Electronic Journal of Statistics
Vol. 13 (2019) 5088–5119
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1639

Hybrid Wasserstein distance and fast distribution clustering*

Isabella Verdinelli and Larry Wasserman

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
e-mail: isabella@stat.cmu.edu; larry@cmu.edu

Abstract: We define a modified Wasserstein distance for distribution clustering which inherits many of the properties of the Wasserstein distance but which can be estimated easily and computed quickly. The modified distance is the sum of two terms. The first term — which has a closed form — measures the location-scale differences between the distributions. The second term is an approximation that measures the remaining distance after accounting for location-scale differences. We consider several forms of approximation with our main emphasis being a tangent space approximation that can be estimated using nonparametric regression and leads to fast and easy computation of barycenters which otherwise would be very difficult to compute. We evaluate the strengths and weaknesses of this approach on simulated and real examples.

MSC 2010 subject classifications: Primary 62G99; secondary 62H30.
Keywords and phrases: Clustering, Wasserstein.

Received December 2018.

1. Introduction

The Wasserstein distance has attracted much attention lately because it has many appealing properties. ([19, 23]). It is especially useful as a tool for clustering a set of distributions $P_1, \ldots, P_N$ because it captures key shape characteristics of the distributions. But the Wasserstein distance is difficult to compute and difficult to estimate from samples. In this paper we introduce a modified Wasserstein distance that can be estimated and computed quickly.

Wasserstein distance If $X \in \mathbb{R}^d$ is a random vector with distribution P and $Y \in \mathbb{R}^d$ is a random vector with distribution Q then, for $p \geq 1$, the p-Wasserstein distance is defined by

$$W_p(P,Q) \equiv W_p(X,Y) = \left(\inf_J \int ||x-y||^p \, dJ(x,y) \right)^{1/p} \tag{1.1}$$

where the infimum is over all joint distributions J for (X,Y) such that X has marginal P and Y has marginal Q. The minimizer J^* is called the *optimal*

*Thanks to the reviewers for helpful suggestions.

transport plan or *the optimal coupling*. In this paper we will focus on the case $p = 2$ and then we write $W(P, Q)$ or $W(X, Y)$ instead of $W_2(P, Q)$ or $W_2(X, Y)$.

The modified distance that we propose is

$$H^2(X, Y) = W^2(Z_X, Z_Y) + W_\dagger^2(\widetilde{X}, \widetilde{Y}) \tag{1.2}$$

where

$$\begin{aligned} Z_X &\sim N(\mu_X, \Sigma_X), \qquad \widetilde{X} = \Sigma_X^{-1/2}(X - \mu_X) \\ Z_Y &\sim N(\mu_Y, \Sigma_Y), \qquad \widetilde{Y} = \Sigma_Y^{-1/2}(Y - \mu_Y), \end{aligned}$$

$\mu_X = \mathbb{E}[X]$, $\mu_Y = \mathbb{E}[Y]$, $\Sigma_X = \mathrm{Var}[X]$ and $\Sigma_Y = \mathrm{Var}[Y]$ and $W_\dagger$ is a distance between the centered and scaled variables $\widetilde{X}$ and $\widetilde{Y}$. We consider several possible choices for $W_\dagger$. We mainly focus on the case where $W_\dagger(\widetilde{X}, \widetilde{Y})$ is a tangent space approximation to $W(\widetilde{X}, \widetilde{Y})$ as defined by [27]. The details of this tangent approximation are given in Section 3. Our version of the tangent space distance is a bit different than the original implementation as we use a combination of density estimation, permutation smoothing and subsampling. We will call H the *hybrid* distance. We will consider other choices for $W_\dagger(\widetilde{X}, \widetilde{Y})$ in Section 6.

The first term in (1.2) measures location-scale differences between the two distributions, is available in closed form (see equation 2.2) and can be estimated at a $n^{-1/2}$ rate where n is the sample size. The second term captures any remaining non-linear differences.

Distribution clustering As mentioned above, our main motivation is "distribution clustering" which requires repeatedly computing distances. Suppose, for example, that we want to cluster a set of distributions $P_1, \ldots, P_N$. Typically, these are empirical distributions corresponding to datasets $\mathcal{D}_1, \ldots, \mathcal{D}_N$. Given a metric d on the set of probability distributions, we can adapt existing methods — such as hierarchical clustering and k-means clustering — to the problem of clustering distributions. But if we use Wasserstein distance then the calculations become onerous since we need to compute many distances. For example, suppose we want to perform agglomerative hierarchical clustering. First we need to compute the $N(N-1)/2$ distances $W_p(P_i, P_j)$. If we decide to cluster, say P_1 and P_2, we need to combine the corresponding datasets $\mathcal{D}_1$ and $\mathcal{D}_2$. Then we need to compute the distance between the new empirical measure corresponding to $\mathcal{D}_1 \bigcup \mathcal{D}_2$ and all the other distributions. Each stage of the hierarchical clustering involves recalculating the distances. Similarly, if we use k-means clustering we need to iterate between assigning points to clusters and computing centroids. Computing the centroid — also known as a barycenter — with respect to the Wasserstein distance is computationally expensive. In fact, there is apparently no off-the-shelf software to compute the barycenter. Replacing W with H significantly reduces the computational burden without sacrificing accurate clustering. In particular, it drastically simplifies the computation of barycenters.

Related work Our works builds on [27] who introduced the idea of using a tangent space approximation to Wasserstein distance. Their motivation was image processing and their implementation of the idea is quite different than our

version. We also make use of subsampling approximations which were suggested in [24]. The Gaussian approximation $W(Z_X, Z_Y)$ is an example of a linear approximation. Optimal linear approximations to Wasserstein distance are studied in [15]. An important reference on clustering distributions with Wasserstein distance is [7] which not only introduces the idea of distribution clustering in this way, but also proposes a trimming procedure to create robust clusterings. We do not consider robustness in this paper, except in the one-dimensional case where the robust Wasserstein distance has a simple form. We note that combining ideas from [7] with the ideas in this paper is an interesting future direction. Wasserstein clustering is also studied in [12] in the context of hierarchical models. That paper not only clusters distributions but, simultaneously, clusters data within each distribution.

Paper outline In Section 2 we review the Wasserstein distance. In Section 3 we give the details of the proposed modified distance. In Section 4 we define several versions of k-means distribution clustering. Section 5 gives some examples. In Section 6 we briefly explain how our ideas can be used for hierarchical clustering and mean-shift clustering. In Section 7 we discuss some different versions of hybridization. Section 8 contains a discussion and concluding remarks.

2. Wasserstein distance

In this section we give a brief review of the Wasserstein distance and we explain why it is useful for distribution clustering. An excellent reference on Wasserstein distance is [26]. Recall that the Wasserstein distance is defined in equation (1.1).

Explicit expressions In general, there is no closed form expression for W_p. There are three notable exceptions.

(i) When $d = 1$, the distance can be written explicitly as

$$W_p(P, Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p \, dz \right)^{1/p}$$

where $F(x) = P(X \leq x)$ and $G(y) = Q(Y \leq y)$. A robust version is

$$W_p(P, Q) = \left(\int_\delta^{1-\delta} |F^{-1}(z) - G^{-1}(z)|^p \, dz \right)^{1/p}$$

where δ is a trimming constant. We use this version when doing one-dimensional clustering.

(ii) If P_n is the empirical distribution of a dataset $X_1, \ldots, X_n$ and Q_n is the empirical distribution of another dataset $Y_1, \ldots, Y_n$ of the same size, then the distance takes the form

$$W_p^p(P_n, Q_n) = \min_\pi \frac{1}{n} \sum_{i=1}^n ||X_i - Y_{\pi(i)}||^p$$

where the minimum is over all permutations. This minimization can be done using various algorithms such as the Hungarian algorithm ([16]) which takes time $O(n^3)$. When $d = 1$ this further simplifies to

$$W_p(P_n, Q_n) = \left(\frac{1}{n} \sum_{i=1}^{n} |X_{(i)} - Y_{(i)}|^p \right)^{1/p} \tag{2.1}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics.

(iii) The distance has a simple expression in the Gaussian case (or more generally, for location-scale families). Suppose that $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$. Then

$$W^2(P, Q) \equiv W^2(X, Y) = ||\mu_X - \mu_Y||^2 + \mathrm{B}^2(\Sigma_X, \Sigma_Y) \tag{2.2}$$

where

$$\mathrm{B}^2(\Sigma_X, \Sigma_Y) = \mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\mathrm{tr}\left[\left(\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2} \right)^{1/2} \right] \tag{2.3}$$

is the Bures distance ([3]) between Σ_X and Σ_Y. See [10] and [20]. From now on, we refer to (2.2) as the *Gaussian Wasserstein distance*. Even for non-Gaussian data, this metric is useful for capturing location-scale effects.

The Monge distance and transport maps A related distance is the Monge distance defined by

$$\left(\inf_T \int ||x - T(x)||^p dP(x) \right)^{1/p} \tag{2.4}$$

where the infimum is over all maps T such that $T(X) \sim Q$. When a minimizer exists, this corresponds to the Wasserstein distance and the map T is called the *optimal transport map*. In this case the optimal coupling J^* is a degenerate distribution on the set $\{(x, T(x))\}$. But, the minimizer might not exist. Consider $P = \delta_0$ and $Q = (1/2)\delta_{-1} + (1/2)\delta_1$ where δ_a denotes a point mass at a. In this case, there is no map T such that $T(X) \sim Q$. In contrast, an optimal coupling always exists and can be thought of as defining a transport plan that allows the mass to be split and assigned to many locations. A sufficient condition for the existence of a unique optimal transport map is that P be absolutely continuous with respect to Lebesgue measure. In the Gaussian case, the optimal transport map is $L(x) = \mu_Y + \Sigma_Y^{1/2} \Sigma_X^{-1/2} (x - \mu_X)$.

Barycenters Given a set of distributions $P_1, \ldots, P_N$, the *barycenter*, with respect to non-negative weights $\lambda_1, \ldots, \lambda_N$, is defined to be the distribution P that minimizes $\sum_j \lambda_j W^2(P, P_j)$. (In this paper, we will always use $\lambda_j = 1/N$.) There is a substantial literature on finding methods to compute the barycenter. In the special case that each P_j is Gaussian, the barycenter takes a special form. Let $P_j = N(\mu_j, \Sigma_j)$ for $j = 1, \ldots, N$. Then the barycenter P is $N(\mu, \Sigma)$ where $\mu = N^{-1} \sum_j \lambda_j \mu_j$ and Σ is the unique, symmetric, positive definite matrix satisfying the fixed point equation

$$\Sigma = \sum_j \lambda_j (\Sigma^{1/2} \Sigma_j \Sigma^{1/2})^{1/2}. \tag{2.5}$$

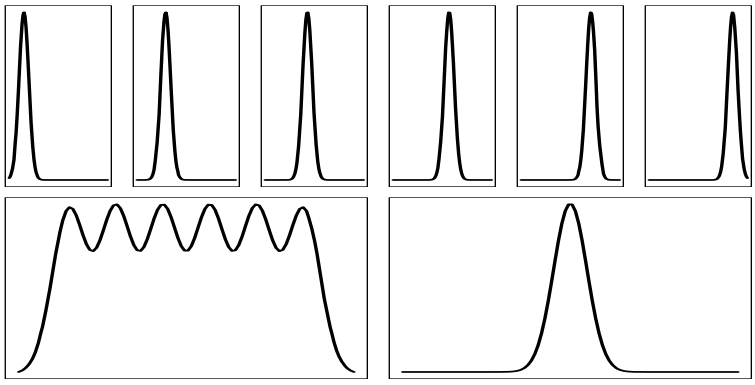

FIG 1. *Top row: Six Gaussian distributions. Bottom left: The usual Euclidean average of the six distributions. Bottom right: The Wasserstein barycenter.*

The barycenter Σ can be found by iterating the equation

$$\Sigma^{(s+1)} \leftarrow (\Sigma^{(s)})^{-1/2} \left(\sum_j \lambda_j \left[(\Sigma^{(s)})^{1/2} \Sigma_j (\Sigma^{(s)})^{1/2} \right]^{1/2} \right)^2 (\Sigma^{(s)})^{-1/2}.$$

The same holds for any location-scale family; see [1]. In one dimension, the barycenter of $P_1, \ldots, P_N$ is the distribution P with cdf F where $F^{-1}(u) = \sum_j \lambda_j F_j^{-1}(u)$.

Key properties There has been a surge of interest in the Wasserstein distance in statistics and machine learning lately. This is because the distance has a number of useful properties. Here we review three key properties, namely: (1) sensitivity to underlying geometry, (2) comparability of discrete and continuous distributions and (3) shape preservation.

1. *The Wasserstein distance is sensitive to the underlying geometry.* Consider the distributions $P_1 = \delta_0, P_2 = \delta_\epsilon$ and $P_3 = \delta_{100}$ and $\epsilon > 0$ is a small positive number. Then $W(P_1, P_2) \approx 0$, $W(P_1, P_3) \approx W(P_2, P_3) \approx 100$. On the other hand, consider the total variation distance d_{TV}. Then $d_{\mathrm{TV}}(P_1, P_2) = d_{\mathrm{TV}}(P_1, P_3) = d_{\mathrm{TV}}(P_2, P_3) = 1$. So the total variation distance fails to capture our intuition that P_1 and P_2 are close while P_3 is far. The same is true for Hellinger distance and Kullback-Leibler distance.
2. *The Wasserstein distance permits direct comparison between discrete and continuous distribution.* If P_1 is continuous and P_2 is discrete, then, for example, $d_{\mathrm{TV}}(P_1, P_2) = 1$. But $W_p(P_1, P_2)$ gives reasonable values. For example, suppose that P_1 is uniform on $[0, 1]$ and P_2 is uniform on $\{1/N, 2/N, \ldots, 1\}$. Then $d_{\mathrm{TV}}(P_1, P_2) = 1$ for all N but $W_p(P_1, P_2) = 1/N$ which again seems quite intuitive.
3. *Shape Preservation.* Suppose we have a set of distributions $P_1, \ldots, P_N$. Recall that the barycenter P minimizes $\sum_j \lambda_j W_2^2(P, P_j)$. The barycenter

P preserves the shape of the distributions. Specifically, if each P_j can be written as a location-scale shift of some distribution P_0, the P is also a location-scale shift of P_0. For example, suppose that $P_1 = N(\mu_1, \Sigma)$ and that $P_2 = N(\mu_2, \Sigma)$. Then the barycenter is $P = N((\mu_1 + \mu_2)/2, \Sigma)$. In contrast, the Euclidean average $(1/2)P_1 + (1/2)P_2$ looks nothing like any of the P_j's. Figure 1 shows a comparison of the Wasserstein barycenter and the usual Euclidean average.

3. The hybrid distance

Let $X \sim P$ and $Y \sim Q$. Define $Z_X \sim N(\mu_X, \Sigma_X)$, $Z_Y \sim N(\mu_Y, \Sigma_Y)$, $\widetilde{X} = \Sigma_X^{-1/2}(X - \mu_X)$, and $\widetilde{Y} = \Sigma_Y^{-1/2}(Y - \mu_Y)$. Our modified distance — which we call the *hybrid distance* — is

$$H^2(X, Y) = W^2(Z_X, Z_Y) + W_\dagger^2(\widetilde{X}, \widetilde{Y}) \tag{3.1}$$

where $W_\dagger$ is described below. The first term $W^2(Z_X, Z_Y)$ has the simple closed form given in (2.2). Given samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$, we can estimate $W^2(Z_X, Z_Y)$ by plugging in sample moments. We then have that $\widehat{W}(Z_X, Z_Y) = W(Z_X, Z_Y) + O_P((n \wedge m)^{-1/2})$; see [20]. We measure the remaining difference by computing the distance between the standardized variables $\widetilde{X}$ and $\widetilde{Y}$ by adapting the method from [27] which we now describe.

Consider a set of distributions $P_1, \ldots, P_N$. Let $\mu_j = \mathbb{E}[X_j]$ and $\Sigma_j = \mathrm{Var}[X_j]$ where $X_j \sim P_j$. Let $Z_j \sim N(\mu_j, \Sigma_j)$ and $\widetilde{X}_j = \Sigma_j^{-1/2}(X_j - \mu_j)$. Let R be a reference measure with density r, (R is defined below.) Define

$$W_\dagger^2(\widetilde{P}_j, \widetilde{P}_k) = \int (\psi_j(z) - \psi_k(z))^2 dz \tag{3.2}$$

where $\psi_j(z) = (T_j(z) - z)\sqrt{r(z)}$, and T_j is the optimal transport map from R to $\widetilde{P}_j$. [27] justify this expression as follows. The set of probability measures endowed with the Wasserstein metric is a Riemannian manifold. Then $\int(\psi_j(z) - \psi_k(z))^2 dz$ is the distance between the projections of $\widetilde{P}_j$ and $\widetilde{P}_k$ onto the tangent space at R; Now

$$\int (\psi_j(z) - \psi_k(z))^2 dz = \int (T_j(z) - T_k(z))^2 dR(z).$$

Hence, if $U_1, \ldots, U_m \sim R$ then

$$\int (\psi_j(z) - \psi_k(z))^2 dz = \frac{1}{m}\sum_{s=1}^{m} (T_j(U_s) - T_k(U_s))^2 + O_P(m^{-1/2}).$$

The hybrid distance is

$$H^2(X_j, X_k) = W^2(Z_j, Z_k) + W_\dagger^2(\widetilde{X}_j, \widetilde{X}_k)$$

$$
\begin{aligned}
&= ||\mu_j - \mu_k||^2 + \mathrm{B}^2(\Sigma_j, \Sigma_k) + \int (\psi_j(z) - \psi_k(z))^2 dz \\
&= ||\mu_j - \mu_k||^2 + \mathrm{B}^2(\Sigma_j, \Sigma_k) + \int (T_j(z) - T_k(z))^2 dR(z) \\
&\approx \underbrace{||\mu_j - \mu_k||^2}_{\text{location}} + \underbrace{\mathrm{B}^2(\Sigma_j, \Sigma_k)}_{\text{scale}} + \underbrace{\frac{1}{m}\sum_i (T_j(U_i) - T_k(U_i))^2}_{\text{shape}} \qquad (3.3)
\end{aligned}
$$

where $U_1, \ldots, U_m \sim R$.

To estimate the distance we need to choose R and estimate T_j. [27], motivated by applications in image processing, suggest using $R = N^{-1}\sum_j P_j$. We take a slightly different approach. Recall that we have datasets $\mathcal{D}_1, \ldots, \mathcal{D}_N$ where $\mathcal{D}_j = \{X_{j1}, \ldots, X_{jn_j}\}$ consists of n_j observations with empirical distribution P_j. Let $\widetilde{D}_j = (\widetilde{X}_{js} : \ 1 \le s \le n_j)$ denote the normalized observations where $\widetilde{X}_{js} = \widehat{\Sigma}_j^{-1/2}(X_{js} - \widehat{\mu}_j)$. The combined dataset $\widetilde{\mathcal{D}} = \bigcup_{j=1}^N \widetilde{\mathcal{D}}_j$ can be regarded as a sample from $\sum_j \pi_j \widetilde{P}_j$ where $\pi_j = n_j / \sum_j n_j$ and $\widetilde{P}_j$ is the distribution of $\widetilde{X}_j$. Let R be the distribution with density r where r is a kernel density estimate obtained from $\widetilde{\mathcal{D}}$ using a simple bandwidth rule such as Silverman's rule [22]. Thus, $R = R_n \star K_h$ (the convolution) where $R_n = \sum_j \pi_j \widetilde{P}_j$ and K_h is a kernel with bandwidth h. This choice of reference measure is simple and smooth.

Remark We have tried a few other reference measures such as Gaussian, Uniform and Cauchy. For the Gaussian and Uniform the results do not change. For the Cauchy, which has thick tails, the results are unstable due to the large outliers and should be avoided. Currently, there is no existing theory about the robustness of the tangent approximation to the choice of reference measure.

Next we have to estimate T_j. Here we use a variation of nonparametric regression that we call *permutation smoothing*. The steps are given in Figure 2. The idea is to sample m observations from each dataset and the reference measure R. The optimal permutation for matching the samples can be found in $O(m^3)$ time. This defines a map T_j from m points drawn from R to m points drawn from $\widetilde{P}_j$. We then extend T_j over the whole space by using b-nearest neighbor regression. It suffices to take $b = 1$ to get a consistent estimator of the transport function. Consistency follows from standard theory ([28]). We can summarize the steps as follows:

Gaussian approximation $\longrightarrow$ subsample $\longrightarrow$ permutation smoothing $\longrightarrow$ tangent approximation.

[27] point out that the tangent space approximation is a well-defined distance and need not be thought of as an approximation to the Wasserstein distance. They show that, even when it does not approximate the Wasserstein distance, it still contains valuable information for comparing distributions.

1. Fix an integer m.
2. Draw a sample $U_1, \ldots, U_m \sim R$ and draw a subsample $\widetilde{X}_{j1}, \ldots, \widetilde{X}_{jm}$ from $\widetilde{\mathcal{D}}_j$.
3. Find the permutation π_j that minimizes $\sum_{i=1}^m ||\widetilde{X}_{ji} - U_{\pi(i)}||^2$. This can be done using the Hungarian algorithm and takes $O(m^3)$ time.
4. Let $V_{j1}, \ldots, V_{jm}$ be the Voronoi tessellation defined by $\widetilde{X}_{j1}, \ldots, \widetilde{X}_{jm}$. Define $\widehat{T}_j(x) = \sum_s U_{\pi_j(s)} I(x \in V_s)$ for all $x \in \{U_1, \ldots, U_m\}$.

FIG 2. *Permutation smoothing algorithm to estimate transport map.*

The idea of using subsamples to approximate Wasserstein distance (rather than using subsamples to estimate the transport map to a reference measure as we are doing) was examined carefully in [24]. For distribution clustering, the subsample size m need not be large. (In our examples we use $m = 100$ but we get similar results even using $m = 20$). This keeps the computation very fast. Also, note that we only ever evaluate $\widehat{T}_j$ on the points $U_1, \ldots, U_m \sim R$.

Remark *[24] suggest estimating Wasserstein distance by averaging over subsamples. Similarly, we could repeat our procedure over several subsamples and average the $\widehat{T}_j$'s. However, we have not found this to be necessary for distribution clustering.* Also, as suggested by a referee, it is possible to average the barycenters over subsamples to improve accuracy although we did not find it necessary in our examples.

Finally, we estimate H^2 by

$$\widehat{H}^2(P_j, P_k) = ||\widehat{\mu}_j - \widehat{\mu}_k||^2 + \mathrm{B}^2(\widehat{\Sigma}_j, \widehat{\Sigma}_k) + \frac{1}{m}\sum_{s=1}^m (\widehat{T}_j(U_s) - \widehat{T}_k(U_s))^2. \quad (3.4)$$

The modified distance retains many properties of Wasserstein distance. The following proposition summarizes these facts. The proof is straightforward and is omitted.

Proposition 1. *The distance H has the following properties.*

1. *H is a metric on the space of distributions with densities and finite second moments. In particular, $H(P, Q) = 0$ if and only if $P = Q$.*
2. *H is exact for Gaussians: if P and Q are Gaussian then $H^2(P, Q) = W^2(P, Q)$.*
3. *If $P_1, \ldots, P_N$ are in a location-scale family then the barycenter is in the same family.*

When clustering we need to repeatedly compute averages. In terms of the Wasserstein distance this corresponds to computing barycenters. To the best of our knowledge, there is no off-the-shelf software to compute Wasserstein barycenters in the multivariate case. But the barycenter with respect to H is easy to compute. Figure 3 summarizes the steps for computing the barycenter.

Barycenters

1. Given $P_1, \ldots, P_N$ and weights $\lambda_1, \ldots, \lambda_N$.
2. Compute the means $\mu_1, \ldots, \mu_N$ and variances $\Sigma_1, \ldots, \Sigma_N$.
3. Use the algorithm in Figure 2 to compute $T_1, \ldots, T_N$.
4. Let $\psi_j(z) = (T_j - z)\sqrt{r(z)}$.
5. Set $\overline{\mu} = \sum_j \lambda_j \mu_j$ and $\overline{\psi} = \sum_j \lambda_j \psi_j$. Compute $\overline{\Sigma}$ from (2.5).
6. Output $(\overline{\mu}, \overline{\Sigma}, \overline{\psi})$.

FIG 3. *Barycenter algorithm.*

Lemma 2. *Let $\overline{P}$ minimize $\sum_j \lambda_j H^2(P, P_j)$. Then $\overline{P}$ can be characterized as follows: $\overline{P}$ is the distribution of the random variable*

$$Y = \overline{\mu} + \overline{\Sigma}^{1/2}\overline{T}(U)$$

where $U \sim R$, $\overline{T}(z) = z + \sum_j \lambda_j (T_j(z) - z)$, $\overline{\mu} = \sum_j \lambda_j \mu_j$ and $\overline{\Sigma}$ is the unique, positive definite matrix satisfying the fixed point equation $\overline{\Sigma} = \sum_j \lambda_j (\overline{\Sigma}^{1/2} \Sigma_j \overline{\Sigma}^{1/2})^{1/2}$.

Proof. Let P be a distribution and let μ be its mean, let Σ be its covariance and let T be the optimal transport map from R to P. From the definition of H we have that

$$\sum_j H^2(P, P_j) = \sum_j ||\mu - \mu_j||^2 + \sum_j B(\Sigma, \Sigma_j) + \sum_j \int (\psi_j(z) - \psi(z))^2 dz$$

where $\psi(z) = (T(z) - z)\sqrt{r(z)}$ and $\psi_j(z) = (T_j(z) - z)\sqrt{r(z)}$. By minimizing each sum separately, the optimal P has mean and variance as stated and its transport map satisfies $\psi = \sum_j \lambda_j \psi_j$ which implies that $\overline{T}(z) = z + \sum_j \lambda_j (T_j(z) - z)$. □

We can regard the triple $(\mu_P, \Sigma_P, \psi_P)$ as a transform ϕ of P where

$$\phi : P \mapsto (\mu_P, \Sigma_P, \psi_P). \tag{3.5}$$

We call ϕ the *hybrid transform*. Note that the transform depends on the reference measure R. Barycenters are computed by averaging each component of the triple separately (with the appropriate fixed point equation used for Σ.) All clustering calculations can be carried out in terms of the representation rather than in terms of the original distribution. The representation is invertible: given a triple $(\mu_P, \Sigma_P, \psi_P)$, the corresponding P is the distribution of the random variable $\mu_P + \Sigma_P^{1/2} T(U)$ with $U \sim R$ and $T(z) = \psi_P(z)/\sqrt{r(z)} + z$. We write $P = \phi^{-1}(\mu_P, \Sigma_P, \psi_P)$.

A further speed-up using hypothesis testing The most expensive step in computing $\widehat{H}^2$ is estimating the transport map T_j. Sometimes $\widetilde{P}_j$ is very close to R and there is no need to compute T_j. We can check this formally by testing $H_0 : \widetilde{P}_j = R$ using any convenient two-sample hypothesis test. If the test rejects, we compute ψ_j. If the test fails to reject we just set $\psi_j = 0$. A convenient nonparametric test with a distribution-free null distribution is the cross-match test ([21]). This idea is illustrated in Section 5.5.

4. *k*-means distribution clustering

Now we are ready to discuss the distribution clustering problem. For concreteness, we focus here on k-means clustering. In Section 6 we briefly consider other clustering methods.

The general outline is as follows. Fix an integer k. Then:

1. Choose k distributions $c_1, \ldots, c_k$ as starting points using k-means^{++} [2].
2. Assign each distribution to the nearest centroid:

$$\mathcal{C}_j = \left\{ P_s : \ d(P_s, c_j) < d(P_s, c_\ell) \text{ for all } \ell \neq j \right\}$$

   where d is a distance.
3. Compute the barycenter c_j of $\mathcal{C}_j$ putting equal weight on each distribution in the cluster.
4. Repeat steps 2 and 3 until convergence.

We consider four versions:

1. **Exact**: Use the Wasserstein distance as the distance at each step. Given the computational burden, we only use the exact method in one-dimensional examples as a point of comparison.
2. **Euclidean**. Compute $W(P_i, P_j)$ for each pair. Use multidimensional scaling to embed the distributions into $\mathbb{R}^a$. Thus we map each P_j to a vector $V_j \in \mathbb{R}^a$. For ease of visualization, we use $a = 2$. We note that the set of distributions equipped with Wasserstein distance is not isometrically embeddable into $\mathbb{R}^a$ and so there will necessarily be some distortion. Once we have the vectors $V_1, \ldots, V_N$ we proceed with k-means clustering as usual.
3. **Gaussian**. We use $W(Z_i, Z_j)$ as an approximation of $W(P_i, P_j)$. Barycenters are computed by averaging the means and by iterating the fixed point equation (2.5) for the variances.
4. **Hybrid.** We use our Hybrid distance $H(P_i, P_j)$. The steps of this approach are in Figure 4.

Of course, the fourth method which uses $d = H$ is the focus of this paper. We include the others for comparison.

Hybrid Wasserstein k-Means

1. Choose starting centroid distributions:
   (a) Choose an integer j at random from $\{1, \ldots, N\}$. Set $\mathcal{C} = \{P_j\}$.
   (b) Compute $H_i^2 = \min_{P \in \mathcal{C}} H^2(P_i, P)$ for $i = 1, \ldots, N$.
   (c) Choose a new $j \in \{1, \ldots, N\}$ with probability proportional to H_j^2. Set $\mathcal{C} = \mathcal{C} \cup \{j\}$.
   (d) Repeat the previous steps until $\mathcal{C} = \{c_1, \ldots, c_k\}$ contains k distributions.
2. Assign each P_j to the closest centroid in $\mathcal{C}$. This defines clusters $C_1, \ldots, C_k$ where $C_j = \{P : \ H(P, c_j) < H(P, c_s), \ s \neq j\}$.
3. Compute the barycenter c_j of C_j.
4. Repeat steps 2 and 3 until convergence.
5. Return $C_1, \ldots, C_k$.

FIG 4. *The hybrid k-means distribution clustering method.*

5. Examples

In this section, we consider some examples. In all the examples that follow, we visualize the set of distributions by applying multidimensional scaling to the pairwise distances. Each point in the plots represents one distribution.

5.1. One dimensional examples

This section presents three one-dimensional examples. These examples are chosen to illustrate the behavior of the four versions of k-means clustering methods, presented in Section 4, and, more specifically, for comparing the performance of the hybrid distance of Section 3 with the procedure based on Wasserstein distance. When $d = 1$, the exact Wasserstein clustering can be easily computed. (The algorithm is in the appendix.)

The first example consists of 15 Normally distributed data sets of size $n = 100$, all with variance $\sigma^2 = 1$. There are three groups of five data sets each with means close together. This specific simple example is chosen to show that all four versions of k-means clustering methods identify the clusters correctly, in straightforward cases. Figure 5 shows the multidimensional scaling projection of the pairwise distances, with different colors used to indicate the clusters.

Our second and third examples are chosen to illustrate that less clean clusters of data-sets cannot be well identified by the Gaussian approximation. Example 2 consists of two clusters. The first cluster has twenty normal distributions with mean 0 and variance 1. The second cluster has twenty distributions each of which has the form $(1/2)\delta_{-1} + (1/2)\delta_1$. This means that the first and second moments are also 0 and 1 so the Gaussian approximation should be unable to find the two clusters.

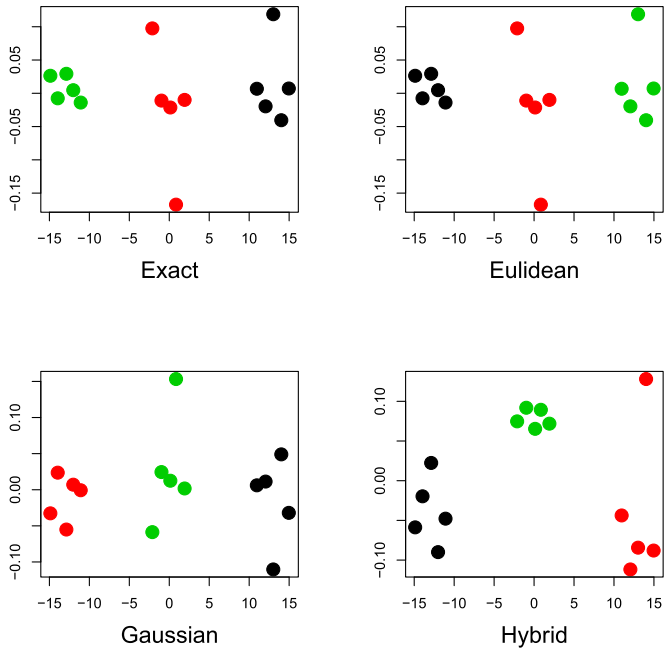

FIG 5. *Example 1. The plots show pairwise distances as represented by multi-dimensional scaling. The plots show three clusters of five normal distribution each. All four methods are accurate.*

Figure 6 shows the distances computed with the Gaussian approximation and the Hybrid method, versus the Wasserstein procedure. The line $x = y$ is included for illustrating the difference between the distances. This plot complements the following Figure 7, showing the clustering obtained in this example, and illustrates that distances from the Gaussian approximation are a poor approximation. Figure 7 shows indeed that, as expected, the Gaussian approximation does not work well.

Our third example consists of three clusters of distributions, each being a mixture of normal distributions. This example also shows that the Gaussian approximation does not identify the clusters. As in Example 2, we first display the plots of the distances, in Figure 8, and then Figure 9 with plots of the multidimensional scaling projections. This confirms the issue noted earlier for the Gaussian approximation. Note that the clusters displayed in some plots are very close together and not easily detectable in the picture. But it is clear that the Gaussian approximation does not fare well.

5.2. Multivariate examples

In this section we consider three artificial data-sets in two dimensions. We no longer consider the exact Wasserstein method which is computationally pro-

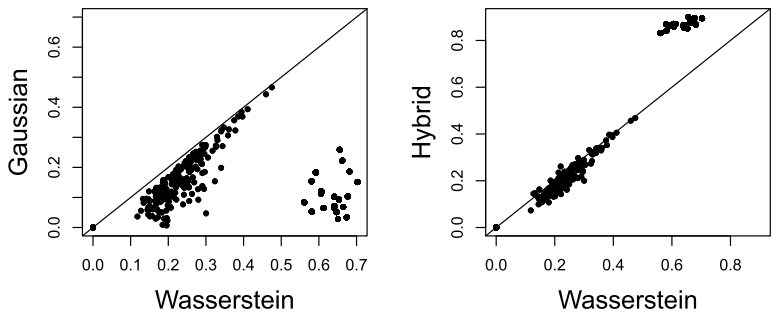

FIG 6. *Example 2. Plots of Gaussian and Hybrid distances of two data sets versus Wasserstein distance. The line across the plots is the line $x = y$. The left hand side plot shows the distances among distributions computed with the normal approximations are far from the Wasserstein distance.*

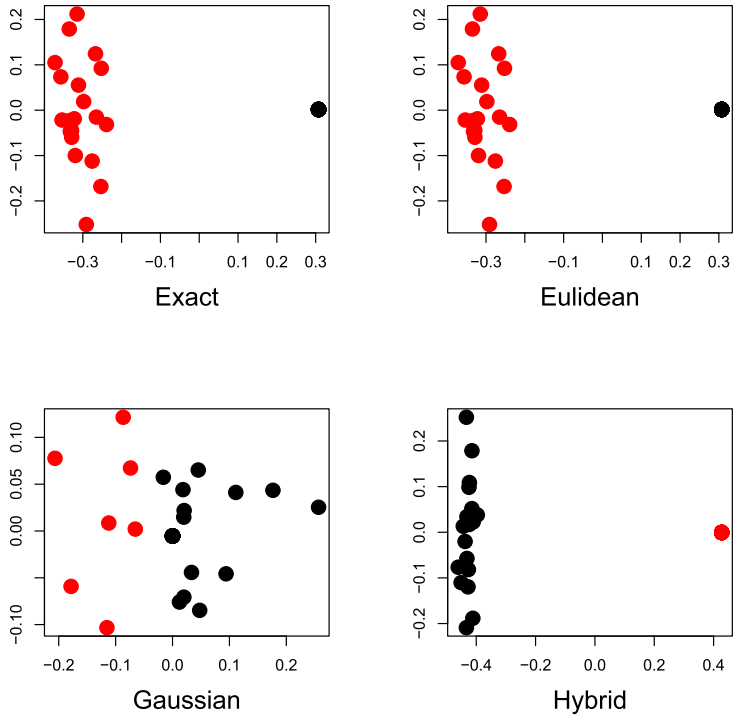

FIG 7. *Example 2. While the Gaussian approximation does detect the two clusters, they are not well identified. The projection plots show random points in the plane. The other three methods, instead, identify the clusters correctly.*

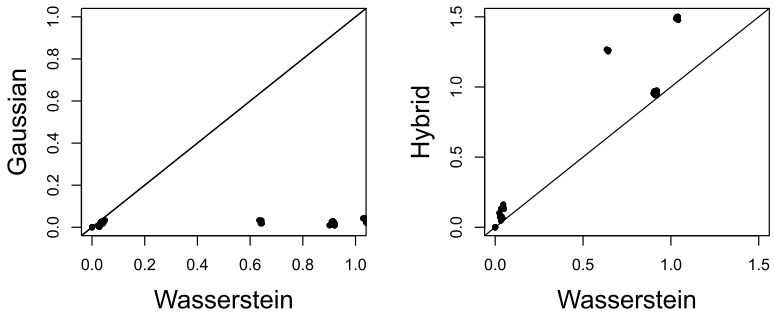

FIG 8. *Example 3. Plots of Gaussian and Hybrid distances of three data sets versus Wasserstein distance. The left hand side plot also shows distances among distributions computed by the Gaussian approximations are far from the Wasserstein distance.*

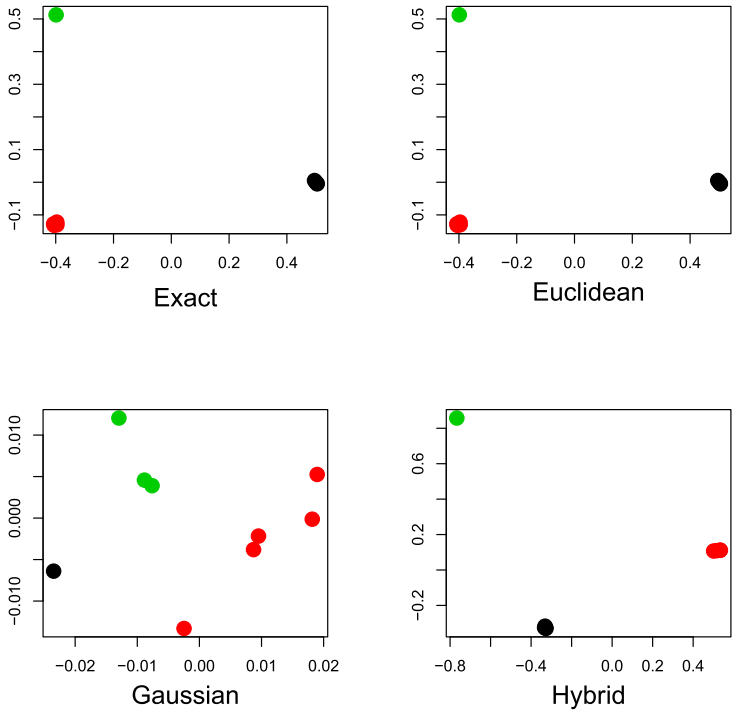

FIG 9. *Example 3. Clusters for mixtures of Normal distributions. The Gaussian approximation does not performs well.*

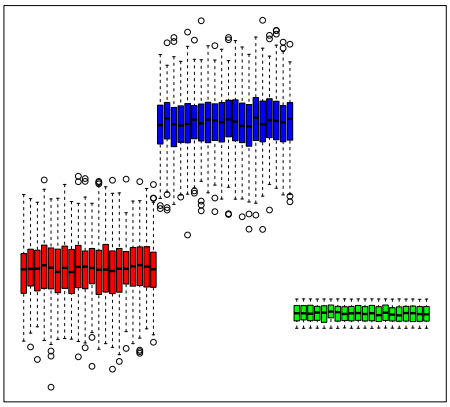

FIG 10. *Boxplots of first coordinates in the bivariate example 1.*

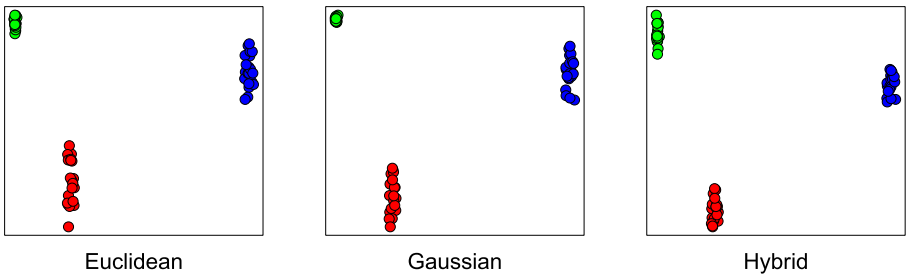

FIG 11. *Bivariate Example 1. Clusters obtained using the procedures Gaussian and Hybrid.*

hibitive. As mentioned earlier, the computation of the barycenter in multivariate cases is still a research problem. We obtain, instead, clustering from the other three methods of Section 4. For each example, we will plot the pairwise distances between datasets using multi-dimensional scaling.

The first example consists of 40 bivariate Normal distributions, with $n = 100$ observations each. Twenty of them have mean $(0, 0)$ and variance I (where I is the identity matrix), the other twenty have mean $(5, 5)$ and variance I. We also include twenty bivariate uniform distributions each with 100 data points, for a total of three distributional clusters.

Figure 10 displays the boxplots of the first coordinates of the three data sets. The clusters appear to be well separated as expected. All three methods performs well, as it can be seen from the clusters obtained from the Euclidean, Gaussian, and Hybrid methods in Figure 11.

The second example consists of four groups of distributions which are uniform on circles, each with $n = 100$ data points. The circles' centers and radii are randomly selected. More specifically the centers are generated from uniform distributions, with various ranges, as are their radii. Figure 12 displays the data

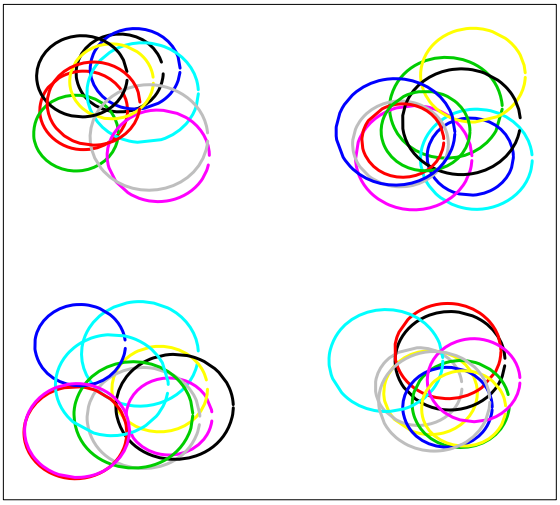

FIG 12. *Bivariate Example 2. The datasets.*

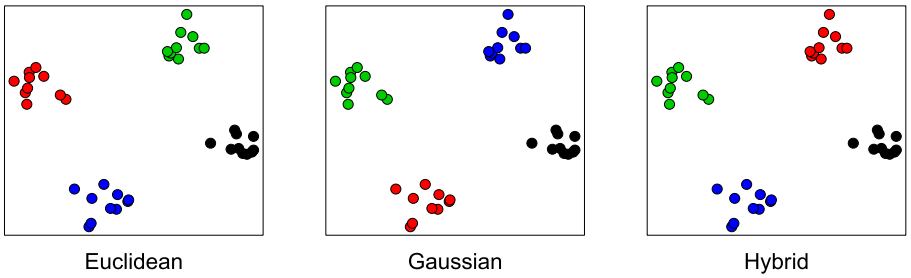

FIG 13. *Bivariate Example 2. Clusters.*

that form four well separated clusters. Figure 13 shows the results for the three methods of clustering. All of them identify the four clusters correctly.

Our third example consists of two groups of distributions. One group of 50 distributions, with 100 observations each, are uniformly distributed on a circle and the 50 distributions in the other group are normal with mean 0 and variance 1. Figure 14 shows the boxplot of their first coordinates. Figure 15 presents the results for the three clustering methods. As expected the Gaussian procedure performs quite poorly.

5.3. Multivariate example: cytometric data

We now apply the Gaussian and Hybrid methods for clustering distributions, presented in Sections 3 and 4 to a collection of data sets from cytometric genetic research by [17]. The authors obtained fluorescence intensity measures

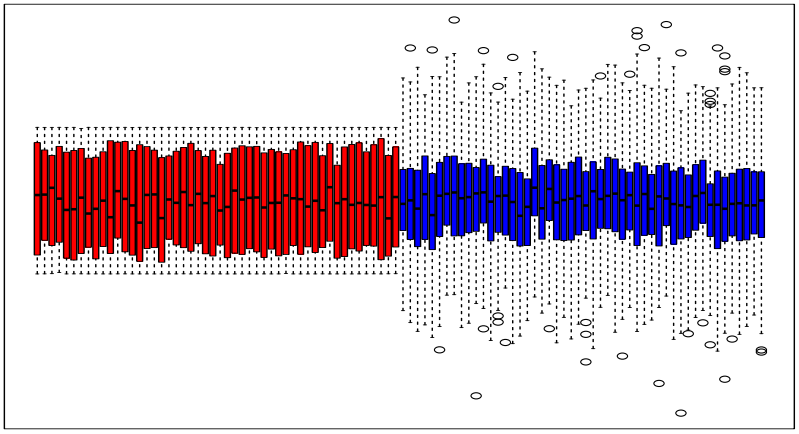

FIG 14. *Bivariate Example 3. Boxplots of first coordinates.*

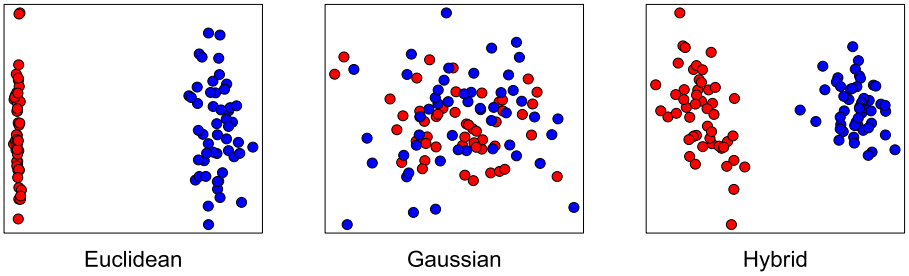

FIG 15. *Bivariate Example 3. Clusters.*

of fluorophore-conjugated reagents on whole blood, stained with an antibody marker. The data record the luminosity of four proteins linked to the T-cells, that are part of the adaptive immune system. Luminosity was measured on the proteins SLP76, ZAP70, CD4, and CD45RA before and after stimulation with the antibody anti-CD3. Two sets of blood samples were collected. One sample consists in 13 four-dimension data-sets, stained prior to anti-CD3 stimulation. The second sample, of 30 more data-sets, stained five minutes after stimulation, for a total of 43 data-sets in four dimension. Figure 16 shows pairwise scatterplots for the first dataset.

In this example, for reasons explained in Section 5.4, we searched for six clusters. Figure 17 show the boxplots with six colors to show the clusters.

The boxplots are color coded by our clustering. Note that the boxplots of the SLP6 and ZAP70 proteins correspond nicely to the six clusters. The third and fourth proteins, CD4 and CD45RA, in Figure 17, instead do not seem to be well clustered. We conclude that the clustering information is deriving mainly from SLP6 and ZAP70.

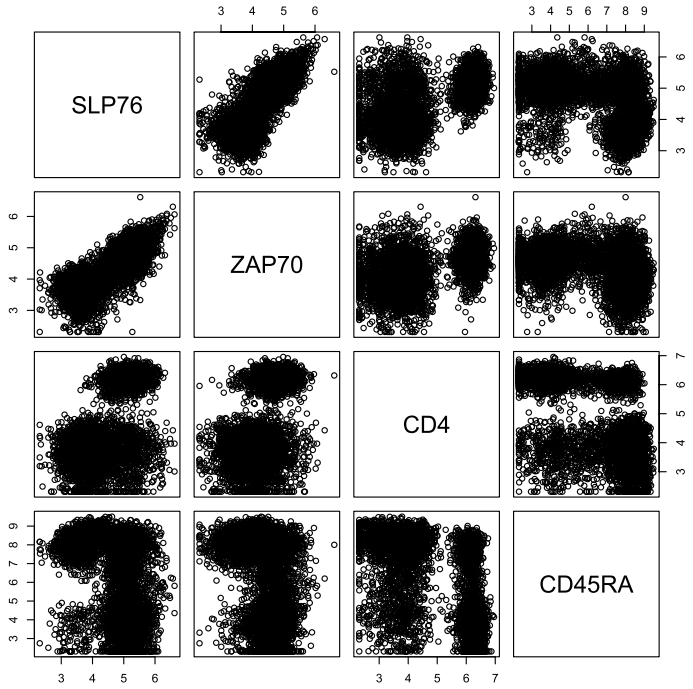

FIG 16. *Cytometric Data. Scatterplots of joint luminosity recorded for the first distribution of 13 blood samples stained prior to the anti-CD3 stimulation.*

The following Figure 18 shows the six clusters obtained with the Gaussian and the Hybrid procedures. Although the clusters appear to be not so well separated, we recall that multidimensional scaling projections from several dimensions to two might not be very accurate. Our choice, in searching for six clusters, has been suggested by the plots in Figure 20 of Section 5.4, where the issue of choosing the number of clusters is examined, and reinforced by the boxplots in Figure 17.

Cytometric data are often modeled as mixtures of Normals. The mixture structure is evident in some of the boxplots. This suggests replacing the first term of the hybrid distance with a metric specially designed for mixtures. Thus we would use $H^2(X,Y) = W^2_{\text{mix}}(X,Y) + W^2_{\dagger}(\widetilde{X},\widetilde{Y})$. Such a hybrid distance would make use of the mixture structure of the cytometric data. But, there is no known simple formula for computing the Wasserstein distance between mixtures. Moreover, it is not clear how to define $\widetilde{X}$ and $\widetilde{Y}$ in this case. Currently, we do not know how to solve this problem. This direction deserves further scrutiny.

5.4. Choosing k: elbows

The issue of choosing k in k-means clustering is a well-studied problem. Perhaps the oldest and simplest method is to plot the sums of squares versus k and look

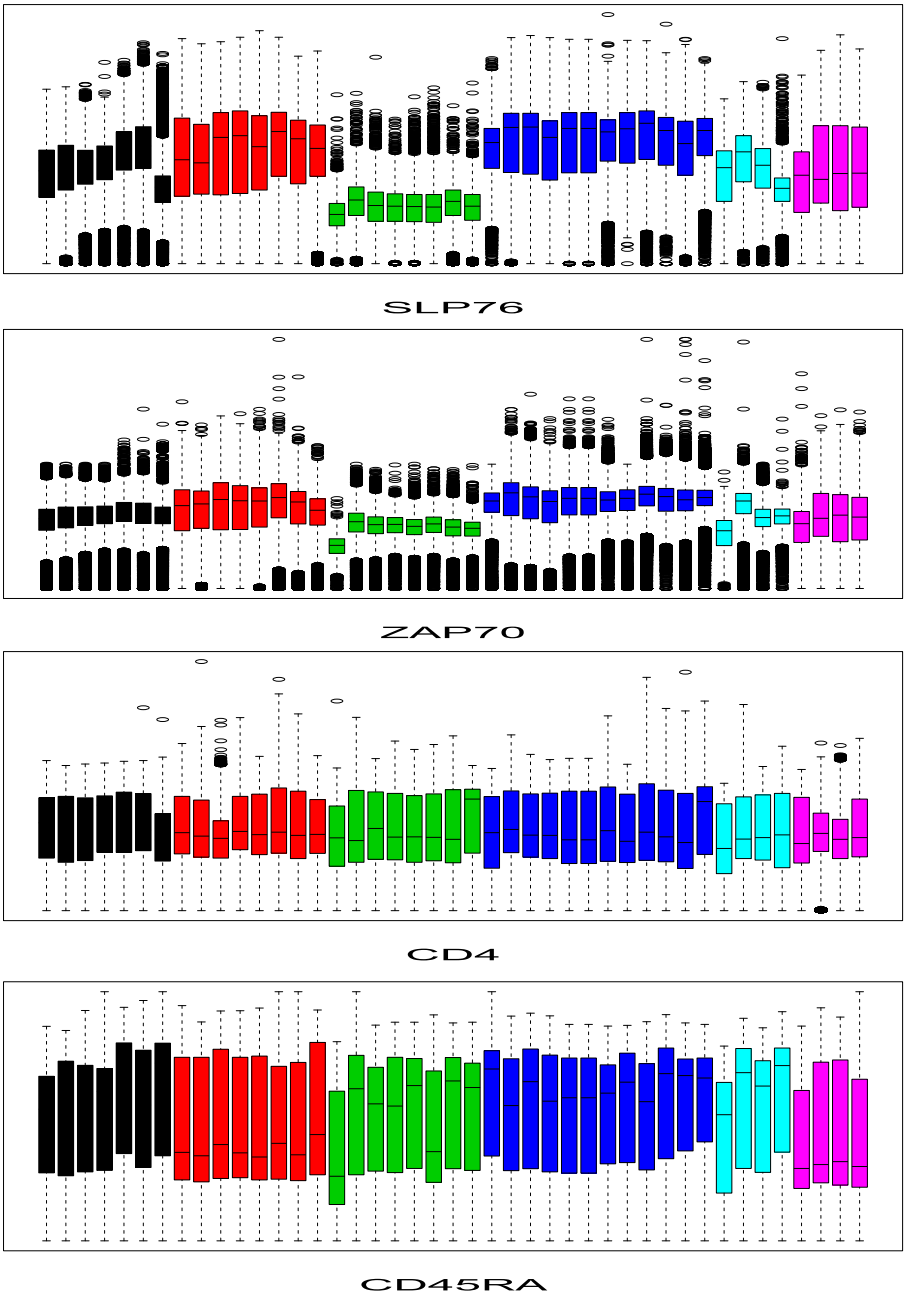

FIG 17. *Cytometric Data. Boxplots of SLP76, ZAP70, CD4 and CD45RA.*

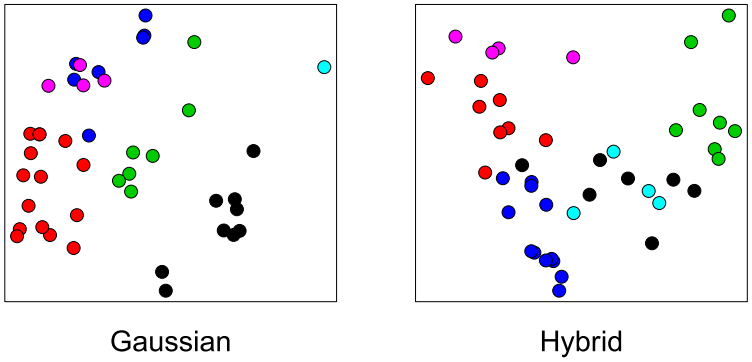

FIG 18. *Cytometric Data. Clusters obtained using the procedures Gaussian and Hybrid.*

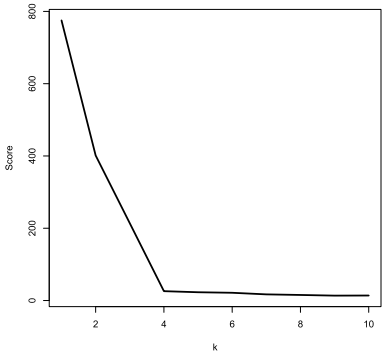

FIG 19. *Plot of* $\sum_{j=1}^{k} \sum_{s \in \mathcal{C}_j} H^2(P_s, c_j)$ *for* $k = 1, \ldots, 10$ *for the example with four groups of circles. Note the pronounced elbow at* $k = 4$.

for an elbow. For distribution clustering, we plot $S_k = \sum_j \sum_{s \in \mathcal{C}_j} H^2(P_s, c_j)$ versus k. In some cases, the elbow is clearer if we plot $1/S_k$ versus k. Figure 19 shows the plot of S_k versus k for the second example in Section 5.2 where data were generated as four groups of circles. We see a clear elbow at $k = 4$. Figure 20 shows another plot of cluster quality versus k. In this case, for the cytometric data sets, we plot $1/\sum_j \sum_{s \in \mathcal{C}_j} H^2(P_s, c_j)$ as the inverse plot shows a clearer signal. The elbow is at $k = 6$.

5.5. The pre-testing speedup

We generated 100 observations from each of 30 multivariate Normal distributions that lie in three distinct groups. After computing the rescaled data $\widetilde{X}$, we apply the cross-match test from [21] to see if the rescaled data differ significantly from the sample from the reference distribution. We used level $\alpha = 0.10$. The test does not reject and we set $\psi_j = 0$ which saves considerable computing time.

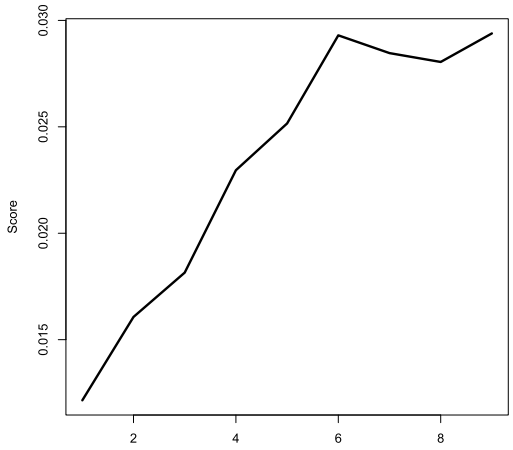

FIG 20. *Plot of* $1/\sum_{j=1}^{k}\sum_{s\in\mathcal{C}_j} H^2(P_s, c_j)$ *for* $k = 1, \ldots, 9$ *for the cytometric data. Note the elbow at* $k = 6$.

Typically, the clustering is perfect (adjusted Rand index of 1). But when the same procedure is applied to the Normal and circles data, the null is almost always rejected (as expected) and we need to do the full computations.

6. Other clustering methods

There are, of course, many other clustering methods besides k-means clustering. In this section we give a sense of how the hybrid distance can be used for two other clustering methods.

Hierarchical clustering The hybrid distance can easily be used for hierarchical clustering. These are the steps for single linkage hierarchical clustering:

1. Compute the transform $(\mu_j, \Sigma_j, \psi_j) = \phi(P_j)$ for each distribution.
2. Compute the pairwise distances $H_{jk} = H(P_j, P_k)$.
3. Find the pair (j, k) that minimizes $H_{j,k}$. Merge P_j and P_k into a single distribution $P_{jk} = \pi_{jk}P_j + (1 - \pi_{jk})P_k$ where $\pi_{jk} = n_j/(n_j + n_k)$ and n_j and n_k are the sample sizes of the two datasets.
4. Compute the representation $(\mu_{jk}, \Sigma_{jk}, \psi_{jk})$ of P_{jk} by computing the barycenter of P_j and P_k with weights π_{jk} and $1 - \pi_{jk}$.
5. Repeat steps 2-4 until all distributions are merged.

By working with the hybrid representations we do not have to keep recomputing Wasserstein distances.

Mean-shift and medoid-shift clustering Next we discuss mean shift clustering and medoid shift clustering [6, 4, 14, 5]. First we review these methods when applied to a single dataset. Give a sample $Y_1, \ldots, Y_n$ from a distribution P

with density p, mean shift clustering works by first computing an estimate $\widehat{p}$ of p. Let $m_1, \ldots, m_k$ be the modes of $\widehat{p}$. Given any point y, if we follow the gradient of $\widehat{p}$ starting at y we will end up at one of the modes. In this way the modes define a partition of the sample space. This partition defined the mode-based clustering. There is a simple iterative algorithm called the mean-shift algorithm to implement this idea. Recently, [8] showed that if we use r-nearest neighbor density estimation, then the iterative algorithm takes the following form. Pick any starting value x. Define $y^{(0)}, y^{(1)}, \ldots$ by $y^{(0)} = y$ and

$$y^{(s)} = \frac{1}{r} \sum_{N_r(y^{(s-1)})} Y_i$$

where $N_r(y)$ denotes the r-nearest neighbors of the point y. This iteration leads to a mode of the estimated density. This assigns any point y to a mode. The iteration can be applied to any set of starting points although these starting points are usually taken to be the data points.

Returning to distribution clustering, we have a set of distributions $P_1, \ldots, P_N$ which we now regard as a sample from a measure Π on the space of distributions. Unfortunately, Π does not have a density in any meaningful sense. Nonetheless, we can formally apply the mean shift clustering algorithm. Given any P, let $N_r(P)$ denote the r closest distributions in $\{P_1, \ldots, P_N\}$ to P under the hybrid distance. Let $\text{Bary}(N_r(P))$ denote the hybrid barycenter of the distributions in $N_r(P)$. We then define $P^{(0)} = P$ and

$$P^{(s)} = \text{Bary}(N_r(P^{(s-1)})).$$

A faster approach, which avoids computing the barycenter, is medoid-shift clustering. In this case, we begin by estimating $\rho(P_j) = 1/d(P, P_r(P))$ where $P_r(P)$ is the r^{th} nearest neighbor. This can be thought of as a pseudo-density. We move P to the distribution in $N_r(P)$ with highest pseudo-density. This is repeated until there is no change. This can be regarded as an approximation to mean-shift clustering.

As an example, we consider a collection of 80, two-dimensional datasets. Each dataset has $n = 100$ observations from a bivariate Normal. We construct the data so that there are four, well-defined clusters. Figure 21 shows medoid-shift clustering using the hybrid distance. The plots correspond to $r = 2, 10, 15$ and 20 nearest neighbors. The plots show the MDS of the data sets and the paths of the data as the clustering proceeds. The red dots show the final destinations, that is, the pseudo-modes. When r reaches 20 we start to oversmooth and end up with 2 clusters. There is a large range of values of r that lead to the correct answer of 4 clusters.

Currently, we do not have a theoretical basis for applying the mean shift algorithm to distributions. We believe that it may be possible to define a pseudo-density on the space of distributions similar to the approach used by [9] in the context of clustering functional data. We conjecture that density clustering in

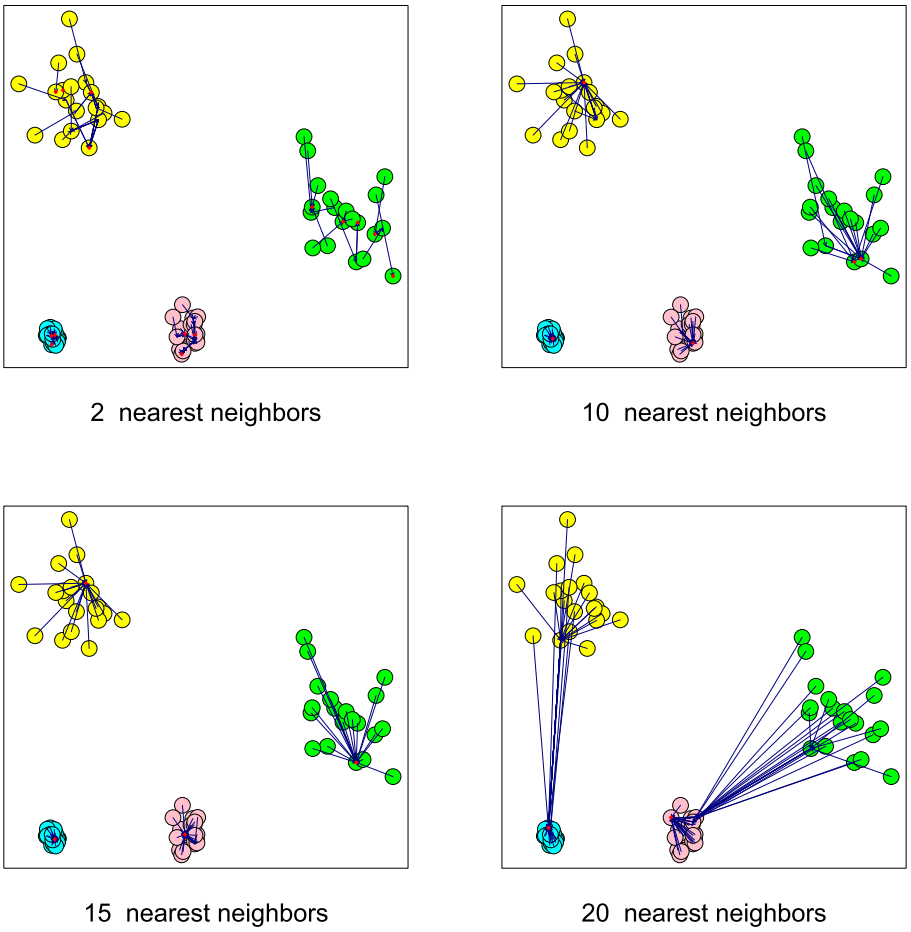

FIG 21. *Medoid-Shift Clustering using the hybrid distance. The plots show the MDS of the data sets and the paths of the data as the clustering proceeds for* $r = 2, 10, 15$ *and 20 nearest neighbors.*

Wasserstein space can be similarly justified by regarding the density estimator as estimating a pseudo-density. For example, if P is a random distribution, one can define the pseudo-density at P_0 by $\mathbb{P}(P \in B(P_0, h))$ where $B(P_0, h) = \{P : H(P_0, P) \leq h\}$. We leave the details for future work.

7. Other hybridizations

The hybrid approach is very flexible. There are many ways to generalize the method. In this section, we consider other hybrid distances of the form,

$$H^2(X, Y) = W^2(Z_X, Z_Y) + W_\dagger^2(\widetilde{X}, \widetilde{Y}).$$

Specifically, we explore other choices besides the tangent distance for $W_\dagger$. In each case, we want an approximate distance that can be computed quickly.

Marginal distance Here we take the average of the easily computed one-dimensional distances

$$W_\dagger^2(X,Y) = \sum_{j=1}^{d} W^2(X(j), Y(j))$$

where $X(j)$ and $Y(j)$ are the j^{th} components of X and Y. It then follows that

$$W_\dagger^2(X,Y) = \sum_{j=1}^{d} \int_0^1 |F_j^{-1}(u) - G_j^{-1}(u)|^2 du$$

where $F_j(t) = P(X(j) \le t)$ and $G_j(t) = Q(Y(j) \le t)$. Hence,

$$H^2(X,Y) = ||\mu_X - \mu_Y||^2 + \mathrm{B}^2(\Sigma_X, \Sigma_Y) + \sum_{j=1}^{d} \int_0^1 |F_j^{-1}(u) - G_j^{-1}(u)|^2 du.$$

The estimate is

$$\widehat{H}^2(X,Y) = ||\widehat{\mu}_X - \widehat{\mu}_Y||^2 + \mathrm{B}^2(\widehat{\Sigma}_X, \widehat{\Sigma}_Y) + \sum_{j=1}^{d} \int_0^1 |\widehat{F}_j^{-1}(u) - \widehat{G}_j^{-1}(u)|^2 du.$$

Assuming we have samples m observations from each dataset, we have the further simplification that

$$\int_0^1 |\widehat{F}_j^{-1}(u) - \widehat{G}_j^{-1}(u)|^2 du = \frac{1}{m} \sum_i (X_{(i)}(j) - Y_{(i)}(j))^2$$

where $X_{(1)}(j), \ldots, X_{(n)}(j)$ and $Y_{(1)}(j), \ldots, Y_{(n)}(j)$ are the order statistics for the j^{th} coordinate of X and Y, respectively.

Next we consider an example. We generate 100 datasets. The first 50 are standard bivariate Normal. The second 50 are uniform on a circle, scaled to have the same mean and covariance as the Normal. The left plot of Figure 22 shows the Gaussian-Wasserstein distances using multidimensional scaling. The colors indicate Normal (black) and uniform on circles (red). The Gaussian-Wasserstein distance cannot distinguish the two types of datasets. The right plot shows the marginal hybrid distance. Here, the two types of datasets are clearly distinguished.

Transformed Gaussian approximation Recall that the Gaussian-Wasserstein distance is

$$||\mu_X - \mu_Y||^2 + B^2(\Sigma_X, \Sigma_Y) \equiv G^2(X,Y).$$

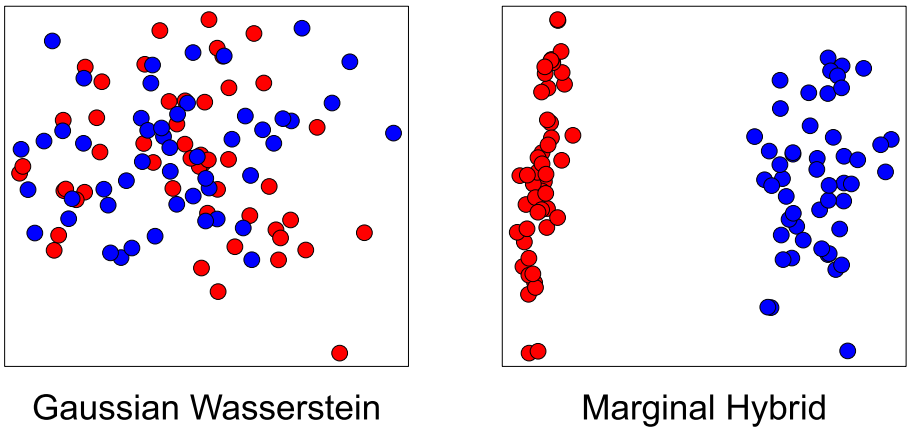

FIG 22. *The left plot shows the Gaussian-Wasserstein distance. As expected, the Gaussian-Wasserstein distance cannot distinguish the two types of datasets. The right plot shows the marginal hybrid distance. Here, the two types of datasets are clearly distinguished.*

A hybrid distance that leverages the simplicity of G can be obtained by repeatedly applying a nonlinear transformation to the variables, using G, standardizing the variables and repeating. Let L be the map that takes X to $\widetilde{X} = \Sigma_X^{-1/2}(X - \mu_X)$ and let Φ be some fixed nonlinear map. Define $H^2(X,Y) = G^2(X,Y) + W_\dagger(X,Y)$ where

$$W_\dagger(X,Y) = \sum_{j=1}^{k} G^2(X_j, Y_j)$$

with

$$X_j = \underbrace{[(\Phi \circ L) \circ \cdots (\Phi \circ L)]}_{k \text{ times}} X, \quad Y_j = \underbrace{[(\Phi \circ L) \circ \cdots (\Phi \circ L)]}_{k \text{ times}} Y.$$

In the case $k = 1$ this simplifies to

$$\begin{aligned} H^2(X,Y) &= G^2(X,Y) + G^2(\Phi(\widetilde{X}), \Phi(\widetilde{Y})) \\ &= ||\mu_X - \mu_Y||^2 + B^2(\Sigma_X, \Sigma_Y) + ||\mu_{\Phi(\widetilde{X})} - \mu_{\Phi(\widetilde{Y})}||^2 \\ &\quad + B^2(\Sigma_{\Phi(\widetilde{X})}, \Sigma_{\Phi(\widetilde{Y})}). \end{aligned}$$

There are many possible choices for Φ. The only requirement is that Φ be nonlinear otherwise the transformation adds no information beyond that already captured by G. A convenient choice is the polynomial transform

$$\Phi(x) = ([x]_2, \ldots, [x]_k)$$

where $[x]_2 = (x_{i_1} x_{i_2} : \ 1 \le i_1 \le i_2 \le d)$, $[x]_3 = (x_{i_1} x_{i_2} x_{i_3} : \ 1 \le i_1 \le i_2 \le i_3 \le d)$, etc.

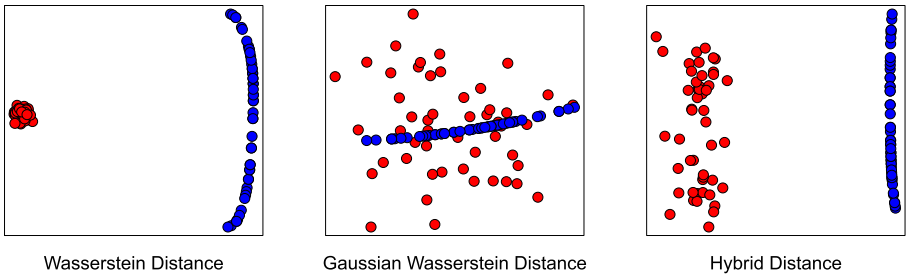

FIG 23. *Left: MDS based on pairwise Wasserstein distance. The two groups are clearly visible. Middle: MDS based on the Gaussian Wasserstein distance. As expected, the mixtures distributions are mixed in with the Normals. Right: Hybrid distance with polynomial transformation. Here we see that again, the distributions are clearly separated.*

As an example, consider a one-dimensional case. We suppose that $P_1, \ldots, P_{50} = N(0,1)$ and $P_j = (1/2)\delta_{-1} + (1/2)\delta_1$ for $j = 51, \ldots, 100$. We take $\Phi(X) = (X^2, X^3, X^4)$. Since the first two moments of all the distributions match, the Gaussian distance is unable to distinguish the 100 distributions. Figure 23 shows the pairwise distances using multidimensional scaling (MDS). The left plot is based on pairwise Wasserstein distances. The two groups are clearly visible. The middle plot used Gaussian Wasserstein distance. As expected, the distributions are mixed up. The right uses hybrid distance with polynomial transformation. Here we see that again, the distributions are clearly separated.

8. Discussion

We have proposed a modified version of the Wasserstein distance called the hybrid distance that can be used for clustering sets of distributions. The distances and the barycenter can be computed quickly. The slowest part of the computation is finding the optimal matching permutation which takes $O(m^3)$ operations. There is large literature on approximate matching in the computer science literature. It would be interesting to incorporate some of those methods to further speed up the computations.

The hybrid distance can be used for other tasks as well such as shrinkage estimation, modeling random effects, domain adaptation, and image processing. We will report on these applications in future work. We conclude by discussing a few issues.

8.1. Energy distance

The energy distance ([25]) is another metric that takes the underlying geometry of the sample space into account. The metric defined by

$$\mathcal{E}(X,Y) \equiv \mathcal{E}(P,Q) = 2\mathbb{E}||X-Y|| - \mathbb{E}||X-X'|| - \mathbb{E}||Y-Y'|| \tag{8.1}$$

where $X, X' \sim P$ and $Y, Y' \sim Q$. The energy distance has many of the desirable properties that the Wasserstein distance has including the ability to compare discrete and continuous distributions.

A sample estimator of the distance based on $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$ is

$$\widehat{\mathcal{E}}(X,Y) = \frac{2}{nm}\sum_{i,j} ||X_i - Y_j|| - \binom{n}{2}\sum_{i\neq j} ||X_i - X_j|| - \binom{m}{2}\sum_{i\neq j} ||Y_i - Y_j||. \quad (8.2)$$

In fact, there are very fast approximations that speed up the calculations; see [13].

To the best of our knowledge, there is no fast way to compute barycenters with this metric. But we can get around this by using k-medoids in place of k-means. This, if $P_1, \ldots, P_r$ is a set of distributions in a cluster, we define the centroid to be P_t where $P_t = \text{argmin}_{1\leq j\leq r}\sum_s \mathcal{E}(P_j, P_s)$. That is, we restrict the search for a centroid to be over the observed distributions.

Hence, if we use k-medoids, distribution clustering with the energy distance is feasible. However, the energy distance is not shape preserving. To see this, consider the following example. $P_1 = \delta_{-a}$ and $P_2 = \delta_a$ where $a > 0$ and δ denotes a point mass. The barycenter P minimizes $\mathcal{E}(P, P_1) + \mathcal{E}(P, P_2)$. Recall the the Wasserstein barycenter is δ_0. But this is not true for the energy distance. To see this, consider distributions of the form $P = (1/2)\delta_{-b} + m(1/2)\delta_b$. It is easy to show that, over such distributions, $\mathcal{E}(P, P_1) + \mathcal{E}(P, P_2)$ is minimized by choosing the mixture $P = (1/2)\delta_{-a} + m(1/2)\delta_a$. This implies that the barycenter cannot be δ_0 as desired.

In summary, we see that when distributions are well separated, the energy barycenter is over-dispersed compared to the Wasserstein barycenter. This leads to situations where the centroid of a cluster might look quite different than the distributions in the cluster. It is tempting to look for a simple modification of the energy distance that fixes this problem. We have tried several approaches without success. Thus, the advantage of energy distance is that it can be computed quickly but the disadvantage is that it does not preserve the shape of the distributions when computing barycenters.

8.2. Optimal preconditioning

The linear transformation that we used, namely, $\widetilde{X} = \Sigma_X^{-1/2}(X - \mu_X)$ matches the first two moments of $\widetilde{X}$ with U where $U \sim R$ is the reference distribution. This transformation is chosen for convenience. An alternative approach is to choose an optimal linear transformation. That is, we could choose a and A to minimize $W^2(a + AX, U)$. Unfortunately, computing the optimal a and A is non-trivial. If we permit linear transformations of both X_j and U then there is a closed form expression for the optimal transformation; see [15]. This requires a different transformation of U for each X_j. Hence, we lose the idea of a single, fixed, reference distribution. It would be possible to construct a composition of

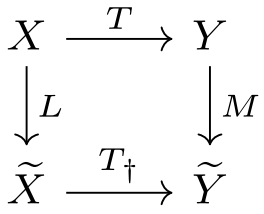

FIG 24. *The optimal linear map, preserves the Wasserstein distance and makes $\widetilde{X}$ and $\widetilde{Y}$ as close as possible.*

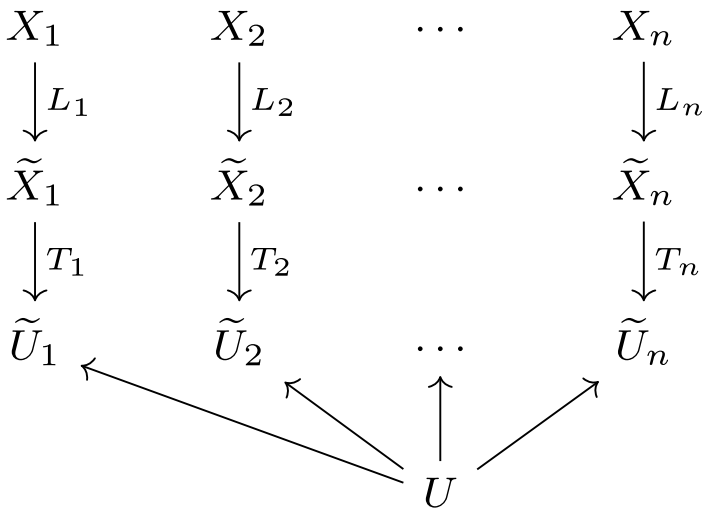

FIG 25. *The optimal linear preconditioning is applied to X_j and U in a pairwise fashion.*

maps $X_j \to \widetilde{X}_j \to \widetilde{Z}_j \to U$ where $X_j \to \widetilde{X}_j$ is the optimal linear map on X_j and $\widetilde{Z}_j \to U$ is the optimal linear map for U. [15] show that the optimal map $T(X)$ is the same as the composition $M^{-1} \circ T_\dagger \circ L$ where $T_\dagger$ is the optimal transport map from $\widetilde{X}$ to $\widetilde{Y}$; see Figures 24 and 25.

This might lead to more accurate approximations at the expense of much more computation. In the interest of keeping the method as simple as possible, we have not use this more involved approach.

8.3. The choice of reference measure and multiple tangents

When we used the tangent approximation, we approximated the distance by projecting on a single tangent space. An alternative, when there are a large number of diverse distributions, is to use several tangent approximations. The datasets could first clustered using some fast approximation, such as the marginal hybrid distance. A separate tangent approximation is computed for each preliminary

cluster. Then each distribution can be represented by its local tangent approximation using the same reference measure.

Another issue is the choice of reference measures. We do not know if there is an optimal reference measure. It is not even clear how to define such a notion of optimality. We conjecture that, at least for distribution clustering, the choice of reference measure is not critical as we mentioned earlier.

8.4. Other applications of the hybrid approach

The hybrid distance can be used for other tasks besides clustering. In this section we outline how the hybrid distance can be used for these tasks.

Nonparametric shrinkage Suppose that $P_1, \ldots, P_N$ are random distributions drawn from a distribution Π. Let $\mathcal{D}_j$ be n_j samples drawn from P_j. Suppose that we want to borrow strength from all the data to estimate each P_j. Let R be a reference measure and let $(\mu_j, \Sigma_j, \psi_j) = \phi(P_j)$ denote the hybrid transform as in (3.5). Let $(\widehat{\mu}_j, \widehat{\Sigma}_j, \widehat{\psi}_j)$ denote the j^{th} estimate. We can apply shrinkage estimation separately to the $\widehat{\mu}_j$'s, the $\widehat{\Sigma}_j$'s and the $\widehat{\psi}_j$'s. Denote these estimates by $\overline{\mu}_j, \overline{\Sigma}_j, \overline{\psi}_j$. Inverting ϕ gives the shrinkage estimator $\overline{P}_j$. We obtain $\overline{\mu}_j$ using standard James-Stein shrinkage, $\overline{\Sigma}_j$ by covariance shrinkage as in [18] and $\overline{\psi}_j$ is obtained by functional shrinkage as in [11].

Multi-sample testing Suppose that we want to test the null hypothesis $H_0 : P_1 = \cdots = P_N$. The hybrid distance allows us to split this null into three different null hypotheses, namely, $H_0 : \mu_1 = \cdots = \mu_N$, $H_0 : \Sigma_1 = \cdots = \Sigma_N$ and $H_0 : \psi_1 = \cdots = \psi_N$. This allows us to separate the deviations from the null in terms of location, scale and shape.

Multi-level clustering After clustering the distributions $P_1, \ldots, P_N$ into clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, we may want to further cluster the datasets $\mathcal{D}_1, \ldots, \mathcal{D}_N$. We could apply any standard clustering algorithm to each dataset. But we may want the distributions in a cluster $\mathcal{C}_j$ to have similar clusterings. For example, we could use Normal mixture clustering on each dataset then apply the shrinkage ideas mentioned above to make the clusterings similar. This is an alternative to simultaneous Wasserstein clustering as in [12] which is quite expensive.

Appendix

This appendix summarizes some algorithm details.

Wasserstein k-means in one dimension Now we give the steps for exact Wasserstein k-means clustering in one dimension. Given datasets $\mathcal{D}_1, \ldots, \mathcal{D}_N$ let $F_1, \ldots, F_N$ be the empirical cdf's. We use the trimmed Wasserstein distance

$$W^2(F_j, F_k) = \frac{1}{1-2\delta} \int_{\delta}^{1-\delta} (F_j^{-1}(s) - F_k^{-1}(s))^2 ds$$

where $\delta > 0$ is some specified positive trimming constant.

First, we use k-means^{++} seeding to get starting centroids:

Wasserstein k-means seeding

---

1. Input: integer k and data sets $\mathcal{D}_1, \ldots, \mathcal{D}_N$.
2. Compute all pairwise Wasserstein distances

$$D_{jk}^2 = W^2(F_j, F_k).$$

3. Let $F_1, \ldots, F_N$ be the empirical cdf's of the data sets.
4. Find the starting values:
   (a) Let $c_1 = F_j$ where j is chosen randomly from $1, \ldots, N$. Set $\mathcal{C} = \{j\}$.
   (b) Choose c_2 randomly from $\{F_1, \ldots, F_N\}$ where F_s is chosen with probability proportional to $\min_{j \in \mathcal{C}} D_{sj}^2$.
   (c) Set $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{j\}$.
   (d) Repeat the last two steps until $\mathcal{C} = \{c_1, \ldots, c_k\}$ has k elements.

The main algorithm is as follows.

Wasserstein k-means

---

1. Run Wasserstein k-means seeding to get centroids $\mathcal{C}$.
2. Compute $D_{js} = W(F_j, c_s)$ for $j = 1, \ldots, N$ and $s = 1, \ldots, k$.
3. Assign each F_j to its nearest centroid: let $\mathcal{C}_j = \{F_r : W(F_r, F_j) < W(F_r, F_t)\ t \neq j\}$.
4. Let c_j be the centroid of $\mathcal{C}_j$ using Wasserstein centroid.
5. Repeat last two steps until convergence.

Wasserstein centroid

---

1. Given one-dimensional cdf's $F_1, \ldots, F_N$.
2. Let $c(s) = \frac{1}{N} \sum_j F_j^{-1}(s)$.
3. Return $F(x) = c^{-1}(x)$.

References

[1] Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. MR3491556

[2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. MR2485254

[3] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018. MR3992484

[4] José E Chacón and Tarn Duong. *Multivariate Kernel Smoothing and Its Applications*. Chapman and Hall/CRC, 2018. MR3822372

[5] José E Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015. MR3432839

[6] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[7] E del Barrio, JA Cuesta-Albertos, C Matrán, and A Mayo-Íscar. Robust clustering tools based on optimal transportation. *Statistics and Computing*, pages 1–22, 2017. MR3905545

[8] Tarn Duong, Gaël Beck, Hanene Azzag, and Mustapha Lebbah. Nearest neighbour estimators of density derivatives, with application to mean shift clustering. *Pattern Recognition Letters*, 80:224–230, 2016.

[9] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006. MR2229687

[10] Clark R Givens, Rae Michael Shortt, et al. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984. MR0752258

[11] Wensheng Guo. Functional mixed effects models. *Biometrics*, 58(1):121–128, 2002. MR1891050

[12] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. *arXiv preprint* arXiv:1706.03883, 2017.

[13] Cheng Huang and Xiaoming Huo. An efficient and distribution-free two-sample test based on energy statistics and random projections. *arXiv preprint* arXiv:1707.04602, 2017.

[14] Heinrich Jiang, Jennifer Jang, and Samory Kpotufe. Quickshift++: Provably good initializations for sample-based mean shift. *arXiv preprint* arXiv:1805.07909, 2018.

[15] Max Kuang and Esteban G Tabak. Preconditioning of optimal transport. *SIAM Journal on Scientific Computing*, 39(4):A1793–A1810, 2017. MR3691729

[16] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. MR0075510

[17] Lisa M Maier, David E Anderson, Philip L De Jager, Linda S Wicker, and David A Hafler. Allelic variant in ctla4 alters t cell phosphorylation patterns. *Proceedings of the National Academy of Sciences*, 104(47):18607–18612, 2007.

[18] Viet Anh Nguyen, Daniel Kuhn, and Peyman Mohajerin Esfahani. Distri-

butionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint* arXiv:1805.07194, 2018.

[19] Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 2018. MR3939527

[20] Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016. MR3545279

[21] Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005. MR2168202

[22] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018. MR0848134

[23] Max Sommerfeld and Axel Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018. MR3744719

[24] Max Sommerfeld, Jörn Schrieber, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *arXiv preprint* arXiv:1802.05570, 2018. MR3990459

[25] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. MR3055745

[26] Cédric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003. MR1964483

[27] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013. MR3021062

[28] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint* arXiv:1707.00087, 2017. MR4003560