, and kaolinite were combined with MGS-1 at concentrations of 1%, 2.5%, 5%, 10%, 20%, and 50%.

The minerals mixed with MGS-1 had grain sizes ranging from 45–75 μm or 125–250 μm, with some grains larger than 250 μm [45]. Fig. 7(a–g) displays the spectra of these minerals and MGS-1. In order to improve the visualization, the spectra have been shifted on the *y*-axis.

4.3. Synthetic database of liquids

A synthetic spectral database of mixtures was created using pure component spectral signatures obtained in the laboratory through the hyperspectral acquisition system, as explained in the acquisition section.

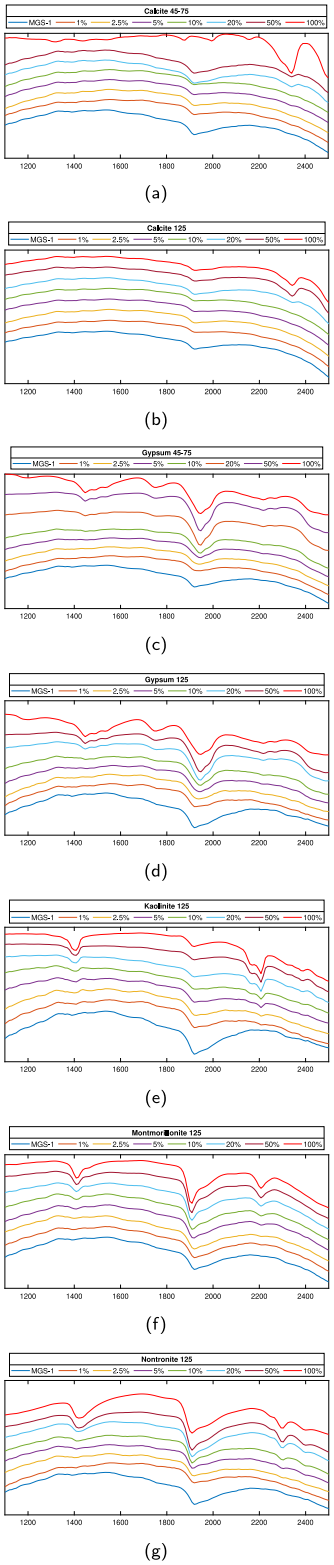

Fig. 7. Spectral signature of MGS-1 and minerals at various concentrations and grain sizes.The minerals analyzed include (a) (b) Calcite with grain sizes of 45–75 and 125 μm (c) (d) Gypsum with grain sizes of 45–75 and 125 μm (e) Kaolinite with a grain size of 125 μm (f) Montmorillonite with a grain size of 125 μm (g) Nontronite with a grain size of 125 μm.

[5] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AOCRZU

The database includes several organic compounds, such as 1-Butanol, 2-Butanol, 1-Propanol, 2-Propanol, Hexane, Octane, Ethyl acetate, Ethyl formate, and Ethanol.

The LQM was used for synthetic imaging because it accurately characterizes the spectral features of mixtures. This model assumes that the signal detected by the sensors comes from the molecular vibrations of the mixture components and takes into account both intramolecular and intermolecular molecular interactions in its formulation.

$$y = SE + \sum_{i=1}^{K} \sum_{j=i}^{K} \gamma_{ij} s_i s_j e_i \odot e_j + \epsilon \tag{13}$$

where K is the number of endmembers and $\odot$ represents the Hadamard product of matrices.

Eq. (13) has linear part, SE, which utilizes two matrices:

- The matrix of abundances $S \in \Re^{PxK} = [s_1, \ldots, s_P]^T$ where $s_i = [a_{i,1}, \ldots, a_{i,K}]$ represents the vector of abundances of size K for each pixel i, where $i = 1 \ldots P$.
- The matrix of spectral signatures $E \in \Re^{KxL} = [e_1, \ldots, e_K]$, where $e_i = [w_{i,1}, \ldots, w_{i,L}]$ represents the vector of wavelengths of size L for each pure component.

The abundance matrix, s_{ij}, was generated randomly using the approach described in [46] through the Dirichlet distribution, verifying that $s_{ij} > 0, \forall i$ and $\sum_j^K s_{ij} = 1, \forall i$.

On the other hand, in the non-linear component where $\sum_{i=1}^{K} \sum_{j=i}^{K} \gamma_{ij} s_i s_j e_i \odot e_j$, all interactions among the endmembers, including their self-interactions through products $e_i \odot e_j$, have been considered with coefficients $\gamma_{ij} \in (0, 1)$.

Specifically, $\gamma_{ii} = 0.01$ for the quadratic terms $e_i \odot e_i$, and $\gamma_{ij} = 0.15$ for the remaining terms $e_i \odot e_j$ have been employed.

The introduction of the non-linear term modifies the value of the vector of abundances s_i, which is then normalized, $s_i = \| s_i \|_1$, so that the sum of abundances becomes 1 again.

The SNR levels of both the synthetic and real images are of the same order (30–40 dB) since the abundance coefficients do not introduce noise.

Fig. 8(a–d) illustrate both real and synthetic spectral signatures in the VNIR and NIR.

4.4. Synthetic database from synthesis tool

An external database was provided by the Computational Intelligence Group of the University of the País Vasco[6].

The database contains 25 synthetic images, 128 × 128 × 431, generated by the 'Synthesis tools' package using 5 endmembers from the USGS spectral library. The database consists of 5 collections: Legendre, Spheric Gaussian Field, Exponential Gaussian Field, Rational Gaussian Field and Matern Gaussian Field.

The collection includes a synthetic hyperspectral image with no noise, while the rest of the images have different signal-to-noise ratios (SNRs) of 20, 40, 60 and 80 dB.

Fig. 9 displays the spectral signatures of each collection.

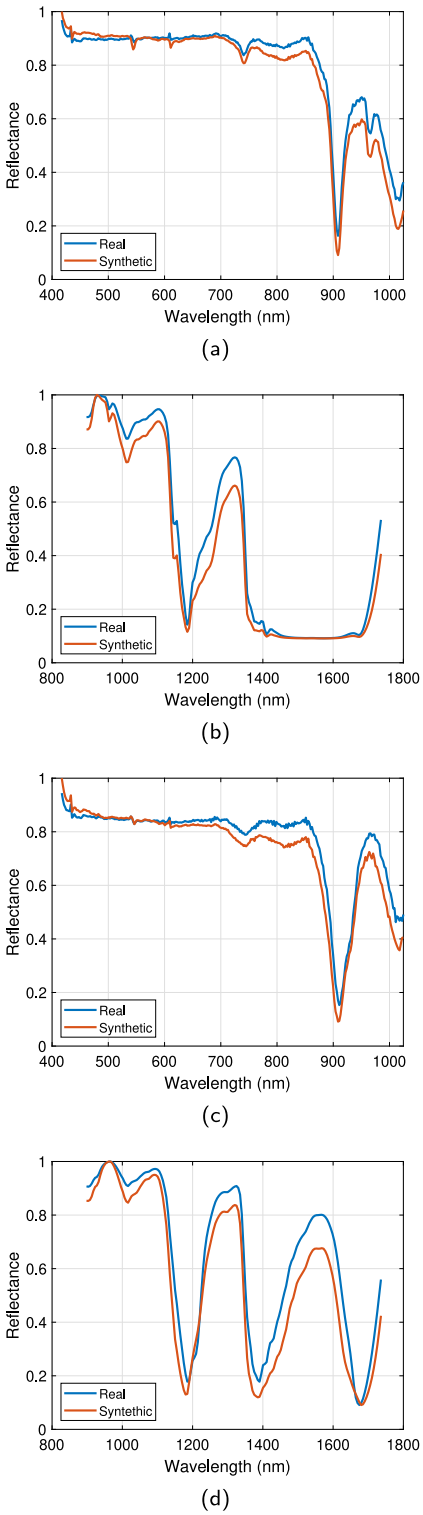

Fig. 8. Spectral signature with a proportion 50%–50% of the: (a) (b) 2-Propanol and 2-Butanol in the VNIR and NIR. (c) (d) Ethyl acetate and Hexane in the VNIR and NIR.

[6] https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Imagery_Synthesis_tools_for_MATLAB

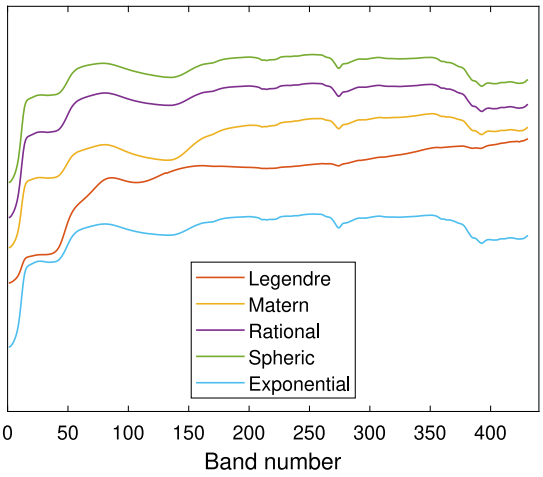

Fig. 9. Spectra of the synthetic collections obtained with the Synthesis tool. Each spectrum has been normalized, is free of noise, and has been shifted on the *y*-axis for better visualization.

5. Experimental results and discussion

5.1. Results of the liquid database

A total of 42 real samples were generated with mixtures of 2 pure components in different proportions.

Table 4 shows the number of endmembers identified using the methods: ILSIA (with a significance level of $\alpha = 0.5$ in the VNIR and $\alpha = 1$ in the NIR), HySime (with Poisson noise in the VNIR and additive noise in the NIR), SAED (with size of superpixel of 2 in the VNIR and 5 in the NIR), UMAP (with distance = 0.6 and neighbors = 30 in the VNIR and distance = 0.7 and neighbors = 15 in the NIR) and NEEC (no parameters required).

The graph in Fig. 10 illustrates the hit percentages derived from the analysis of 42 real mixtures with varying proportions.

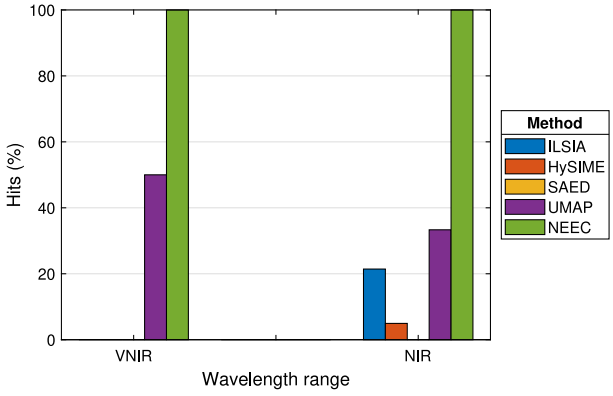

Fig. 10. The hit rate of the number of endmembers achieved using the ILSIA, HySime, SAED, UMAP and NEEC methods for 2 endmembers in both the VNIR and NIR.

In the VNIR range, the ILSIA method consistently overestimated the number of endmembers by one or two, while the HySIME method overestimated by one in 98% of cases. Although the HySIME method is not accurate, its error tends to be consistent and predictable. Regarding the SAED method, it did not accurately estimate the number of endmembers and consistently overestimated by approximately two in 83% of cases.

On the other hand, the UMAP method showed greater efficiency with an accuracy of 50%. However, it occasionally overestimated the number of endmembers by one.

Finally, the NEEC method is highly accurate, correctly identifying the number of endmembers 100% of the time without any instances of overestimation.

When transitioning to the NIR range, the ILSIA method correctly identified the number of endmembers 21.43%. However, when it failed, it tended to overestimate the number by adding an additional endmember in most cases.

In contrast, the HySIME method showed significantly lower accuracy, correctly identifying the number of endmembers only 4.95% of the time. This method also tended to overestimate to a greater extent, adding three additional endmembers when it failed to identify the correct amount.

The SAED method was incorrect in all cases. It overestimated the number of endmembers by two in about 79% of the cases. This indicates a significant limitation in its ability to accurately determine the number of endmembers.

The UMAP method was more effective than ILSIA, HySIME, and SAED, being correct in 33.33% of cases. However, the performance of the method was irregular as it tended to overestimate the number of endmembers by one to three when it failed to achieve accurate identification.

The NEEC method stood out significantly from the others, achieving 100% accuracy in identifying the number of endmembers. This method did not show any cases of overestimation, proving to be the most effective of all those evaluated.

In general, the methods, except for the NEEC method, tended to overestimate the number of endmembers when they failed to identify the correct number.

5.2. Results of the mineral database

Table 5 details the evaluation outcomes for the five methods applied to the binary mineral mixture database utilizing the MGS-1.

The table highlights the precision of each method and its tendency to either overestimate or underestimate the endmember counts. The best performing parameters were used to evaluate the methods.

The methods evaluated include ILSIA (with a significance level of $\alpha = 0.5$), HySime (with additive noise), SAED (with a superpixel size of 5), UMAP (with a distance of 0.5 and 15 neighbors), and NEEC (which requires no parameters). Fig. 11 displays the hit rate for the methods under review.

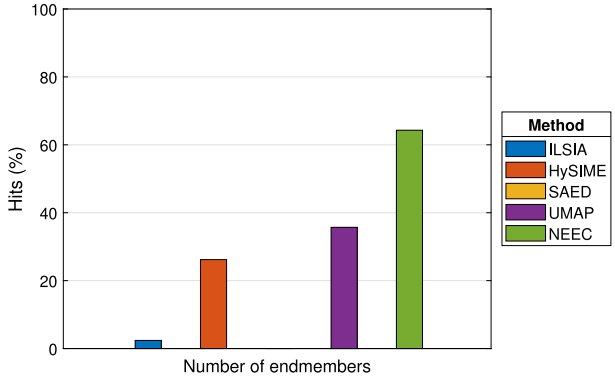

Fig. 11. The hit rate of the number of endmembers achieved using the ILSIA, HySime, SAED, UMAP and NEEC methods for binary mixtures of minerals and MGS-1.

The ILSIA method was always unsuccessful. It underestimated the number of endmembers by one in 98% of the cases.

In contrast, HySIME demonstrated a moderate ability to correctly identify the number of endmembers, with a hit rate of 22.2%. However, its performance was inconsistent in error cases, indicating variability in its ability to handle different types of mixtures.

Table 4
Number of components identified in 2 endmembers mixtures using ILSIA, HySime, SAED, UMAP and NEEC methods.

		Method									
Mixture	Proportion	ILSIA		HySIME		SAED		UMAP		NEEC	
		VNIR	NIR	VNIR	NIR	VNIR	NIR	VNIR	NIR	VNIR	NIR
	15–85	4	3	7	6	4	4	3	4	2	2
	30–70	4	3	3	5	4	4	3	4	2	2
Ehtyl Acetate	45–55	3	5	3	2	4	4	2	4	2	2
and	50–50	4	3	3	5	5	4	2	2	2	2
2 Propanol	55–45	4	2	3	5	4	4	3	2	2	2
	70–30	3	3	3	5	4	4	3	2	2	2
	85–15	4	3	3	5	4	4	1	2	2	2
	15–85	3	3	3	5	7	4	3	5	2	2
	30–70	3	2	3	5	4	5	2	4	2	2
Ethyl Acetate	45–55	3	5	3	2	4	4	4	2	2	2
and	50–50	3	3	3	5	4	5	3	3	2	2
Hexane	55–45	4	2	3	5	4	4	3	5	2	2
	70–30	4	5	3	5	4	4	2	3	2	2
	85–15	3	3	3	5	5	7	2	3	2	2
	15–85	3	2	3	5	4	4	2	5	2	2
	30–70	4	2	3	5	4	4	3	3	2	2
Ethyl Acetate	45–55	3	2	3	5	4	4	3	2	2	2
and	50–50	3	3	3	5	4	4	3	5	2	2
2 Butanol	55–45	3	3	3	5	4	4	4	2	2	2
	70–30	4	3	3	5	5	4	3	2	2	2
	85–15	4	3	3	5	4	4	2	5	2	2
	15–85	3	3	3	5	4	4	2	4	2	2
	30–70	3	3	3	5	5	4	3	5	2	2
2 Propanol	45–55	3	3	3	5	4	4	2	3	2	2
and	50–50	4	2	3	5	4	5	1	2	2	2
Hexane	55–45	4	4	3	5	4	4	2	4	2	2
	70–30	4	3	3	5	4	4	3	6	2	2
	85–15	3	3	3	5	4	4	2	2	2	2
	15–85	4	3	3	5	4	4	2	6	2	2
	30–70	3	2	3	5	4	4	2	3	2	2
2 Propanol	45–55	5	3	3	5	4	4	2	3	2	2
and	50–50	4	3	3	5	4	4	2	2	2	2
2 Butanol	55–45	3	4	3	5	4	4	2	3	2	2
	70–30	4	5	3	5	4	4	3	2	2	2
	85–15	4	3	3	5	5	6	2	4	2	2
	15–85	4	3	3	5	4	5	2	2	2	2
	30–70	4	3	3	5	4	4	1	2	2	2
Hexane	45–55	5	3	3	5	5	4	3	3	2	2
and	50–50	4	3	3	5	4	4	3	5	2	2
2 Butanol	55–45	4	3	3	5	4	5	2	3	2	2
	70–30	5	5	3	5	4	4	2	3	2	2
	85–15	4	2	3	5	4	4	2	3	2	2

SAED was consistently incorrect, with a tendency to overestimate the number of endmembers by two in 66.7% of cases, indicating a tendency to interpret variations or noise in the data as additional endmembers.

The UMAP method had a success rate of 35.7% in correctly identifying the number of endmembers, but its performance was inconsistent in cases of error, often overestimating the number of endmembers.

However, the NEEC method had an accuracy of 64.3%. When NEEC failed, it consistently overestimated by one endmember. This result demonstrates not only its ability to correctly identify the number of endmembers, but also the consistency of its errors, which is valuable for future improvements.

5.3. Results from the database of synthetic liquids.

A total of 120 synthetic mixtures samples were prepared for the VNIR and NIR spectrums, including 2, 3, and 4 pure components of varying proportions.

Tables 6, 7, and 8 present a comparison of the number of endmembers identified from the diverse mixtures and proportions through the use of the ILSIA algorithm (with a significance level of $\alpha = 0.5$), HySime (using Poisson noise in the VNIR region and additive noise in the NIR region), SAED (with a superpixel size of 5 in both ranges), UMAP (with a distance of 0.1 and 30 neighbors in VNIR and a distance of 0.7 and 15 neighbors in NIR) and NEEC (no parameters are required).

Table 6 shows the results for mixtures of two endmembers, demonstrating the effectiveness of both the HySime and NEEC methods in accurately identifying endmembers in all cases for both the VNIR and NIR. In contrast, the ILSIA, SAED, and UMAP approaches consistently overestimate the number of endmembers across both spectral ranges.

The results of mixtures of three endmembers are shown in Table 7. ILSIA achieves its best performance with a 65% success rate in the VNIR range.

HySIME is inconsistent. It correctly identifies endmembers in the VNIR but fails in the NIR. Both SAED and UMAP tend to overestimate the number of endmembers. In contrast, NEEC correctly identifies endmembers in both ranges.

Table 8 displays the results of mixtures of 4 endmembers. The ILSIA and HySime algorithms systematically underestimate the number of endmembers when they fail, while the UMAP method tends to overestimate the number of endmembers in both ranges.

The SAED method demonstrates optimal performance with a 95% accuracy rate in both ranges, and the NEEC method achieves a success rate of 90% in both spectral ranges.

Fig. 12(a) represents the percentage of hits for the 120 synthetic mixtures evaluated in the VNIR region, while Fig. 12(b) illustrates the

Table 5
Number of components identified in 2 endmember mixtures using ILSIA, HySime, SAED, UMAP and NEEC methods in the mineral database.

		Method				
Mixture	Percentage	ILSIA	HySIME	SAED	UMAP	NEEC
	1%	1	2	4	4	3
Calcite	2.5%	1	0	4	4	2
45–75	5%	1	8	0	4	3
and	10%	1	2	5	2	3
MGS-1	20%	1	4	4	5	3
	50%	1	17	9	5	2
	1%	1	21	5	4	2
Calcite	2.5%	1	2	6	6	3
125	5%	1	7	5	7	3
and	10%	1	2	10	7	3
MGS-1	20%	1	0	4	4	3
	50%	1	0	4	4	2
	1%	1	3	4	6	3
Gypsum	2.5%	1	4	0	6	3
45–75	5%	1	3	4	9	3
and	10%	1	6	7	2	2
MGS-1	20%	1	0	4	6	3
	50%	1	3	5	4	2
	1%	1	2	4	4	2
Gypsum	2.5%	1	2	4	2	2
125	5%	1	2	4	2	2
and	10%	1	2	4	5	2
MGS-1	20%	1	21	4	5	2
	50%	1	6	4	4	2
	1%	1	2	4	4	2
Kaolinite	2.5%	1	0	4	2	2
125	5%	1	0	4	2	2
and	10%	1	23	4	2	2
MGS-1	20%	1	11	4	2	2
	50%	1	0	4	2	2
	1%	1	16	6	5	3
Montmorillonite	2.5%	1	0	4	6	2
125	5%	2	29	5	4	3
and	10%	1	6	4	2	3
MGS-1	20%	1	7	7	2	2
	50%	1	26	4	6	2
	1%	1	2	7	2	2
Nontronite	2.5%	1	0	4	2	2
125	5%	1	21	4	4	2
and	10%	1	2	4	2	2
MGS-1	20%	1	0	4	2	2
	50%	1	0	4	1	2

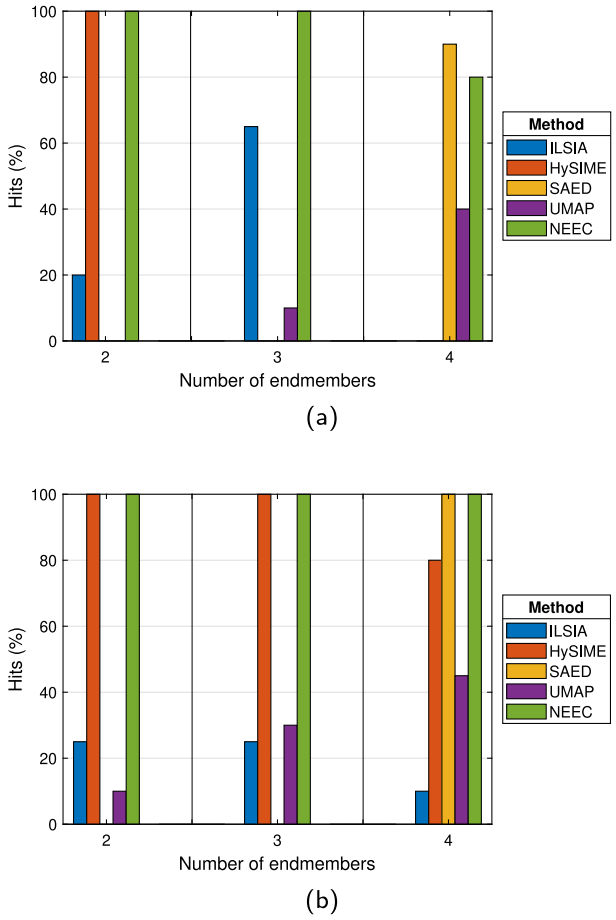

Fig. 12. The hit rate of the number of endmembers achieved using the ILSIA, HySime, SAED, UMAP and NEEC methods for 2, 3, and 4 endmembers in both the (a) VNIR and (b) NIR ranges.

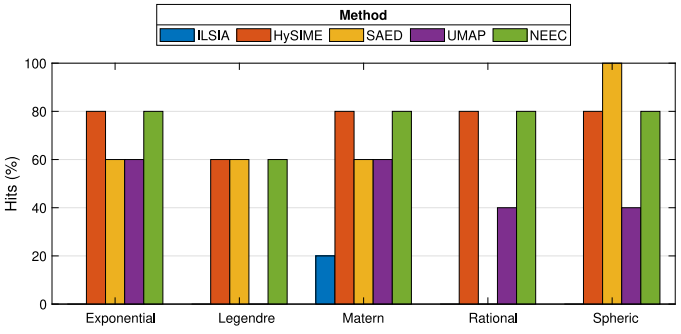

Fig. 13. The hit rate of the number of endmembers achieved using the ILSIA, HySime, SAED, UMAP and NEEC methods for 5 endmembers.

percentage of hits for the NIR region. The graphs demonstrate the stability and efficiency advantages of the NEEC method over the other two methods, even under conditions where the number of components and their proportions in the mixtures increase.

5.4. Result of synthetic database from synthesis tool

Table 9 shows the number of endmembers identified, while Fig. 13 illustrates the success rates achieved across the five collections generated by the Synthesis tool.

Ilsia correctly identified the number of endmembers only once, underestimating the number in all other evaluations.

HySime was 76% accurate, but tended to overestimate the number of endmembers in situations where it failed to get it right.

SAED was 56% accurate, but showed relatively inconsistent performance on errors. It showed variability in its ability to adapt to the specific characteristics of each mixture, with no clear tendency toward overestimation or underestimation.

UMAP had a success rate of 44%. It was the method that most often overestimated the number of endmembers. Its performance was notably erratic, which may reflect excessive sensitivity to variations in the data.

NEEC was 76% accurate, but underestimated the number of endmembers in cases where it missed.

The evaluation shows that HySime and NEEC are the most accurate and stable methods in this database, with a success rate of 76%. However, HySime tends to overestimate in failure cases, while NEEC underestimates, reflecting differences in their approaches to handling uncertainty in determining the number of endmembers.

The other methods, especially UMAP, showed greater irregularity.

Table 6
Evaluation of ILSIA, HySIME, SAED, UMAP, and NEEC methods for determining the number of endmembers in synthetic liquid mixtures using 2 endmembers in the VNIR and NIR spectral ranges.

Proportion	2 endmembers mixtures - VNIR																								
	1 Butanol and 2 Butanol					Ethyl acetate and Ethanol					Ethyl acetate and Hexane					Hexane and Octane					Ethyl formate and Hexane				
	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC
30–70	2	2	4	7	2	3	2	4	7	2	3	2	5	7	2	2	2	4	5	2	3	2	4	5	2
50–50	3	2	4	6	2	2	2	5	4	2	3	2	4	7	2	3	2	4	6	2	3	2	5	8	2
70–30	3	2	4	5	2	3	2	4	7	2	3	2	5	5	2	3	2	4	4	2	3	2	4	6	2
80–20	2	2	4	5	2	3	2	4	6	2	3	2	5	7	2	3	2	4	6	2	3	2	5	7	2
	2 endmembers mixtures - NIR																								
30–70	3	2	5	6	2	2	2	4	2	2	3	2	4	7	2	3	2	4	4	2	2	2	5	7	2
50–50	3	2	4	7	2	3	2	4	8	2	3	2	4	5	2	3	2	4	4	2	3	2	4	6	2
70–30	3	2	5	7	2	3	2	4	5	2	2	2	4	9	2	2	2	5	7	2	3	2	4	7	2
80–20	2	2	4	6	2	3	2	4	8	2	3	2	4	2	2	3	2	5	7	2	3	2	4	9	2

Table 7
Evaluation of ILSIA, HySIME, SAED, UMAP, and NEEC methods for determining the number of endmembers in synthetic liquid mixtures using 3 endmembers in the VNIR and NIR spectral ranges.

Proportion	3 endmembers mixtures - VNIR																								
	1 Propanol, Ethanol, Octane					2 Propanol, Hexane, Octane					Ethyl acetate, Ethanol, Ethyl formate					Ethyl acetate, Ethyl formate, Hexane					Ethyl formate, 1 Butanol, Hexane				
	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC
10–40–50	3	2	4	7	3	3	2	4	6	3	3	2	6	3	3	3	2	4	7	3	3	2	4	5	3
20–40–40	3	2	4	6	3	2	2	6	8	3	3	2	4	7	3	2	2	4	5	3	3	2	4	5	3
20–70–10	2	2	5	6	3	3	2	4	5	3	3	2	4	5	3	2	2	4	8	3	2	2	4	6	3
31–33–36	2	2	4	9	3	3	2	4	5	3	3	2	4	5	3	3	2	5	4	3	4	2	4	3	3
	3 endmembers mixtures - NIR																								
	1 Butanol, Ethyl acetate, Hexane					2 Propanol, Ethyl acetate, Ethyl formate					Ethyl acetate, Ethyl formate, Hexane					Ethanol, Hexane, Octane					Ethyl formate, Hexane, Octane				
10–40–50	3	3	5	4	3	3	3	4	4	3	2	3	4	4	3	2	3	4	6	3	2	3	4	8	3
20–40–40	3	3	4	3	3	2	3	4	3	3	2	3	4	4	3	3	3	4	7	3	2	3	4	5	3
20–70–10	2	3	4	3	3	3	3	4	6	3	2	3	6	6	3	2	3	4	3	3	2	3	4	6	3
31–33–36	2	3	4	4	3	2	3	4	3	3	2	3	4	4	3	2	3	4	8	3	2	3	6	3	3

Table 8
Evaluation of ILSIA, HySIME, SAED, UMAP, and NEEC methods for determining the number of endmembers in synthetic liquid mixtures using 4 endmembers in the VNIR and NIR spectral ranges.

Proportion	4 endmembers mixtures - VNIR																								
	1 Butanol, 1 Propanol, 2 Butanol, 2 Propanol					1 Butanol, 2 Butanol, Ethyl acetate, Ethyl formate					2 Butanol, Ethyl acetate, Ethyl formate, Octane					Ethyl acetate, Ethanol, Ethyl formate, Hexane					Ethyl acetate, Ethanol, Hexane, Octane				
	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC
20–15–25–40	3	2	4	9	5	3	2	4	5	4	2	2	4	4	4	2	2	4	4	4	3	2	4	7	4
25–20–30–25	2	2	4	3	5	3	2	4	8	4	3	2	4	4	4	3	2	6	4	4	2	2	6	6	4
25–35–20–20	2	2	4	6	5	2	2	4	5	4	3	2	4	6	4	3	2	4	4	4	2	2	4	4	4
35–25–15–25	3	2	4	5	5	2	2	4	5	4	2	2	4	5	4	2	2	4	4	4	3	2	4	4	4
	4 endmembers mixtures - NIR																								
	1 Butanol, 1 Propanol, 2 Butanol, 2 Propanol					1 Butanol, 2 Butanol, Ethyl acetate, Ethyl formate					1 Butanol, 2 Butanol, Hexane, Octane					1 Butanol, Ethyl acetate, Ethyl formate, Octane					Ethyl acetate, Ethanol, Ethyl formate, Hexane				
20–15–25–40	3	3	4	6	4	2	4	4	4	4	3	4	4	4	4	2	4	4	5	4	2	4	4	5	4
25–20–30–25	2	3	4	8	4	3	4	4	4	4	4	4	4	4	4	3	4	4	6	4	3	4	4	3	4
25–35–20–20	3	3	4	7	4	4	4	4	4	4	3	4	4	4	4	3	4	4	6	4	3	4	4	4	4
35–25–15–25	3	3	4	6	4	2	4	4	6	4	3	4	4	4	4	2	4	4	7	4	3	4	4	4	4

Table 9
Evaluation of ILSIA, HySIME, SAED, UMAP, and NEEC methods for determining the number of endmembers in synthetic mixtures produced by the Synthesis tool using 5 endmembers.

SNR	5 endmembers mixtures																									
	Exponential Gaussian Field					Legendre					Matern Gaussian Field					Rational Gaussian Field					Spheric Gaussian Field				
	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC	ILSIA	HySIME	SAED	UMAP	NEEC
default	2	5	5	5	5	2	5	5	16	5	3	5	5	5	5	3	5	4	5	5	3	5	5	10	5
20	3	5	4	5	3	4	4	8	5	2	5	5	7	5	3	3	5	4	5	3	4	5	5	5	3
40	3	6	4	6	5	3	6	6	9	4	3	6	7	6	5	3	6	4	6	5	3	6	5	7	5
60	3	5	5	5	5	2	5	5	14	5	2	5	5	5	5	3	5	4	6	5	3	5	5	5	5
80	3	5	5	7	5	2	5	5	10	5	3	5	5	6	5	3	5	4	6	5	3	5	5	6	5

6. Conclusions and future work

This work introduces the NEEC method as an innovative unsupervised technique for dimensionality reduction in HSI of homogeneous mixtures. It is positioned as a critical initial step in both linear and non-linear unmixing processes. The NEEC method employs eigenvalue analysis and incorporates a criterion based on the ratio of consecutive values in a sequence derived from a non-linear transformation of the eigenvalues from the sample correlation matrix.

Both real and simulated images from our own collections and external databases were used in the experiments of this study. Simulated spectra were generated using the LQM to reflect the non-linear aspects of the experiments. An extensive comparison with other methods was performed to evaluate the efficiency and robustness of the proposed method. Results from analyses on both synthetic and real samples underscore the exceptional accuracy of NEEC, surpassing selected methods. Specifically, NEEC achieved weighted average accuracy of 86.6% in real experiments and 93.1% in synthetic experiments.

Future research efforts will focus on integrating the NEEC method as a preliminary step in non-linear unmixing processes. This strategic development aims to improve the efficiency of the method and broaden its application scope, which is a promising direction for the advancement of hyperspectral imaging techniques.

CRediT authorship contribution statement

Karina Baños: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Julio Esclarín:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Juan Ortega:** Supervision, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the Ministry of Science and Innovation of the Government of Spain for their financial support [CTQ2015-68428-P].

Karina Baños thanks the Ministry of Science and Innovation of the Government of Spain for her predoctoral contract [BES-2016-079438]. Furthermore, Karina Baños gratefully acknowledges the Center for Research in Applied Mathematics (CMLA) at the University of Paris-Saclay for their generous hospitality during her research stay, which greatly aided the progress of this work. A special thanks to Jean-Michel Morel for his valuable contributions.

References

[1] O.Y. Rodionova, L.P. Houmøller, A.L. Pomerantsev, P. Geladi, J. Burger, V.L. Dorofeyev, A.P. Arzamastsev, NIR spectrometry for counterfeit drug detection: A feasibility study, Anal. Chim. Acta 549 (1–2) (2005) 151–158.

[2] P.Y. Sacré, C. De Bleye, P.F. Chavez, L. Netchacovitch, P. Hubert, E. Ziemons, Data processing of vibrational chemical imaging for pharmaceutical applications, J. Pharm. Biomed. Anal. 101 (2014) 123–140.

[3] R.E. Correa Pabón, C.R. de Souza Filho, W.J. de Oliveira, Reflectance and imaging spectroscopy applied to detection of petroleum hydrocarbon pollution in bare soils, Sci. Total Environ. 649 (2019) 1224–1236.

[4] A. ul Rehman, S.A. Qureshi, A review of the medical hyperspectral imaging systems and unmixing algorithms' in biological tissues, Photodiagnosis Photodyn. Ther. 33 (September 2020) (2021) 102165.

[5] V. Kopačková, L. Hladíková, Applying spectral unmixing to determine surface water parameters in a mining environment, Remote Sens. 6 (11) (2014) 11204–11224.

[6] B. Ma, L. Wu, X. Zhang, X. Li, Y. Liu, S. Wang, Locally adaptive unmixing method for lake-water area extractionbased on MODIS 250 m bands, Int. J. Appl. Earth Obs. Geoinf. 33 (1) (2014) 109–118.

[7] E. Alcântara, C. Barbosa, J. Stech, E. Novo, Y. Shimabukuro, Improving the spectral unmixing algorithm to map water turbidity Distributions, Environ. Model. Softw. 24 (9) (2009) 1051–1061.

[8] M. van der Meijde, N.M. Knox, S.L. Cundill, F. Noomen, H.M. van der Werff, C. Hecker, Detection of hydrocarbons in clay soils: A laboratory experiment using spectroscopy in the mid-and thermal infrared, Int. J. Appl. Earth Obs. Geoinf. 23 (1) (2013) 384–388.

[9] S. Kritchman, B. Nadler, Determining the number of components in a factor model from limited noisy data, Chemometr. Intell. Lab. Syst. 94 (1) (2008) 19–32.

[10] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.

[11] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (6) (1974) 716–723.

[12] M.W. Graham, D.J. Miller, Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection, IEEE Trans. Signal Process. 54 (4) (2006) 1289–1303.

[13] A. Halimi, P. Honeine, M. Kharouf, C. Richard, J.-Y. Tourneret, Estimating the intrinsic dimension of hyperspectral images using a noise-whitened eigengap approach, IEEE Trans. Geosci. Remote Sens. 54 (7) (2016) 3811–3821.

[14] E. Terreaux, J.-P. Ovarlez, F. Pascal, New model order selection in large dimension regime for complex elliptically symmetric noise, in: 2017 25th European Signal Processing Conference, EUSIPCO, IEEE, 2017, pp. 1090–1094.

[15] S. Das, J.N. Kundu, A. Routray, Estimation of number of endmembers in a Hyperspectral image using Eigen thresholding, in: 12th IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control: (E3-C3), INDICON 2015, IEEE, 2016, pp. 1–5.

[16] P.R. Peres-Neto, D.A. Jackson, K.M. Somers, How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Comput. Statist. Data Anal. 49 (4) (2005) 974–997.

[17] J.M. Bioucas-Dias, J.M. Nascimento, Hyperspectral subspace identification, IEEE Trans. Geosci. Remote Sens. 46 (8) (2008) 2435–2445.

[18] C.I. Chang, Q. Du, Estimation of number of spectrally distinct signal sources in hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 42 (3) (2004) 608–619.

[19] S. Das, J.N. Kundu, A. Routray, Estimation of number of endmembers in a hyperspectral image using eigen thresholding, in: 2015 Annual IEEE India Conference, INDICON, 2015, pp. 1–5, http://dx.doi.org/10.1109/INDICON.2015.7443556.

[20] X. Zhu, Y. Kang, J. Liu, Estimation of the Number of Endmembers via Thresholding Ridge Ratio Criterion, IEEE Trans. Geosci. Remote Sens. 58 (1) (2020) 637–649.

[21] S. Das, A. Routray, A.K. Deb, Noise robust estimation of number of endmembers in a hyperspectral image by eigenvalue based gap index, in: 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS, IEEE, 2016, pp. 1–5.

[22] A. Ambikapathi, T.-H. Chan, C.-Y. Chi, K. Keizer, Hyperspectral data geometry-based estimation of number of endmembers using p-norm-based pure pixel identification algorithm, IEEE Trans. Geosci. Remote Sens. 51 (5) (2012) 2753–2769.
[23] X. Wang, Y. Zhong, Y. Xu, L. Zhang, Y. Xu, Saliency-based endmember detection for hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 56 (7) (2018) 3667–3680.
[24] K. Cawse-Nicholson, S.B. Damelin, A. Robin, M. Sears, Determining the intrinsic dimension of a hyperspectral image using random matrix theory, IEEE Trans. Image Process. 22 (4) (2013) 1301–1310.
[25] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, 2018, arXiv preprint arXiv:1802.03426.
[26] M. Graña, I. Villaverde, J.O. Maldonado, C. Hernandez, Two lattice computing approaches for the unsupervised segmentation of hyperspectral images, Neurocomputing 72 (10–12) (2009) 2111–2120.
[27] M.A. Veganzones Bodon, Contributions to Hyperspectral Image Processing from Lattice Computing and Computational Intelligence (Ph.D. thesis), Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2012.
[28] G.X. Ritter, G. Urcid, A lattice matrix method for hyperspectral image unmixing, Inform. Sci. 181 (10) (2011) 1787–1803.
[29] K.T. Shahid, I.D. Schizas, Unsupervised hyperspectral unmixing via nonlinear autoencoders, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–13.
[30] F. Xiong, J. Zhou, S. Tao, J. Lu, Y. Qian, SNMF-Net: Learning a deep alternating neural network for hyperspectral unmixing, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–16.
[31] C. Zhou, M.R.D. Rodrigues, ADMM-based hyperspectral unmixing networks for abundance and endmember estimation, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–18.
[32] E. Myasnikov, Comparison of spectral dissimilarity measures and dimension reduction techniques for hyperspectral images, Pattern Recognit. Image Anal. 31 (2021) 454–465.
[33] X. Wang, Y. Zhong, C. Cui, L. Zhang, Y. Xu, Autonomous endmember detection via an abundance anomaly guided saliency prior for hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 59 (3) (2021) 2336–2351.
[34] N. Keshava, J.F. Mustard, Spectral unmixing, IEEE Signal Process. Mag. 19 (1) (2002) 44–57.
[35] A. Picon, O. Ghita, P.F. Whelan, P.M. Iriondo, Fuzzy spectral and spatial feature integration for classification of nonferrous materials in hyperspectral data, IEEE Trans. Ind. Inform. 5 (4) (2009) 483–494.
[36] R. Heylen, M. Parente, P. Gader, A review of nonlinear hyperspectral unmixing methods, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (6) (2014) 1844–1868.
[37] A. Halimi, Y. Altmann, N. Dobigeon, J.Y. Tourneret, Nonlinear unmixing of hyperspectral images using a generalized bilinear model, IEEE Trans. Geosci. Remote Sens. 49 (11 PART 1) (2011) 4153–4162.
[38] I. Meganem, P. Déliot, X. Briottet, Y. Deville, S. Hosseini, Linear–quadratic mixing model for reflectances in urban environments, IEEE Trans. Geosci. Remote Sens. 52 (1) (2014) 544–558.
[39] G. Shaw, D. Manolakis, Signal processing for hyperspectral image exploitation, IEEE Signal Process. Mag. 19 (1) (2002) 12–16.
[40] Y.F. Huang, Statistical signal processing, in: The Electrical Engineering Handbook, 2005, pp. 921–932, http://dx.doi.org/10.1016/B978-012170960-0/50066-9.
[41] X. Zhu, X. Guo, T. Wang, L. Zhu, Dimensionality determination: A thresholding double ridge ratio approach, Comput. Statist. Data Anal. 146 (2020) 106910.
[42] Q. Xia, W. Xu, L. Zhu, Consistently determining the number of factors in multivariate volatility modelling, Statist. Sinica 25 (3) (2015) 1025–1044.
[43] K.M. Cannon, D.T. Britt, T.M. Smith, R.F. Fritsche, D. Batcheldor, Mars global simulant MGS-1: A rocknest-based open standard for basaltic martian regolith simulants, Icarus 317 (2019) 470–478.
[44] J. Tarnas, Hyperspectral Images for "Successes and Challenges of Factor Analysis/target Transformation Application to Visible-To-Near-Infrared Hyperspectral Data, Harvard Dataverse, 2021, http://dx.doi.org/10.7910/DVN/AOCRZU.
[45] J. Tarnas, J. Mustard, X. Wu, E. Das, K. Cannon, C. Hundal, A. Pascuzzo, J. Kellner, M. Parente, Successes and challenges of factor analysis/target transformation application to visible-to-near-infrared hyperspectral data, Icarus 365 (2021) 114402.
[46] J.M. Nascimento, J.M. Dias, Does independent component analysis play a role in unmixing hyperspectral data? IEEE Trans. Geosci. Remote Sens. 43 (1) (2005) 175–187.