IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10641170

HYPERSPECTRAL IMAGE CLUSTERING VIA LEARNED REPRESENTATION IN WASSERSTEIN SPACE

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

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

ABSTRACT

Hyperspectral images (HSI) capture rich information of large spatial scenes, yet generating labeled training data can be expensive and time-consuming. Unsupervised clustering of HSI allows for segmentation in the absence of labels and is an important problem in processing rapidly collected HSI. In order to accurately cluster noisy and high-dimensional HSI, meaningful data representations that capture latent intrinsic structure must be developed. We propose to leverage regularized dictionary learning in Wasserstein space to efficiently and accurately cluster HSI by modeling HSI pixels as probability distributions. We characterize pixels as similar if they can be synthesized as entropic Wasserstein barycenters with a common set of learned reference distributions. Our approach learns representations that preserve the geometry of the space of HSI spectra and our barycentric coding spectral clustering algorithm, which leverages these learned features, shows promise on benchmark HSI data.

Index Terms— Hyperspectral imaging, unsupervised clustering, dictionary learning, Wasserstein space, optimal transport

1. INTRODUCTION

Hyperspectral images (HSI) capture material reflectance over a wide span of spectral wavelengths, which generates powerful and discriminatory data on the scene surveyed via airborne or spaceborne instruments. When large quantities of labeled training data are available, supervised machine learning methods such as support vector machines [1], random forests [2], and deep learning [3] provide tools to accurately label the material class of pixels in an HSI. However, the collection of labeled HSI pixels is often expensive and time-consuming, particularly when the regions surveyed are remote from large human settlements. The difficulty in generating training data has motivated semisupervised and unsupervised learning methods for HSI that require few or no labeled training data points to provide a material segmentation map of the scene [4]. The challenge of learning in these low-label settings is compounded by the high dimensionality of HSI (typically over 100 spectral recordings per pixel), which problematizes classical statistical learning approaches.

In this paper, we approach the problem of clustering high-dimensional hyperspectral images through a novel method based on representation learning in the Wasserstein space, which models data as probability measures and makes comparisons between data points using entropic Wasserstein distances [5]. Unlike Euclidean or graph-based approaches to comparing HSI pixels, our Wasserstein approach crucially leverages the geometry of the HSI pixels by first performing nonlinear dictionary learning of HSI spectra in Wasserstein space, then performing spectral clustering on the learned coefficients. Promising empirical results are shown on the Salinas A HSI, which suggests the viability of our Wasserstein dictionary learning approach to HSI clustering.

The remainder of this paper is organized in the following manner. In Section 2, we overview HSI clustering methods before providing the necessary background on data analysis in Wasserstein space. In Section 3, we provide a detailed discussion of our approach to HSI clustering via Wasserstein dictionary learning. Section 4 demonstrates the efficacy of our approach on the Salinas A HSI, and we conclude and lay out directions for future research in Section 5.

2. BACKGROUND

Given HSI pixels $\{\mu_i\}_{i=1}^n \subset \mathbb{R}^D$, unsupervised clustering algorithms produce labels $\{y_i\}_{i=1}^n$ with each y_i lying in in the label set $\{1, 2, \ldots, K\}$ corresponding to K classes or material types present in the image; typically K is a user input though it may be learned. This is done without any training labels, and typical methods explain the geometric and statistical patterns in the dataset $\{\mu_i\}_{i=1}^n$ to infer clusters [6]. An important class of clustering approaches that have had success in labeling HSI scenes are graph-based approaches called *spectral clustering* [7, 8], which leverage the structural properties of a latent data graph to determine internally coherent and externally well-separated clusters [9, 10, 11, 12]. For some metric $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$, we define a graph weight matrix $W \in \mathbb{R}^{n \times n}$ with edge weight between μ_i and μ_j given

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

by $W_{ij} := \exp(-d(\mu_i, \mu_j)^2/\sigma^2)$ for some tuning parameter $\sigma^2 > 0$ so that W_{ij} is large if and only if $d(\mu_i, \mu_j)$ is small; alternatively, W can be set as a k-nearest neighbors graph with respect to the metric d for some suitable k. Natural community structure in the associated graph can be learned by analyzing the lowest frequency eigenvectors of the normalized graph Laplacian $L := I - D^{-1/2}WD^{-1/2}$ where $D \in \mathbb{R}^{n\times n}$ is the diagonal degree matrix with $D_{ii} := \sum_{\ell=1}^{n} W_{i\ell}$ for all $i = 1, \ldots, n$; see [13]. In the typical spectral clustering setting, k-means is run on the k eigenvectors of L with the smallest eigenvalues [8].

A crucial question when running spectral clustering is, what is an appropriate metric d [14, 15]? We propose to learn representations of HSI in Wasserstein space so that pixels with similar geometry will be represented similarly and then consider the usual Euclidean distance on these representations. To learn good representations, we consider *dictionary learning in the Wasserstein space* of probability distributions as follows. Let $\Delta^D := \{(x_1, \ldots, x_D) \in \mathbb{R}^D \mid \sum_{k=1}^{D} x_k = 1, x_k \geq 0, \ \forall k\}$ be the space of probability distributions in $\mathbb{R}^D$. Given $\mu, \nu \in \Delta^D$, define $\Pi(\mu, \nu) \subset \Delta^{D\times D}$ as the set of *couplings* between μ and ν, that is the collection of $\pi \in \mathbb{R}^{D\times D}_{\geq 0}$ such that $\forall j, \ \sum_{i=1}^{D} \pi_{ij} = \nu_j$, and $\forall i, \ \sum_{j=1}^{D} \pi_{ij} = \mu_i$. Suppose μ, ν are supported on a common set of points $\{b_k\}_{k=1}^{D} \subset \mathbb{R}$; this will be the case for HSI spectra from a common scene. For $p \geq 1$, the *entropic p-Wasserstein distance* [16] between μ and ν is

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

where $\epsilon > 0$ is a regularization parameter. The solution π^* to (1) aligns the mass in distribution μ with that in ν in an efficient way (first term) while ensuring the mass is smoothly distributed (second term); for a detailed discussion of entropic optimal transport, we refer to [5] and [17].

For a set of m 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* [18, 19, 20, 21] to be

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

Figure 1 shows $m = 2$ probability distributions: ν_1 a uniform distribution on $[20, 80]$ with small support added to make it bounded away from 0 over the domain considered; and $\nu_2 \sim$ Laplace$(140, 4)$. We show the corresponding family of linear mixtures $(1 - t)\mu_1 + t\mu_2$ and a family of entropic 2-Wasserstein barycenters corresponding to weights $(1-t, t) \in \Delta^2$ for $t \in \{0, .05, .1, .15, \ldots, .95, 1\}$. The entropic Wasserstein barycenters smoothly deform one unimodal distribution to another, unlike the linear mixtures which generate bimodal intermediate distributions.

Given observed probability distributions $\{\mu_i\}_{i=1}^{n}$ (interpreted in our context as HSI pixels after an appropriate normalization), we can learn meaningful representations useful for spectral clustering 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 μ_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 [22, 23]:

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

where $\rho > 0$ is a regularization parameter that balances the two terms. This non-convex optimization problem tries to reconstruct each observation well as a entropic 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

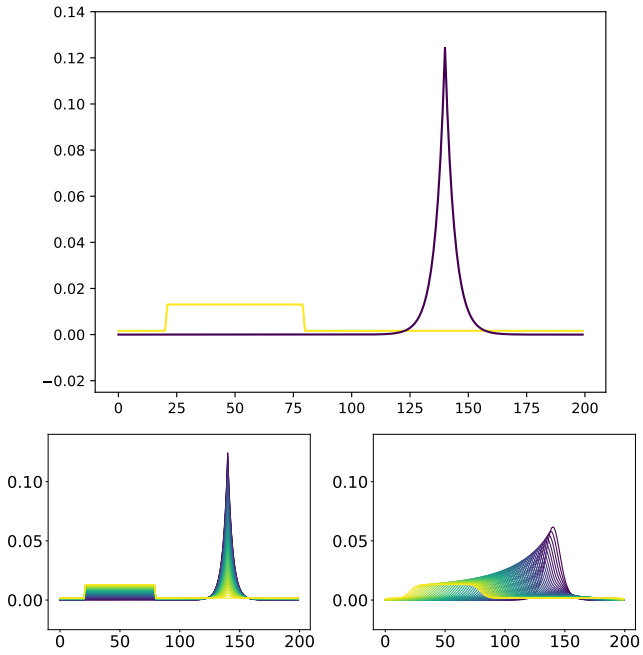


Fig. 1. (Top) A nearly uniform distribution and a Laplace distribution, represented by their densities. (Bottom) *Left*: linear mixture. *Right:* entropic Wasserstein barycenters. While linear mixtures show bi-modal behavior, entropic Wasserstein barycenters show geometry preservation. In particular, all intermediate distributions are uni-modal, just like the generating distributions. We note that the entropic regularization causes some smoothing of the two distributions, which could be mitigated by decreasing ϵ at the cost of an increase in runtime.

jointly optimize the atoms and weights. For more comprehensive information on the optimization, we refer the reader to [22] and [23].

We now proceed to a detailed discussion of how to approximately solve (3) and utilize the learned coefficients for spectral clustering.

3. CLUSTERING HYPERSPECTRAL IMAGES IN WASSERSTEIN SPACE

Given HSI pixels $\{\mu_i\}_{i=1}^n$ to cluster, we first solve (3) to acquire coefficients $\{w_i\}_{i=1}^n$. Intuitively, if two pixels μ_i and μ_j have similar coefficients w_i and w_j, then they use similar learned atoms in their barycentric reconstructions. This implies meaningful similarity of *the original data points themselves via similarity of the learned coefficients*. In the context of linear signal processing, this observation is the basis for a range of clustering methods [24, 25, 26], and we apply it in the context of Wasserstein dictionary learning. Specifically, we consider the metric $d(\mu_i, \mu_j) = \|w_i - w_j\|_2$ as the metric for spectral clustering. Our overall approach to *barycentric coding spectral clustering (BCSC)* therefore has two components: first, learn weights via approximately solving (3), second use the learned weights via the metric $\|w_i - w_j\|_2$ in spectral clustering.

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. The basic procedure is to: (i) initialize the weights $\{w_i\}_{i=1}^n$ uniformly at random from Δ^m and the atoms via Wasserstein k-means over the training data (ii) iteratively update via automatic differentiation on $\mathcal{G}$. This is outlined in Algorithm 1.

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 $\sigma(\boldsymbol{\alpha}) = (\nu_1 \mid \nu_2 \mid \ldots \mid \nu_m)^\top$ and $\sigma(\boldsymbol{\beta}) = (w_1 \mid w_2 \mid \ldots \mid w_n)^\top$ where σ acts row-wise on a matrix and on a vector as:

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

Once weights $\{w_i\}_{i=1}^n$ have been learned, they are used for spectral clustering. The overall approach we call *barycentric coding spectral clustering (BCSC)* and is detailed in Algorithm 2. In order to improve runtime, a random sample of pixels can be used for learning weights and spectral clustering based on this *spectral information* and all remaining points can be labeled via *spatial inpainting*. Specifically, the 10 ℓ^1 spatial labeled nearest neighbors of each unlabeled pixel are computed and the majority label among those labeled pixels is used for the unlabeled pixels.

Algorithm 1: Geometric Wasserstein Dictionary Learning

Input: HSI spectra: $\{\mu_i\}_{i=1}^n \subset \mathbb{R}^D$; Wasserstein parameter: p; entropic regularization: ϵ; locality regularization: ρ; number iterations: L; number of atoms: m.
Normalize each pixel μ_i to lie in Δ^D.
Initialize variables $\boldsymbol{\alpha}^{(0)} \in \mathbb{R}^{m\times N}$, $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{n\times m}$.
for $k \leftarrow 1, \ldots, L$ **do**
  $\{\nu_j^{(k)}\}_{j=1}^m \leftarrow \sigma(\boldsymbol{\alpha}^{(0)}), \{w_i^{(k)}\}_{i=1}^n \leftarrow \sigma(\boldsymbol{\beta}^{(0)})$.
  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: $\{w_i\}_{i=1}^n \leftarrow \sigma(\boldsymbol{\beta}^{(k)})$.

Algorithm 2: Barycentric Coding Spectral Clustering (BCSC)

Input: $\{\mu_i\}_{i=1}^n$: HSI pixels; N: number of pixels for which to solve (3); number of nearest neighbors for graph: NN; K: number of clusters
Select a random subset of size N among the $\{\mu_i\}_{i=1}^n$ to train on, call them $\{\tilde{\mu}_i\}_{i=1}^N$.
Run Algorithm 1 on $\{\tilde{\mu}_i\}_{i=1}^N$ to learn barycentric weights $\{w_i\}_{i=1}^N \subset \Delta^m$.
Run K-means on the K lowest frequency eigenvectors of the symmetric normalized Laplacian associated to the NN-nearest neighbors graph with respect to $d(\mu_i, \mu_j) = \|w_i - w_k\|_2$, to acquire labels $\{\tilde{y}\}_{i=1}^N$.
Assign the $n - N$ unlabeled pixels labels via majority vote among the 10 ℓ^1 spatial nearest neighbors that are among the $\{\tilde{y}\}_{i=1}^N$; call the resulting inpainted labels $\{y_i\}_{i=1}^n$.
Output: Cluster labels: $\{y_i\}_{i=1}^n$.

Code implementing Algorithms 1 and 2 using the Python OT library [27] and all experiments are public[1].

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

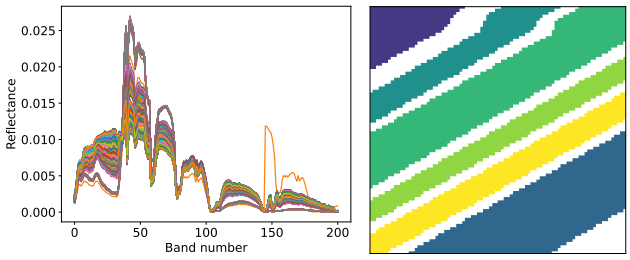


Fig. 2. *Left:* Salinas A spectra. *Right:* Salinas A ground truth; white classes are unlabelled.

4. EXPERIMENTAL CLUSTER ANALYSIS OF SALINAS A

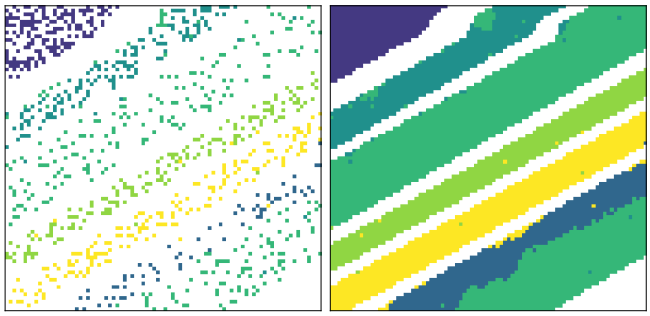
The Salinas A hyperspectral image was generated in 1998 by the AVIRIS sensor. The imaged scene is an agricultural region in Salinas Valley, CA, USA. It consists of 224 spectral bands ranging from 380-2500 nm across 224 bands. There are 6 material classes in the image: broccoli greens; corn green weeds; and romaine lettuce at 4 different growth times, namely 4, 5, 6, and 7 weeks. The full scene is 83×86, totaling 7138 pixels; see Figure 2 for an image of Salinas A spectra and its spatial ground truth with unlabeled pixels colored in white.

We run Algorithm 2 using 1002 randomly selected pixels (167 for each of the 6 labeled material classes) to run the full Algorithm 1 on, followed by spatial inpainting for the rest. Figure 3 shows the results before and after inpainting. To quantitatively assess the performance of Algorithm 2, we use overall accuracy (OA) as a metric. Overall accuracy is the total number of correctly labeled pixels divided by the total number of pixels in the ground truth. To compute this, we utilize the Hungarian assignment algorithm to match cluster labels with the ground truth labels. We perform 10 experiments with different random samples of 1002 training pixels, and achieved an average OA of .75 before inpainting. Results were bimodal, with results either in the mid 80s or mid 60s; representative examples are in Figure 3. In this experiment, parameters $p = 1$, $\epsilon = .1$, $\rho = .001$, $NN = 25$, $L = 400$, and $m = 32$ were used.

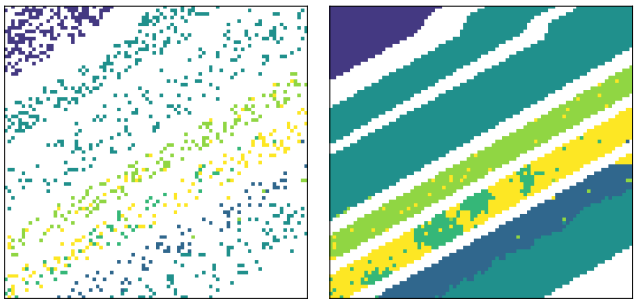
5. CONCLUSIONS AND FUTURE WORK

This paper presented a novel method for HSI clustering based on Wasserstein dictionary learning and spectral clustering. The learned coefficients provide a useful, dimension-reduced representation of the original HSI that captures intrinsic geometric information as parametrized by entropic Wasserstein barycenters.

The impact of the random training set is significant in the performance on Salinas A, and increasing the robustness of the dictionary learning approach to both noise and outliers is a topic of ongoing work. The key idea of using learned barycentric coefficients for downstream labeling tasks need not be constrained to unsupervised learning. Indeed, these features are natural and interpretable for semisupervised learning paradigms. Extending our approach in this direction is a topic of ongoing research. Note, the errors in Figure 3 (a) are mostly due to the bottom right class being incorrectly split. This could be corrected with a few carefully chosen training labels via *active learning* [29]; developing a criteria for which pixels to query for labels based on the learned atoms and coefficients will be pursued in future work. Further, we aim to compare the proposed algorithm with NMF employing simplex constraints [30] and the Wasserstein NMF algorithm proposed in [31]. In this paper, having optimized our algorithm on partial random samples, we propagated labels via nearest neighbours using the ℓ_1 metric. Considering that the weights in our framework are distributions, we plan to investigate using the Wasserstein distance between these weights for inpainting.



(a) OA = 86% before inpainting.



(b) OA = 65% before inpainting.

Fig. 3. In (a), a result with OA 86% before inpainting and 84% after, which is competitive with state-of-the-art graph-based methods for HSI clustering [28]. The only major error is in splitting the bottom right cluster, which is a common error in unsupervised clustering of Salinas A. In (b), we see a less competitive result, achieving 65% OA before inpainting. Inpainting lowers accuracy further to 53%.

6. REFERENCES

[1] Farid Melgani and Lorenzo Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[2] Jisoo Ham, Yangchi Chen, Melba M Crawford, and Joydeep Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492–501, 2005.

[3] 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.

[4] 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.

[5] 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.

[6] Rui Xu and Donald Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[7] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[8] Andrew Ng, Michael Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

[9] Nathan D Cahill, Wojciech Czaja, and David W Messinger, "Schroedinger eigenmaps with nondiagonal potentials for spatial-spectral clustering of hyperspectral imagery," in *Algorithms and technologies for multispectral, hyperspectral, and ultraspectral imagery XX*. SPIE, 2014, vol. 9088, pp. 27–39.

[10] Wei Zhu, Victoria Chayes, Alexandre Tiard, Stephanie Sanchez, Devin Dahlberg, Andrea L Bertozzi, Stanley Osher, Dominique Zosso, and Da Kuang, "Unsupervised classification in hyperspectral imagery with nonlocal total variation and primal-dual hybrid gradient algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2786–2798, 2017.

[11] Rong Wang, Feiping Nie, and Weizhong Yu, "Fast spectral clustering with anchor graph for large hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2003–2007, 2017.

[12] 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.

[13] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[14] Geoffrey Schiebinger, Martin J Wainwright, and Bin Yu, "The geometry of kernelized spectral clustering," *Annals of Statistics*, vol. 43, no. 2, pp. 819–846, 2015.

[15] Anna Little, Mauro Maggioni, and James M Murphy, "Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms," *Journal of Machine Learning Research*, vol. 21, no. 6, pp. 1–66, 2020.

[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] 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.

[23] 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.

[24] Ehsan Elhamifar and René Vidal, "Sparse manifold clustering and embedding," *Advances in neural information processing systems*, vol. 24, 2011.

[25] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[26] 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.

[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] 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.

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

[30] Maryam Abdolali and Nicolas Gillis, "Simplex-structured matrix factorization: Sparsity-based identifiability and provably correct algorithms," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 2, pp. 593–623, 2021.

[31] Antoine Rolet, Marco Cuturi, and Gabriel Peyré, "Fast dictionary learning with a smoothed wasserstein loss," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Arthur Gretton and Christian C. Robert, Eds., Cadiz, Spain, 09–11 May 2016, vol. 51 of *Proceedings of Machine Learning Research*, pp. 630–638.