Clustering, factor discovery and optimal transport

Hongkang Yang* Esteban G. Tabak†

Oct 25, 2019

Abstract

The clustering problem, and more generally, latent factor discovery –or latent space inference– is formulated in terms of the Wasserstein barycenter problem from optimal transport. The objective proposed is the maximization of the variability attributable to class, further characterized as the minimization of the variance of the Wasserstein barycenter. Existing theory, which constrains the transport maps to rigid translations, is extended to affine transformations. The resulting non-parametric clustering algorithms include k-means as a special case and exhibit more robust performance. A continuous version of these algorithms discovers continuous latent variables and generalizes principal curves. The strength of these algorithms is demonstrated by tests on both artificial and real-world data sets.

Keywords: Clustering, optimal transport, Wasserstein barycenter, factor discovery, explanation of variability, principal curve

AMS Subject classification: 62H30, 62H25, 49K30

1 Introduction

Clustering a data set $\{x_i\}$ consists of assigning to each sample $x_i \in X$ a label $z_i \in \{1, \dots K\}$, based on some notion of similarity among data points. In the broader context of factor discovery, the label or latent variable z_i can be drawn from more general spaces, such as $\mathbb{R}^d$ or smooth manifolds. One way to conceptualize factor discovery, alternative to the notion of similarity, is through the latent variable's capability of explaining the variability— or reducing the uncertainty— in the data.

Given raw data $\{x_i\}$ or the underlying probability $\rho(x)$, a natural way to characterize its uncertainty is through the variance

$$U^- = \int \|x - \bar{x}\|^2 d\rho(x),$$

where $\bar{x}$ is the mean. The class assignment can be characterized as a joint distribution $\rho(x, z)$ between x and z that decomposes $\rho(x)$. Then, factor discovery can be posed as finding an assignment $\rho(x, z)$ that minimizes the amount of remaining uncertainty. This remaining uncertainty can be quantified, for instance, as the expected value of the conditional variance

$$U^+ = \int \|x - \bar{x}(z)\|^2 d\rho(x, z),$$

where each $\bar{x}(z) = \mathbb{E}_{\rho(x|z)}[x]$ is the conditional mean of the conditional distribution $\rho(x|z)$.

An alternative is to remove from x the variability due to z through a map

$$y = T(x; z),$$

so that the resulting y is independent of z, and characterize the remaining variability as the variance of the distribution μ of y:

$$U^+ = \int \|y - \bar{y}\|^2 d\mu(y).$$

*Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, hy1194@nyu.edu

†Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, tabak@cims.nyu.edu

Such transformations of the data are customarily performed to eliminate the effect of confounding factors, such as batch effects in biostatistics. Then the resulting variable y can be thought of as the original x cleaned of the effect of z. Following the preceding work [30], we propose to seek latent factors z such that their removal minimizes the uncertainty left in the data. Such twist in the objective function of factor discovery has a number of conceptual and practical advantages:

1. **Existing factors.** Often the data includes, in addition to x_i, known factors z_i, such as a patient's age, that explain part of the variability in x. Including these known factors in the proposed removal of variability allows one to determine hidden factors that do not overlap with them, as explaining variability already accounted for by known factors would serve little purpose.

2. **Robust treatment of the remaining variability.** One can further analyze the filtered data $\{y_i\}$, which contains the variability in x not explained by the z. In particular, one can perform density estimation, a further clustering or factor discovery, all of which become easier because the data has been cleaned of the variability due to z. This is similar to the search for further, more biologically meaningful patterns in medical data, after having eliminated the batch effects.

3. **More general notion of uncertainty.** In the considerations above, variability was quantified in terms of the variance (more precisely, in terms of the trace of the covariance matrix.) Yet this is not always the most natural way to measure variability. By pushing the measurement of variability to the representative distribution μ, one can adopt much more general characterizations. As shown in [37], the measurement of variability and the characterization of the "optimal" maps $y = T(x; z)$ are closely linked.

The problem of removing from data the variability associated with class assignment adopts a natural formulation in terms of optimal transport. Suppose we are given a joint distribution $\rho(x, z)$ that decomposes the data $\rho(x)$, or equivalently, a labeled sample set $\{x_i, z_i\}$. Removing the variability attributable to factor z amounts to mapping the clusters or conditional distributions $\rho(x|z) \approx \{x_i|z_i = z\}$ through class-dependent maps $y = T(x; z)$ onto a common distribution

$$\mu \approx \{y_i = T(x_i; z_i)\}$$

which can be seen as a common representative of all $\rho(x|z)$. The variable y must be independent of z, for otherwise the variability explainable by class would not have been completely removed. Moreover, one should seek maps $T(\cdot; z)$ that, while satisfying the aforementioned condition, deform the data minimally, so that only the variability due to class is removed. Quantifying the deformation due to T with the expected value of a pointwise "transportation" cost $c\big(x, T(x, z)\big)$, one arrives at the Wasserstein barycenter problem [37]:

$$\min_{\mu,T} \iint c\big(x, T(x; z)\big)\ d\rho(x, z)$$

where $T(x; z)$ ranges over all maps such that each $T(\cdot; z)$ transports the conditional distribution $\rho(x|z)$ to some μ. Specifically, for clustering, the assignment $\rho(x, z)$ reduces to a collection of clusters $\rho_k(x)$ with weights P_k (that is, the proportion of samples contained in class k), and the Wasserstein barycenter problem becomes:

$$\min_{\mu,T_k} \sum_{k=1}^{K} P_k \int c\big(x, T_k(x)\big)\ d\rho_k(x),$$

where T_k ranges over all maps that transports ρ_k to μ.

As described, this procedure comes after clustering, as it assumes that each sample x_i has already been assigned to a class k_i. However, it can also be invoked to define an objective function for the clustering process itself: one should assign the samples to classes so as to minimize the unexplained variability, i.e. the variability left in μ.

When posed in this framework, the clustering problem contains two levels of optimization. The lower level seeks a barycenter μ for the clusters ρ_k or equivalently the conditional distributions $\rho(x|z)$, when a class assignment $\rho(x, z)$ is given. Transporting the clusters onto a barycenter corresponds to filtering out

class distinctions, and minimizing the transportation cost corresponds to avoiding data deformation. So this level corresponds to the real-world problems of "removing batch effects", "pooling data' or "supervised unlearning" [30].

The higher level of optimization determines the class assignment $\rho(z|x)$, an unsupervised learning problem. The effectiveness of learning is measured by the amount of uncertainty removed, calculated as the difference between the variance before and after the "supervised unlearning" of the inner level. This approach can be compared to the task of data compression.

This framework was first proposed in [30], which showed that, if one quantifies the variability in μ through the trace of the covariance matrix, uses as cost $c(x, y)$ the squared Euclidean distance, and restricts the maps T_k to rigid displacements, this formulation becomes equivalent to k-means. Yet, this idea can be extended much further, for instance by quantifying the variability in μ in alternative ways, using cost functions different from the square distance, and allowing more general transport maps. This article develops in detail the next natural step, which replaces the rigid translations of [30] by general affine maps.

The literature on clustering data sets is vast and rich, including a broad array of diverse methodologies, which we cannot possibly summarize here. The purpose of this article is not so much to build one new method for clustering, but rather to develop a conceptual paradigm that characterizes clustering as a procedure for the reduction of variability in data, which can be formulated naturally in terms of the Wasserstein barycenter problem with hidden class assignment. The main contribution is to extend the proposal in [30], which derives k-means using rigid translation as transport maps, to incorporate the more general affine maps. As such, this article opens the way to considering more general characterizations of variability, as well as to moving from discrete to continuous class assignments. Along the way, it develops an optimal-transport-based numerical procedure for clustering that is shown to compare favorably with more standard approaches.

The plan of this article is as follows. After this introduction, Section 2 summarizes the elements that we need from the Wasserstein barycenter problem. Section 3 introduces the affine optimal transport maps. Section 4 describes factor discovery (i.e. class assignment or clustering) with affine maps, its solution via gradient descent, its relation to k-means and its reduction and simplification in various scenarios. Section 5 introduces its continuous extension and discuss its relation to PCA and principal surfaces. Section 6 compares the performance of the various variants of the method, k-means and fuzzy k-means on synthetic and real world data. Finally section 7 summarizes the work and sketches some directions of current research.

2 The Wasserstein Barycenter problem

Denote by $\mathcal{P}_2(\mathbb{R}^d)$ the space of Borel probability measures in $\mathbb{R}^d$ with finite second moments, and by $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ the subspace of absolutely continuous measures. Given any $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, we say that $\{\rho_k, P_k\}_{k=1}^K$ is a clustering plan for ρ if $\rho_k \in \mathcal{P}_2(\mathbb{R}^d)$, the P_k are positive with $\sum_{k=1}^K P_k = 1$, and

$$\rho = \sum_{k=1}^{K} P_k \rho_k. \tag{1}$$

We seek to remove the variability in ρ attributable to the classes $1, \dots K$ by transporting these ρ_k to a weighted barycenter μ.

The notion of transportation can be visualized as follows: given a random variable X and a map $T : \mathbb{R}^d \to \mathbb{R}^d$, the transport of X by T is simply $T(X)$. If ρ is the distribution of X, then we say that T pushes forward ρ to μ, and write $\mu = T\#\rho$, if μ is the distribution of $T(X)$. Equivalently, for all measureable subset $A \subseteq \mathbb{R}^d$,

$$\mu(A) = \rho(T^{-1}(A)).$$

Monge [23] introduced the optimal transport problem

$$I(\rho, \mu) := \inf_{T\#\rho=\mu} \mathbb{E}_{\rho(x)}\big[c(x, T(x))\big],$$

where $c(x, y)$ represented the cost of moving a moving a unit of mass from x to y, such as the Euclidean distance between the two points.

Kantorovich [17] generalized the transport maps to couplings, proposing the relaxation

$$I(\rho,\mu) := \inf_{\pi\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \mathbb{E}_{\pi(x,y)}\big[c(x,y)\big], \text{ such that } \pi_X=\rho, \pi_Y=\mu,$$

where π_X, π_Y denote the two marginals of π. When the cost $x(x,y)$ is given by the squared Euclidean distance $||x-y||^2$, the optimal transport cost $I(\rho,\mu)$, denoted $W_2^2(\rho,\mu)$, defines the 2-Wasserstein metric on $\mathcal{P}_2(\mathbb{R}^d)$.

Given a clustering plan $\{\rho_k, P_k\}_{k=1}^K$, its Wasserstein barycenter is the minimizer of the total transport cost

$$\mu = \text{argmin}_{\tilde{\mu}\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{k=1}^K P_k W_2^2(\rho_k,\tilde{\mu}). \tag{2}$$

The barycenter always exists [2], and if at least one ρ_k belongs to $\mathcal{P}_{2,ac}(\mathbb{R}^d)$, then the barycenter is unique and absolutely continuous [2, 18]. If all $\rho_k \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, then we have Monge solutions for the transport $T_k\#\rho_k = \mu$, with transport maps of the form [32],

$$T_k = x - \nabla\phi_k(x), \tag{3}$$

where ϕ_k is a convex function.

In the context of factor discovery, the class labels $k \in Z = \{1,\dots K\}$ are latent variables underlying $\rho(x)$. The clustering plan $\{\rho_k, P_k\}_{k=1}^K$ can be seen as a joint distribution $\rho(x,z)$ over $\mathbb{R}^d \times Z$, such that the ρ_k are the conditional distributions $\rho(x|z=k)$ and $\{P_k\}$ becomes the marginal distribution over Z. It is natural to extend the definition of Wasserstein barycenter to joint distributions $\rho(x,z) = \rho(x|z)v(z)$ over more general latent spaces Z [37]:

$$\mu = \text{argmin}_{\tilde{\mu}\in\mathcal{P}_2(\mathbb{R}^d)} \int W_2^2(\rho(x|z),\tilde{\mu})dv(z). \tag{4}$$

(where the conditional measures $\rho(x|z)$ are defined by disintegration [**?**]). Such generalized Wasserstein barycenters were studied recently in [24, 18], which showed existence and regularity. A treatment of barycenters with more general cost functions c, data spaces, and latent spaces Z can be found in [37].

3 Optimal affine transport

Linear problems are often amenable to closed-form solutions and efficient computations. In particular, the Wasserstein barycenter problem would become much simpler if all the optimal transport maps T_k were affine. Intuitively, a sufficient condition for this desirable property is that the clusters ρ_k have similar shapes (e.g. they are all Gaussians), which implies that the barycenter μ as a representative of ρ_k should also have that shape. Consequently, the map T_k only needs to translate and dilate the ρ_k to transform them into μ. This reasoning can be formalized using location-scale families [4] that generalize the Gaussian family:

Definition 3.1. *Fix some arbitrary distribution $\mathbb{P}_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. The location-scale family $\mathcal{F}(\mathbb{P}_0)$ induced by $\mathbb{P}_0$ is defined as*

$$\mathcal{F}(\mathbb{P}_0) := \big\{L\#\mathbb{P}_0 \mid L(x) = \Sigma x + \overline{x},\ \Sigma\in\mathcal{S}_d^+,\ \overline{x}\in\mathbb{R}^d\big\} \tag{5}$$

where $\mathcal{S}_d^+$ is the space of symmetric positive-definite matrices. For convenience, we can set $\mathbb{P}_0$ to have zero mean and identity covariance. Then, $\overline{x}, \Sigma$ become the mean and covariance of $L\#\mathbb{P}_0$, and we denote each element of $\mathcal{F}(\mathbb{P}_0)$ by $\mathbb{P}_{\Sigma,\overline{x}}$.

Theorem 3.1. *Given any location-scale family $\mathcal{F}(\mathbb{P}_0)$, any measureable space Z, and any joint distribution*

$$\rho(x,z) = \rho(x|z)v(z) \in \mathcal{P}(\mathbb{R}^d \times Z)$$

such that each conditional distribution belongs to the family

$$\rho(x|z) = \mathbb{P}_{\Sigma(z),\overline{x}(z)} \in \mathcal{F}(\mathbb{P}_0),$$

and the marginal $\rho(x)$ has finite second moment,

$$\mathbb{E}_{\rho(x)}\left[||x||^2\right] \leq \infty,$$

then there exists a unique barycenter μ, it belongs to the family: $\mu = \mathbb{P}_{\Sigma_y, \overline{y}} \in \mathcal{F}(\mathbb{P}_0)$, and it satisfies

$$\overline{y} = \int \overline{x}(z) dv(z) = \mathbb{E}_{\rho(x)}[x] \tag{6}$$

$$\Sigma_y = \int \left(\Sigma_y^{\frac{1}{2}} \cdot \Sigma(z) \cdot \Sigma_y^{\frac{1}{2}}\right)^{\frac{1}{2}} \, dv(z) \tag{7}$$

where $\Sigma^{\frac{1}{2}}$ is the principal matrix square root.

Proof. See Appendix A. □

Remark 3.1. If we are only concerned with the traditional setting of clustering (1) when the latent space $Z = \{1, \dots K\}$ is finite, then the above result has been given by Corollary 4.5 of [4]. Nevertheless, in Section 5, our clustering algorithms will be extended to the continuous case with $Z = \mathbb{R}$, and we need the full strength of Theorem 3.1.

Corollary 3.1.1. *If we further assume that each $\rho(x|z)$ is isotropic: $\Sigma(z) = \sigma^2(z) \cdot Id$, then the unique barycenter is also isotropic, with standard deviation*

$$\sigma = \int \sigma(z) dv(z). \tag{8}$$

Next, we verify that the optimal transport maps T_k are affine. Denote the means and covariances of ρ_k by $\overline{x}_k, \Sigma_k$, so that $\rho_k = \mathbb{P}_{\overline{x}_k, \Sigma_k} \in \mathcal{F}(\mathbb{P}_0)$. Theorem 2.1 of [9] shows that the T_k are given by

$$\begin{aligned} T_k(x) &= \alpha_k \cdot x + \beta_k \\ \alpha_k &= \Sigma_k^{-\frac{1}{2}} (\Sigma_k^{\frac{1}{2}} \Sigma_y \Sigma_k^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_k^{-\frac{1}{2}} \\ \beta_k &= \overline{y} - \alpha_k \overline{x_k}. \end{aligned} \tag{9}$$

Remark 3.2. We offer an alternative, short derivation of (9) based on preconditioning arguments [20], which essentially follow from the convexity in (3). With a suitable preconditioning $\rho_k^* = F \# \rho_k$ and $\mu^* = G \# \mu$, the composite map

$$T_k = G^{-1} \circ T_k^* \circ F$$

will be optimal if T_k^* is the optimal transport map from ρ_k^* to μ^*. By [20], the following is an admissible pair

$$F(x) = \Lambda^{1/4} Q^T \Sigma_k^{-\frac{1}{2}} (x - \overline{x}_k), \quad G(y) = \Lambda^{-1/4} Q^T \Sigma_k^{\frac{1}{2}} (y - \overline{y})$$

where $Q \Lambda Q^T$ is an eigendecomposition of $\Sigma_k^{\frac{1}{2}} \Sigma_y \Sigma_k^{\frac{1}{2}}$. Then, ρ_k^* and μ^* have the same mean and covariance. Since they belong to a location-scale family, T_k^* must be the identity. It follows that

$$T_k(x) = G^{-1} \circ F = \Sigma_k^{-\frac{1}{2}} (\Sigma_k^{\frac{1}{2}} \Sigma_y \Sigma_k^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_k^{-\frac{1}{2}} (x - \overline{x}_k) + \overline{y}$$

□

From now on, we will always model the clusters ρ_k (or the conditional distributions $\rho(x|z)$) as members of a location-scale family $\mathcal{F}(\mathbb{P}_0)$. Besides the benefit that the optimal transport maps T_k are simplified into affine maps, the unique barycenter μ also belongs to $\mathcal{F}(\mathbb{P}_0)$, with mean and covariance determined by (6) and (7). The restriction to location-scale families is not a stringent requirement in practice. Common clustering methods such as EM algorithm [25, 3] often model the clusters as Gaussian, while the "standard data" for k-means consists of spherical clusters with equal radii [28].

4 Factor discovery with affine maps

In the introduction, we have characterized clustering as a latent factor discovery problem that seeks to maximize the information gained after class assignment. If information gain is measured by the reduction in uncertainty, then clustering maximizes $Var(\rho(x)) - Var(\mu)$, where $\rho(x)$ is the unlabeled data and the barycenter μ represents the labeled data $\rho(x,z)$ produced by class assignment. Equivalently, clustering becomes

$$\min_{\rho(x,z)\in\mathcal{P}(\mathbb{R}^d\times Z)} Var\big(\mu\big) = Tr[\Sigma_y] \tag{10}$$

with the constraint that the x-margin of $\rho(x,z)$ must be the unlabeled data $\rho(x)$. This minimization effectively searches through all (pointwise) class assignments $\rho(z|x)$ over the class labels $Z = \{1,\dots K\}$.

In practice, we are given a sample $\{x_i\}_{i=1}^N$ from $\rho(x)$, and any joint distribution $\rho(x,z)$ in (10) becomes an $N\times K$ matrix. Since the x-margin of $\rho(x_i,z)$ is uniformly N^{-1}, we define the matrix

$$P = [P_k^i] = N\cdot[\rho(x,k)]$$

Then, each entry $P_k^i = \rho(k|x_i)$ is the membership probability for x_i to belong to cluster ρ_k, and the probability vector P^i is the "soft" class assignment $\rho(z|x_i)$. The z-margin of P/N yields the weights $\{P_k\}$.

Denote the K-dimensional simplex by Δ^K. Then, P^i lives on Δ^K, so the stochastic matrix $P\in\prod_i\Delta^K$. A hard assignment k_i implies that P^i is the unit vector $\vec{e}_{k_i}$, and the domain of hard assignments is the set of extremal points of $\prod_i\Delta^K$ [5].

Once a tentative class assignment $P = N\rho(x,z)$ is given, the clusters $\rho_k(x)$ can be estimated using Bayes' formula,

$$\rho_k(x_i) = \frac{\rho(k|x_i)\rho(x_i)}{\rho(k)} = \frac{P_k^iN^{-1}}{P_k} = \frac{P_k^i}{\sum_{j=1}^N P_k^j}$$

Then, the cluster means $\overline{x}_k$ and covariances Σ_k can be estimated by

$$\overline{x}_k = \frac{\sum_i P_k^i x_i}{\sum_i P_k^i},\ \Sigma_k = \frac{\sum_i P_k^i(x_i-\overline{x}_k)\cdot(x_i-\overline{x}_k)^T}{\sum_i P_k^i}. \tag{11}$$

By Theorem 3.1, the covariance of the barycenter μ is given by the discrete version of (7):

$$\Sigma_y = \sum_{k=1}^K P_k\big(\Sigma_y^{\frac{1}{2}}\Sigma_k\Sigma_y^{\frac{1}{2}}\big)^{\frac{1}{2}} \tag{12}$$

This is a non-linear matrix equation that admits a unique positive-definite solution Σ_y if all conditional covariances Σ_k are positive-definite [4]. Then Σ_y can be calculated through the following iteration scheme [4]:

$$\Sigma(n+1)\leftarrow\Sigma(n)^{-\frac{1}{2}}\Big(\sum_{k=1}^K P_k\Sigma(n)^{\frac{1}{2}}\Sigma_k\Sigma(n)^{\frac{1}{2}}\Big)^2\Sigma(n)^{-\frac{1}{2}} \tag{13}$$

where the initialization $\Sigma(0)$ is an arbitrary positive-definite matrix.

Since the objective function of clustering in (10) is $Tr[\Sigma_y]$, which from (11) and (12) is a function of the class assignment matrix P, we will derive below the gradient

$$\nabla_P Tr[\Sigma_y]$$

and use it to build clustering algorithms.

Remark 4.1. To gain insight into the functioning of (10), we can analyze its behavior in a much simplified setting. Since a location-scale family is an abstraction of probability distributions up to their second moment, a simpler set-up has all clusters ρ_k belonging to a "location family", i.e. differing only in their means $\overline{x}_k$.

Then Theorem 2.1 of [9] implies that the optimal transport maps T_k are rigid translations that align $\overline{x}_k$ to $\overline{y}$. In that case, [30] shows that the barycenter's variance is reduced to the sum of within-class variances (or equivalently, sum of squared errors, SSE)

$$Var(\mu) = \sum_{k=1}^{K} \sum_{i=1}^{N} P_k^i ||x_i - \overline{x}_k||^2 \tag{14}$$

which is exactly the objective function of k-means. Hence, by upgrading to location-scale families and affine transport maps, one can expect more robust algorithms that make use also of second moments.

4.1 Gradient descent solution

Our goal is to perform gradient descent on $P \in \prod_i \Delta^K$ to minimize $Tr[\Sigma_y]$. Even though the implicit nonlinear matrix equation (12) determines Σ_y uniquely, it is not clear whether the solution Σ_y is differentiable. Thus, we prove in Appendix B that the partial derivatives $\partial\Sigma_y/\partial P_k^i$ always exist. Then, in Appendix C, we derive explicit formulae for these derivatives. We find that the gradient of $Tr[\Sigma_y]$ with respect to each sample point x_i's probability vector P^i is given by

$$\nabla_{P^i} Tr[\Sigma_y] = \sum_{k=1}^{K} vec(I)^T \cdot W_k \cdot vec\big[(x_i - \overline{x_k}) \cdot (x_i - \overline{x_k})^T + \Sigma_k\big] \vec{e_k} \tag{15}$$

where vec is vectorization, and the W_k are the weight matrices

$$\begin{aligned} W_k := \big(\Sigma_y^{\frac{1}{2}} \otimes \Sigma_y^{\frac{1}{2}}\big) \Big[\sum_{h=1}^{K} P_h (U_h \otimes U_h)(D_h^{\frac{1}{2}} \otimes I + I \otimes D_h^{\frac{1}{2}})^{-1} (D_h^{\frac{1}{2}} \otimes D_h^{\frac{1}{2}})(U_h^T \otimes U_h^T) \Big]^{-1} \\ \Big[(U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1} (U_k^T \otimes U_k^T) \Big] (\Sigma_y^{\frac{1}{2}} \otimes \Sigma_y^{\frac{1}{2}}). \end{aligned} \tag{16}$$

Here $\otimes$ is the Kronecker product, and the U are the orthonormal and D the diagonal matrices in the eigendecompositions

$$\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}} = U_k D_k U_k^T \text{ and } \Sigma_y = U_y D_y U_y^T.$$

The update rule at each time of a gradient descent step t is given by

$$P(t+1) = Proj_{\prod_i \Delta^K} \Big(P(t) - \eta \cdot \nabla_P Tr[\Sigma_y] \Big),$$

where η is the learning rate and $Proj$ is the projection onto the closest stochastic matrix in $\prod_i \Delta^K$, which can be computed efficiently as described in [34]. The step size η can be either fixed at a small value, or determined at each step via backtracking line search [6], using a threshold $\alpha \in (0, 1/2)$ and a shortening rate $\beta \in (0, 1)$, and reducing η into $\beta\eta$ if the amount of descent is not enough:

$$Tr[\Sigma_y]\big(P(t+1)\big) - Tr[\Sigma_y]\big(P(t)\big) > \alpha \; vec(\nabla_P Tr[\Sigma_y])^T \cdot vec\big[P(t+1) - P(t)\big].$$

This descent-based clustering algorithm is summarized below, with initialization based on that of k-means.

Data: Sample $\{x_i\}$ and number of classes K
Initialize the means $\{\overline{x}_k\}$ randomly
Initialize assignment matrix P either randomly or set each P^i to be the one-hot vector corresponding to the mean $\overline{x}_k$ closest to x^i
while *not converging* **do**
  Compute the barycenter's covariance Σ_y by iteration (13)
  Compute weight matrices $W_1, \dots W_K$ by (16)
  Compute the gradient
  $\nabla_P Tr[\Sigma_y] = (\nabla_{P^i} Tr[\Sigma_y])_i = \sum_{i,k} vec(I)^T W_k vec\big[(x_i - \overline{x_k}) \cdot (x_i - \overline{x_k})^T + \Sigma_k\big] \vec{e}_{ik}$
  (Optimize step size η by backtracking)
  Update $P \leftarrow Proj_{\prod_i \Delta^K}\big(P - \eta \cdot \nabla_P Tr[\Sigma_y]\big)$
  Update cluster means $\{\overline{x}_k\}$ and covariances $\{\Sigma_k\}$ by (11)
end
return Assignment P

Algorithm 1: Barycentric clustering

Remark 4.2. It might appear at first sight that Algorithm 1 performs an alternating descent, similarly to k-means, alternating between optimizing the assignment P and updating the means $\overline{x}_k$. Nevertheless, the derivation of (15) in Appendix C does not treat $\overline{x}_k$ as constants. Instead, it directly solves for the gradient $\nabla_P Tr[\Sigma_y]$, incorporating the derivatives $\nabla_P \overline{x}_k$. Hence, Algorithm 1 is simply a gradient descent on the objective (15), which with sufficiently small step size η necessarily converges to a local minimum.

Recall that k-means minimizes the sum of squared errors (14), whose partial derivatives are simply

$$\partial_{P_k^i} SSE = ||x_i - \overline{x}_k||^2 + 2 \sum_j P_k^j (x_j - \overline{x}_k) \frac{\partial \overline{x}_k}{\partial P_k^i} = ||x_i - \overline{x}_k||^2$$

Instead of performing gradient descent, k-means directly assigns x_i to the closest cluster k_i, that is,

$$k_i = \text{argmin}_k ||x_i - \overline{x}_k||^2 = \text{argmin}_k \partial_{P_k^i} SSE$$

We can interpret this hard assignment as equivalent to a gradient descent on P^i with arbitrarily large step size, followed by the projection $Proj_{\Delta^K}$, so that P^i arrives at an extremal point of Δ^K, which is a one-hot vector.

In exactly the same way, we can simplify Algorithm 1 into a hard assignment algorithm, such that each x_i is assigned to the cluster k_i with the smallest gradient term in $\partial_{P^i} Tr[\Sigma_y]$.

Data: Sample $\{x_i\}$ and number of classes K
Initialize the means $\{\overline{x}_k\}$ randomly and the labels k_i by the closest mean
while *not converging* **do**
  Compute the gradient $\nabla_P Tr[\Sigma_y]$
  for x_i *in sample* **do**
    $k_i \leftarrow \text{argmin}_k\big(\partial_{P_k^i} Tr[\Sigma_y]\big)$
  end
  Update cluster means $\{\overline{x}_k\}$ and covariances $\{\Sigma_k\}$
  (Possibly apply an update rate c to smooth the update: $\overline{x}_k \leftarrow c$ new $\overline{x}_k + (1-c)$ old $\overline{x}_k$)
end
return Labels $\{k_i\}$

Algorithm 2: Hard barycentric clustering

4.2 Relation to *k*-means

We show next that the barycentric clustering algorithms reduce to k-means in the latter's setting. The "standard data" for k-mean consist of spherical clusters with identical radii and proportions [28], which

implies that $P_1 = \cdots = P_K = 1/K$ and $\Sigma_1 = \cdots = \Sigma_K = \frac{\sigma^2}{d} I$ for some common variance σ^2. Then the gradient (15) simplifies into

$$\partial_{P_k^i} Tr[\Sigma_y] = \frac{\sigma^2}{d} Tr\big[(x_i - \overline{x_k}) \cdot (x_i - \overline{x_k})^T + \Sigma_k\big] = \frac{\sigma^2}{d}(||x_i - \overline{x}_k||^2 + \sigma^2).$$

Since each P^i lies in the simplex Δ^K, the direction of gradient descent must be parallel to Δ^K. Hence the term σ^2, shared by all entries of $\nabla_{P^i} Tr[\Sigma_y]$, is eliminated by the projection map $Proj_{\prod_i \Delta^K}$ of Algorithm 1, while for Algorithm 2, it is eliminated by the argmin_k step. The resulting gradient

$$\nabla_{P^i} Tr[\Sigma_y] = \sum_{k=1}^{K} ||x_i - \overline{x}_k||^2 \vec{e_k}$$

is precisely the gradient of the sum of squared errors (14), the objective function of k-means. It is straightforward to check that Algorithm 2 reduces to k-means, and thus k-means can be seen as a special case of barycentric clustering.

4.3 Isotropic solution

Section 6 will demonstrate that the barycentric clustering algorithms can recognize clusters that deviate from the "standard data", for which k-means and fuzzy k-means would fail, but this robustness comes at the expense of the complexity of gradients in (15) and (16). Here we explore a situation in between, making hypotheses weaker than the "standard data", yet strong enough to yield solutions that, while more robust than k-means, are at the same time simpler than (15) and easier to interpret.

Since the complexity of (15) results mostly from the non-commutativity of the matrix product, we can impose the assumption that all covariances are of the form

$$\Sigma_k = \frac{\sigma_k^2}{d} I$$

where σ_k^2 is the variance of cluster ρ_k. This assumption can be seen as a generalization of the "standard data"'s requirement that all clusters be radial with equal variances.

From Corollary 3.1.1, the barycenter's covariance becomes

$$\Sigma_y = \frac{\sigma_y^2}{d} I, \ \sigma_y = \sum_{k=1}^{K} P_k \sigma_k \tag{17}$$

and the gradient (15) reduces to

$$\partial_{P_k^i} Tr[\Sigma_y] = \frac{\sigma_y^3}{d}\Big(\frac{||x_i - \overline{x}_k||^2}{\sigma_k} + \sigma_k\Big).$$

Since the algorithms are only concerned with the gradient's direction, the gradient is effectively

$$\nabla_{P^i} Tr[\Sigma_y] = \sum_k \Big(\frac{||x_i - \overline{x}_k||^2}{\sigma_k} + \sigma_k\Big)\vec{e}_k. \tag{18}$$

Remark 4.3. Alternatively, we can obtain the gradient (18) directly differentiating the weighted sum of standard deviations (17). Note that the standard deviation can also be calculated via

$$\sigma_k = \sqrt{Var(\rho_k)} = \Big(\frac{1}{2}\iint ||x - y||^2 d\rho_k(x) d\rho_k(y)\Big)^{\frac{1}{2}} = \frac{\Big(\sum_{i,j=1}^{N} P_i^k P_j^k ||x_i - x_j||^2\Big)^{\frac{1}{2}}}{\sqrt{2}\sum_{i=1}^{N} P_i^k}.$$

Then the computation of the gradient

$$\frac{\partial \sigma_y}{\partial P_i^k} = \frac{\sum_{j=1}^N P_j^k ||x_i - x_j||^2}{\left(2\sum_{i,j=1}^N P_k^j P_k^l ||x_j - x_l||^2\right)^{\frac{1}{2}}} \tag{19}$$

involves the samples only through the pairwise distances $||x_i - x_j||^2$, which can be computed at the onset of the algorithm. This is helpful when the data space $\mathbb{R}^d$ has very large dimension, so that computing the means $\overline{x}_k$ and distances $\|x_i - \overline{x}_k\|^2$ of (18) at each iteration becomes prohibitive. Moreover, we can replace $\|x_i - x_j\|^2$ with any "dissimilarity measure" such as Riemannian distance, graph distance or kernel functions (even though naive substitution might not be justified by our barycenter model). Nevertheless, computing (19) takes $O(N^2)$ time, while (18) takes $O(N \cdot d)$, so the latter is more efficient for large sample sets of low-dimensional data.

In the isotropic scenario, barycentric clustering (Algorithms 1 and 2) can be modified via (18) into the following:

Initialize the means $\{\overline{x}_k\}$ randomly and the stochastic matrix P (by the closest $\overline{x}_k$);
while *not converging* **do**
  Compute and normalize the gradient $\nabla_P Tr[\Sigma_y] = \sum_{i,k} \left(\sigma_k + \frac{||x_i - \overline{x}_k||^2}{\sigma_k}\right)\vec{e}_{ik}$
  Optimal step size η by backtracking
  Update $P \leftarrow Proj_{\prod_i \Delta_i^K}\left(P - \eta \cdot \nabla_P Tr[\Sigma_y]\right)$
  Update the cluster means $\overline{x}_k$ and standard deviations σ_k
end
return Assignment P

Algorithm 3: Isotropic barycentric clustering

Initialize the means $\{\overline{x}_k\}$ (randomly) and labels k_i (by the closest $\overline{x}_k$);
while *not converging* **do**
  Update cluster means $\overline{x}_k$ and standard deviations σ_k
  for x_i *in sample* **do**
    $k_i \leftarrow \operatorname{argmin}_k\left(\frac{||x_i - \overline{x}_k||^2}{\sigma_k} + \sigma_k\right)$;
  end
end
return Labels $\{k_i\}$

Algorithm 4: Barycentric k-means

(We name Algorithm 4 "Barycentric k-means", as it closely resembles k-means.) Section 6 will confirm the expectation that these algorithms are more robust than k-means under varying proportions and radii (P_k and σ_k^2), but are more vulnerable to non-isotropy (Σ_k not of the form $\sigma_k^2 I$) than Algorithms 1 and 2.

4.4 Relation to Mahalanobis distance

Barycentric clustering is not the first clustering algorithm that deals with non-isotropic clusters using second moment information. A series of clustering methods [8, 12, 19] based on k-means, measure the distance between sample points and clusters by the Mahalanobis distance:

$$d^2(x_i, \overline{x}_k) = (x_i - \overline{x}_k)^T \Sigma_k^{-1} (x_i - \overline{x}_k), \tag{20}$$

which reduces the distance along the directions corresponding to the large eigenvalues of the covariance Σ_k. However, as pointed out in [19], applying (20) to k-means has the problem that the objective function (14) becomes trivial:

$$SSE = \sum_{i,k} P_k^i (x_i - \overline{x}_k)^T \Sigma_k^{-1} (x_i - \overline{x}_k) = \sum_k P_k Tr[\Sigma_k \Sigma_k^{-1}] \equiv Tr[I].$$

The Gustafson–Kessel algorithm [12, 19] remedies this problem by modifying (20) into

$$d^2(x_i, \overline{x}_k) = \det(\Sigma_k)^{\frac{1}{d}} (x_i - \overline{x}_k)^T \Sigma_k^{-1} (x_i - \overline{x}_k). \tag{21}$$

To compare barycentric clustering and the Mahalanobis distance-based algorithms, note that (20) is dimensionless, in the sense that any shrinkage or dilation of cluster ρ_k (with respect to the mean $\overline{x}_k$) would be completely cancelled out in (20), which explains how its objective function becomes trivial. Meanwhile, both the squared Euclidean distance and the modified Mahalanobis distance (21) have dimension $[l]^2$. For our algorithms, if we assume that the weight P_k is small so that the influence of Σ_k on Σ_y is small, then it is routine to check that the gradient (15, 16) has dimension $[l]$.

This comparison becomes explicit in the isotropic setting:

$$\frac{||x_i - \overline{x}_k||^2}{\sigma_k^2},\ ||x_i - \overline{x}_k||^2,\ \frac{||x_i - \overline{x}_k||^2}{\sigma_k} + \sigma_k$$

The first term is the Mahalanobis distance (20), the second term is squared Euclidean distance or equivalently the modified Mahalanobis (21), and the third term is the gradient (18), the isotropic version of (15). Hence, our algorithms can be seen as a balanced solution between the Euclidean case that completely ignores second moment information and the Mahalanobis case where too much normalization nullifies the problem.

As a side remark, the EM algorithm [25, 3] is also a clustering method that can recognize non-isotropic clusters, typically modeling ρ_k as Gaussian distributions. The Gaussians are essentially an exponential family built from the Mahalanobis distance (20), indicating a possible connection between EM and our gradients (18, 15), and thus a connection between maximum-likelihood-based method and the Wasserstein barycenter framework.

4.5 An alternative approach

We describe briefly here a different approach to solving the clustering problem (10), which avoids dealing directly with the barycenter μ. By the Variance Decomposition Theorem of [37], we have the identity

$$Var\big(\rho(x)\big) - Var(\mu) = \int W_2^2\big(\rho(x|z), \mu\big) dv(z),$$

i.e. the decrease in variance after clustering matches exactly the total transport cost from $\rho(x|z)$ to the barycenter. It follows that the clustering problem (10) is equivalent to

$$\max_{\rho(z|x)} \int W_2^2\big(\rho(x|z), \mu\big) dv(z) = \sum_{k=1}^{K} P_k W_2^2(\rho_k, \mu). \tag{22}$$

By Theorem 2.1 of [9], if ρ_k belong to a location-scale family, then the cost W_2^2 can be computed via

$$W_2^2(\rho_k, \mu) = ||\overline{x}_k - \overline{y}||^2 + Tr[\Sigma_k] + Tr[\Sigma_y] - 2Tr\big[(\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}})^{\frac{1}{2}}\big], \tag{23}$$

and Lemma 2.4 of [26] provides a formula for the partial derivatives of (23) with respect to $\overline{x}_k, \Sigma_k$. Hence one could optimize the alternative formulation (22) using gradient descent. A drawback, however, is that the gradient terms $\nabla_P W_2^2(\rho_k, \mu)$ involve the gradient of the barycenter's covariance Σ_y, suggesting that this approach would take at least as much effort as computing (13) and (15).

To circumvent the computation of Σ_y, we look for inspiration in the Euclidean space, as often geometric identities in $\mathbb{R}^d$ can be lifted to the Wasserstein space $\big(\mathcal{P}_2(\mathbb{R}^d), W_2\big)$ (This is not surprising, since $\mathbb{R}^d$ can be isometrically embedded in $\big(\mathcal{P}_2(\mathbb{R}^d), W_2\big)$ via $x \mapsto \delta_x$, and $\mathcal{P}_2(\mathbb{R}^d)$ is exactly the closed convex hull of δ_x [33].) Specifically, we consider the following identity

$$\forall \rho \in \mathcal{P}_2(\mathbb{R}^d),\ \int ||x - \overline{x}||^2 d\rho(x) = \frac{1}{2} \iint ||x - y||^2 d\rho(x) d\rho(y),$$

which computes the variance without involving the mean $\overline{x}$. Similarly, the objective (22) can be seen as a "variance", though not in $\mathbb{R}^d$ but in $\mathcal{P}_2(\mathbb{R}^d)$, so omitting the mean $\overline{x}$ corresponds to omitting the barycenter μ. The following theorem verifies this intuition.

Theorem 4.1. *For any measureable space Z and joint distribution $\rho(x,z) = \rho(x|z)v(z) \in \mathcal{P}(\mathbb{R}^d \times Z)$, let μ be any Wasserstein barycenter of $\rho(x,z)$. Then, the total transport cost can be written as*

$$\int W_2^2\big(\rho(x|z),\mu\big)dv(z) = \frac{1}{2}\iint W_2^2\big(\rho(x_1|z_1),\rho(x_2|z_2)\big)dv(z_1)dv(z_2). \tag{24}$$

Proof. See Appendix D. □

It follows that the clustering problem (10, 22), is equivalent to

$$\max_{\rho(z|x)} \frac{1}{2}\sum_{k,h=1}^{K} P_k P_h W_2^2(\rho_k,\rho_h).$$

The gradient $\nabla_P W_2^2(\rho_k,\rho_h)$ only involves the terms $\nabla_P \overline{x}_k$ and $\nabla_P \Sigma_k$, so we can apply the formulas of [26] avoiding the complexity of Σ_y. Although this approach avoids computing the iteration (13), a drawback is that its objective function contains $O(K^2)$ terms.

5 Continuous extension

This section shows how formulation (10) for clustering and the algorithms of Section 4 can be extended to latent spaces Z more general than the discrete class labels $\{1,\dots K\}$. For brevity, we focus on the simple case when $Z = \mathbb{R}$ and all clusters $\rho(x|z)$ are isotropic, as it is straightforward to extend the discussion to more general cases. Hence, given a joint distribution $\rho(x,z)$, we assume that the conditional distributions $\rho(x|z)$ belong to some location-scale family and are isotropic, and denote their means and variances by $\overline{x}(z)$ and $\sigma^2(z)$.

In practice, we are only given a finite sample set $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$. If one would use hard assignment $\rho(z|x_i) = \delta_{z_i}$ and model each label $z_i \in \mathbb{R}$ as an independent variable to be optimized, it would be impossible to evaluate the integral (8): since the sample set is finite whereas there are infinitely many $z \in \mathbb{R}$, almost all conditional distributions $\rho(x|z)$ will be represented by zero or at most one sample point.

Our solution is inspired by human vision. For the image below, it is evident that $\rho(x|z_1)$ has greater variance than $\rho(x|z_0)$, even though there is no sample point whose assignment is exactly z_0 or z_1. The key is that we can estimate $\rho(x|z_1)$ using the points nearby, $\{x_i|z_i \approx z_1\}$.

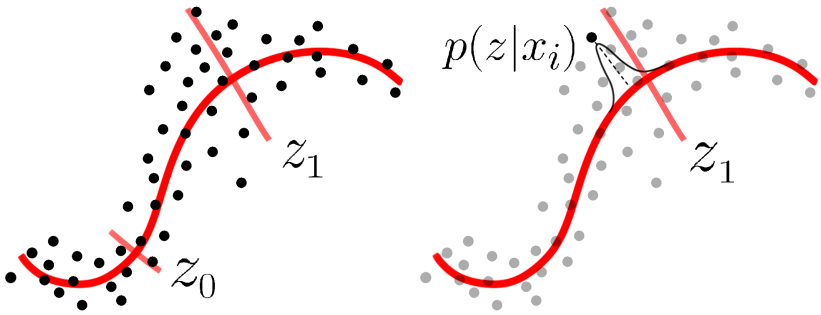

Figure 1: A two-dimensional sample set. Left: assignment of z by orthogonal projection onto the red curve. Right: soft assignment $\rho(z|x_i)$.

Hence, it is natural to use a soft assignment $\rho(z|x_i)$, which is concentrated around some z_i and decays for z far away from z_i. Effectively, the latent distribution $\{z_i\}$ is smoothed into

$$v(z) = \frac{1}{N}\sum_{i=1}^{N}\rho(z|x_i)$$

Given any z, the conditional density $\rho(x|z)$ can be estimated using Bayes' formula,

$$\rho(x_i|z) = \frac{\rho(z|x_i)\rho(x_i)}{v(z)} = \frac{\rho(z|x_i)}{\sum_{j=1}^N \rho(z|x_j)}, \tag{25}$$

and Corollary 3.1.1 implies that the barycenter's standard deviation σ is given by

$$\sigma = \int \sigma(z) dv(z) = \int \Big[\frac{\sum_{i=1}^N ||x_i - \overline{x}||^2 \rho(z|x_i)}{\sum_{i=1}^N \rho(z|x_i)}\Big]^{\frac{1}{2}} dv(z). \tag{26}$$

The objective of clustering (10) is equivalent to

$$\min_{\rho(z|x)} \sigma. \tag{27}$$

For simplicity, we parameterize the soft assignment by a Gaussian, $\rho(z|x_i) = \mathcal{N}(z|\theta_i, \epsilon^2)$, where θ_i are the means and ϵ^2 is the common variance (Notice that this does not impose any particular form on the conditional distributions $\rho(x|z)$.) The means θ_i are parameters by which we minimize (26), and we define the vector $\theta = [\theta_i]$. When the sample set $\{x_i\}$ is large, rather than having an independent variable θ_i for each x_i, one can replace θ_i by a parameterized function $\theta(x_i)$, for instance through a neural network.

Note that ϵ^2 should adapt to the set $\{\theta_i\}$. Otherwise, a fixed ϵ^2 would lead to the trivial solution where the θ_i are arbitrarily far apart, so that by (25) each conditional distribution $\rho(x|z)$ would be concentrated at some x_i and $\sigma(z)$ would go to zero. Hence, we choose ϵ^2 so as to make the distributions $\rho(z|x_i)$ close to each other for nearby x_i. Intuitively, for some fixed $0 < \alpha < 1$ (e.g. $\alpha = 10\%$), we want that for each $\rho(x|z)$, roughly a fraction α of the $\{x_i\}$ participate in $\rho(x|z)$. The trivial solution corresponds to $\alpha \approx 0$, and one should not set $\alpha \approx 1$ either, for otherwise each x_i would have significant presence in each $\rho(x|z)$, contrary to the goal of clustering as a partition of the $\{x_i\}$. Hence we set the following objective for ϵ^2, based on maximum likelihood:

$$\max_{\epsilon^2} \prod_{i,j=1}^N \mathcal{N}(\theta_j|\theta_i, (\epsilon/\alpha)^2),$$

with optimal solution given by

$$\epsilon^2 = \frac{\alpha^2}{2N^2} \sum_{i,j}^N ||\theta_i - \theta_j||^2 = \frac{\alpha^2}{N} \sum_{i=1}^N ||\theta_i - \overline{\theta}||^2 = \alpha^2 Var(\{\theta_i\}), \tag{28}$$

where $\overline{\theta}$ is the sample mean. This choice of ϵ^2 dilates $\rho(z|x_i)$ proportionally to the spread of the $\{\theta_i\}$, thus preventing the trivial solution.

To fix θ_i, we can further require that their mean should not drift away from 0. Adding for this an extra term $\overline{\theta}^2$ to the penalty yields the simpler formula

$$\epsilon^2 = \alpha^2(Var(\{\theta_i\}) + \overline{\theta}^2) = \alpha^2 \frac{||\theta||^2}{N}, \tag{29}$$

so we propose

$$\rho(z|x_i) = \mathcal{N}\Big(\theta_i, \frac{\alpha^2||\theta||^2}{N}\Big). \tag{30}$$

Next, we derive the gradient of the barycenter's standard deviation σ. By Appendix E, we can differentiate under the integral sign in (26) to obtain the following gradient:

$$\frac{\partial \sigma}{\partial \theta_i} = \mathbb{E}_{v(z)}[G_i(z)] \tag{31}$$

$$G_i(z) = \frac{1}{2N \cdot v(z)} \Big[C(z)\theta_i + \frac{z - \theta_i}{\epsilon^2} \rho_i(z) \Big(\sigma(z) + \frac{||x_i - \overline{x}(z)||^2}{\sigma(z)}\Big)\Big] \tag{32}$$

$$C(z) = \frac{1}{||\theta||^2} \sum_{j=1}^{N} \Big[\sigma(z) + \frac{||x_j - \overline{x}(z)||^2}{\sigma(z)}\Big] \cdot \Big[\frac{||z-\theta_j||^2}{\epsilon^2} - \frac{1}{||\theta||}\Big] \rho_j(z). \tag{33}$$

The computation of the integrand $G_i(z)$ for any z takes linear time $O(N)$. Estimating the expectation (31) by random sampling from the latent distribution $v(z)$, we obtain the following stochastic gradient descent algorithm, which will be tested in Section 6.4:

Input: Sample $\{x_i\}_{i=1}^N$, learning rate η, proportion constant α.
Initialize each θ_i either randomly in $[-1, 1]$ or proportionally to the principal component of the sample.
while *not converging* **do**
  Randomly sample a latent variable z from $v(z) = \frac{1}{N}\sum_{i=1}^N \rho(z|x_i)$.
  Compute the conditional mean $\overline{x}(z)$ and standard deviation $\sigma(z)$
  Compute the constant $C(z)$ by (33)
  Update each θ_i by the gradient (32): $\theta_i \leftarrow \theta_i - \eta G_i(z)$
end
return θ_i

Algorithm 5: Affine Factor Discovery. It can be seen as a continuous version of Algorithm 3.

Remark 5.1. Whenever we obtain a joint distribution $\rho(x, z)$, the conditional mean $\overline{x}(z)$ can be seen as a curve, parameterized by $z \in \mathbb{R}$, that summarizes the data $\rho(x)$. If, as in Remark 4.1, we make the simplifying assumption that all conditional distributions $\rho(x|z)$ are equivalent up to translations, Corollary 3.1.1 implies that the variance of the barycenter is given by

$$Var(\mu) = \int Var(\rho(x|z))dv(z) \approx \frac{1}{N}\sum_{i=1}^{N} ||x_i - \overline{x}(z_i)||^2$$

Or, in terms of soft assignments similar to (26),

$$Var(\mu) = \int \frac{1}{N}\sum_{i=1}^{N} ||x_i - \overline{x}(z_i)||^2 d\rho(z|x_i)$$

Since the objective of clustering, (10) and (27), is to minimize $Var(\mu)$, an immediate corollary is that, given any curve $\overline{x}(z)$, the sample x_i should be assigned to the closest point on $\overline{x}(z)$:

$$z_i = \text{argmin}_z ||x_i - \overline{x}(z)||^2$$

or the soft assignment $\rho(z|x_i)$ should be concentrated around this z_i. It follows that our formulation of clustering is reduced to an alternating descent algorithm that takes turns updating the conditional means $\overline{x}(z)$ and reassigning z_i. Yet, this procedure is exactly the principal curve algorithm [13, 14]. Hence, problem (27) is a generalization of principal curves (and principal surfaces if we set $Z = \mathbb{R}^k$).

6 Performance

Sections 6.1 and 6.2, compare k-means and the barycentric clustering algorithms on artificial data that deviate from the "standard data" of k-means, and section 6.3 tests them on real-world classification data sets. Finally, section 6.4, tests Affine factor discovery (Algorithm 5) on earthquake data to discover meaningful continuous latent variables.

6.1 Comparison of soft assignments

We design two families of artificial data. The "expansion test" is a collection of three spherical Gaussian distributions in $\mathbb{R}^2$:

$$100 \text{ samples from } \mathcal{N}\big([0,0]^T, \tfrac{1}{10}I\big),$$
$$100(1+t) \text{ samples from } \mathcal{N}\big([0,2+t]^T, \tfrac{(1+t)^2}{10}I\big),$$

$$100(1+2t) \text{ samples from } \mathcal{N}\Big(\big[\tfrac{t+1}{t+2}\sqrt{12(2t+1)}, \tfrac{2(1-t^2)}{t+2}\big]^T, \tfrac{(1+2t)^2}{10} I\Big).$$

The means and variances are designed so that, for all $t \geq 0$, the three samples are roughly contained in three pairwise adjacent balls of radii 1, $1+t$ and $1+2t$. As t increases, the sample sizes and radii grow in distinct rates. The "dilation test" is given by

$$100 \text{ samples from } \mathcal{N}\Big(\begin{bmatrix}0\\1\end{bmatrix}, \tfrac{1}{25}\begin{bmatrix}(1+t)^2 & 0\\ 0 & 1\end{bmatrix}\Big),$$

$$100 \text{ samples from } \mathcal{N}\Big(\begin{bmatrix}0\\0\end{bmatrix}, \tfrac{1}{25}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\Big),$$

$$100 \text{ samples from } \mathcal{N}\Big(\begin{bmatrix}0\\-1\end{bmatrix}, \tfrac{1}{25}\begin{bmatrix}(1+t)^2 & 0\\ 0 & 1\end{bmatrix}\Big),$$

which are Gaussians stretched horizontally at different rates. The expansion test challenges the "standard data" [28] in its first two assumptions: similar radii and similar proportions, while the dilation test challenges the assumption of isotropy. In both cases, the amount of deviation from the standard data is parametrized by $t \geq 0$.

The performance of each algorithm is measured by its correctness rate, the percentage of overlap between the true labeling and the labeling produced by the algorithm, maximized over all identifications between the proposed clusters and the true clusters: given the true labeling $\{z_i\}$ and either the labeling $\{k_i\}$ or the stochastic matrix P produced by algorithm, we define the correctness rate as

$$\max_{g \in S_K} \sum_i \mathbf{1}_{z_i = g(k_i)} \text{ or } \max_{g \in S_K} \sum_i P^i_{g(z_i)} \tag{34}$$

where g ranges over the permutation group S_K.

We first compare the soft assignment algorithms: Fuzzy k-means, barycentric clustering (Algorithm 1), and isotropic barycentric clustering (Algorithm 3). Note that k-means' objective (14) is approximately a linear function in the assignment $P = [P^k_i]$, and thus the optimal solutions are the extremal points of $Dom(P) = \prod_i \Delta^K$, which are hard assignments. Hence, in order to obtain a valid comparison among soft assignments, we use the the fuzzy k-means algorithm [5], which generalizes k-means, minimizing the following objective function:

$$J_c(P, \{\overline{x}_k\}) = \sum_i \sum_k (P^k_i)^c ||x_i - \overline{x}_k||^2.$$

This is a generalization of the sum of squared errors (14), with an exponent $c > 1$ that makes J_c strictly convex in P^i_k, and therefore yields soft assignments. Here we adopt the common choice $c = 2$.

Data: Sample $\{x_i\}$, exponent $c = 2$
Initialize the means $\overline{x}_k$ and stochastic matrix $P = (P^k_i)$ randomly
while *not converging* **do**
  for x_i *in sample and* k *= 1 to* K **do**
    $P^k_i \leftarrow (||x_i - \overline{x}_k||^2)^{1-c} / \sum_j (||x_j - \overline{x}_k||^2)^{1-c}$
  end
  for k *= 1 to* K **do**
    $\overline{x}_k \leftarrow \sum_k (P^i_k)^c x_i / \sum_k (P^i_k)^c$
  end
end

Algorithm 6: Fuzzy k-means

Since each algorithm starts with a random initialization, we stabilize performance by running each algorithm 100 times over the same sample set and selecting the result that minimizes the algorithm's objective function (J_c for fuzzy k-means, (10) for barycentric clustering, and (17) for isotropic barycentric clustering.)

The experimental results are plotted below. The first row corresponds to the expansion test with $t = 2.2$, and the second row to the dilation test with $t = 3.0$. The class assignment displayed is given by the maximum probability, $k_i \leftarrow \text{argmax}_k P^i_k$.

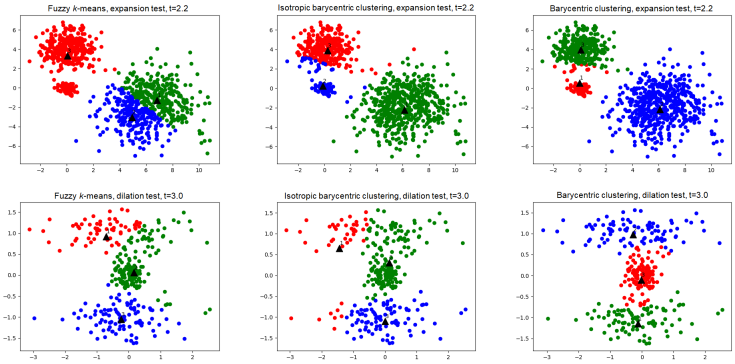

Figure 2: Clusters produced by fuzzy k-means (left), isotropic barycentric clustering (middle) and barycentric clustering (right). The black arrows indicate the clusters' means.

For the expansion test, fuzzy k-means merged the two smaller clusters and split the largest one, whereas the barycentric clustering algorithms only made a few errors on the periphery. For dilation test, the correct clusters are produced only by Barycentric clustering, whereas fuzzy k-means and Isotropic barycentric clustering split the clusters. These results are not surprising, since Section 4.2 shows that k-means is an approximation to Barycentric clustering that assumes clusters with identical sizes and radii, while Isotropic barycentric clustering, by design, assumes isotropic clusters.

Below are the plots of correct rates (34) for $t \in [0, 4]$. Fuzzy k-means, with a steady decline, is dominated by the barycentric clustering algorithms, while the difference between the latter two is small. Eventually, as $t \to \infty$, all algorithms deviate from the true labeling, since for very large t the Gaussians become so disparate that the true labeling no longer minimizes $Tr[\Sigma_y]$ or yields reasonable clusters that agree with human perception.

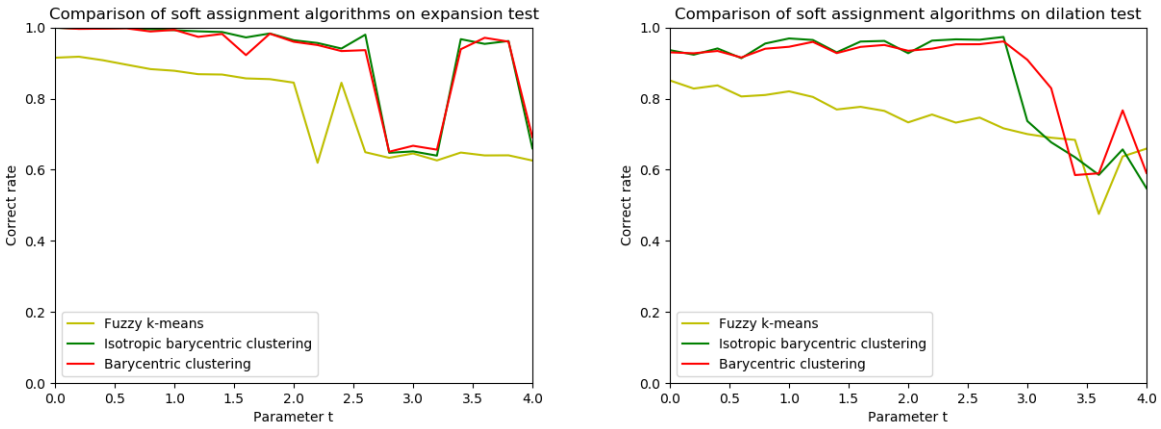

Figure 3: Correctness rates. Left: expansion test. Right: dilation test.

In fact, for the dilation test, the contrast can be seen well before $t = 3.0$. The following is the result for $t = 1.6$, with the shaded regions representing the convex hulls containing the "core points" of each class, defined as $C_k = \{x_i, P_k^i > 1/3\}$. The soft clusters produced by fuzzy k-means exhibit significant overlap, indicating that many sample points are assigned with highly ambiguous probability vectors P^i.

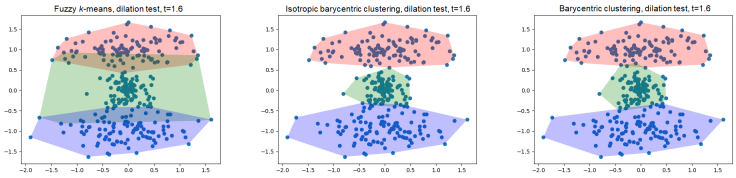

Figure 4: Convex hulls of core points. Only the barycentric clustering algorithms correctly cover each cluster.

6.2 Comparison of hard assignments

We compare next the hard assignment algorithms: k-means [3], Hard barycentric clustering (Algorithm 2), and Barycentric k-means (Algorithm 4). The results on the expansion test and dilation test are plotted below. Again, for each algorithm on each sample set, the objective-minimizing result over 100 trials is selected. The performance comparison is analogous to that of soft assignment.

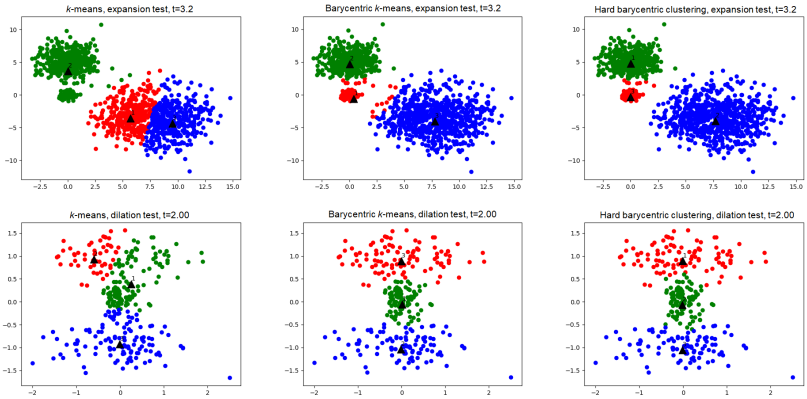

Figure 5: First row: Expansion test with $t = 3.2$. Second row: Dilation test with $t = 2.0$. Left: k-means. Middle: Barycentric k-means. Right: Hard barycentric clustering.

Nevertheless, Hard barycentric clustering is a simplified version of Barycentric clustering, replacing the latter's gradient descent, which moves by small steps, by class reassignment, which hops among the extremal points of $\prod_i \Delta^K$. The correctness rate curves of dilation test indicate that, whereas Barycentric k-means has similar performance as Isotropic barycentric clustering, Hard barycentric clustering is more unstable than Barycentric clustering.

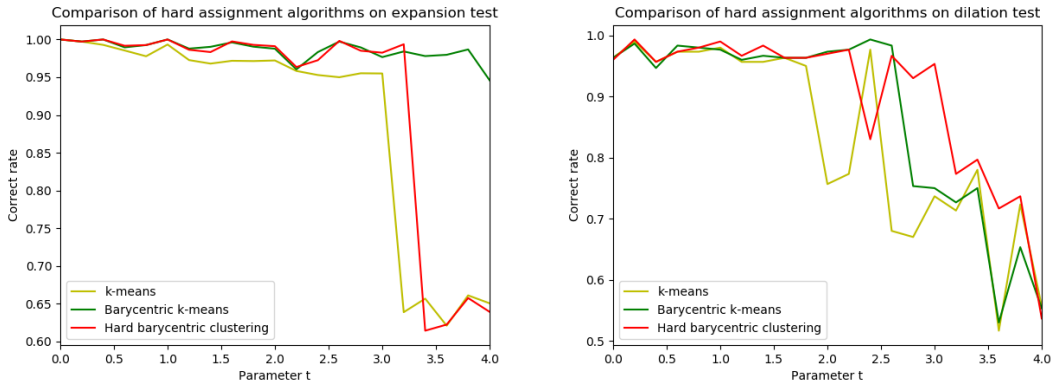

Figure 6: For the expansion test, we have the ranking: k-means < Hard barycentric clustering < Barycentric k-means, and for the dilation test: k-means < Barycentric k-means < Algorithm Hard barycentric clustering.

6.3 Clustering on real-world data

To compare the performance of k-means and our algorithms on real-world problems, we use data sets from the online UCI Machine Learning Repository [11]. These data sets, intended for classification, are provided with labels, which we use to calculate the correctness rates (34). The "Wine" [1] data set classifies wines based on chemical compositions, "Seeds" [7] classifies wheats by the shapes of wheat kernels, "Breast cancer (original)" [36] classifies benign/malign cancers by the shapes of cell nuclei, while "Breast cancer (diagnostic)" [29] classifies them by other cell statistics, "Parkison's" [22] diagnoses the disease by the patients' voice and speech, and "E.coli" [16] classifies proteins by their sequences and structures.

Since the setting for our clustering problem is for data in $\mathbb{R}^d$, the data's categorical attributes as well as entries with missing data are removed. The samples are normalized along each dimension before clustering, since their attributes are often on disparate scales. Again, each algorithm is run 100 times on each sample set, and the objective-minimizing result is selected.

In the following table, for each sample set, the marked entries are the ones with maximum correctness rates among the hard and soft assignment groups.

	Wine	Seeds	Breast cancer (original)	Breast cancer (diagnostic)	Parkinson's	E.coli
Number of classes K	3	3	2	2	2	8
Dimension d	13	7	9	30	22	6
Sample size	178	210	683	569	197	336
Correct rates %						
k-means	96.63	91.90	95.75	91.04	54.36	55.65
Algorithm 4	97.19	91.90	96.34	89.46	53.33	59.82
Algorithm 2	97.19	92.86	96.49	90.69	60.00	59.82
Fuzzy k-means	60.92	74.76	87.19	73.79	54.21	34.01
Algorithm 3	94.34	89.56	96.51	88.78	53.25	57.41
Algorithm 1	91.71	88.73	96.29	89.94	50.91	52.67

Table 1: Our algorithms outperformed (fuzzy) k-means on the majority of data sets. For hard assignment, we have the ranking: k-means < Barycentric k-means < Hard barycentric clustering. For soft assignment we have: fuzzy k-means < Barycentric clustering < Isotropic barycentric clustering, although the influence of the isotropic simplification seems minuscule.

One notable difference between our synthetic tests and these real-world data is that the latter have higher dimensions, which negatively influence fuzzy k-means' performance. Previous studies [35] have shown that,

as dimension increases, the pairwise distances of the sample points become homogeneous, and fuzzy k-means tends to produce uniform assignments: $P_k^i \approx 1/K$. Nevertheless, the barycentric clustering algorithms remain robust, as shown below.

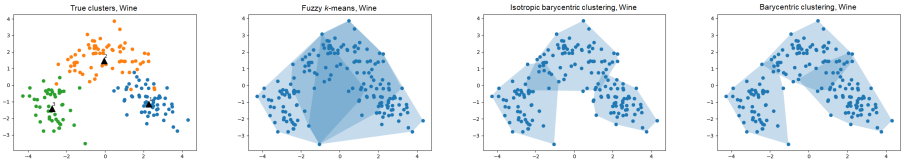

Figure 7: Soft assignment on the "Wine" data set. The sample is projected onto its principal 2-plane, with the true labeling given in the first panel. The following three panels correspond to fuzzy k-means, Isotropic barycentric clustering, and Barycentric clustering. The shaded polygons represent the convex hulls spanned by the "core points" of each cluster, $C_k = \{x_i, P_k^i > 1/4\}$. Fuzzy k-means assigned each x_i with ambiguous probabilities ($P_k^i \approx 1/3$), and many sample points belong to the "cores" of two or more clusters, whereas the assignments produced by the barycentric clustering algorithms are relatively "hard".

6.4 Continuous latent variable and seismic data

Finally, we test Affine factor discovery (Algorithm 5). As discussed in Remark 5.1, the continuous factor discovery problem (27) generalizes principal curves, so it is natural to evaluate Algorithm 5 in terms of its "principal curve", that is, the conditional mean $\overline{x}(z)$. Given data that appears to cluster around one or several curves, the curve $\overline{x}(z)$ should discover these patterns.

We use the earthquake data from [31], which covers more than two thousand earthquakes in the 20th century in the Southeast Asia earthquake zone. The sample $\{x_i\}$ is two dimensional, recording the latitude and longitude of the earthquakes. We apply Affine factor discovery with a fixed number of iterations $T = 50000$, proportion constant $\alpha = 2.5\%$, and learning rate $\eta = 5 \times 10^{-1}$, and we initialize θ_i to be proportional to longitude, which is evidently far from the optimal solution. The curve of conditional means $\overline{x}(z)$ is plotted below.

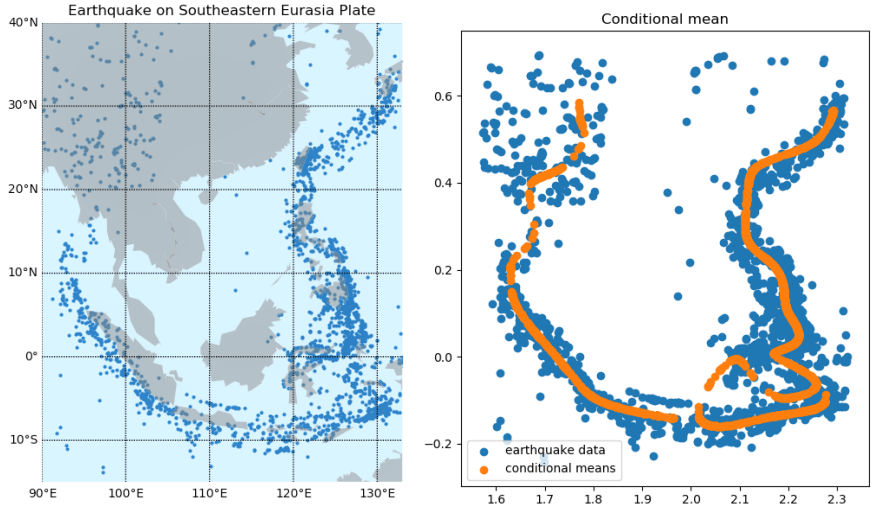

Figure 8: Left: Plot of earthquake data $\{x_i\}$. Right: The conditional means $\overline{x}(z)$ with z sampled from $v(z)$.

The Southeast Asia earthquake zone lies at the intersection of the Australian Plate, Eurasian Plate, Indian Plate, Philippine Plate, Yangtze Plate, Amur Plate, and numerous minor plates and microplates. The tectonic boundaries are complex and cannot be represented by a single curve. Affine factor discovery

automatically solved this problem using piecewise principal curves. Note that even though the latent space $Z = \mathbb{R}$ is connected, the support of the latent distribution, $\text{supp}v(z)$, consists of several disjoint clusters, giving rise to piecewise continuous $\overline{x}(z)$.

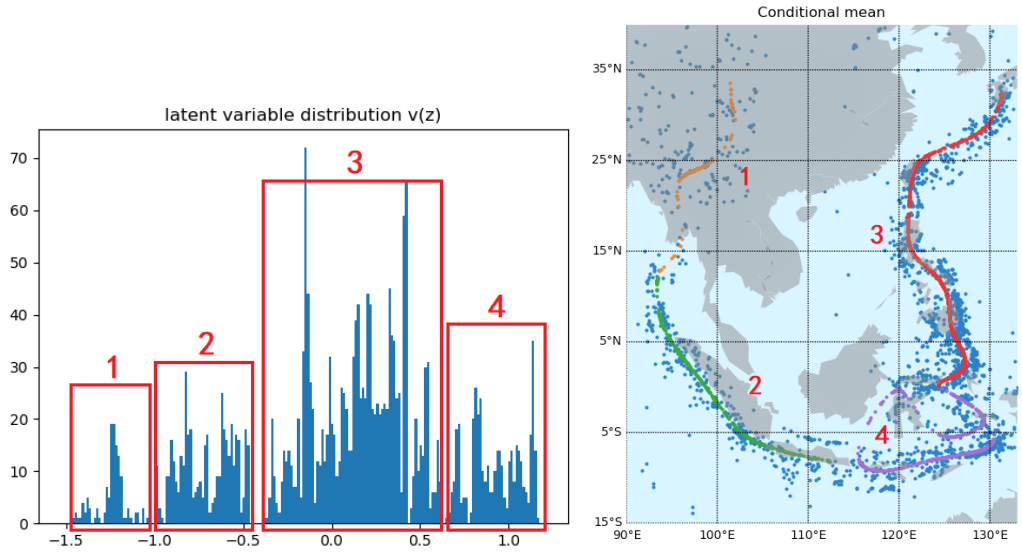

Figure 9: Left: Latent variable distribution $v(z)$, represented by the conditional label means $\{\theta_i\}$. We can roughly identify four disjoint components of $\text{supp}v(z)$, or clusters. Right: The curve $\overline{x}(z)$ of each cluster corresponds to an earthquake belt.

7 Conclusions

This paper developed a novel formalization of clustering in terms of Wasserstein barycenters, exteding ideas proposed in [30]. Clustering can be seen as a special case of the latent variable discovery problem: given data $\sim \rho(x) \in \mathcal{P}(\mathbb{R}^d)$ and a latent variable space Z such as the class labels $\{1, \dots K\}$, find a joint distribution $\rho(x, z) \in \mathcal{P}(\mathbb{R}^d \times Z)$ so that each data point x is given a latent variable assignment $\rho(z|x)$. The resulting conditional distributions $\rho(x|z)$ can be regarded as the clusters.

Conventionally, the fitness of the assignment $\rho(z|x)$ is assessed by some "dissimilarity measure": similar points should be grouped together in one cluster. Instead, we start from the intuition that the goal of learning is to obtain the most information. Specifically, we maximize the information gained from the assignment of z in $\rho(x, z)$. If information gain is characterized by the reduction in uncertainty, and if uncertainty is measured by variance, then the objective of clustering, as well as latent variable discovery in general, is to minimize the x-variance of $\rho(x, z)$, given $\rho(x)$.

To properly define the x-variance of a joint distribution $\rho(x, z)$, we choose the distribution μ that best represents all clusters $\rho(x|z)$, and then define the variance of $\rho(x, z)$ by $Var(\mu)$. Moving from $\rho(x)$ to the representative μ eliminates the data's variability attributable to the latent variable z. The representative μ can be naturally defined as the Wasserstein barycenter from optimal transport theory, which minimizes a transport cost between each cluster $\rho(x|z)$ and μ, which is our setting corresponds to a measure of data-deformation.

Hence, the optimal clustering plan is defined as the latent variable assignment $\rho(z|x)$ that minimizes the variance of the barycenter, $Var(\mu)$. It was shown in [30] that, in the simple setting when all clusters $\rho(x|z)$ are identical up to rigid translations, this formulation becomes exactly k-means. This article,

1. examined the more general scenario when the clusters belong to a location-scale family, and derived the barycentric clustering algorithms (Algorithms 1 and 2) capable of recognizing non-isotropic clusters with varying radii and sizes,

2. devised simplified versions, Isotropic barycentric clustering and Barycentric k-means (Algorithms 3 and 4), which can run efficiently with little loss in robustness, and

3. based on new findings in Wasserstein barycenters with infinitely many marginals, constructed Affine factor discovery (Algorithm 5), which uncovers continuous latent variables and generalizes principal curves and surfaces.

The barycentric clustering algorithms were tested and compared with the more standard k-means and fuzzy k-means on real-world data, as well as artificial data designed to demonstrate the effects of relaxing the various hypotheses underlying the "standard data" [28] for k-means. In nearly all cases, the new procedure in its most general version shows better and more robust performance. Similarly, Affine factor discovery yielded principal curves that reliably summarize the data, automatically generating piecewise continuous curves when needed.

The methodology developed in this article, in addition to its value as a practical tool for clustering and continuous factor discovery, opens the way to further inquiry into various directions:

1. Moving from the standard Euclidean distance to other problem-specific metrics. In particular, one can apply the (squared) geodesic distance of the manifold that underlies the data, such as the Fermat distance introduced in [27].

2. The uncertainty or variability of the data, quantified as variance in this paper, could be generalized to other variability measures. Clustering methods similar to barycentric clustering may stem from different variability measures.

Acknowledgments

The work of E. G. Tabak was partially supported by NSF grant DMS-1715753 and ONR grant N00014-15-1-2355.

References

[1] Aeberhard, S., Coomans, D., and De Vel, O. Comparative analysis of statistical pattern recognition methods in high dimensional settings. *Pattern Recognition 27*, 8 (1994), 1065–1077.

[2] Agueh, M., and Carlier, G. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis 43*, 2 (2011), 904–924.

[3] Alpaydin, E. *Introduction to machine learning*. MIT press, 2009.

[4] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications 441*, 2 (2016), 744–762.

[5] Bezdek, J. C. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[6] Boyd, S., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

[7] Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., and Żak, S. Complete gradient clustering algorithm for features analysis of x-ray images. In *Information technologies in biomedicine*. Springer, 2010, pp. 15–24.

[8] Chernoff, H., et al. *Metric considerations in cluster analysis*. Stanford University. Department of Statistics, 1979.

[9] Cuesta-Albertos, J. A., Matrán-Bea, C., and Tuero-Diaz, A. On lower bounds for the L2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability 9*, 2 (1996), 263–283.

[10] Del Moral, P., and Niclas, A. A taylor expansion of the square root matrix function. *Journal of Mathematical Analysis and Applications 465*, 1 (2018), 259–266.

[11] Dheeru, D., and Taniskidou, E. K. UCI machine learning repository. `https://archive.ics.uci.edu/ml/index.php`. Accessed: 2019-02-15.

[12] Gustafson, D. E., and Kessel, W. C. Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE conference on decision and control including the 17th symposium on adaptive processes* (1979), IEEE, pp. 761–766.

[13] Hastie, T., and Stuetzle, W. Principal curves. *Journal of the American Statistical Association 84*, 406 (1989), 502–516.

[14] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer 27*, 2 (2005), 83–85.

[15] Horn, R. A., and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

[16] Horton, P., and Nakai, K. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb* (1996), vol. 4, pp. 109–115.

[17] Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)* (1942), vol. 37, pp. 199–201.

[18] Kim, Y.-H., and Pass, B. Wasserstein barycenters over riemannian manifolds. *Advances in Mathematics 307* (2017), 640–683.

[19] Krishnapuram, R., and Kim, J. A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms. *IEEE Transactions on Fuzzy systems 7*, 4 (1999), 453–461.

[20] Kuang, M., and Tabak, E. G. Preconditioning of optimal transport. *SIAM Journal on Scientific Computing 39*, 4 (2017), A1793–A1810.

[21] Lang, S. *Undergraduate analysis*. Springer Science & Business Media, 2013, ch. XIII.3 Interchanging Derivatives and Integrals.

[22] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., and Moroz, I. M. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical engineering online 6*, 1 (2007), 23.

[23] Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[24] Pass, B. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis 264*, 4 (2013), 947–963.

[25] Redner, R. A., and Walker, H. F. Mixture densities, maximum likelihood and the em algorithm. *SIAM review 26*, 2 (1984), 195–239.

[26] Rippl, T., Munk, A., and Sturm, A. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis 151* (2016), 90–109.

[27] Sapienza, F., Groisman, P., and Jonckheere, M. Weighted geodesic distance following Fermat's principle. *6th International Conference on Learning Representations* (2018).

[28] Selim, S. Z., and Alsultan, K. A simulated annealing algorithm for the clustering problem. *Pattern recognition 24*, 10 (1991), 1003–1008.

[29] Street, W. N., Wolberg, W. H., and Mangasarian, O. L. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization* (1993), vol. 1905, International Society for Optics and Photonics, pp. 861–870.

[30] Tabak, E. G., and Trigila, G. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics 71*, 1 (2018), 163–199.

[31] USGS. Centennial earthquake catalog. `https://earthquake.usgs.gov/data/centennial/centennial_Y2K.CAT`, 2008. Accessed: 2019-07-30.

[32] Villani, C. *Topics in optimal transportation*. No. 58. American Mathematical Soc., 2003.

[33] Villani, C. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

[34] Wang, W., and Carreira-Perpinán, M. A. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541* (2013).

[35] Winkler, R., Klawonn, F., and Kruse, R. Fuzzy c-means in high dimensional spaces. *International Journal of Fuzzy System Applications (IJFSA) 1*, 1 (2011), 1–16.

[36] Wolberg, W. H., and Mangasarian, O. L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences 87*, 23 (1990), 9193–9196.

[37] Yang, H., and Tabak, E. G. BaryNet: Conditional density estimation and factor discovery through optimal transport. Submitted to CPAM, available in arXiv.

Appendices:

A Proof of Theorem 3.1

The special case when $\mathcal{F}(\mathbb{P}_0)$ is the Gaussian family and when the labels are finite $Z = \{1, \dots K\}$ is proven by Theorem 6.1 of [2]. Theorem 5 of [37] extends it to the case with general Z, provided that the marginal $\rho(x)$ has finite second moment. Essentially, it uses an approximation argument to go from the finite latent space $Z = \{1, \dots K\}$ to general measurable spaces.

Since Corollary 4.5 of [4] has proved Theorem 3.1 in the simplified setting with the finite latent space $Z = \{1, \dots K\}$, the approximation argument of [37] can be applied to generalize Z to general measurable spaces.

B Existence of derivative

Here we establish that $\partial \Sigma_y / \partial P_k^i$, the partial derivatives of the barycenter's covariance implicitly defined by (12) with respect to the membership probabilities $\rho(k|x_i)$, always exist. Theorem 1.1 from [10] is a useful result on matrix derivatives which we restate below. Denote by $\mathcal{S}_d, \mathcal{S}_d^+ \subseteq \mathcal{M}_d$ the linear subspace of symmetric matrices and the cone of positive-definite matrices.

Theorem B.1. *The principal matrix square root function $S \in \mathcal{S}_d^+ \to S^{\frac{1}{2}} \in \mathcal{S}_d^+$ is Fréchet differentiable to any order, and the first order derivative is given by the operator*

$$(\nabla S^{\frac{1}{2}})(H) := \nabla S^{\frac{1}{2}}|_H = \int_0^\infty e^{-S^{\frac{1}{2}}t} \cdot H \cdot e^{-S^{\frac{1}{2}}t} dt$$

such that for any $H \in \mathcal{S}_d$ and $S + hH \in \mathcal{S}_d^+$

$$\lim_{h \to 0} \frac{1}{h} [(S + hH)^{\frac{1}{2}} - S^{\frac{1}{2}} - h(\nabla S^{\frac{1}{2}})(H)] = 0.$$

Now we prove the existence of the partial derivatives through the Implicit Function Theorem.

Theorem B.2. *For $\{\Sigma_1, \dots \Sigma_K\} \subseteq \mathcal{S}_d^+$, the solution Σ_y to the covariance formula (12) depends differentiably on $\Sigma_1, \dots \Sigma_K$.*

Proof. For convenience, define the function

$$F(\Sigma_y, \Sigma_1, \dots \Sigma_K) = \sum_{k=1}^{K} P_k (\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}})^{\frac{1}{2}} - \Sigma_y$$

It is a composition of C^1 functions on $\prod_{i=1}^{K+1} S_d^+$ and thus is C^1. To confirm that the gradient $\nabla_{\Sigma_y} F$ is non-singular, perturb Σ_y along an arbitrary direction $S \in \mathcal{S}_d$,

$$\begin{aligned}
(\nabla_{\Sigma_y} F)(S) =& \sum_{k=1}^{K} P_k \Big(\nabla \big(\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}} \big)^{\frac{1}{2}} \Big) \Big(\big(\nabla \Sigma_y^{\frac{1}{2}} \big)(S) \Sigma_k \Sigma_y^{\frac{1}{2}} + \Sigma_y^{\frac{1}{2}} \Sigma_k \big(\nabla \Sigma_y^{\frac{1}{2}} \big)(S) \Big) - S \\
=& \sum_{k=1}^{K} \int_0^\infty e^{-(\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}})^{\frac{1}{2}} t} \Big[\Big(\int_0^\infty e^{-\Sigma_y^{\frac{1}{2}} u} S e^{-\Sigma_y^{\frac{1}{2}} u} du \Big) \Sigma_k \Sigma_y^{\frac{1}{2}} \\
& + \Sigma_y^{\frac{1}{2}} \Sigma_k \Big(\int_0^\infty e^{-\Sigma_y^{\frac{1}{2}} u} S e^{-\Sigma_y^{\frac{1}{2}} u} du \Big) \Big] e^{-(\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}})^{\frac{1}{2}} t} dt - S
\end{aligned}$$

To evaluate integrals of the form $\int_0^\infty e^{-\Sigma^{\frac{1}{2}} t} S e^{-\Sigma^{\frac{1}{2}} t} dt$, we can apply the eigendecomposition $\Sigma = U D U^T$

$$\int_0^\infty e^{-(U D U^T)^{\frac{1}{2}} t} S e^{-(U D U^T)^{\frac{1}{2}} t} dt = U \Big(\int_0^\infty e^{-D^{\frac{1}{2}} t} U^T S U e^{-D^{\frac{1}{2}} t} dt \Big) U^T = U (T \circ U^T S U) U^T$$

where $\circ$ is Hadamard product and $T_{ij} = \frac{1}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}$.

Thus, using $\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}} = U_k D_k U_k^T$, $\Sigma_y = U_y D_y U_y^T$ and the corresponding T_k, T_y, we obtain

$$(\nabla_{\Sigma_y} F)(S) = \sum_{k=1}^{K} P_k U_k \Big[T_k \circ U_k^T \Big[U_y \Big(T_y \circ U_y^T S U_y \Big) U_y^T \Sigma_k \Sigma_y^{\frac{1}{2}} + \Sigma_y^{\frac{1}{2}} \Sigma_k U_y \Big(T_y \circ U_y^T S U_y \Big) U_y^T \Big] U_k \Big] U_k^T - S$$

To check non-singularity, we set $(\nabla_{\Sigma_y} F)(S) = 0$ and vectorize the equation to disentangle S. We apply the identity that $vec(AXB) = (B^T \otimes A) vec(X)$ where $\otimes$ is the Kronecker product [15]. Meanwhile, vectorizing the Hadamard product yields

$$vec(T \circ X) = diag(vec(T)) vec(X) = (D^{\frac{1}{2}} \otimes I + I \otimes D^{\frac{1}{2}})^{-1} vec(X).$$

Then, the equation $(\nabla_{\Sigma_y} F)(S) = 0$ becomes,

$$\begin{aligned}
vec(S) =& vec \Bigg\{ \sum_{k=1}^{K} P_k U_k \Big[T_k \circ U_k^T \Big[U_y \Big(T_y \circ U_y^T S U_y \Big) U_y^T \Sigma_k \Sigma_y^{\frac{1}{2}} + \Sigma_y^{\frac{1}{2}} \Sigma_k U_y \Big(T_y \circ U_y^T S U_y \Big) U_y^T \Big] U_k \Big] U_k^T \Bigg\} \\
=& \sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1} (U_k^T \otimes U_k^T) \big(\Sigma_y^{\frac{1}{2}} \Sigma_k \otimes I + I \otimes \Sigma_k \Sigma_y^{\frac{1}{2}} \big) \\
& (U_y \otimes U_y)(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1} (U_y^T \otimes U_y^T) vec(S)
\end{aligned}$$

Splitting the term $\Sigma_y^{\frac{1}{2}} \Sigma_k \otimes I = (\Sigma_y^{\frac{1}{2}} \Sigma_k \Sigma_y^{\frac{1}{2}} \otimes I) \cdot (\Sigma_y^{-\frac{1}{2}} \otimes I)$, we get

$$\begin{aligned}
vec(S) = \Bigg\{ & \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1} \big(D_k \otimes I \big) (U_k^T \otimes U_k^T) \Big] \cdot \\
& \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes I)(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1} (U_y^T \otimes U_y^T) \Big] + \\
& \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1} \big(I \otimes D_k \big) (U_k^T \otimes U_k^T) \Big] \cdot
\end{aligned}$$

$$\Big[(U_y \otimes U_y)(I \otimes D_y^{-\frac{1}{2}})(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big]\Big\} vec(S).$$

Applying the three identities

$$D_k \otimes I = (D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})(D_k^{\frac{1}{2}} \otimes I - I \otimes D_k^{\frac{1}{2}}) + I \otimes D_k,$$
$$I \otimes D_k = (D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^2 - D_k^{\frac{1}{2}} \otimes I - 2D_k^{\frac{1}{2}} \otimes D_k^{\frac{1}{2}},$$
$$D_y^{-\frac{1}{2}} \otimes D_y^{-\frac{1}{2}} = (D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(D_y^{-\frac{1}{2}} \otimes I + I \otimes D_y^{-\frac{1}{2}}),$$

we obtain

$$\begin{aligned}
vec(S) = \Big\{ & \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I - I \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \\
& \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes I)(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big] \\
& + \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1}\big(I \otimes D_k\big)(U_k^T \otimes U_k^T)\Big] \cdot \\
& \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes D_y^{-\frac{1}{2}})(U_y^T \otimes U_y^T)\Big]\Big\} vec(S) \\
= \Big\{ & \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes D_y^{-\frac{1}{2}})(U_y^T \otimes U_y^T)\Big] \\
& - \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I)(U_k^T \otimes U_k^T)\Big] \cdot \Big[(U_y \otimes U_y)(I \otimes D_y^{-\frac{1}{2}})(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big] \\
& - \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(I \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes I)(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big] \\
& - \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1}(D_k^{\frac{1}{2}} \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \\
& \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes D_y^{-\frac{1}{2}})(U_y^T \otimes U_y^T)\Big]\Big\} vec(S) \\
= \Big\{ & \big(\Sigma_y \otimes I\big) \cdot \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes I)(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big] \\
& + \big(I \otimes \Sigma_y\big) \cdot \Big[(U_y \otimes U_y)(I \otimes D_y^{-\frac{1}{2}})(D_y^{\frac{1}{2}} \otimes I + I \otimes D_y^{\frac{1}{2}})^{-1}(U_y^T \otimes U_y^T)\Big] \\
& - \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1}(D_k^{\frac{1}{2}} \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \\
& \Big[(U_y \otimes U_y)(D_y^{-\frac{1}{2}} \otimes D_y^{-\frac{1}{2}})(U_y^T \otimes U_y^T)\Big]\Big\} vec(S) \\
= \Big\{ & I - \Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1}(D_k^{\frac{1}{2}} \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \big(\Sigma_y^{-\frac{1}{2}} \otimes \Sigma_y^{-\frac{1}{2}}\big)\Big\} vec(S)
\end{aligned}$$

Hence, it follows that

$$\Big[\sum_{k=1}^{K} P_k (U_k \otimes U_k)(D_k^{\frac{1}{2}} \otimes I + I \otimes D_k^{\frac{1}{2}})^{-1}(D_k^{\frac{1}{2}} \otimes D_k^{\frac{1}{2}})(U_k^T \otimes U_k^T)\Big] \cdot \big(\Sigma_y^{-\frac{1}{2}} \otimes \Sigma_y^{-\frac{1}{2}}\big) vec(S) = O$$

Denote the lengthy matrix by $\big[\sum_{k=1}^{K} P_k Y_k\big] Y$. Since each Y_k and Y are positive-definite, the entire matrix is positive-definite and the equation holds if and only if $S = O$. We can conclude that the gradient

$(\nabla_{\Sigma_y} F)$ is always non-singular, and the implicit function theorem implies that Σ_y depends differentiably on $\{\Sigma_1, \dots \Sigma_K\} \subseteq \prod_{i=1}^K \mathcal{S}_d^+$. □

It follows that since Σ_k depends differentiably on P_k^i, the derivatives $\partial\Sigma_y/\partial P_k^i$ exist.

C Computation of derivative

To solve for the gradient $\nabla_{P^i} Tr[\Sigma_y]$, set $\Lambda_k^i = \partial\Sigma_y/\partial P_k^i \in \mathcal{S}_d$ as an unknown variable. Rather artificially, define the term

$$\begin{aligned}\Omega_k^i &:= \frac{1}{P_k}\frac{\partial (P_k)^2\Sigma_k}{\partial P_k^i} = \frac{1}{P_k}\frac{\partial P_k \sum_{i=1}^N P_k^i (x^i - \overline{x}_k)\cdot(x^i-\overline{x}_k)^T}{\partial P_k^i} \\ &= \Sigma_k + (x_i - \overline{x_k})\cdot(x_i-\overline{x_k})^T + 2\sum_{i=1}^N P_k^i (x^i-\overline{x}_k)\cdot\Big(-\frac{\partial \overline{x}_k}{\partial P_k^i}\Big)^T \\ &= \Sigma_k + (x_i - \overline{x_k})\cdot(x_i-\overline{x_k})^T\end{aligned}$$

Taking partial derivative ∂P_k^i on both sides of the covariance formula (12), we obtain

$$\begin{aligned}\Lambda_k^i &= \sum_{h\neq k}\Big(\nabla\big(\Sigma_y^{\frac12}\big((P_h)^2\Sigma_h\big)\Sigma_y^{\frac12}\big)^{\frac12}\Big)\Big(\big(\nabla\Sigma_y^{\frac12}\big)(\Lambda_k^i)\big((P_h)^2\Sigma_h\big)\Sigma_y^{\frac12} + \Sigma_y^{\frac12}\big((P_h)^2\Sigma_h\big)\big(\nabla\Sigma_y^{\frac12}\big)(\Lambda_k^i)\Big) \\ &\quad + \Big(\nabla\big(\Sigma_y^{\frac12}\big((P_k)^2\Sigma_k\big)\Sigma_y^{\frac12}\big)^{\frac12}\Big)\Big(\big(\nabla\Sigma_y^{\frac12}\big)(\Lambda_k^i)\big((P_k)^2\Sigma_k\big)\Sigma_y^{\frac12} + \Sigma_y^{\frac12}P_k\Omega_k^i\Sigma_y^{\frac12} + \Sigma_y^{\frac12}\big((P_k)^2\Sigma_k\big)\big(\nabla\Sigma_y^{\frac12}\big)(\Lambda_k^i)\Big) \\ &= \sum_{h=1}^K U_h\Big[(P_h)^{-1}T_h\circ U_h^T\Big[U_y\Big(T_y\circ U_y^T\Lambda_k^iU_y\Big)U_y^T(P_h)^2\Sigma_h\Sigma_y^{\frac12} + \Sigma_y^{\frac12}(P_h)^2\Sigma_hU_y\Big(T_y\circ U_y^T\Lambda_k^iU_y\Big)U_y^T\Big]U_h\Big]U_h^T \\ &\quad + U_k\Big[(P_k)^{-1}T_k\circ U_k^T\big(\Sigma_y^{\frac12}P_k\Omega_k^i\Sigma_y^{\frac12} + \Sigma_y^{\frac12}\big)U_k\Big]U_k^T \\ &= \sum_{h=1}^K P_hU_h\Big[T_h\circ U_h^T\Big[U_y\Big(T_y\circ U_y^T\Lambda_k^iU_y\Big)U_y^T\Sigma_h\Sigma_y^{\frac12} + \Sigma_y^{\frac12}\Sigma_hU_y\Big(T_y\circ U_y^T\Lambda_k^iU_y\Big)U_y^T\Big]U_h\Big]U_h^T \\ &\quad + U_k\Big[T_k\circ U_k^T\big(\Sigma_y^{\frac12}\Omega_k^i\Sigma_y^{\frac12} + \Sigma_y^{\frac12}\big)U_k\Big]U_k^T\end{aligned}$$

Vectorize and simplify it by the previous computations,

$$\begin{aligned}vec(\Lambda_k^i) = \Big\{I - \Big[\sum_{h=1}^K P_h(U_h\otimes U_h)(D_h^{\frac12}\otimes I + I\otimes D_h^{\frac12})^{-1}(D_h^{\frac12}\otimes D_h^{\frac12})(U_h^T\otimes U_h^T)\Big]\cdot\big(\Sigma_y^{-\frac12}\otimes\Sigma_y^{-\frac12}\big)\Big\} \\ \cdot vec(\Lambda_k^i) + (U_k\otimes U_k)(D_k^{\frac12}\otimes I + I\otimes D_k^{\frac12})^{-1}(U_k^T\otimes U_k^T)(\Sigma_y^{\frac12}\otimes\Sigma_y^{\frac12})vec(\Omega_k^i)\end{aligned}$$

Therefore,

$$\begin{aligned}vec(\Lambda_k^i) = \big(\Sigma_y^{\frac12}\otimes\Sigma_y^{\frac12}\big) \\ \Big[\sum_{h=1}^K P_h(U_h\otimes U_h)(D_h^{\frac12}\otimes I + I\otimes D_h^{\frac12})^{-1}(D_h^{\frac12}\otimes D_h^{\frac12})(U_h^T\otimes U_h^T)\Big]^{-1} \\ \Big[(U_k\otimes U_k)(D_k^{\frac12}\otimes I + I\otimes D_k^{\frac12})^{-1}(U_k^T\otimes U_k^T)\Big](\Sigma_y^{\frac12}\otimes\Sigma_y^{\frac12})\cdot vec(\Omega_k^i)\end{aligned}$$

Denote the solution by $\Lambda_k^i = vec^{-1}(W_k vec(\Omega_k^i))$, we obtain an expression for the gradient of the objective function

$$\nabla_{P^i} Tr[\Sigma_y] = \sum_{k=1}^K Tr[\Lambda_k^i]\vec{e_k} = \sum_{k=1}^K vec(I)^T\cdot W_k\cdot vec(\Omega_k^i)\vec{e_k}$$

D Proof of Theorem 4.1

By Theorem 1 of [37], the left side of (24) can be written as

$$\int W_2^2\big(\rho(x|z),\mu\big)dv(z) = \min_{\pi\in\Pi_1}\int ||x-y||^2 d\pi(x,y,z) \tag{35}$$
$$\Pi_1 := \big\{\pi\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d\times Z)\mid \pi_{XZ}=\rho(x,z),\pi_{YZ}=\pi_Y\otimes\pi_Z\big\}$$

and the minimum is attained by some π. Similarly, Kantorovich theorem [33] shows that for each $z_1,z_2\in Z$, there exists a coupling $\tilde{\pi}(x_1,x_2|z_1,z_2)$ between $\rho(x_1|z_1),\rho(x_2|z_2)$ that achieves the optimal transport cost

$$W_2^2\big(\rho(x_1|z_1),\rho(x_2|z_2)\big) = \int ||x_1-x_2||^2 d\tilde{\pi}(x_1,x_2|z_1,z_2)$$

By the standard measureable selection theorems (e.g. Corollary 5.22 of [33]), there exists a measureable selection of optimal couplings $\tilde{\pi}(x_1,x_2|z_1,z_2)$ so that we can collect them into a joint measure

$$\tilde{\pi}(x_1,x_2|z_1,z_2)v(z_1)v(z_2) = \tilde{\pi}(x_1,x_2,z_1,z_2)\in\mathcal{P}(\mathbb{R}^d\times Z\times\mathbb{R}^d\times Z)$$

Moreover, it can be made to satisfy

$$\tilde{\pi}(x_1,x_2|z_1,z_2) = \tilde{\pi}(x_2,x_1|z_2,z_1) \tag{36}$$

for all z_1,z_2. It follows that the right side of (24) becomes

$$\frac{1}{2}\iint W_2^2\big(\rho(x_1|z_1),\rho(x_2|z_2)\big)dv(z_1)dv(z_2) = \frac{1}{2}\min_{\tilde{\pi}\in\tilde{\Pi}_2}\int ||x_1-x_2||^2 d\tilde{\pi}(x_1,x_2,z_1,z_2) \tag{37}$$
$$\Pi_2 := \big\{\tilde{\pi}\in\mathcal{P}(\mathbb{R}^d\times Z\times\mathbb{R}^d\times Z)\mid \tilde{\pi}_{X_1Z_1}=\tilde{\pi}_{X_2Z_2}=\rho(x,z),$$
$$\tilde{\pi}_{Z_1Z_2}=v(z)\otimes v(z),\tilde{\pi}(x_1,x_2|z_1,z_2)=\tilde{\pi}(x_2,x_1|z_2,z_1)\big\}$$

Given the solution π of (35), define the joint measure

$$\pi'(x_1,z_1,x_2,z_2,y) := \pi(x_1,z_1|y)\otimes\pi(x_2,z_2|y)\pi_Y(y)$$

Then, its marginal $\pi'' := \pi'_{X_1Z_1X_2Z_2}$ belongs to Π_2. It follows that

$$\begin{aligned}(35) &= \int ||x-y||^2 d\pi\\ &= \frac{1}{2}\int ||x_1-y||^2+||x_2-y||^2-2\langle x_1-y,x_2-y\rangle d\pi'\\ &= \frac{1}{2}\int ||x_1-x_2||^2 d\pi''\\ &\geq (37)\end{aligned}$$

Conversely, given the solution $\tilde{\pi}$ of (37), define the random variables $(X_1,Z_1,X_2,Z_2)\sim\tilde{\pi}$, as well as $X_{z_1}\sim\tilde{\pi}_{X_1Z_1}(x_1|z_1)$ and $X_{z_2}\sim\tilde{\pi}_{X_2Z_2}(x_2|z_2)$. The symmetry (36) implies that we can define

$$Y := \int X_{z_1}dv(z_1) = \int X_{z_2}dv(z_2)$$

Then, define the joint measure $\tilde{\pi}'(x_1,z_1,x_2,z_2,y)$ as the distribution of (X_1,Z_1,X_2,Z_2,Y). It follows that the margin $\tilde{\pi}'' := \tilde{\pi}'_{X_1YZ_1}$ belongs to Π_1 and

$$\begin{aligned}(37) &= \frac{1}{2}\int ||x_1-x_2||^2 d\tilde{\pi} = \frac{1}{2}\int ||x_1-y||^2+||x_2-y||^2-2\langle x_1-y,x_2-y\rangle d\tilde{\pi}'\\ &= \int ||x-y||^2 d\tilde{\pi}'' \geq (35)\end{aligned}$$

Hence, we have the equality (35) = (37).

E Derivation of gradients in Section 5

For each z, the partial derivative of the conditional standard deviation $\sigma(z)$ times the marginal latent distribution $v(z)$ with respect to the conditional latent distribution $\rho_i = \rho(z|x_i)$ is given by

$$\begin{aligned}\frac{\partial\big(\sigma(z)v(z)\big)}{\partial\rho_i} &= \frac{\sum_{j=1}^N ||x_j-\overline{x}(z)||^2\rho_j + ||x_i-\overline{x}(z)||^2\cdot\sum_{j=1}^N\rho_j - 2\sum_{j=1}^N\rho_j(x_j-\overline{x})^T\cdot\frac{\partial\overline{x}(z)}{\partial\rho_i}}{2N\Big(\sum_{j=1}^N\rho_j\cdot\sum_{j=1}^N||x_j-\overline{x}(z)||^2\rho_j\Big)^{\frac{1}{2}}}\\ &= \frac{1}{2N}\Big[\sigma(z)+\frac{||x_i-\overline{x}(z)||^2}{\sigma(z)}\Big]\end{aligned}$$

Meanwhile, for each ρ_i, the partial derivatives of (30) are given by

$$\frac{\partial\rho_i(z)}{\partial\theta_j} = \Big[-\frac{\theta_j}{||\theta||^3}+\frac{N}{\alpha^2}\frac{\theta_j||z-\theta_i||^2}{||\theta||^4}+\mathbf{1}_{i=j}\frac{N}{\alpha^2}\frac{z-\theta_i}{||\theta||^2}\Big]\rho_i(z)$$

So the Jacobian matrix $J_\theta p$ of the vector $p := [\rho_i]$ with respect to the parameter vector $\theta := [\theta_j]$ is

$$\begin{aligned}J_\theta^T p(z) &= \theta\cdot\Big[\big(-\frac{1}{||\theta||^3}+\frac{N}{\alpha^2}\frac{||z-\theta_i||^2}{||\theta||^4}\big)\rho_i(z)\Big]_i^T+\mathrm{diag}\Big(\frac{N}{\alpha^2}\frac{z-\theta_i}{||\theta||^2}\rho_i(z)\Big)\\ &= \theta\cdot\Big[\big(-\frac{1}{||\theta||}+\frac{||z-\theta_i||^2}{\epsilon^2}\big)\frac{\rho_i(z)}{||\theta||^2}\Big]_i^T+\mathrm{diag}\Big(\frac{z-\theta_i}{\epsilon^2}\rho_i(z)\Big)\end{aligned}$$

To show that we can differentiate under the integral sign of σ in (26), we rewrite the standard deviation $\sigma(z)$ into

$$\sigma(z) = \Big[\frac{1}{2}\iint||x-y||^2d\rho(x|z)d\rho(y|z)\Big]^{\frac{1}{2}} = \Big(\frac{1}{2}\sum_{i,j=1}^N||x_i-x_j||^2\rho_i\rho_j\Big)^{\frac{1}{2}} = \frac{\Big(\sum_{i,j=1}^N||x_i-x_j||^2\rho_i\rho_j\Big)^{\frac{1}{2}}}{\sqrt{2}\sum_{i=1}^N\rho_i}$$

Define the matrix $D = [D_{ij}] = [||x_i-x_j||^2]$. Then, the distribution term $dv(z)$ in σ can be cancelled, and we obtain a simpler formula for σ:

$$\sigma = \frac{1}{\sqrt{2}N}\int\Big(\sum_{i,j=1}^N D_{ij}\rho_i\rho_j\Big)^{\frac{1}{2}}dz = \frac{1}{\sqrt{2}N}\int||\sqrt{D}p||dz$$

It is straightforward to show that for each $\theta_0\in\mathbb{R}^N$ ($\theta\neq\mathbf{0}$), there exists some compact neighborhood $\overline{U}$ ($\theta_0\in U^o\subseteq\overline{U}\subseteq\mathbb{R}^N-\{\mathbf{0}\}$) such that the integrand $||\sqrt{D}p||$ and its θ gradient are uniformly bounded by some integrable function. Thereby, Theorem 3.2 of [21] shows that at θ_0, we are allowed to take derivatives under the integral sign:

$$\begin{aligned}\nabla_\theta\sigma &= \int\nabla_\theta(\sigma(z)v(z))dz = \int\frac{\nabla_\theta\big(\sigma(z)v(z)\big)}{v(z)}dv(z)\\ &= \int\frac{1}{v(z)}J_\theta^T p(z)\cdot\nabla_p\big(\sigma(z)v(z)\big)dv(z)\\ \frac{\partial\sigma}{\partial\theta_i} &= \int\frac{1}{2N\cdot v(z)}\Big\{\frac{\theta_i}{||\theta||^2}\sum_{j=1}^N\Big[\sigma(z)+\frac{||x_j-\overline{x}(z)||^2}{\sigma(z)}\Big]\cdot\Big[\frac{||z-\theta_j||^2}{\epsilon^2}-\frac{1}{||\theta||}\Big]\rho_j(z)\\ &\quad+\frac{z-\theta_i}{\epsilon^2}\rho_i(z)\Big[\sigma(z)+\frac{||x_i-\overline{x}(z)||^2}{\sigma(z)}\Big]\Big\}dv(z)\end{aligned}$$

In particular, since the Jacobian $J_\theta p$ consists of a rank-one matrix and a diagonal matrix, computing the above integrand for any z takes only linear time, $O(N)$.