A deep disentangled approach for interpretable hyperspectral unmixing

Ricardo Augusto Borsoi, Tales Imbiriba, Deniz Erdoğmuş

A DEEP DISENTANGLED APPROACH FOR INTERPRETABLE HYPERSPECTRAL UNMIXING

Ricardo Augusto Borsoi[†]*, Tales Imbiriba*[‡]*, Deniz Erdoğmuş*[‡]

[†] Université de Lorraine, CNRS, CRAN, Vandoeuvre-lès-Nancy, France

[‡] Dept. of Electrical & Computer Engineering, Northeastern University, Boston, MA, USA

ABSTRACT

Deep learning-based frameworks have been recently applied to hyperspectral umixing due to their flexibility and powerful representation capabilities. However, such techniques either use black-box models which are not physically interpretable, or fail to address the non-idealities of the unmixing problem. In this paper, we propose a physically interpretable deep learning method for hyperspectral unmixing accounting for nonlinearity and the variability of the endmembers. The proposed method is based on a probabilistic variational deep learning framework which employs semi-supervised disentanglement learning to properly separate the abundances and endmembers. A self-supervised strategy is used to generate labeled training data, and the model is learned end-to-end using stochastic backpropagation. Experimental results on both synthetic and real datasets illustrate the performance of the proposed method compared to state-of-the-art algorithms.

***Index Terms*—** Hyperspectral data, hyperspectral unmixing, neural networks, disentanglement learning, deep learning.

1. INTRODUCTION

Hyperspectral unmixing (HU) consists in estimating the spectral signatures of pure materials in a scene (i.e., *endmembers* – EMs) and the proportions with which they are contained in each pixel (i.e., *abundances*) directly from a hyperspectral image (HI) [1]. Due to the unsupervised nature of HU, adequately exploring the physics of the problem when devising modeling strategies is paramount for obtaining stable and accurate EM and abundance estimations. Traditional methods considered the interaction between light and the EMs to be linear. These strategies were based on frameworks including, e.g., nonnegative matrix factorization, Bayesian estimation and sparse regression [1]. However, such approaches disregard non-idealities such as nonlinear interactions between light and the materials [2] and the variability of the EMs in different HI pixels [3]. Extensions of the linear model were also proposed to deal with EM variability. However, such models are still over-simplified, motivating machine learning approaches capable of allying both flexibility and performance [4]. Nonetheless, interpretability remains a key point when leveraging machine learning strategies for HU [5, 6].

Recently, physically-motivated machine learning approaches have been successfully applied to HU [7–10]. The advantage of such models with respect to fully black-box strategies lies in the interpretability of the estimated EMs and abundances, which is a requirement for meaningful unmixing results. When deep learning strategies come into play, autoencoder (AEC) architectures are of special interest due to the intrinsic low-dimensionality of the abundance space with respect to the pixels, and to the connection between such strategies and hyperspectral mixing models [11]. Thus, several approaches using AECs were proposed to solve HU addressing phenomena such as nonlinearity [5, 8, 12] and EM variability [9, 13].

Although deep learning methods presented relevant solutions for HU reaching high levels of accuracy while retaining physical interpretation, such strategies fail to provide a separation between EMs and abundances that is both physical interpretable and accounts for existing spectral variability and nonlinear effects. Recently, supervised disentanglement learning has become a popular approach to separate latent variables in deep learning models into different factors of variation that can have a physical interpretation [14]. Disentangled decompositions have been considered for different applications (e.g., separating content from style in images [14]), and its potential will be explored in this work to aid the separation between abundance and EM variations.

In this paper, an interpretable probabilistic deep disentanglement learning framework that is based on self- and semi-supervised learning is proposed. The proposed method, which is named PhI-Net (*Physically Interpretable disentangled neural Networks for HU*), accounts for both nonlinearities in the mixing model and EM variability. Differently from traditional latent variable models, the proposed framework leverages disentanglement learning and physically-inspired neural networks (NNs) to provide more interpretable EM and abundance estimates. HU is performed as a fully unsupervised inference problem, where self-supervised learning strategies are leveraged to generate the training data. The parameters of the model are learned by maximizing a lower bound to the log-likelihood of training and test data. Importance sampling and Monte Carlo sampling are employed in order to approximate intractable distributions in the form of an optimization objective that can be optimized more efficiently. Differently from related black-box strategies, the proposed method addresses the challenges in HU by using clearly defined statistical models and hypotheses, and disentanglement is achieved during inference through independence assumptions. Moreover, PhI-Net augments physically motivated models with deep NNs. This leads to flexible but interpretable model, in which the influence of, e.g., nonlinearity and EM variability can be adjusted through the use of appropriate regularization strategies. Experimental results with synthetic and real data illustrate the performance of PhI-Net.

2. PROBLEM FORMULATION AND PROPOSED METHOD

HU aims at recovering P endmembers, $\boldsymbol{M} \in \mathbb{R}^{L\times P}$, and corresponding abundances $\boldsymbol{a}_n \in \mathbb{R}^P$ for each pixel $n \in \{1, \dots, N_U\}$ in the HI with L bands and N_U pixels. The most simplistic and widely used model used to describe the interaction between light and materials in the scene is the linear mixing model (LMM) [1]. The LMM, however, fails to accurately represent many scenarios where nonidealities such as nonlinearity [2] and EM variability [2] become non-

This work was supported in part by the National Geographic Society under Grant NGS-86713T-21.

negligible, requiring more sophisticated models capable of modeling such phenomena. In general, these models can be represented as:

$$\boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{a}_n, \boldsymbol{M}, \boldsymbol{\xi}_n) + \boldsymbol{e}_n \,, \tag{1}$$

where function $\boldsymbol{f}$ describes the mixing process, $\boldsymbol{\xi}_n$ is a vector of parameters which account for nonlinearity and spectral variability, and $\boldsymbol{e}_n \in \mathbb{R}^L$ denotes additive noise.

Although we address unsupervised HU, the proposed methodology will be developed in two steps. First, the statistical modeling and variational inference process are formulated in a semi-supervised learning process, for which we consider a dataset $\mathcal{D}$ with N pixels, which is partitioned in labeled $\mathcal{D}_S$ and unlabeled $\mathcal{D}_U$ portions with N_S and N_U pixels, respectively, such that $\mathcal{D} = \mathcal{D}_U \bigcup \mathcal{D}_S$. Later, we provide a self-supervised learning strategy is used to generate the training data $\mathcal{D}_S$ and provide a fully unsupervised HU pipeline.

Next, we propose a mixing model that incorporates nonlinearity and variability existing in real scenarios. Then, we formulate the HU problem using a variational inference framework followed by a discussion of the cost function for the proposed semi-supervised objective. We conclude the section with a description of the NN architectures and the self-supervised strategy. In the following, we will drop the pixel index n to lighten the notation.

2.1. The mixing model

Abundance prior: The Dirichlet distribution is a natural choice for modeling abundance vectors since it enforces the non-negativity and sum-to-one physical constraints of the model, being supported at the unit simplex. We consider a flat Dirichlet distribution:

$$p(\boldsymbol{a}) = \mathrm{Dir}(\boldsymbol{a}; \mathbf{1}_P) \,, \tag{2}$$

where $\mathbf{1}_P$ is a vector of ones which contains its concentration parameters. The flatness indicates the lack of a priori knowledge over the abundances other than its physical constraints.

Endmember model: Recently proposed EM models include the use of additive [15] or multiplicative [16–18] perturbations of reference EM signatures, or a combination of both [19]. Nonetheless, EM variability can be complex and specifying a probability distribution function (PDF) $p(\boldsymbol{M})$ analytically is difficult. In this work we consider a deep generative EM model to provide a flexible representation of spectral variability while accounting for their low intrinsic dimension [9]. We model $\boldsymbol{M}$ as a random variable which follows a Gaussian distribution when conditioned on latent variables $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_P] \in \mathbb{R}^{H \times P}$ of dimension $H \ll L$ (i.e., $p(\boldsymbol{M}|\boldsymbol{Z})$ is Gaussian). The variable $\boldsymbol{Z}$ control the variability of the EMs. Thus, although the conditional distribution is tractable, the marginal PDF $p(\boldsymbol{M}) = \int p(\boldsymbol{M}|\boldsymbol{Z}) p(\boldsymbol{Z}) d\boldsymbol{Z}$ can be arbitrarily flexible. Considering the EMs to be conditionally independent leads to:

$$p_\theta(\boldsymbol{M}|\boldsymbol{Z}) = \prod_{k=1}^{P} p_\theta(\boldsymbol{m}_k|\boldsymbol{z}_k) \,, \tag{3}$$

with $p_\theta(\boldsymbol{m}_k|\boldsymbol{z}_k) = \mathcal{N}\big(\boldsymbol{m}_k; \boldsymbol{\mu}_\theta^{m,k}(\boldsymbol{z}_k), \mathrm{diag}(\boldsymbol{\sigma}_\theta^{m,k}(\boldsymbol{z}_k))\big)$, where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian PDF in $\boldsymbol{x}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{m}_k$ is the k-th column of $\boldsymbol{M}$. Functions $\boldsymbol{\mu}_\theta^{m,k}(\boldsymbol{z}_k)$ and $\boldsymbol{\sigma}_\theta^{m,k}(\boldsymbol{z}_k)$ return the mean and the diagonal of the covariance matrix of the PDF, and θ contains all parameters of the mixing model. We assign the following prior for $p(\boldsymbol{Z})$:

$$p(\boldsymbol{Z}) = \prod_{k=1}^{P} p(\boldsymbol{z}_k) \,, \quad p(\boldsymbol{z}_k) = \mathcal{N}(\boldsymbol{z}_k; \mathbf{0}, \boldsymbol{I}) \,. \tag{4}$$

Mixing model: The complexity of the mixing process motivated the development of HU method based on nonparametric models where the nonlinearity is learned from the data, such as in kernel-based methods [7, 20] and nonlinear AEC networks [5, 8, 12]. In this work, we consider a decomposition of the nonlinear mixing process as the sum of a linear contribution (i.e., the LMM) and a nonparametric nonlinear function of $\boldsymbol{a}$ and $\boldsymbol{M}$ [5, 7, 21]:

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{a} + \boldsymbol{\mu}_\theta^y(\boldsymbol{a}, \boldsymbol{M}) + \boldsymbol{e} \,, \tag{5}$$

where function $\boldsymbol{\mu}_\theta^y$ denotes the contribution of nonlinear mixing in the model. Although the connection between (5) and models such as the bilinear mixing model and Hapke's model is not direct, it has the important property that the amount of nonlinearity can be explicitly controlled by penalizing the norm of $\boldsymbol{\mu}_\theta^y$ during the learning process. Considering white Gaussian noise, the likelihood becomes:

$$p_\theta(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{M}) = \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{M}\boldsymbol{a} + \boldsymbol{\mu}_\theta^y(\boldsymbol{a}, \boldsymbol{M}), \sigma_\theta^y \boldsymbol{I}\big) \,, \tag{6}$$

where σ_θ^y is the noise variance. Using independence between the abundances, EMs and noise, the joint distribution factorizes as:

$$p_\theta(\boldsymbol{y}, \boldsymbol{a}, \boldsymbol{M}, \boldsymbol{Z}) = p_\theta(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{M}) p_\theta(\boldsymbol{a}) p_\theta(\boldsymbol{M}|\boldsymbol{Z}) p(\boldsymbol{Z}) \,. \tag{7}$$

2.2. The unmixing problem

The HU problem consists of finding the posterior distribution $p(\boldsymbol{a}, \boldsymbol{M}, \boldsymbol{Z}|\boldsymbol{y})$. However, due to the choice of distributions in (7) this PDF cannot be computed analytically. Therefore, we propose to use a variational approximation to the posterior. This consists in specifying a parametric distribution $q_\phi(\boldsymbol{a}, \boldsymbol{M}, \boldsymbol{Z}|\boldsymbol{y})$ from a sufficiently flexible family, and finding the parameters ϕ that make it as close as possible to the true posterior PDF $p(\boldsymbol{a}, \boldsymbol{M}, \boldsymbol{Z}|\boldsymbol{y})$. We use the following factorization for this PDF:

$$q_\phi(\boldsymbol{a}, \boldsymbol{M}, \boldsymbol{Z}|\boldsymbol{y}) = q_\phi(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{M}, \boldsymbol{Z}) q_\phi(\boldsymbol{M}|\boldsymbol{Z}, \boldsymbol{y}) q_\phi(\boldsymbol{Z}|\boldsymbol{y}). \tag{8}$$

We additionally simplify the problem by assuming that $\boldsymbol{a}$ is independent of $\boldsymbol{Z}$ conditioned on $\boldsymbol{y}$ and $\boldsymbol{M}$, and that $\boldsymbol{M}$ is independent of $\boldsymbol{y}$ conditioned on $\boldsymbol{Z}$:

$$q_\phi(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{M}, \boldsymbol{Z}) = q_\phi(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{M}) \,, \tag{9}$$

$$q_\phi(\boldsymbol{M}|\boldsymbol{Z}, \boldsymbol{y}) = q_\phi(\boldsymbol{M}|\boldsymbol{Z}) \,. \tag{10}$$

Such assumption *disentangles* $\boldsymbol{a}$ and $\boldsymbol{M}$ from $\boldsymbol{Z}$ and from $\boldsymbol{y}$, respectively. In the following, we define specific forms for each of those factors. We consider $q_\phi(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{M})$ as a Dirichlet distribution:

$$q_\phi(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{M}) = \mathrm{Dir}(\boldsymbol{a}; \breve{\boldsymbol{\gamma}}_\phi^a(\boldsymbol{y}, \boldsymbol{M})) \,, \tag{11}$$

where function $\breve{\boldsymbol{\gamma}}_\phi^a$ computes its concentration parameters.

Many HU works consider black-box models to compute the abundances from the mixed pixels. On the other hand, it is important to introduce a priori information about the model to help interpret the results when defining the functional form of the posterior PDF of the abundances in (11). In this case, it was shown in [5] that for AEC networks the amount of nonlinearity in the mixing and inference models are closely related. Thus, it is helpful to split the encoder network into linear and nonlinear parts, in such a way that the amount of nonlinearity in both the encoder and decoder can be adjusted. Thus, $\breve{\boldsymbol{\gamma}}_\phi^a(\boldsymbol{y}, \boldsymbol{M})$ can be written as:

$$\breve{\boldsymbol{\gamma}}_\phi^a(\boldsymbol{y}, \boldsymbol{M}) = s_{\mathrm{ReLU}}\left(\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{lin}}(\boldsymbol{y}, \boldsymbol{M}) + \breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{nlin}}(\boldsymbol{y}, \boldsymbol{M})\right) \,, \tag{12}$$

where $\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{lin}}(\boldsymbol{y},\boldsymbol{M})$ is a piecewise-linear function that estimates the abundance concentration parameters using a model-inspired NN architecture, while $\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{nlin}}(\boldsymbol{y},\boldsymbol{M})$ is a deep NN that can compensate nonlinearities in the model. $s_{\mathrm{ReLU}}(\boldsymbol{x}) = \max(\boldsymbol{0},\boldsymbol{x})$ represents the ReLU activation, which ensures the nonnegativity of $\breve{\boldsymbol{\gamma}}_\phi^{a}(\boldsymbol{y},\boldsymbol{M})$.

We assume distribution $q_\phi(\boldsymbol{Z}|\boldsymbol{y})$ to factorize as $q_\phi(\boldsymbol{Z}|\boldsymbol{y}) = \prod_{k=1}^{P} q_\phi(\boldsymbol{z}_k|\boldsymbol{y})$, where each $q_\phi(\boldsymbol{z}_k|\boldsymbol{y})$ is a Gaussian PDF:

$$q_\phi(\boldsymbol{z}_k|\boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{z}_k; \breve{\boldsymbol{\mu}}_\phi^{z,k}(\boldsymbol{y}), \mathrm{diag}(\breve{\boldsymbol{\sigma}}_\phi^{z,k}(\boldsymbol{y}))\big)\,, \tag{13}$$

in which functions $\breve{\boldsymbol{\mu}}_\phi^{z,k}$ and $\breve{\boldsymbol{\sigma}}_\phi^{z,k}$ compute its mean and the elements of its diagonal covariance matrix, respectively.

We also assume that $q_\phi(\boldsymbol{M}|\boldsymbol{Z})$ can be factorized as $q_\phi(\boldsymbol{M}|\boldsymbol{Z}) = \prod_{k=1}^{P} q_\phi(\boldsymbol{m}_k|\boldsymbol{z}_k)$, with:

$$q_\phi(\boldsymbol{m}_k|\boldsymbol{z}_k) = \mathcal{N}\big(\boldsymbol{m}_k; \breve{\boldsymbol{\mu}}_\phi^{m,k}(\boldsymbol{z}_k), \mathrm{diag}(\breve{\boldsymbol{\sigma}}_\phi^{m,k}(\boldsymbol{z}_k))\big)\,. \tag{14}$$

where $\breve{\boldsymbol{\mu}}_\phi^{m,k}$ and $\breve{\boldsymbol{\sigma}}_\phi^{m,k}$ are functions that compute the mean and the elements of the diagonal covariance matrix of this PDF, respectively. Note that to reduce the freedom of the model and simplify the inference process, we consider conditional EM distributions in the generative and inference models to be equal, i.e., $q_\phi(\boldsymbol{M}|\boldsymbol{Z}) = p_\theta(\boldsymbol{M}|\boldsymbol{Z})$.

2.3. Objective function

To learn the parameters of the model and perform unmixing, we maximize a cost function composed of three terms, defined as:

$$\mathcal{L}_{\mathrm{T}}(\theta,\phi;\mathcal{D}) = \mathcal{L}_{\mathrm{Data}}(\mathcal{D}) - \mathcal{L}_{\mathrm{Sparse}}(\mathcal{D}) - \mathcal{R}\,. \tag{15}$$

The first term aims to maximize a regularized lower bound to the log-likelihood of the labelled and unlabelled dataset $\log p(\mathcal{D})$, which is given by [14, 22]:

$$\mathcal{L}_{\mathrm{Data}}(\mathcal{D}) = L_U(\mathcal{D}_U) + \lambda\Big[L_S(\mathcal{D}_S) + \beta \sum_{(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})\in\mathcal{D}_S} \log q_\phi(\boldsymbol{a},\boldsymbol{M}|\boldsymbol{y})\Big]\,,$$

where $L_U(\mathcal{D}_U)$ and $L_S(\mathcal{D}_S)$ are the evidence lower bounds (ELBOs) to $\sum_{\boldsymbol{y}\in\mathcal{D}_U} \log p(\boldsymbol{y})$ and $\sum_{(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})\in\mathcal{D}_S} \log p(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})$, respectively, and are given by [14, 22]:

$$L_U(\mathcal{D}_U) = \sum_{\boldsymbol{y}\in\mathcal{D}_U} \mathbb{E}_{q_\phi(\boldsymbol{a},\boldsymbol{M},\boldsymbol{Z}|\boldsymbol{y})}\Big\{\log \frac{p_\theta(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M},\boldsymbol{Z})}{q_\phi(\boldsymbol{a},\boldsymbol{M},\boldsymbol{Z}|\boldsymbol{y})}\Big\}\,, \tag{16}$$

$$L_S(\mathcal{D}_S) = \sum_{(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})\in\mathcal{D}_S} \mathbb{E}_{q_\phi(\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})}\Big\{\log \frac{p_\theta(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M},\boldsymbol{Z})}{q_\phi(\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})}\Big\}\,. \tag{17}$$

Parameter λ balances the contribution of labelled and unlabelled data in the cost function, and β weights an additional regularization term that maximizes the likelihood of the posterior of $\boldsymbol{a}$ and $\boldsymbol{M}$ [22].

The second term is a regularization that promotes sparsity of the estimated abundances by penalizing the concentration parameters of the abundance posterior distribution (11) using the $\mathrm{L}_{1/2}$ norm:

$$\begin{aligned}\mathcal{L}_{\mathrm{Sparse}}(\mathcal{D}) = \tau \sum_{(\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})\in\mathcal{D}_S} \big\|\breve{\boldsymbol{\gamma}}_\phi^{a}(\boldsymbol{y},\boldsymbol{M})\big\|_{1/2} \\ +\tau \sum_{\boldsymbol{y}\in\mathcal{D}_U} \mathbb{E}_{q_\phi(\boldsymbol{M}|\boldsymbol{Z})q_\phi(\boldsymbol{Z}|\boldsymbol{y})}\Big\{\|\breve{\boldsymbol{\gamma}}_\phi^{a}(\boldsymbol{y},\boldsymbol{M})\|_{1/2}\Big\}\,,\end{aligned} \tag{18}$$

where τ is a regularization parameter. The $\mathrm{L}_{1/2}$ norm has shown good performance to promote abundance sparsity in HU [23]. The last term is a regularization on the model parameters:

$$\mathcal{R} = \varsigma_1\big\|\boldsymbol{\mu}_\theta^{y}\big\| + \varsigma_2\big\|\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{nlin}}\big\|\,, \tag{19}$$

where ς_1 and ς_2 are regularization parameters and the norm depends on the parameters in θ and ϕ.

In order to optimize the cost function $\mathcal{L}_{\mathrm{T}}(\theta,\phi;\mathcal{D})$, there are a few challenges related to the presence of the expectations, and of PDFs which are not directly accessible (e.g., $q_\phi(\boldsymbol{Z}|\boldsymbol{y},\boldsymbol{a},\boldsymbol{M})$). To proceed, we first use self-normalized importance sampling and Jensen's inequality to write all terms of $\mathcal{L}_{\mathrm{T}}(\theta,\phi;\mathcal{D})$ as a function of the PDFs available in the right-hand-side of the selected factorizations in (7) and (8). Then, the reparametrization trick [24] is used to write the expectations in terms of random variables that do not depend on θ and ϕ, and Monte Carlo sampling is used to approximate them. Due to space limitations, we omit details of these derivations, which will be available in an extended version of this work.

2.4. Neural network architectures

We parametrize the models as follows. All NNs use the ReLU activation function in the hidden layers and linear output activations, unless noted otherwise. $\breve{\boldsymbol{\mu}}_\phi^{m,k}(\boldsymbol{z}_k) = \boldsymbol{\mu}_\theta^{m,k}(\boldsymbol{z}_k)$ are represented using a fully connected multilayer perceptron (MLP) with the same architecture as in [9]. The EM covariances were set as $\breve{\boldsymbol{\sigma}}_\phi^{m,k}(\boldsymbol{z}_k) = \boldsymbol{\sigma}_\theta^{m,k}(\boldsymbol{z}_k) = \sigma^{m,k}\boldsymbol{1}_L$, with $\sigma^{m,k}$ being a learnable constant and $\boldsymbol{1}_L \in \mathbb{R}^L$ a vector of ones. For $\breve{\boldsymbol{\mu}}_\phi^{z,k}(\boldsymbol{y})$, we considered an MLP with four fully connected layers with L, $5H$, $2H$ and H neurons, whereas for $\breve{\boldsymbol{\sigma}}_\phi^{z,k}(\boldsymbol{y})$, we considered six layers with L, $5H$, $2H$, $2H$, $2H$ and H neurons. The parameters of the first two hidden layers of $\breve{\boldsymbol{\mu}}_\phi^{z,k}(\boldsymbol{y})$ and $\breve{\boldsymbol{\sigma}}_\phi^{z,k}(\boldsymbol{y})$ are shared. The nonlinear contribution in the mixing model $\boldsymbol{\mu}_\theta^{y}(\boldsymbol{a},\boldsymbol{M})$ was a fully connected MLP with the same architecture as in [5]. The noise variance σ_θ^{y} was set as a positive learnable constant. The piecewise linear abundance NN $\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{lin}}(\boldsymbol{y},\boldsymbol{M})$ was selected by unrolling a sparse regression architecture with 11 layers [10]. The nonlinear part $\breve{\boldsymbol{\gamma}}_\phi^{a,\mathrm{nlin}}(\boldsymbol{y},\boldsymbol{M})$ was a fully connected MLP with L, $2L$, $\lfloor L/2\rceil$, $\lfloor L/4\rceil$, $4P$ and P neurons, where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer.

2.5. Self-supervised training strategy

To obtain supervised training data $\mathcal{D}_{\mathrm{S}}$ for the semi-supervised framework, we propose to use a self-supervised strategy based on methods which extract different samples from endmembers directly from an observed HI (i.e., in the form of pure pixels) as done in [3, 25].

In this paper, we consider a two step procedure. In the first step, we create a dictionary $\mathcal{D}_{\mathrm{ppx}}$ of pure pixels (i.e., EM samples) extracted from the HI as described in [9]. The second step generates synthetic data incorporating the variability in $\mathcal{D}_{\mathrm{ppx}}$. At each iteration k we generate tuples $\{(\boldsymbol{y}_j,\boldsymbol{a}_j,\boldsymbol{M}_k)\}_{j=1}^{P}$ where $\boldsymbol{M}_k \in \mathbb{R}^{L\times P}$ is an EM matrix sampled from the pure pixel dictionary. For each k we generate P abundance vectors $\boldsymbol{a}_j$, $j \in \{1,\ldots,P\}$, with the elements $a_{j,i}$ of the j-th vector satisfying $a_{j,i} = 1$ if $i = p$ and $a_{j,i} = 0$ otherwise (i.e., a one-hot encoding of each EM). The mixed pixels are generated according to $\boldsymbol{y}_j = \boldsymbol{M}_k\boldsymbol{a}_j + \boldsymbol{e}$, with $\boldsymbol{e}$ being white Gaussian noise with a signal-to-noise ratio (SNR) of ϑ.

3. EXPERIMENTS

We compared PhI-Net with the FCLS (with EMs extracted using the VCA method [26]), the PLMM [15], and the GLMM [17], which account for EM variability, and with recent deep learning-based strategies that address both EM variability and nonlinearity, namely, DeepGUn [9], and RBF-AEC [27]. For best performance, the parameters of all methods were adjusted for each experiment.

For PhI-Net, we optimize the cost function using the Adam optimizer for up to 30 epochs, with a batch size of 16, and initial

Table 1. Simulation results with synthetic data.

	NRMSE$_{\boldsymbol{A}}$	NRMSE$_{\boldsymbol{M}}$	SAM$_{\boldsymbol{M}}$	NRMSE$_{\boldsymbol{Y}}$	Time
Data Cube 1 – DC1					
FCLS	0.3706	–	–	0.0931	1.2
PLMM	0.3370	0.1656	0.0885	0.0349	842.9
GLMM	0.3453	0.0851	0.0319	0.0505	91.0
DeepGUn	0.2460	0.0681	0.0201	0.0582	607.4
RBF-AEC	0.4582	0.1289	0.0670	0.0681	82.0
PhI-Net	0.1697	0.0398	0.0162	0.0420	436.2
Data Cube 2 – DC2					
FCLS	0.5109	–	–	0.2746	1.1
PLMM	0.5066	0.6245	0.4874	0.0483	1050.6
GLMM	0.4371	0.4855	0.1972	0.1976	64.5
DeepGUn	0.2399	0.3072	0.0914	0.1193	678.5
RBF-AEC	0.5138	0.2919	0.1338	0.1450	78.3
PhI-Net	0.1910	0.3133	0.0838	0.1593	725.5

learning rate of 0.001 decreased by 10% per epoch until the 10-th epoch. The labeled dataset $\mathcal{D}_S$ was constructed using the described self-supervised strategy, with VCA [26] being used to extract 100 pure pixels to build $\mathcal{D}_{\text{ppx}}$, and $\vartheta = 30$dB. We set the EM latent space dimension as $H = 2$. The regularization parameters were selected within the intervals $\lambda \in [0.01, 1000]$, τ was either as zero or in $\tau \in [0.001, 1]$, and $\varsigma_1, \varsigma_2 \in [10^{-4}, 10^5]$. The performance of the algorithms was evaluated through the normalized root mean squared error (NRMSE), defined as $\text{NRMSE}_{\boldsymbol{X}} = \|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F / \|\boldsymbol{X}\|_F$ computed between a matrix $\boldsymbol{X}$ and its estimated version $\widehat{\boldsymbol{X}}$. We also compute the spectral angle mapper (SAM) between the true and the estimated EMs at each pixel: $\text{SAM}_{\boldsymbol{M}} = \frac{1}{N_U}\sum_{n=1}^{N_U}\sum_{j=1}^{P} \arccos(\boldsymbol{m}_{n,j}^\top \widehat{\boldsymbol{m}}_{n,j} / \|\boldsymbol{m}_{n,j}\| \|\widehat{\boldsymbol{m}}_{n,j}\|)$, where $\boldsymbol{m}_{n,j}$ and $\widehat{\boldsymbol{m}}_{n,j}$ are the true and estimated signatures of the j-th EM in the n-th pixel, respectively.

Synthetic data: We created two synthetic data cubes, DC1 and DC2, simulating nonlinear effects and EM variability, respectively. To generate the first datacube (DC1), with $N_U = 2500$ pixels, we considered synthetic abundance maps with $P = 3$ EM signatures with $L = 224$ spectral bands selected from the USGS Spectral Library. The reflectance of each pixel was generated according to the bilinear mixing model $\boldsymbol{y}_n = \boldsymbol{M}\boldsymbol{a}_n + \sum_{i=1}^{P}\sum_{j=i+1}^{P} a_{n,i}a_{n,j}\boldsymbol{m}_i \odot \boldsymbol{m}_j + \boldsymbol{e}_n$, with $\boldsymbol{e}_n$ being white Gaussian noise with an SNR of 30dB. To generate the second datacube (DC2), with $N_U = 2500$ pixels, we considered synthetic abundance maps and to incorporate EM variability in this dataset, sets of EM signatures from $P = 5$ pure materials (roof, metal, dirt, tree and asphalt) with realistic variability were first manually extracted from a real HI. Then, to generate each pixel $\boldsymbol{y}_n$, spectral signatures for each EM were then randomly sampled from these sets and used in a generalized version of the LMM $\boldsymbol{y}_n = \boldsymbol{M}_n\boldsymbol{a}_n + \boldsymbol{e}_n$ with pixelwise EM matrices, where $\boldsymbol{e}_n$ was white Gaussian noise with an SNR of 30dB.

The quantitative results for both data cubes are presented in Table 1 (visual results are omitted due to space limitations). It can be seen that the PhI-Net obtained the best NRMSE$_{\boldsymbol{A}}$ for both data cubes, with improvements of 31% and 20% over the results given by DeepGUn. In terms of EM estimation performance, PhI-Net obtained the best SAM$_{\boldsymbol{M}}$ for both datacubes, and the best NRMSE$_{\boldsymbol{M}}$ for DC1, with RBF-AEC obtaining the best results for DC2. RBF-AEC did not obtain quantitatively good abundance reconstructions, which is likely due to its tendency to mark most pixels as being pure that will be illustrated in the next example. PLMM obtained the smallest NRMSE$_{\boldsymbol{Y}}$ due to its large flexibility to represent the HI pixels. The execution times of PhI-Net were comparable to those of the PLMM and DeepGUn methods.

Table 2. Simulation results with the Samson dataset.

	FCLS	PLMM	GLMM	DeepGUn	RBF-AEC	PhI-Net
NRMSE$_{\boldsymbol{Y}}$	0.0687	0.0239	0.0011	0.0750	0.2062	0.0569
Time	2.45	225.92	96.78	676.43	230.91	1606.34

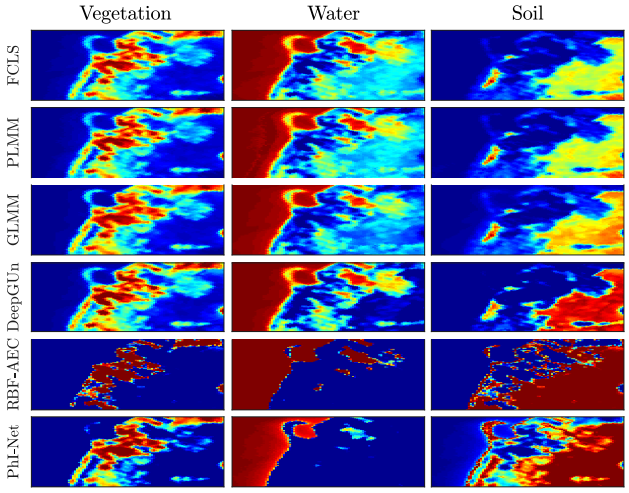

Fig. 1. Estimated abundances for the Samson HI.

Real data: To evaluate the performance of the methods on real data, we considered the Samson HI. This HI was acquired by AVIRIS, which captures 224 spectral bands. Water absorption regions and bands with low SNR were removed, resulting in $L = 156$ bands. The Samson HI was previously shown to have $P = 3$ EMs [9]. The estimated abundances are shown in Figure 1, and the quantitative are show in Table 2.

It can be seen that the methods based on deep learning (DeepGUn, RBF-AEC and PhI-Net) provide a better separation between the different EMs in the HI when compared to the other methods, which is observed more clearly for the Soil EM. Moreover, PhI-Net obtained better abundance reconstructions in regions with more heavily mixed EMs when compared to RBF-AEC, for which most pixels are assigned only a single material. The quantitative results in Table 2 show that the GLMM obtained the smallest reconstruction error, which is expected due to its large number of degrees of freedom, whereas PhI-Net obtained intermediate results. Note, however, that small NRMSE$_{\boldsymbol{Y}}$ does not necessarily imply accurate abundance reconstructions. The execution time of PhI-Net was larger than those of the other algorithms, but still comparable with that of more complex methods such as DeepGUn. Future work will explore more efficient inference algorithms to reduce the complexity of HU.

4. CONCLUSIONS

In this paper we proposed a deep disentangled variational inference framework for HU that accounts for both nonlinearity and EM variability. To cope with the unlabeled nature of HI datasets, we employed a self-supervised learning strategy, allowing for the exploitation of semi-supervised learning algorithms. Furthermore, we disentangled endmembers and abundances by assuming conditional independence on the variational posterior. Experiments with synthetic and real data show that the proposed PhI-Net leads to more accurate abundance and EM estimates compared to state-of-the-art methods.

5. REFERENCES

[1] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. G., and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[2] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: Models and algorithms," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 82–94, Jan 2014.

[3] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, C. Richard, J. Chanussot, L. Drumetz, J.-Y. Tourneret, A. Zare, and C. Jutten, "Spectral variability in hyperspectral data unmixing: A comprehensive review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 223–270, 2021.

[4] B. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Blind hyperspectral unmixing using autoencoders: A critical comparison," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1340–1372, 2022.

[5] H. Li, R. A. Borsoi, T. Imbiriba, P. Closas, J. C. Bermudez, and D. Erdoğmuş, "Model-based deep autoencoder networks for nonlinear hyperspectral unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[6] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, J. Chanussot, and X. Zhu, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 52–87, 2021.

[7] J. Chen, C. Richard, and P. Honeine, "Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model," *IEEE Transactions on Signal Processing*, vol. 61, pp. 480–492, Jan 2013.

[8] D. Hong, L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, and B. Zhang, "Endmember-guided unmixing network (EGU-Net): A general deep learning framework for self-supervised hyperspectral unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[9] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Deep generative endmember modeling: An application to unsupervised spectral unmixing," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 374–384, 2019.

[10] Y. Qian, F. Xiong, Q. Qian, and J. Zhou, "Spectral mixture model inspired network architectures for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7418–7434, 2020.

[11] B. Palsson, J. Sigurdsson, J. R. Sveinsson, and M. O. Ulfarsson, "Hyperspectral unmixing using a neural network autoencoder," *IEEE Access*, vol. 6, pp. 25 646–25 656, 2018.

[12] M. Wang, M. Zhao, J. Chen, and S. Rahardja, "Nonlinear unmixing of hyperspectral data via deep autoencoder networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1467–1471, 2019.

[13] S. Shi, M. Zhao, L. Zhang, Y. Altmann, and J. Chen, "Probabilistic generative model for hyperspectral unmixing accounting for endmember variability," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[14] N. Siddharth, B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] P.-A. Thouvenin, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 525–538, Feb. 2016.

[16] L. Drumetz, M.-A. Veganzones, S. Henrot, R. Phlypo, J. Chanussot, and C. Jutten, "Blind hyperspectral unmixing using an extended linear mixing model to address spectral variability," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3890–3905, 2016.

[17] T. Imbiriba, R. A. Borsoi, and J. C. M. Bermudez, "Generalized linear mixing model accounting for endmember variability," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 1862–1866.

[18] R. A. Borsoi, T. Imbiriba, P. Closas, J. C. M. Bermudez, and C. Richard, "Kalman filtering and expectation maximization for multitemporal spectral unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[19] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2019.

[20] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, and C. Richard, "A blind multiscale spatial regularization framework for kernel-based spectral unmixing," *IEEE Transactions on Image Processing*, vol. 29, pp. 4965–4979, 2020.

[21] M. Zhao, M. Wang, J. Chen, and S. Rahardja, "Hyperspectral unmixing via deep autoencoder networks for a generalized linear-mixture/nonlinear-fluctuation model," *arXiv preprint: 1904.13017*, 2019.

[22] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. 27th International Conference on Neural Information Processing Systems-Volume 2*, 2014, pp. 3581–3589.

[23] Y. Qian, S. Jia, J. Zhou, and A. Robles-Kelly, "Hyperspectral unmixing via $L_{1/2}$ sparsity-constrained nonnegative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4282–4297, 2011.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. 2nd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., Banff, AB, Canada, 2014.

[25] B. Somers, M. Zortea, A. Plaza, and G. P. Asner, "Automated extraction of image-based endmember bundles for improved spectral unmixing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 396–408, 2012.

[26] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex Component Analysis: A fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, April 2005.

[27] K. T. Shahid and I. D. Schizas, "Unsupervised hyperspectral unmixing via nonlinear autoencoders," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.