IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10642450

# NONLINEAR UNMIXING OF HYPERSPECTRAL IMAGES VIA REGULARIZED WASSERSTEIN DICTIONARY LEARNING

*Scott Fullenbaum, Marshall Mueller, Abiy Tasissa, James M. Murphy*

Tufts University, Department of Mathematics
Medford, MA 02155, USA

## ABSTRACT

Hyperspectral images consist of large numbers of pixels across hundreds of spectral bands, making statistical analysis computationally challenging. However, these images often exhibit intrinsic structure that can be leveraged for efficient statistical and machine learning. We propose a novel nonlinear method for unmixing hyperspectral images. In contrast to classical methods which consider an additive linear model, we propose to represent hyperspectral spectra as probability distributions in Wasserstein space and characterize pure spectra as those that allow for typical observations to be reconstructed as entropic Wasserstein barycenters. This allows for the analysis and synthesis of hyperspectral spectra in a geometry-preserving fashion. Results on synthetic data and real HSI show important geometric features of hyperspectral spectra are preserved when utilizing our nonlinear Wasserstein unmixing scheme.

***Index Terms*—** Hyperspectral images, unmixing, optimal transport, Wasserstein space

## 1. INTRODUCTION

Hyperspectral images (HSI) record reflectance across a range of electromagnetic wavelengths, providing a powerful, high spectral resolution characterization of imaged scenes. While its rich information is useful for a range of remote sensing tasks (e.g., land cover classification [1, 2], land change detection [3], precision agriculture [4], and anomaly detection [5, 6]), HSI data is very high-dimensional. Typically hundreds of spectral reflectance ranges are recorded for millions of pixels. The curse of dimensionality for HSI challenges naive statistical and machine learning methods and demands approaches that leverage intrinsic low-dimensional structures in HSI [7].

Dimension reduction methods such as principal component analysis (PCA) [8], non-negative matrix factorization (NMF) [9, 10], and sparse dictionary learning [11] have proven useful for reducing the dimension of HSI while retaining properties that allow for downstream applications such as material labeling or image segmentation. A related notion in HSI analysis is that of *unmixing*, which models individual pixels in HSI as consisting of reflectances from multiple "pure" material spectra, due to the low spatial resolution of typical HSI sensors (e.g., on the order of tens of square meters). Important in its own right for understanding the context of a scene captured by an HSI, unmixing also provides useful features for classification and segmentation methods [12]. HSI unmixing not only provides useful information about the constituent material contents of an individual pixel (when interpreting the learned spectra as pure materials) but also provides an efficient parameterization of the data when the number of learned pure spectra is smaller than the number of spectral bands in the full HSI.

We approach the problem of HSI unmixing via a novel dictionary learning method based on *optimal transport* [13]. While most HSI unmixing methods posit an additive mixture model, we consider a nonlinear mixture model based on entropic Wasserstein barycenters, which preserves the intrinsic geometry of the spectral signatures more effectively than existing methods. We apply our novel geometric Wasserstein dictionary learning scheme to synthetic and real hyperspectral spectra to demonstrate the efficient and interpretable components learned by our method.

The rest of this paper is organized as follows. Section 2 provides necessary background on unmixing and optimal transport. Section 3 details our Wasserstein Hyperspectral Unmixing (WaHU) algorithm, which is then validated on synthetic and real HSI data in Section 4. We conclude and discuss future work in Section 5.

## 2. BACKGROUND

Let $\{\mu_i\}_{i=1}^n \subset \mathbb{R}^D$ denote a collection of hyperspectral pixels with $n$ being the total number of pixels imaged and $D$ the number of spectral bands. Typical approaches to HSI unmixing attempt to learn pure spectra — which we will refer to as atoms in what follows — $\{z_j\}_{j=1}^m \subset \mathbb{R}^D$ and mixture weights $\{w_i\}_{i=1}^n \subset \mathbb{R}^m$ such that $\frac{1}{n}\sum_{i=1}^n \|\mu_i - \sum_{j=1}^m [w_i]_j z_j\|_2$ is small (i.e., a typical pixel is well-reconstructed by a linear combination of the learned atoms) where $m \ll n$ and $\|\cdot\|_2$

This research was partially supported by the NSF through grants DMS 1924513, DMS-2208392, DMS-2309519, and DMS-2318894.

denotes the $\ell^2$ norm. Various constraints on the atoms (e.g., non-negativity; lie on an $D$-simplex) and weights (e.g., non-negativity; lie on an $m$-simplex; small $\ell^p$ norm; sparsity) may be imposed to generate interpretable and useful atoms and weights. While well-studied [14, 15], approaches to HSI unmixing based on a linear reconstruction model (i.e., approximating true pixel $\mu_i$ via a linear combination of the $\{z_j\}_{j=1}^m$) may fail to capture intrinsic geometry for even simple datasets. See for example Figure 2 (b) and (c) for an illustration of how two linear approaches for unmixing — principal component analysis (PCA) and nonnegative matrix factorization (NMF) — may fail to efficiently represent a family of translated and rescaled Gaussians.

Our approach abandons the linear unmixing model and considers a novel approach to HSI unmixing based on *optimal transport* [13] between probability measures. Let $\Delta^D := \{(x_1, \ldots, x_D) \in \mathbb{R}^D \mid \sum_{k=1}^D x_k = 1, x_k \geq 0, \ \forall k\}$ denote the probability simplex in $\mathbb{R}^D$. For two probability distributions $\mu, \nu \in \Delta^D$, let $\Pi(\mu, \nu) \subset \Delta^{D\times D}$ denote the space of *couplings* between $\mu$ and $\nu$, namely the non-negative matrices $\pi \in \mathbb{R}^{D\times D}$ such that $\forall j$, $\sum_{i=1}^D \pi_{ij} = \nu_j$, and $\forall i$, $\sum_{j=1}^D \pi_{ij} = \mu_i$. We suppose $\mu, \nu$ are supported on a common set of points $\{b_k\}_{k=1}^D \subset \mathbb{R}$. For $p \geq 1$, the *entropic* $p$*-Wasserstein distance* [16] between $\mu$ and $\nu$ is

$$W^p_{p,\epsilon}(\nu, \mu) := \min_{\pi\in\Pi(\mu,\nu)} \sum_{k=1}^{D}\sum_{\ell=1}^{D} (\pi_{k\ell}|b_k - b_\ell|^p + \epsilon \pi_{k\ell} \log(\pi_{k\ell})), \tag{1}$$

where $\epsilon > 0$ is a regularization parameter. At an intuitive level, the solution $\pi^*$ to (1) transports the mass in distribution $\mu$ with that in $\nu$ in a distance-minimizing manner (first term) while ensuring the mass is smoothly distributed (second term); see [13] and [17] for a thorough overview of the computational and theoretical aspects of entropic optimal transport, respectively.

We can now define a notion of *averaging with respect to entropic* $p$*-Wasserstein distance*, namely entropic Wasserstein barycenters [18, 19, 20, 21]. For a set of $m$ probability distributions $\{\nu_j\}_{j=1}^m \subset \Delta^D$ and a vector of weights $w \in \Delta^m$, we define the entropic $p$-Wasserstein barycenter to be

$$\text{Bary}(\{\nu_j\}_{j=1}^m; w) := \underset{\mu\in\Delta^D}{\arg\min} \sum_{j=1}^{m} w_j W^p_{p,\epsilon}(\mu, \nu_j). \tag{2}$$

Figure 1 shows $m = 2$ Gaussians $\nu_1 \sim \mathcal{N}(50, 5)$, $\nu_2 \sim \mathcal{N}(130, 10)$ as well as a family of linear mixtures $(1 - t)\nu_1 + t\nu_2$ and a family of entropic 2-Wasserstein barycenters corresponding to weights $(1 - t, t) \in \Delta^2$ for $t \in \{0, .02, .04, \ldots, .98, 1\}$. Unlike the linear mixtures, the barycenters smoothly deform from one Gaussian to another. In fact, the associated path of probability measures is related to geodesic paths in the space of probability measures [22].

Given observed probability distributions $\{\mu_i\}_{i=1}^n$ (interpreted in our context as HSI pixels after normalization so they lie in $\Delta^D$), we can perform unmixing via a learning scheme that finds probability distributions $\{\nu_j\}_{j=1}^m$ and weights $\{w_i\}_{i=1}^n \subset \Delta^m$ such that each data point $\mu_i$ is close to an entropic barycenter with reference measures $\{\nu_j\}_{j=1}^m$ and weights $w_i$. Specifically, we solve the following regularized Wasserstein dictionary learning problem [23, 24]:

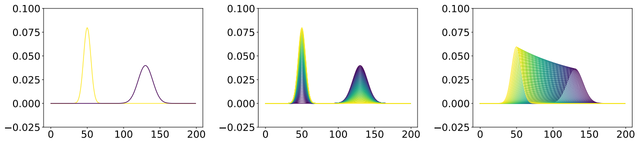


**Fig. 1**. *Left*: two Gaussian distributions, represented by their densities. *Middle*: linear mixtures of these two Gaussians, showing bi-modal behavior. *Right:* Barycenters of the two Gaussians, showing geometry preservation; in particular, entropic 2-Wasserstein barycenters of Gaussians remain Gaussian [20].

$$\begin{aligned}
&(\{\nu_j^*\}_{j=1}^m, \{w_i^*\}_{i=1}^n) \\
&= \underset{\{\nu_j\}_{j=1}^m, \{w_i\}_{i=1}^n}{\arg\min} \sum_{i=1}^{n} W^p_{p,\epsilon}(\text{Bary}(\{\nu_j\}_{j=1}^m; w_i), \mu_i) \qquad (3) \\
&+\rho \sum_{i=1}^{n}\sum_{j=1}^{m} [w_i]_j W^p_{p,\epsilon}(\mu_i, \nu_j),
\end{aligned}$$

where $\rho > 0$ is a regularization parameter. This formulation generalizes linear unmixing models by (i) replacing linear reconstruction with entropic Wasserstein barycenter reconstruction and (ii) penalizing the use of non-local atoms in the reconstructions. A linear formulation of (3) was considered in [11] and shown to be effective for unsupervised clustering of HSI.

The non-convex optimization problem (3) tries to reconstruct each observed data point well as a barycenter (first term), subject to a locality regularizer that promotes representing using nearby atoms (second term). This program can be approximately optimized using first-order methods that jointly optimize the atoms and weights; we refer to [23] and [24] for details. Our focus is on learning meaningful atoms for HSI unmixing via solving (3) as described in Section 3. We note that our approach differs from existing approaches to HSI unmixing that leverage entropic optimal transport [25], in that our synthesis model for combining atoms is non-linear, based on entropic Wasserstein barycenters.

## 3. THE WASSERSTEIN HYPERSPECTRAL UNMIXING (WAHU) ALGORITHM

We consider HSI pixels as probability distributions and solve (3) to learn generators that reconstruct the observed data well under the Wasserstein barycenter synthesis model. The approach is laid out in Algorithm 1. The key idea for approximately solving (3) is to consider the loss function

$$\begin{aligned}&\mathcal{G}(\{\nu_j\}_{j=1}^m, \{w_i\}_{i=1}^n, \{\mu_i\}_{i=1}^n) := \\ &\sum_{i=1}^n W_{p,\epsilon}^p(\mathrm{Bary}(\{\nu_j\}_{j=1}^m; w_i), \mu_i) \\ &+\rho \sum_{i=1}^n \sum_{j=1}^m [w_i]_j W_{p,\epsilon}^p(\mu_i, \nu_j),\end{aligned} \tag{4}$$

and to use automatic differentiation to iteratively update the weights $\{w_i\}_{i=1}^n$ and atoms $\{\nu_j\}_{j=1}^m$; we note that the training data $\{\mu_i\}_{i=1}^n$ is fixed in the learning process. First, we initialize the weights uniformly at random from $\Delta^m$ and the atoms with $k$-means++ [26] in Wasserstein space over the training data. From these initializations, we iteratively update via automatic differentiation on $\mathcal{G}$.

**Algorithm 1:** Wasserstein Hyperspectral Unmixing (WaHU)

```
Input: HSI spectra: $\{\mu_i\}_{i=1}^n \subset \mathbb{R}^D$; Wasserstein parameter: $p$; entropic regularization parameter: $\epsilon$; locality regularization parameter: $\rho$; # iterations: $L$; number of atoms: $m$
Normalize each pixel $\mu_i$ to lie in $\Delta^D$.
Initialize variables $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{m\times N}$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{n\times m}$.
for $k \leftarrow 1, \dots, L$ do
    $\{\nu_j^{(k)}\}_{j=1}^m \leftarrow \sigma(\boldsymbol{\alpha}^{(\mathbf{0})}), \{w_i^{(k)}\}_{i=1}^n \leftarrow \sigma(\boldsymbol{\beta}^{(\mathbf{0})})$.
    Compute the objective function
    loss $\leftarrow \mathcal{G}(\{\nu_j^{(k)}\}_{j=1}^m, \{w_i^{(k)}\}_{i=1}^n, \{\mu_i\}_{i=1}^n)$.
    Compute the gradients with automatic differentiation: loss.backward().
    Update $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}$.
end for
Output: Learned atoms: $\{\nu_j\}_{j=1}^m \leftarrow \sigma(\boldsymbol{\alpha}^{(k)})$; learned weights: $\{w_i\}_{i=1}^n \leftarrow \sigma(\boldsymbol{\beta}^{(k)})$.
```

In order to perform our optimization over arbitrary matrices $\boldsymbol{\alpha} \in \mathbb{R}^{m\times D}, \boldsymbol{\beta} \in \mathbb{R}^{n\times m}$ instead of $\{\nu_j\}_{j=1}^m, \{w_i\}_{i=1}^n$ which are constrained to be non-negative and sum-to-1, we use the softmax change of variables function $\sigma(\boldsymbol{\alpha}) = (\nu_1 \mid \nu_2 \mid \dots \mid \nu_m)^\top$ and $\sigma(\boldsymbol{\beta}) = (w_1 \mid w_2 \mid \dots \mid w_n)^\top$ where $\sigma$ acts row-wise on a matrix and acts on a vector $(x_1, x_2, \dots, x_n)$ as:

$$\sigma(x_1, \dots, x_n) := \left( \frac{\exp(x_1)}{\sum_{i=1}^n \exp(x_i)}, \dots, \frac{\exp(x_n)}{\sum_{i=1}^n \exp(x_i)} \right).$$

Code implementing Algorithm 1 using the Python OT library [27] as well as all experiments in Section 4 is publicly available[1].

## 4. EXPERIMENTS ON SYNTHETIC AND REAL HSI

To demonstrate the efficacy of WaHU for HSI unmixing, we consider two datasets: synthetic Gaussians and real Salinas A spectra. We contrast our method with two classical unmixing methods: PCA, which puts no positivity constraints and seeks only to minimize $\ell^2$ reconstruction error; and NMF, which enforces non-negativity constraints on atoms and weights.

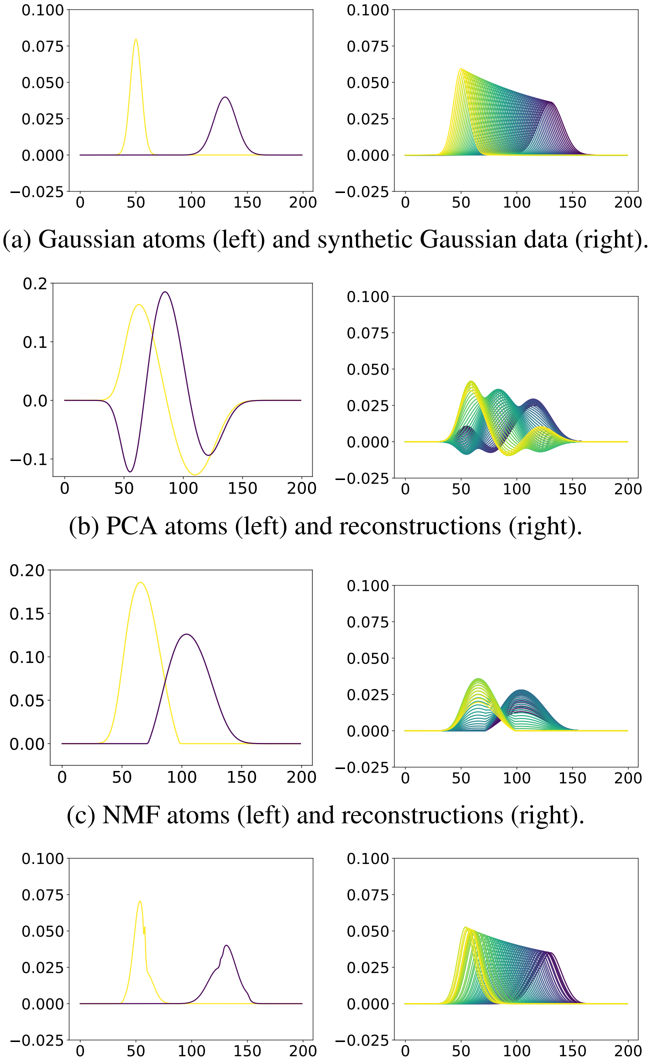
**Synthetic Gaussian Data**: We compare WaHU on synthetic Gaussian data in Figure 2. The training data is the same as in the right plot in Figure 1. As we see in Figure 2 (b), PCA learns atoms that *are not probability distributions*, owing to the lack of positivity constraints. We see in Figure 2 (c) that while NMF learns positive atoms that resemble the shapes of the true atoms (albeit they are off by a translation), the reconstructions are poor owing to the underlying additive linear reconstruction model. Such a model cannot adequately account for the particular form of intrinsic low-dimensionality in this data, namely that all observed Gaussians are smooth deformations between two reference Gaussians (more precisely, that they all lay near the geodesic between the generating reference measures in Wasserstein space). On the other hand as seen in Figure 2 (d), WaHU not only learns decent approximations to the generating atoms, but reconstructs the data faithfully. For this experiment, parameters $p = 2$, $\epsilon = .001$, $\rho = 0$, $m = 2$, and $L = 400$ were used for WaHU.

**Salinas A Data**: Salinas A is a hyperspectral image captured by the AVIRIS sensor in 1998 of an agricultural region in Salinas Valley, CA, USA. It ranges from 380-2500 nm across 224 bands. While the full scene is 83 × 86, totaling 7138 pixels, we randomly sample 1002 pixels (167 for each of the 6 labeled material classes) for our unmixing experiments. Figure 3 shows unmixing results with $m = 4$, which allows for easy visualization and contrast with PCA and NMF. As with the synthetic data, PCA fails to even preserve the non-negativity of the training data. NMF learns non-negative atoms as expected, albeit none resemble the observed training points. WaHU learns suitable atoms that indeed resemble typical observations in the data, owing to the use of the locality regularizer in (3). We note that all methods learn atoms that, with respect to their associated synthesis models (linear for PCA and NMF and nonlinear for WaHU), reconstruct the training data well, albeit PCA and NMF appear more affected by outlier pixels than WaHU. In particular, WaHU yields smoother reconstructions. In this experiment, parameters $p = 1$, $\epsilon = .05$, $\rho = .01$, $m = 4$, and $L = 400$ were used.

## 5. CONCLUSIONS AND FUTURE WORK

We established a novel unmixing paradigm for hyperspectral images based on optimal transport, specifically entropic $p$-

[1] https://github.com/fullenbs/WDL_HSI



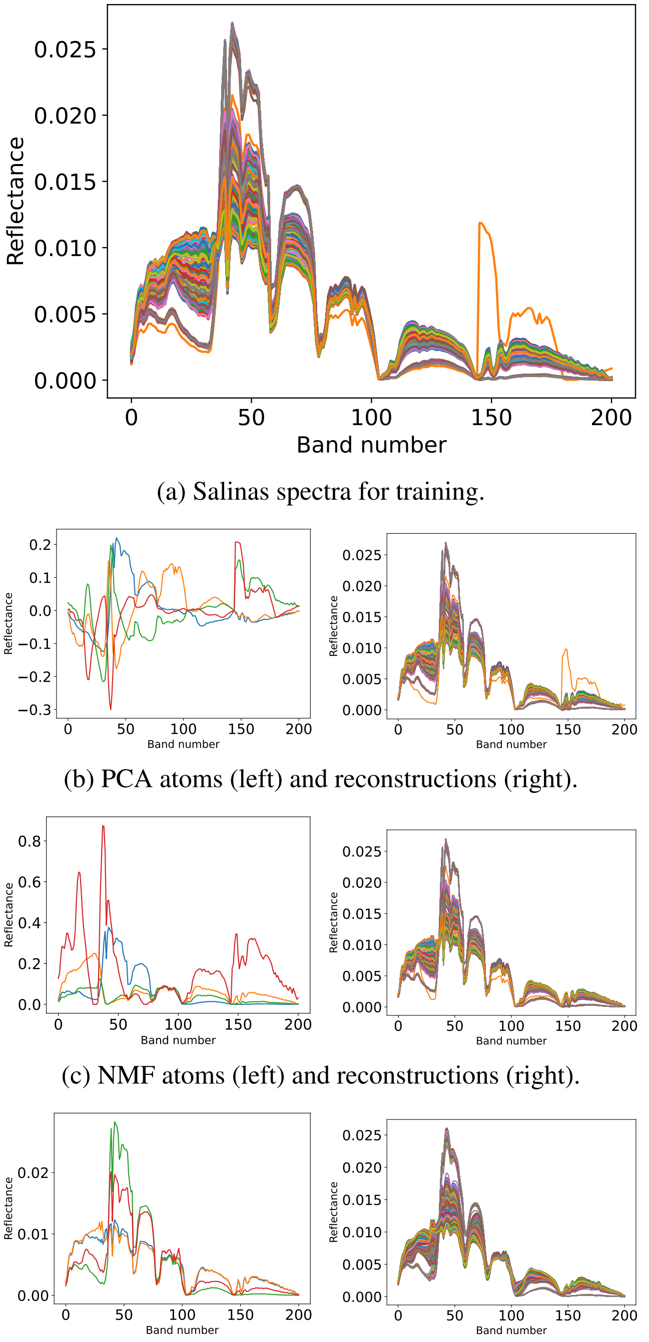
(a) Gaussian atoms (left) and synthetic Gaussian data (right).

(b) PCA atoms (left) and reconstructions (right).

(c) NMF atoms (left) and reconstructions (right).

(d) WaHU atoms (left) and reconstructions (right).

**Fig. 2**. Unmixing results for Gaussian mixtures. PCA and NMF fail to capture the smooth deformations of the pair of Gaussians. This is because a linear synthesis model is inefficient for this data.



(a) Salinas spectra for training.

(b) PCA atoms (left) and reconstructions (right).

(c) NMF atoms (left) and reconstructions (right).

(d) WaHU atoms (left) and reconstructions (right).

**Fig. 3**. Unmixing results for Salinas A spectra. PCA, NMF, and WaHU all reconstruct well with respect to their respective synthesis models, but only WaHU learns atoms that resemble the training data.

Wasserstein distances and barycenters. By capturing smooth variations between data understood as probability measures, more efficient and interpretable learning is achieved compared to linear benchmarks. In the future, we will consider an unmixing formulation in terms of *unbalanced optimal transport* [28, 29] that allows us to consider spectra without the need to force them to be probability measures. While this paper has focused on the question of learning good atoms, the learned weights $w_i \in \Delta^m$ provide efficient features for downstream learning tasks (e.g., pixel classification and scene segmentation [30, 31, 12, 32]) that may allow for the curse of dimensionality to be broken when $m \ll D$. Developing uses of the weights in supervised and unsupervised labeling tasks will be pursued in future work.

## 6. REFERENCES

[1] Han Zhai, Hongyan Zhang, Pingxiang Li, and Liangpei Zhang, "Hyperspectral image clustering: Current achievements and future lines," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 35–67, 2021.

[2] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[3] Sicong Liu, Daniele Marinelli, Lorenzo Bruzzone, and Francesca Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140–158, 2019.

[4] Chunying Wang, Baohua Liu, Lipeng Liu, Yanjun Zhu, Jialin Hou, Ping Liu, and Xiang Li, "A review of deep learning used in the hyperspectral image analysis for agriculture," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 5205–5253, 2021.

[5] Hongjun Su, Zhaoyue Wu, Huihui Zhang, and Qian Du, "Hyperspectral anomaly detection: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 64–90, 2021.

[6] Stefania Matteoli, Marco Diani, and Giovanni Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 5–28, 2010.

[7] Dalton Lunga, Saurabh Prasad, Melba M Crawford, and Okan Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 55–66, 2013.

[8] John A Lee, Michel Verleysen, et al., *Nonlinear dimensionality reduction*, vol. 1, Springer, 2007.

[9] Nicolas Gillis, *Nonnegative matrix factorization*, SIAM, 2020.

[10] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[11] Abiy Tasissa, Pranay Tankala, James M. Murphy, and Demba Ba, "K-deep simplex: Manifold learning via local dictionaries," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3741–3754, 2023.

[12] Sam L Polk, Kangning Cui, Aland HY Chan, David A Coomes, Robert J Plemmons, and James M Murphy, "Unsupervised diffusion and volume maximization-based clustering of hyperspectral images," *Remote Sensing*, vol. 15, no. 4, pp. 1053, 2023.

[13] Gabriel Peyré, Marco Cuturi, et al., "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[14] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn 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.

[15] Rob Heylen, Mario Parente, and Paul Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 1844–1868, 2014.

[16] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[17] Marcel Nutz, "Introduction to entropic optimal transport," *Lecture notes, Columbia University*, 2021.

[18] Martial Agueh and Guillaume Carlier, "Barycenters in the wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[19] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–11, 2015.

[20] Hicham Janati, Marco Cuturi, and Alexandre Gramfort, "Debiased sinkhorn barycenters," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4692–4701.

[21] Matthew Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, and James M Murphy, "Measure estimation in the barycentric coding model," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23781–23803.

[22] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2005.

[23] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck, "Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 643–678, 2018.

[24] Marshall Mueller, Shuchin Aeron, James M Murphy, and Abiy Tasissa, "Geometrically regularized wasserstein dictionary learning," in *Topological, Algebraic and Geometric Learning Workshops 2023*. PMLR, 2023, pp. 384–403.

[25] Sina Nakhostin, Nicolas Courty, Rémi Flamary, and Thomas Corpetti, "Supervised planetary unmixing with optimal transport," in *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2016, pp. 1–5.

[26] David Arthur and Sergei Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.

[27] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al., "Pot: Python optimal transport," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 3571–3578, 2021.

[28] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, "Scaling algorithms for unbalanced optimal transport problems," *Mathematics of Computation*, vol. 87, no. 314, pp. 2563–2609, 2018.

[29] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, "Unbalanced optimal transport: Dynamic and kantorovich formulations," *Journal of Functional Analysis*, vol. 274, no. 11, pp. 3090–3123, 2018.

[30] Mauro Maggioni and James M Murphy, "Learning by active nonlinear diffusion," *Foundations of Data Science*, vol. 1, no. 3, pp. 271–291, 2019.

[31] James M Murphy and Mauro Maggioni, "Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 3, pp. 1829–1845, 2019.

[32] Kangning Cui, Ruoning Li, Sam L Polk, Yinyi Lin, Hongsheng Zhang, James M Murphy, Robert J Plemmons, and Raymond H Chan, "Superpixel-based and spatially-regularized diffusion learning for unsupervised hyperspectral image clustering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, no. 4405818, pp. 1–18, 2024.