Deep Nonnegative Dictionary Factorization for Hyperspectral Unmixing

Wenhong Wang¹,* and Hongfu Liu²

¹ College of Computer Science, Liaocheng University, Liaocheng 252059, China
² Volen National Center for Complex Systems, Departments of Computer Science, Brandeis University, Waltham, MA 02453, USA; hongfuliu@brandeis.edu
* Correspondence: wangwenhong@lcu.edu.cn

Received: 25 July 2020; Accepted: 3 September 2020; Published: 5 September 2020

Abstract: As a powerful blind source separation tool, Nonnegative Matrix Factorization (NMF) with effective regularizations has shown significant superiority in spectral unmixing of hyperspectral remote sensing images (HSIs) owing to its good physical interpretability and data adaptability. However, the majority of existing NMF-based spectral unmixing methods only adopt the single layer factorization, which is not favorable for exploiting the complex and structured representation relationship of endmembers implied in HSIs. In order to overcome such an issue, we propose a novel two-stage Deep Nonnegative Dictionary Factorization (DNDF) approach with a sparseness constraint and self-supervised regularization for HSI unmixing. Beyond simply extending one-layer factorization to multi-layer, DNDF conducts fuzzy clustering to tackle the mixed endmembers of HSIs. Moreover, self-supervised regularization is integrated into our DNDF model to impose an effective constraint on the endmember matrix. Experimental results on three real HSIs demonstrate the superiority of DNDF over several state-of-the-art methods.

Keywords: deep nonnegative dictionary factorization; hyperspectral unmixing; self-supervised learning; sparse coding

1. Introduction

Hyperspectral remote sensing images (HSIs) capture hundreds of narrow spectral channels, and provide detailed spectral information and spatial distribution status about the target ground objects. With their rich spectral information for distinguishing different objects, HSIs have received many successful applications, such as segmentation [1], classification [2], target detection [3], tracking [4], and recognition [5]. Due to the low spatial resolution of sensors and the complexity of the terrain, HSIs generally comprise the so-called “mixed pixel” [6], which causes spectral information of different materials presents in a single pixel of HSIs [7]. Larger spectral variance in mixed pixels makes them difficult to interpret [8], leading negative impacts to the quantitative analysis of HSIs.

To tackle this problem, hyperspectral unmixing (HU) plays a fundamental role in extracting a set of constituent signatures (i.e., endmembers) from the target HSIs and estimates the corresponding proportions (i.e., abundances) of them in each mixed pixel [6]. Nonnegative Matrix Factorization (NMF), as an effective tool for blind source separation, has drawn much attention in the HU applications owing to its good interpretability and data adaptability [9,10]. To improve the unmixing performance of basic NMF, introducing effective constraints into NMF based on the problem-dependent knowledge has been witnessed as one of the effective measures, such as the manifold constraint [11], sparseness constraint [12], low-rank constraint [13], smooth constraint [14], as well as local spectral similarity preserving constraint [15]. Unfortunately, the majority of existing NMF-based unmixing methods only adopt the single layer factorization, which limits their learning ability for hierarchical
structure information of complex and mixed HSI data. Recently, inspired by the success of deep learning, deep NMF has been developed and shown promising unmixing performance owing to its multi-layer learning mechanism [16–19]. However, existing deep NMF-based methods commonly focus on factorizing the coefficient matrix to explore the abstract features of the data [17], which is not favorable for efficiently utilizing the complex hierarchical and multi-layers structured representation information between the endmembers and the mixed pixels included in HSIs.

To overcome the above issues, we propose a novel Deep Nonnegative Dictionary Factorization (DNDF) model in this paper, by which the single-layer NMF is used in succession to decompose the learned dictionary layer by layer within the pre-training stage, then a fine-tuning stage is employed to further optimize all the matrix factors obtained in the previous stage by reconstructing the original data via the production of these factors. As a result, a set of dictionaries that can be mutually represented in a hierarchical relationship are learned in a deep learning fashion in our method, by which the multi-layer structured representation property of HSIs is expected to be fully utilized to accurately extract the target endmembers. In addition, we introduce the $L_{1/2}$ sparseness regularization and self-supervised regularization into the DNDF model. Specifically, the sparseness regularization plays a role in enforcing the sparseness of coefficient matrices during each factorization, while the self-supervised regularization aims to keep the similarity between the atoms of dictionaries and the approximate endmember signatures learned by conducting fuzzy C-means clustering on the target HSI. The performance of the proposed method is evaluated on three real HSI datasets, where our DNDF demonstrates the effectiveness and superiority over several state-of-the-art algorithms. For the sake of clarity, the major contributions of this paper are highlighted as follows.

- We propose a novel Deep Nonnegative Dictionary Factorization method for hyperspectral unmixing. By decomposing the learned dictionary layer by layer in our method, the multi-layer structured representation information implied in HSIs is expected to be captured and used to more accurately extract the target endmembers.
- Our method incorporates $L_{1/2}$ sparse constraint and self-supervised regularization into the deep factorization framework, by which the abundance sparseness prior, the sparsity-representation property of dictionaries, as well as the endmember signatures information presented in the data are jointly exploited to guide the unmixing.
- Our method consists of two stages, where the pre-training stage is designed to decompose the learned basis matrix layer by layer, and the fine-tuning stage are implemented to further optimize all the matrix factors obtained in the previous stage by reconstructing the original data via the production of these factors. Consequently, a set of dictionaries that capture the hierarchical representation relationship of the data are learned in a deep learning way.

2. Related Work

Here we briefly review HU methods based on constrained NMF and deep NMF. Then, the differences between our method and the existing methods are highlighted.

NMF has been demonstrated to be a promising method for HU under the linear spectral mixture model (LMM) [6]. It can decompose a nonnegative HSI matrix into a production of nonnegative endmember and abundance matrices, such that each pixel in an HSI can be represented as a linear combination of endmember spectra in the endmember matrix weighted by its fractional abundances in the abundance matrix [20]. This makes NMF not only possible to simultaneously unmix the endmembers and the corresponding abundances, but also ensure the nonnegativity of the estimated endmembers and abundances, and thus the unmixing results have good interpretability [12]. However, the basic NMF is easy to get into local minima, and also suffers from the problem of non-unique solution [21].

In order to enhance the performance of basic NMF for HU, many improvements have been conducted by developing effective regularizations for NMF, such as the geometry-based regularization and sparseness-based regularization. For example, minimum volume constrained NMF (MVCNMF) [22] is developed by introducing geometry-based regularization in attempt to minimize the volume
of the simplex formed by the candidate endmembers. Aiming at exploiting the sparse prior of abundance, $L_{1/2}$ sparsity-constrained NMF ($L_{1/2}$-NMF) [20] is proposed by employing an effective $L_{1/2}$ regularizer. On this basis, graph-regularized $L_{1/2}$-NMF (GLNMF) [23] introduces a graph structure-based regularization that can leverage the intrinsic manifold structure of HSIs. Recently, spatial group sparsity regularized NMF (SGSNMF) is proposed by utilizing a spatial group-structured prior information of HSIs [24]. Although existing works have improved the performance of NMF-based unmixing methods, these methods only adopt the single layer factorization, which is not favorable for efficiently utilizing the multi-layers structured information implied in HSIs [17].

Recently, deep NMF-based methods have shown their potentials for unmixing [19,25,26]. The advantages of deep NMF lie in that unlike the single layer NMF decomposes the original data into the production of two matrix factors, it learns the hierarchical and abstract features of the original data by factorizing the coefficient matrix layer by layer [16]. As a result, more complex abstract features from data can be captured and used to improve unmixing performance [19]. However, the existing deep NMF-based unmixing methods mainly focus on factorizing the coefficient matrix neglecting to capture the multi-layer representation relationship among basis matrices.

Unlike these methods, we propose the two-stage deep dictionary factorization to effectively exploit the hierarchical representation relationship of dictionaries in a deep learning way, and design effective regularizations for the hyperspectral unmixing purpose, by which the multi-layers structured information included in HSIs, the abundance sparsity property of data, as well as the endmember signatures information implied in the data are expected to be jointly utilized to more accurately estimate the endmembers and abundances.

3. Methodology

3.1. Notation and Problem Definition

In this paper, we use boldface uppercase letters to represent matrices, and boldface lowercase letters to denote the vectors. For a matrix $X \in \mathbb{R}^{m \times n}$, $X^\top$ denotes the transpose of $X$, boldface lowercase $x_n$ stands for the $n$-th column vector of $X$, and $[X]_{mn}$ represents the entry of $X$ in the $m$-th row and $n$-th column. We use $\|X\|_F$ to denote the Frobenius norm of $X$. $X \succeq 0$ signifies that each entry of $X$ is equal to or larger than zero. For $I \in \mathbb{N}$, we use $I_l$ to denote the $l \times l$ identity matrix, and $1_l$ to represent an all-one $l \times l$ matrix. The symbol "$\odot$" and "$\odot^\prime$" are used to represent the element-wise multiplication and division operation of the matrix, respectively.

Logically, an HSI can be regarded as an image cube $X \in \mathbb{R}^{M \times C \times B}$, where $B$ denotes the number of bands involved in each pixel, whilst $M$ and $C$ denote the number of rows and columns of pixels, respectively. For the sake of computational convenience, this cube can be represented in a matrix form as $Y = [y_1, y_2, ..., y_N]$, where $Y \in \mathbb{R}^{B \times N}$, $N = M \times C$, and $y_n \in \mathbb{R}^B$ denotes the $n$-th pixel spectrum. According to LMM, given that $Y$ is known, the HU problem can be solved from the perspective of blind source separation, by which $Y$ is approximately reconstructed by $Y \approx AS$, where $A = [a_1, a_2, ..., a_K] \in \mathbb{R}^{B \times K}$ is the nonnegative endmember matrix with each column $a_k$ denoting an endmember; $S = [s_1, s_2, ..., s_N] \in \mathbb{R}^{K \times N}$ represents the nonnegative abundance matrix with $s_n = (s_{n1}, s_{n2}, ..., s_{nk})^\top$ standing for the abundance vector of the pixel spectrum $y_n$. In other words, we need to estimate both $A$ and $S$ given that only $Y$ is known in actual HU applications.

3.2. DNDF Model

The unmixing performance of constrained NMF-based methods is limited by their single layer factorization ability. Although the existing deep NMF-based unmixing methods have overcome this issue to some extent by decomposing the coefficient matrix in succession, this manner of factorization is unfavorable to exploit the complex multi-layer representation relationship between the endmembers and the mixed pixels included in HSIs. To deal with this challenge, the proposed DNDF model decomposes the basis matrix layer by layer instead of the coefficient matrix. Specifically, the data
matrix $Y$ is first factorized into $A_1$ and $S_1$, then $A_1$ is further decomposed into $A_2$ and $S_2$, next the factorization procedure is continued until the $L$-th layer is reached. This can be formally expressed as

$$
Y \approx A_1 S_1, \\
A_1 \approx A_2 S_2, \\
\ldots, \\
A_{l-1} \approx A_l S_l, \\
A_L \approx A_L S_L.
$$

where $A_l$ and $S_l$ respectively represents the basis and coefficient matrices in the $l$-th layer. Based on this multi-layer factorization schema, the data matrix $Y$ can be approximately reconstructed by

$$
Y \approx A_L S_L S_{L-1} \cdots S_1,
$$

where $A_L$ is regarded as the endmember matrix, and $S_1 S_{L-1} \cdots S_1$ is used to generate the abundances.

As shown in Figure 1, our proposed DNDF unmixing model consists of two parts: pre-training model and fine-tuning model. By the pre-training model, we aim to learn multi-layer dictionaries by decomposing the basis matrix layer by layer, so that the endmembers can be better learned from the dictionaries extracted from the HSI data in a deep learning fashion. With the guidance of fine-tuning model, the obtained matrices by the previous model are expected to be optimized jointly by reconstructing the original data as far as possible. Formally, the objective functions of the pre-training $Q_p$ and fine-tuning $Q_f$ can be respectively expressed as follows:

$$
\begin{align*}
\min_{A_l, S_l} & \; Q_p(A_l, S_l) \\
= & \|A_{l-1} - A_l S_l\|^2_F + \lambda \sum_{k=1}^{K_l} \|A_l - E_l\|^2_F + \beta \|S_l\|^2_1, \quad \text{s.t. } A_l, S_l \succeq 0
\end{align*}
$$

$$
\begin{align*}
\min_{A_l, S_L, \ldots, S_1} & \; Q_f(A_L, S_L, \ldots, S_1) \\
= & \|Y - \prod_{l=L}^{1} S_l\|^2_F + \lambda \sum_{k=1}^{K_l} \|A_l - S_l E_l\|^2_F + \beta \sum_{l=1}^{L} \|S_l\|^2_1, \quad \text{s.t. } A_l, S_L, \ldots, S_1 \succeq 0
\end{align*}
$$

where $Y$ denotes the known HSI matrix; $A_{l-1}$ is the basis matrix of the $(l - 1)$-th layer and $A_{l-1} = Y$ in case of $l = 1$, while $A_l$ and $S_l$ represent the basis and coefficient matrix of the $l$-th layer, respectively; $\lambda$ and $\beta$ are regularization parameters. Specifically, the first term in $Q_p$ plays the same role of single layer NMF for decomposing the dictionary $A_{l-1}$, whereas the function of first term in $Q_f$ is to reconstruct $Y$ using the product of the matrix factors learned based on $Q_p$. The second and third terms in both $Q_p$ and $Q_f$ are the self-supervised regularization and sparseness regularization, respectively. Since weakly supervised information for the basis matrix helps improve the performance of NMF [27], we employ the self-supervised fashion to learn effective weakly supervised information as guidance for extracting accurate endmembers. Specifically, before each factorization in the layer $l$, where $l = 1, 2, \ldots, L$, the Fuzzy C-means (FCM) clustering method is conducted on $Y$, then the obtained cluster centers are used to construct the approximate endmember matrix $E_l$, which functions as the weakly supervised information for the endmember matrix. Note that the number of clusters of FCM equals the number of column vectors in the basis matrix of the $l$-th layer; $E_l = [e_1, e_2, \ldots, e_{K_l}]$ denotes
the approximate endmember matrix with each \( e_k, k = 1, 2, ..., K \) representing a clustering center given by FCM, whereas \( \bar{A}_k \) and \( \bar{E}_k \) are respectively defined by

\[
\bar{A}_k = [a_{k1}, a_{k2}, ..., a_{kL-1}] \\
\bar{E}_k = [e_1, e_2, ..., e_{k-1}, e_{k+1}, ..., e_K].
\]

It is worth pointing out that the numerator of proposed self-supervised regularization term plays the role of keeping the achieved \( A_l \) as similar as the approximate endmember matrix \( E_l \) whilst the denominator aims to make each estimated endmember \( a_k \) has the maximum difference with the other clustering centers except for \( e_k \).

To learn a sparse representation between dictionaries, sparsity constraint is also imposed on the coefficient matrix in the factorization of each layer. The effectiveness of this measures lies in that the sparsity of abundance distribution is an intrinsic property of HSIs [20,23,24], as well as sparse representation enforces the learned endmember signatures to be more similar to the real endmember. Previous studies have demonstrated that \( L_{1/2} \) regularization has the superiority for enforcing the sparseness of decomposed abundance matrix. Thus, the \( L_{1/2} \) regularizer \[ \|S\|_2 = \sum_{i,j}(s_{ij})^{1/2} \] is introduced into \( Q_p \) and \( Q_f \) to impose sparsity constraint on the \( S_l \).

\[ \text{Figure 1. Illustration of the proposed model of deep nonnegative dictionary factorization that includes two parts: pre-training and fine-tuning stages. By the pre-training stage, we aim to learn a set of dictionaries } A_1, ..., A_L \text{ with hierarchical representation relationship by decomposing the basis matrix layer by layer, so that endmembers can be better learned from the dictionaries extracted from the HSI data in a deep leaning fashion. With the guidance of fine-tuning model, the obtained matrix factors } A_L, S_L, \dots, S_1 \text{ in the previous stage are expected to be further optimized jointly by reconstructing the HSI matrix } Y \text{ as far as possible.} \]

3.3. Optimization and Implementation

Since the objective functions in Equations (3) and (4) are both non-convex w.r.t two or more matrix factors simultaneously, we deduce the multiplicative update rules (MURs) of the proposed unmixing model by alternating optimization, by which each matrix factor is updated with the others being fixed.

(i) MURs for the pre-training stage. Considering the objective function \( Q_p \) in Equation (3), when \( A_l \) remains unchanged, the MUR of \( S_l \) via [20] can be given as follows:

\[ S_l \leftarrow S_l \odot A_l^\top A_{l-1} \odot ((A_l^\top A_l)S_l + \beta(S_l)^{-\frac{1}{2}}), \] (5)

where \[ ((S_l)^{-\frac{1}{2}})_{ij} = \frac{1}{(s_{ij})^{1/2}} \]. As for the MUR of \( A_l \), we first give the gradient of function \( Q_p \) w.r.t \( A_l \) by

\[ \frac{\partial Q_p}{\partial A_l} = A_l S_l^\top S_l - A_{l-1} S_l^\top - \lambda G(A_l, E_l) + \lambda H(A_l, E_l), \]
where matrix $W = I_K - I_K$; The functions $G$ and $H$ are respectively defined by

$$
G(A_l, E_l) = \frac{E_l (\sum_{k=1}^{K_l} \| \hat{A}_k^l - E_k^l \|^2 + \| A_l - E_l \|^2 W)}{(\sum_{k=1}^{K_l} \| \hat{A}_k^l - E_k^l \|^2)^2},
$$

$$
H(A_l, E_l) = \frac{(\sum_{k=1}^{K_l} \| \hat{A}_k^l - E_k^l \|^2 + (K_l - 1) \| A_l - E_l \|^2) A_l}{(\sum_{k=1}^{K_l} \| \hat{A}_k^l - E_k^l \|^2)^2}.
$$

Based on the gradient descent principle, the update rule for $A_l$ can be given by

$$
[A]_{mi} \leftarrow [A]_{mi} - \eta_{mi} \frac{\partial Q_p}{\partial [A]_{mi}}.
$$

Then, we set $\eta_{mi} = A_l \otimes (A_l S_l^T S_l + \lambda H(A_l, E_l))$ in Equation (6). The MUR for $A_l$ can be deduced as

$$
A_l \leftarrow A_l \otimes (A_l \otimes (A_l S_l^T S_l + \lambda G(A_l, E_l)) \otimes (A_l S_l^T S_l + \lambda H(A_l, E_l))).
$$

(ii) MURs for the fine-tuning stage. For the objective function $Q_f$ in Equation (4), when $S_L, \ldots, S_1$ keep fixed and are regarded as a whole matrix, the MUR of $A_L$ can be obtained in terms of the deduction procedure of MUR in Equation (7). It can be expressed as

$$
A_L \leftarrow A_L \otimes (Y (\prod_{l=L}^{1} S_l)^T + \lambda G(A_L, E_L)) \otimes (A_L \otimes (\prod_{l=L}^{1} S_l)^T (A_L \otimes (\prod_{l=L}^{1} S_l)) S_1 + \beta (S_1)^{-\frac{1}{2}}).
$$

Similarly, when $A_L S_L S_{L-1} \cdots S_2$ is considered as a fixed matrix, the MUR of $S_1$ can be given as

$$
S_1 \leftarrow S_1 \otimes (A_L \prod_{l=L}^{2} S_l)^T Y \otimes ((A_L \prod_{l=L}^{2} S_l)^T (A_L \prod_{l=L}^{2} S_l) S_1 + \beta (S_1)^{-\frac{1}{2}}),
$$

For the derivation of the MUR of $S_l$, where $l = 2, 3, \ldots, L$, $Q_f$ in Equation (4) can be re-expressed as

$$
Q_f(S_l) = \frac{1}{2} \| Y - \Phi S_l \Psi \|_F^2 + \frac{\lambda}{2} \sum_{k=1}^{K_l} \| \hat{A}_k^l - E_k^l \|^2 + \beta \sum_{l=1}^{L} \| S_l \|_F^2,
$$

where $\Phi = A_L S_L \ldots S_{l+1}$ and $\Psi = S_{l-1} \ldots S_1$, respectively. The additive update for $S_l$ can be given as

$$
[S]_{uv} \leftarrow [S]_{uv} - \eta_{uv} \frac{\partial Q_d}{\partial [S]_{uv}}.
$$

The gradient of $Q_f$ w.r.t $S_l$ can be expressed by

$$
\frac{\partial Q_f}{\partial [S]_{uv}} = -\Phi \Psi^T + \Phi \Psi S_l \Psi^T + \frac{\lambda}{2} S_l^{-\frac{1}{2}}.
$$

By substituting Equation (12) into Equation (11) and setting $\eta_{uv} = S_l \otimes (\Phi \Psi S_l \Psi^T + \lambda S_l^{-\frac{1}{2}})$, the MUR of $S_l$ can be derived as

$$
S_l \leftarrow S_l \otimes (\Phi \Psi S_l \Psi^T) \otimes (\Phi \Psi S_l \Psi^T + \lambda S_l^{-\frac{1}{2}}).
$$
Based on the above analysis, the pseudo-code of the proposed DNDF algorithm for HU is summarized in Algorithm 1. It mainly consists of two parts: the pre-training part (Line 2–10) and the fine-tuning part (Line 12–18). For the pre-training part, it is worth noting that $A_l$ is initialized based on the cluster centers given by the FCM method, while the initial values of $S_l$ are provided by the fully constrained least squares algorithm [28].

**Algorithm 1: DNDF Algorithm For HU**

**Input:** HSI matrix $Y \equiv [y_1, y_2, \ldots, y_B] \in \mathbb{R}^{N \times B}$; 
Number of endmembers $K$; 
Algorithm parameters $\lambda$ and $\beta$.

**Output:** Endmember matrix $A$ and abundance matrix $S$.

1. #Pre-training stage
2. \hspace{1em} for $l = 1$ to $L$ do
3. \hspace{2em} Execute FCM algorithm on $Y$ and obtain the clustering centers $e_1, e_2, \ldots, e_K$;
4. \hspace{2em} Set $A_l = [e_1, e_2, \ldots, e_K] \top$ and $E_l = [e_1, e_2, \ldots, e_K] \top$;
5. \hspace{2em} Initialize $S_l$ by the FCLS method;
6. \hspace{2em} while the stopping criteria are not satisfied do
7. \hspace{3em} Update $A_l$ by Equation (7);
8. \hspace{3em} Update $S_l$ by Equation (5);
9. \hspace{2em} end
10. end
11. #Fine-tuning stage
12. repeat
13. \hspace{1em} Update $A_L$ by Equation (8);
14. \hspace{1em} for $l = L$ to 2 do
15. \hspace{2em} Update $S_l$ by Equation (13);
16. \hspace{1em} end
17. \hspace{1em} Update $S_1$ by (9);
18. until the stopping criteria are not satisfied;
19. Return $A = A_L$ and $S = S_L \cdots S_1$.

4. Experiment

Here, we first briefly introduce the datasets used in the experiments, the baseline methods and algorithm settings, as well as the performance metrics. Then, the quantitative evaluation, factors exploration and parameter analysis are addressed.

4.1. Experiment Setting

**Datasets.** Table 1 summarizes the statistics of three real datasets used in the experiments and Figure 2 shows their visual pseudo-color images. (1) *Samson* comprises $952 \times 952$ pixels and has 156 bands covering the wavelengths from 0.401 to 0.889 $\mu$m. For computational convenience, a subimage that has the size of $95 \times 95$ pixels and contains three types of endmembers (Soil, Tree, and Water) is chosen from the original data and used in the experiment [19,29]. (2) *Jasper Ridge* was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. The original data has $512 \times 614$ pixels, each of which is recorded with 224 bands with the wavelength ranging from 0.380 to 2.5 $\mu$m. Following the setting of [30], a subimage with $100 \times 100$ pixels is adopted. The target scene mainly has four kinds of endmembers: Soil, Water, Road, and Tree. To eliminate the impact of the dense water vapor and atmospheric effects, the spectral bands 1–3, 108–112, 154–166 and 220–224 are removed from the data, leaving 198 bands. (3) *Washington DC Mall* is captured by the airborne HYperspectral Digital Imagery Collection Experiment (HYDICE) sensor over the Washington DC Mall area, USA.
The original data consists of 210 spectral bands with wavelengths from 0.4 to 2.5 µm. A subimage with 150 × 150 pixels from the original HSI is used in the experiment [31]. There are five types of endmembers in the selected subimage including Grass, Tree, Street, Roof, and Water. We deleted the bands corrupted by noise seriously, as well as the water vapor absorption bands (including bands 103–106, 138–148 and 207–210) from the data and the remaining 191 bands were used.

| Dataset       | #Endmember | #Band | #Pixel    | Wavelength | Sensor-Type |
|---------------|------------|-------|-----------|------------|-------------|
| Samson        | 3          | 156   | 95 × 95   | 0.401–0.889 µm | SAMSON      |
| Jasper Ridge  | 4          | 198   | 100 × 100 | 0.38–2.5 µm   | AVIRIS      |
| Wash. DC Mall | 5          | 191   | 150 × 150 | 0.4–2.5 µm    | HYDICE      |

# means the number of something. # Band means the number of band.

Figure 2. Pseudo-Color images of three real hyperspectral image datasets. (a) Samson; (b) Jasper Ridge; (c) Wash. DC Mall.

Baseline Methods. We compare the proposed DNDF with four kinds of NMF-based unmixing methods. (1) Two sparsity-constrained NMF-based algorithms such as L_{1/2}-NMF [20] and SGSNMF [24]; (2) An NMF-based method with spectral geometry related constrain, i.e., MVCNMF [22]; (3) A representative graph-regularization based NMF, named GLNMF [23]; (4) A state-of-the-art deep NMF-based unmixing approach, i.e. deep NMF with total variation (SDNMF-TV) [19]. The algorithm parameters for all baseline methods adopt the same values as those in their original works. To make a fair comparison, all the algorithms initialize the endmember and abundance matrices by vertex component analysis (VCA) [32] and FCLS [28], respectively, except for SGSNMF, which initializes the endmember matrix with the help of superpixel segmentation. Additionally, we test our method under the condition that the number of factorization layers is set to 1 and 2, respectively, and the obtained algorithms are referred to as DNDF-L1 and DNDF-L2 accordingly. The maximum number of iterations is set to 500 for pre-training, and 4000 for the fine-tuning stage. We set the experience values of regularization parameters to λ = 0.1 and β = 0.1, respectively. All the algorithms are tested 20 times and the average performance is reported.

Performance Metrics. To evaluate the performance of our method, Spectral Angle Distance (SAD) [19,20], which can measure the similarity between an estimated endmember and the reference one, is adopted. Given that \( \hat{a}_k \) and \( a_k \) respectively denote the estimated and the reference endmember, SAD can be defined as

\[
SAD(a_k, \hat{a}_k) = \arccos \left( \frac{a_k^\top \hat{a}_k}{\|a_k\| \|\hat{a}_k\|} \right),
\]

where \( \| \cdot \| \) denotes the Euclidean norm of vectors. Since SAD represents the angular distance between two vectors, a smaller SAD value indicates better performance. According to its definition, the SAD measure has the unit of “radians”.


4.2. Performance Evaluation

To test the performance of our method, three experiments are conducted on different datasets. The performance evaluation is presented from three aspects: quantitative evaluation, factor exploration, and algorithm parameter analysis.

Quantitative Evaluation. The quantitative comparisons about the experimental results on three datasets are reported in Table 2 including the SAD values of each endmember and their averages of all the endmembers provided by each baseline algorithm. Each row corresponding to the individual endmember shows the performance of all methods w.r.t this endmember, whereas the rows indicated by “Average” denote the average SAD values of all the endmembers of a specific dataset. The value with red color denotes the best performance, whilst the blue value indicates the second best. As can be seen from the average SAD respectively, the proposed DNMF-L2 performs best on all three datasets. It reduces 9.9% mean SAD on the Samson dataset compared with the second best method, i.e., DNMF-L1. According to the results on the Jasper Ridge and Washington DC Mall datasets, DNMF-L2 decreases the mean SAD of 9.7% and 8.6% respectively than the second best method, i.e., GLNMF. As for the performance on individual endmember, DNDF-L2 has the largest number of second-best instances on Samson, and provides a best and a second best instances among four endmembers on Jasper Ridge. On Washington DC Mall, DNDF-L2 shows its superiority by giving one best and two second best instances within five endmembers.

Table 2. Performance comparison w.r.t the metric of spectral angle distance (SAD) on three datasets (in radians), where a smaller SAD value indicates better performance signifying the better similarity between the estimated endmember and the reference one. Red bold and blue italic denote the best and second-best results, respectively.

| Dataset     | DNDF-L1 | DNDF-L2 | GLNMF | SDNMF-TV | SGSNMF | MVCNMF | L_{1/2}-NMF |
|-------------|---------|---------|-------|----------|--------|--------|-------------|
| Samson      |         |         |       |          |        |        |             |
| Soil        | 0.0220  | 0.0160  | 0.0275| 0.1761   | 0.0097 | 0.3148 | 0.0279      |
| Tree        | 0.0290  | 0.0316  | 0.0423| 0.1310   | 0.0459 | 0.0734 | 0.0426      |
| Water       | 0.1758  | 0.1567  | 0.2294| 0.0922   | 0.2289 | 0.3204 | 0.3015      |
| Average     | 0.0756  | 0.0681  | 0.0997| 0.1331   | 0.0949 | 0.2365 | 0.1240      |
| Jasper Ridge|         |         |       |          |        |        |             |
| Tree        | 0.0609  | 0.0729  | 0.1270| 0.1026   | 0.1415 | 0.1752 | 0.1187      |
| Water       | 0.0694  | 0.0659  | 0.1872| 0.1088   | 0.2236 | 0.2980 | 0.2022      |
| Soil        | 0.1173  | 0.2010  | 0.1170| 0.1069   | 0.1385 | 0.1406 | 0.1324      |
| Road        | 0.3754  | 0.1123  | 0.0697| 0.4662   | 0.0483 | 0.1458 | 0.0690      |
| Average     | 0.1558  | 0.1130  | 0.1252| 0.1961   | 0.1380 | 0.1899 | 0.1306      |
| Wash. DC Mall|        |         |       |          |        |        |             |
| Tree        | 0.0095  | 0.0321  | 0.1738| 0.0993   | 0.1606 | 0.2372 | 0.1738      |
| Grass       | 0.1128  | 0.1404  | 0.2585| 0.1682   | 0.1425 | 0.2571 | 0.2688      |
| Street      | 0.7073  | 0.5293  | 0.3887| 0.4520   | 0.6963 | 0.5859 | 0.4195      |
| Roof        | 0.2435  | 0.2632  | 0.1942| 0.3645   | 0.2525 | 0.1663 | 0.1720      |
| Water       | 0.0753  | 0.0462  | 0.0920| 0.2509   | 0.1001 | 0.1117 | 0.0892      |
| Average     | 0.2297  | 0.2023  | 0.2214| 0.2670   | 0.2704 | 0.2716 | 0.2247      |

Factor Exploration. To provide an intuitive evaluation about DNDF, we provide some visual results. Figure 3 shows an example of the learned hierarchical dictionaries on the Samon by the obtained basis matrices $A_1$ and $A_2$ in the pre-training stage of DNDF-L2. The atoms of the dictionary learned by the factorization of first layer are respectively plotted in Figure 3a–c according to their categories, whereas Figure 3d–f show the atoms of dictionary of the second layer. Here, the red solid lines denote the reference endmember signatures while the dotted lines represent the atoms. As can be seen in Figure 3a–c, the learned atoms have good similarity to the reference signatures except for a particular learned signature with yellow color in Figure 3c. It also can be seen from Figure 3d–f, the learned atoms have good consistency with the corresponding reference signatures, revealing the effectiveness of our
method. In order to qualitatively interpret the abundance estimations, we illustrate some instances of the abundance maps unmixed by all the methods in Figure 4. All the abundance maps are represented by a heatmap manner, where the pixel with high color temperature denotes a big presence proportion of corresponding endmember.

Next, we investigate the effectiveness of each term in our proposed model. Table 3 shows the performance comparison of on three datasets, where DNDF denotes the proposed method without any regularization, DNDF-SP, DNDF-SS, and DNDF-L2 represent the proposed method with sparse regularization, self-supervised regularization and both regularizations, respectively. We can see that the integrated model outperforms the partial components, which demonstrates the effectiveness of the joint work of different terms.

Figure 3. Comparison between the reference endmembers (red solid line) and the atoms (dotted line) of the learned dictionaries given by our DNDF-L2 on Samson. (a–c) plots the reference endmembers and the atoms of the dictionary learned in the first layer factorization, and (d–f) compares the reference endmembers and the atoms of the dictionary of second layer. (a) Soil; (b) Tree; (c) Water; (d) Soil; (e) Tree; (f) Water.
Figure 4. Abundance maps estimated by DNDF-L1, DNDF-L2, GLNMF, SDNMF-TV, SGSNMF, MVCNMF, $L_{1/2}$-NMF on Washington DC Mall. (a) Grass; (b) Tree; (c) Roof; (d) Street; (e) Water.
Table 3. Performance comparison w.r.t the SAD values on three datasets (in radians), where DNDF denotes the proposed method without any regularization, DNDF-SP, DNDF-SS, and DNDF-L2 represent the proposed method with sparse regularization, self-supervised regularization and both regularizations, respectively. Red bold and blue italic denote the best and second-best results, respectively.

| Dataset       | DNDF | DNDF-SP | DNDF-SS | DNDF-L2 |
|---------------|------|---------|---------|---------|
| Samson        |      |         |         |         |
| Soil          | 0.0169 | 0.0165 | 0.0164 | 0.0160 |
| Tree          | 0.0325 | 0.0319 | 0.0321 | 0.0316 |
| Water         | 0.1672 | 0.1635 | 0.1610 | 0.1567 |
| Average       | 0.0722 | 0.0706 | 0.0698 | 0.0681 |
| Jasper Ridge  |      |         |         |         |
| Tree          | 0.0772 | 0.0742 | 0.0735 | 0.0729 |
| Water         | 0.0686 | 0.0621 | 0.0666 | 0.0659 |
| Soil          | 0.1979 | 0.2066 | 0.1764 | 0.2010 |
| Road          | 0.1481 | 0.1327 | 0.1423 | 0.1123 |
| Average       | 0.1230 | 0.1189 | 0.1147 | 0.1130 |
| Wash. DC Mall |      |         |         |         |
| Tree          | 0.0427 | 0.0339 | 0.0447 | 0.0321 |
| Grass         | 0.1483 | 0.1424 | 0.1465 | 0.1404 |
| Street        | 0.5958 | 0.5686 | 0.5570 | 0.5293 |
| Roof          | 0.2586 | 0.2594 | 0.2579 | 0.2632 |
| Water         | 0.0478 | 0.0456 | 0.0384 | 0.0462 |
| Average       | 0.2186 | 0.2100 | 0.2089 | 0.2023 |

Parameter Analysis. Here, we analyze the impact of two algorithm parameters. We first fix $\beta = 0.1$, then choose different values for $\lambda$ from the set $\{1 \times 10^{-4}, 1 \times 10^{-3}, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Figure 5a shows the performance of DNDF-L2 on Jasper Ridge dataset. As can be seen, the performance our algorithm ascends in the first steps gets better SAD values when $\lambda \in [0.01, 0.2]$, and the performance degrades when $\lambda \geq 0.3$. $\lambda = 0.1$ is the optimal selection according to SAD criteria. Next, the impact of $\beta$ on our algorithm is tested on the condition that $\lambda$ is fixed to 0.1, whilst selecting different values for $\beta$ in the same collection as $\lambda$. As Figure 5b shows that our algorithm has better performance when $\beta \in [0.01, 0.3]$, and $\beta = 0.1$ is the best choice.

Running Time. To evaluate the performance of our methods on running time, we tested all the seven methods on three datasets. The average running time of them are reported in Table 4. From Table 4, it can be seen that DNDF-L2 needs shorter time than SDNMF-TV and MVCNMF on all the the datasets. According to the results on the Samson dataset, we can see from Table 4 that the running time of DNDF-L1 is the least, and SDNMF-TV needs the most time. When all the methods
are tested on Jasper Ridge dataset, $L_{1/2}$-NMF consumes the least time, followed by DNDF-L1. As to the results on Wash. DC Mall dataset, $L_{1/2}$-NMF performs best, and DNDF-L1 is better than SGSNMF, DNDF-L2, SDNMF-TV, and MVCNMF.

Table 4. Running time on three real data sets (in seconds).

| Data Sets         | DNDF-L1 | DNDF-L2 | GLNMF | SDNMF-TV | SGSNMF | MVCNMF | $L_{1/2}$-NMF |
|-------------------|---------|---------|-------|----------|--------|--------|---------------|
| Samson            | 25.77   | 113.17  | 48.20 | 462.18   | 42.49  | 262.04 | 42.41         |
| Jasper Ridge      | 60.82   | 169.68  | 64.56 | 558.40   | 68.28  | 583.24 | 54.52         |
| Wash. DC Mall     | 130.81  | 385.52  | 102.13| 1062.78  | 161.88 | 1229.91| 89.43         |

5. Conclusions

We proposed a novel Deep Nonnegative Dictionary Factorization framework with $L_{1/2}$ sparseness constraint and self-supervised regularization for hyperspectral unmixing task, which integrated the pre-training stage and fine-tuning stage into a unified model. By imposing sparse constraints on the coefficient matrices, DNDF jointly exploited the sparseness prior of abundance, as well as the sparsity-representation property of dictionaries. In addition, the self-supervised regularization based on the guidance of fuzzy clustering helped to learn hierarchical dictionaries that captured the latent endmembers. Experiments conducted on three real-world datasets demonstrated the effectiveness and superiority of our approach compared with several state-of-the-art methods based on single-layer NMF and deep NMF.

Author Contributions: Conceptualization, W.W. and H.L.; Data curation, W.W.; Formal analysis, W.W.; Funding acquisition, W.W.; Investigation, W.W.; Methodology, W.W.; Project administration, H.L.; Resources, H.L.; Software, W.W.; Supervision, H.L.; Validation, W.W.; Visualization, W.W.; Writing—original draft, W.W.; Writing—review & editing, W.W. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the project ZR2017MF028 supported by Shandong Provincial Natural Science Foundation.

Acknowledgments: We would like to thank Yanfei Zhong (State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China) and Hairong Qi (Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA) for sharing the source codes.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| Abbreviation | Description                              |
|--------------|------------------------------------------|
| HSI          | Hyperspectral remote sensing image       |
| HU           | Hyperspectral unmixing                   |
| NMF          | Nonnegative matrix factorization         |
| DNDF         | Deep nonnegative dictionary factorization|
| VCA          | Vertex component analysis                |
| FCLS         | Fully constrained least squares          |
| LMM          | Linear spectrum mixture model            |
| SAD          | Spectral angle distance                  |
| RMSE         | Root mean square error                   |
| MUR          | Multiplicative update rule               |

References

1. Nalepa, J.; Myller, M.; Kawulok, M. Validating Hyperspectral Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* 2019, 16, 1264–1268. [CrossRef]
1. Lee, D.D.; Seung, H.S. Algorithms for Non-negative Matrix Factorization. In Advances in Neural Information Processing Systems; Leen, T.K., Dietterich, T.G., Tresp, V., Eds.; MIT Press: Cambridge, MA, USA, 2001; pp. 556–562.

2. Cai, D.; He, X.; Han, J.; Huang, T.S. Graph Regularized Nonnegative Matrix Factorization for Data Representation. IEEE Trans. Pattern Anal. Mach. Intell. 2011, 33, 1548–1560.

3. Zhang, W.; Lu, X.; Li, X. Similarity Constrained Convex Nonnegative Matrix Factorization for Hyperspectral Anomaly Detection. IEEE Trans. Geosci. Remote Sens. 2019, 57, 4810–4822. [CrossRef]

4. Xiong, F.; Zhou, J.; Qian, Y. Material Based Object Tracking in Hyperspectral Videos. IEEE Trans. Image Process. 2020, 29, 3719–3733. [CrossRef]

5. Chen, Q.; Sun, J.; Palade, V.; Shi, X.; Liu, L. Hierarchical Clustering Based Band Selection Algorithm for Hyperspectral Face Recognition. IEEE Access 2019, 7, 24333–24342. [CrossRef]

6. Bioucas-Dias, J.M.; Plaza, A.; Dobigeon, N.; Parente, M.; Du, Q.; Gader, P.; Chanussot, J. Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2012, 5, 354–379. [CrossRef]

7. Wang, N.; Du, B.; Zhang, L. An Endmember Dissimilarity Constrained Non-Negative Matrix Factorization Method for Hyperspectral Unmixing. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2013, 6, 554–569. [CrossRef]

8. Zare, A.; Ho, K.C. Endmember Variability in Hyperspectral Analysis: Addressing Spectral Variability During Spectral Unmixing. IEEE Signal Process. Mag. 2014, 31, 95–104. [CrossRef]

9. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. Nature 1999, 401, 788–791. [CrossRef]

10. Lee, D.D.; Seung, H.S. Algorithms for Non-negative Matrix Factorization. In Advances in Neural Information Processing Systems; Leen, T.K., Dietterich, T.G., Tresp, V., Eds.; MIT Press: Cambridge, MA, USA, 2001; pp. 560–562.
25. Rajabi, R.; Ghassemian, H. Spectral Unmixing of Hyperspectral Imagery Using Multilayer NMF. *IEEE Geosci. Remote Sens. Lett.* 2015, 12, 38. [CrossRef]

26. Fang, H.; Li, A.; Wang, T.; Xu, H. Hyperspectral unmixing using double-constrained multilayer NMF. *Remote Sens. Lett.* 2019, 10, 224–233. [CrossRef]

27. Choo, J.; Lee, C.; Reddy, C.K.; Park, H. Weakly supervised nonnegative matrix factorization for user-driven clustering. *Data Min. Knowl. Disc.* 2015, 29, 1598–1621. [CrossRef]

28. Heinz, D.C.; Chang, C.I. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* 2001, 39, 529–545. [CrossRef]

29. Zhu, F.; Wang, Y.; Fan, B.; Xiang, S.; Meng, G.; Pan, C. Spectral Unmixing via Data-Guided Sparsity. *IEEE Trans. Image Process.* 2014, 23, 5412–5427. [CrossRef]

30. Zhu, F. Hyperspectral Unmixing: Ground Truth Labeling, Datasets, Benchmark Performances and Survey. *arXiv* 2017, arXiv:1708.05125.

31. Jia, S.; Qian, Y. Constrained Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* 2009, 47, 161–173. [CrossRef]

32. Nascimento, J.M.P.; Bioucas-Dias, J.M. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* 2005, 43, 898–910. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).