Relating Distributions with Optimal Transport

Kevin O'Connor

Abstract

In this document, we give a brief introduction to optimal transport and Wasserstein distances. Then we review a recent application of optimal transport to the clustering of distributions. We conclude with two experiments and a discussion of directions for future research.

1 Introduction

The optimal transport problem has been well-studied by the mathematical community. It was originally posed by Monge in the form,

$$\inf_{T} \int c(x, T(x)) d\mu(x) \tag{1.1}$$

where the infimum is taken over μ-measurable $T : \mathcal{X} \to \mathcal{Y}$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a cost function. Any map T that achieves the infimum quantity above is referred to as an *optimal transport map*. Initial motivations for this problem were for finding the cheapest way to transport dirt from several locations to some set of destinations. When c is chosen to be some metric, d, we can interpret the optimal transport map as minimizing the expected distance over which mass must be transported.

Equation (1.1) is useful because it admits a clear interpretation. However, it does not always exist. Intuitively, this is because our plan can only send mass at a point $x \in \mathcal{X}$ to a single other point $T(x) \in \mathcal{Y}$. If one were to try to transport a point mass on $x \in \mathcal{X}$ to a continuous distribution like Unif$[0, 1]$, we would find that no such T exists. Kantorovich circumvented this issue by modifying the problem to allow mass to be split and sent to multiple locations. This is achieved via *couplings*:

Definition 1.1. *A* ***coupling*** *of two measure spaces* $(\mathcal{X}, \mu)$ *and* $(\mathcal{Y}, \nu)$ *is a joint measure* π *on* $\mathcal{X} \times \mathcal{Y}$ *such that* $\pi(\mathcal{X} \times B) = \nu(B)$, $\forall B \in \mathcal{B}(\mathcal{Y})$ *and* $\pi(A \times \mathcal{Y}) = \mu(A)$, $\forall A \in \mathcal{B}(\mathcal{X})$. *The set of couplings of* $(\mathcal{X}, \mu)$ *and* $(\mathcal{Y}, \nu)$ *will be denoted* $\Pi(\mu, \nu)$.

For those unfamiliar with measure theory, a coupling can be equivalently characterized as a random variable $Z \in \mathcal{X} \times \mathcal{Y}$ whose $\mathcal{X}$-marginal is μ-distributed and $\mathcal{Y}$-marginal is ν-distributed.

Now notice that a coupling defines a kind of transport plan that allows for mass to be split. For a coupling $\pi \in \Pi(\mu, \nu)$, $\pi(A, B)$ tells us how much of the mass at $A \in \mathcal{B}(\mathcal{X})$ should be send to $B \in \mathcal{B}(\mathcal{Y})$. Kantorovich used this notion to define the modified optimal transport problem:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int c(x, y) d\pi(x, y) \tag{1.2}$$

It is clear that a coupling always exists since the independent coupling, $\pi := \mu \otimes \nu$, always exists. Furthermore, for well-behaved cost functions, a minimizer of Equation (1.2) always exists. This allows us to find transport plans between discrete and continuous distributions.

1.1 Metric on the Space of Distributions

One can use the optimal transport distance in Equation (1.2) to define a distance between probability distributions.

Definition 1.2. *We define the Wasserstein-2 distance between distributions $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{X})$ as*

$$W_2^2(P, Q) := \inf_{\pi \in \Pi(P,Q)} \int \|x - y\|^2 d\pi(x, y) \tag{1.3}$$

Note that we could use any metric on the underlying space $\mathcal{X}$. We can think of the Wasserstein distance as *lifting* an underlying metric up to the space of probability distributions on that space. This highlights one of the primary advantages of the OT distance over other metrics on probability distributions. Other common choices like Jensen-Shannon or total variation ignore any underlying geometry and therefore fail to capture many simple and important differences between distributions.

Remark 1.3. *We note that $W_2^2(P, Q)$ is not truly a metric on the space $\mathcal{P}(\mathcal{X})$ since it may be infinite. One can restrict to the set of distributions $\mathcal{P}_2(\mathcal{X})$ where W_2^2 distance is finite and show that it is a metric on this restricted space. A curious reader should refer to [2] for (many) more details.*

1.2 Closed Forms

At this point, we discuss two cases where a closed form is available for the Wasserstein distance.

Proposition 1.4. *If $P \sim \mathcal{N}_d(\mu_P, \Sigma_P)$ and $Q \sim \mathcal{N}_d(\mu_Q, \Sigma_Q)$, i.e. P and Q are multivariate Gaussian measures, then*

$$W_2^2(P, Q) = \|\mu_P - \mu_Q\|^2 + tr(\Sigma_P) + tr(\Sigma_Q) - 2tr((\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}) \tag{1.4}$$

The Gaussian case gives us some intuition for what the Wasserstein distance is measuring. It incorporates not only the location, μ, of the distribution but also the scale, Σ.

Remark 1.5. *Note that we define the Bures distance,*

$$B^2(\Sigma_P, \Sigma_Q) = tr(\Sigma_P) + tr(\Sigma_Q) - 2tr((\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}) \tag{1.5}$$

Thus for brevity, we will write Equation (1.4) *as*

$$W_2^2(P, Q) = \|\mu_P - \mu_Q\|^2 + B^2(\Sigma_P, \Sigma_Q) \tag{1.6}$$

Next we consider the case of the expected cost of the optimal transport map (Equation (1.1)) between two empirical measures.

Proposition 1.6. *If $P = 1/n \sum_{i=1}^n \delta_{x_i}$ and $Q = 1/n \sum_{i=1}^n \delta_{y_i}$, then the expected cost associated with the optimal transport map is given by $x_i \mapsto y_{\gamma(i)}$ where*

$$\gamma = \arg\min_{\gamma \in S(n)} \sum_{i=1}^n c(x_i, y_{\gamma(i)}) \tag{1.7}$$

where the minimum is taken over permutations, γ, in the set of permutations of $[n]$, $S(n)$.

Again, this closed form gives us some intuition for the OT distance. In the case of the optimal transport map, we assign mass at x_i to be transported to some other y_j so as to minimizes the expected cost. Since each point has mass $1/n$ and mass cannot be split, this amounts to finding optimal pairs $\{(x_i, y_j)\}$, or equivalently a permutation $\{(x_i, y_{\gamma(i)})\}$ as in Equation (1.7).

1.3 Computational Optimal Transport

Here we briefly discuss how the optimal transport distance is computed in practice. In Prop. 1.6, we saw how the optimal transport map could be computed for two empirical distributions. [1] suggest using the Hungarian algorithm for finding the optimal permutation which runs in $\mathcal{O}(n^3)$ time. In other settings, it is typically much more difficult to find an optimal transport map.

For discrete measures, the optimal transport plan can found as a linear program. Indeed if $P = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $Q = \sum_{j=1}^m \beta_i \delta_{y_j}$, then any coupling $\pi \in \Pi(P, Q)$ has the form

$$\pi = \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} \delta_{(x_i, y_j)}, \qquad \text{s.t.} \quad \sum_{i=1}^n \gamma_{ij} = \beta_j, \quad \sum_{j=1}^m \gamma_{ij} = \alpha_i, \quad \gamma_{ij} \geqslant 0, \forall i, j \tag{1.8}$$

which yields an expected cost of

$$\int c(x, y) d\pi(x, y) = \sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} c(x_i, y_j) \tag{1.9}$$

Then minimization of the expected cost with the marginal constraints in Equation (1.8) amounts to minimization of a linear objective subject to linear constraints. Thus, it can be solved using standard linear programming algorithms like the simplex method.

1.4 Organization of the Paper

The remainder of the paper will be organized as follows. Section 2 will review a method proposed by [1] which defines a hybrid wasserstein distance that allows for efficient computation and straightforward application of commonly used clustering algorithms to the clustering of distributions. Section 3 provides two computational experiments to illustrate how the hybrid wasserstein distance can be used to compare distributions. Finally, we provide some further discussion and avenues for future research in Section 4.

2 Hybrid Wasserstein Distance

Recent work by Verdinelli and Wasserman [1] studies a hybrid Wasserstein distance to obviate the computational burden of computing Wasserstein distances in distributional clustering. Specifically, they propose

$$H(P, Q) := H(X, Y) = W_2^2(Z_X, Z_Y) + W_\dagger^2(\tilde{X}, \tilde{Y}) \tag{2.1}$$

where $Z_X \sim \mathcal{N}(\mathbb{E}X, \mathrm{Cov}(X))$, $Z_Y \sim \mathcal{N}(\mathbb{E}Y, \mathrm{Cov}(Y))$, $\tilde{X} = \mathrm{Cov}(X)^{-1/2}(X - \mathbb{E}X)$ and $\tilde{Y} = \mathrm{Cov}(Y)^{-1/2}(Y - \mathbb{E}Y)$ with $W_\dagger^2$ some other distance which is easier to compute. Note that the first term is a Wasserstein distance between Gaussian measures and thus has the closed form expression,

$$W_2^2(Z_X, Z_Y) = \|\mu_X - \mu_Y\|^2 + B^2(\Sigma_X, \Sigma_Y) \tag{2.2}$$

where

$$B^2(\Sigma_X, \Sigma_Y) := \mathrm{tr}(\Sigma_X) + \mathrm{tr}(\Sigma_Y) - 2\mathrm{tr}((\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2})^{1/2}) \tag{2.3}$$

$B^2(\Sigma_X, \Sigma_Y)$ is referred to as the *Bures distance.*

The authors leave the choice of $W_\dagger^2$ somewhat flexible but focus on a specific choice which we will discuss now.

2.1 How to define $W_\dagger^2$?

The idea behind $W_\dagger^2(\tilde{X}, \tilde{Y})$ comes from [3]. They propose to: choose a reference measure R with density r, map $\tilde{X}$ and $\tilde{Y}$ to R via the optimal transport maps T_X and $T_Y$and then compute the average squared distance in this space. Specifically, let $\psi_j(z) = (T_j(z) - z)\sqrt{r(z)}$. Then

$$W_\dagger^2(\tilde{P}_X, \tilde{P}_Y) := \int (\psi_X(z) - \psi_Y(z))^2 dz = \int (T_X(z) - T_Y(z))^2 dR(z) \tag{2.4}$$

For real data, we have only estimates of the true underlying distributions P_X and P_Y, so T_X and T_Y cannot be directly computed. Furthermore, a reference measure R must still be chosen. The authors choose $R = R_n \star K_h$, i.e. the kernel density estimate with bandwidth h and discrete weights,

$$R_n := \frac{1}{n} \sum_{i=1}^{m} n_i \tilde{P}_i \tag{2.5}$$

Then the authors use the finite sample approximation,

$$\int (\psi_X(z) - \psi_Y(z))^2 dz = \frac{1}{m} \sum_{i=1}^{m} (T_X(U_i) - T_Y(U_i))^2 + \mathcal{O}_p(m^{-1/2}) \tag{2.6}$$

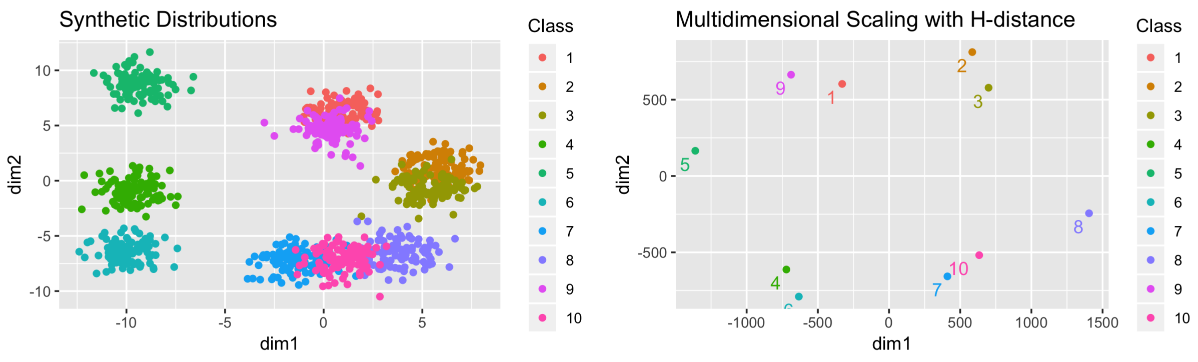
where $U_1, ..., U_m \sim R$. This suggests the finite sample approximation,

$$W_\dagger^2(\tilde{P}_X, \tilde{P}_Y) \approx \frac{1}{m} \sum_{i=1}^{m} (T_X(u_i) - T_Y(u_i))^2 \tag{2.7}$$

for a random draw of $u_1, ..., u_m \sim R$.

2.2 Estimating T_X and T_Y

Since the distributions of $\tilde{X}$ and $\tilde{Y}$ are not known in practice, T_X and T_Y, the optimal transport maps to the reference distribution R are not known. The following estimation procedure is recommended: Sample $X_1, ..., X_m \sim \tilde{P}_X$ and $R_1, ..., R_m \sim R$. Then find the optimal transport map, $\hat{T}_X$ from $\{X_i\}_{i=1}^m$ to $\{R_i\}_{i=1}^m$. For points outside of $\{X_i\}_{i=1}^m$, let T_X map to the output of the 1-nearest neighbor.



(a) Synthetic data.

(b) Multidimensional scaling with H-distance.

Figure 2.1: Hybrid Wasserstein distance for Gaussian data.

2.3 Complete Hybrid Wasserstein Distance

Combining the results above with the closed form expression for the Wasserstein distance between two Gaussians, we get the complete expression for the sample hybrid Wasserstein distance:

$$\hat{H}^2(P_X, P_Y) = \|\hat{\mu}_X - \hat{\mu}_Y\|^2 + B^2(\hat{\Sigma}_X, \hat{\Sigma}_Y) + \frac{1}{m}\sum_{i=1}^{m}(\hat{T}_X(u_i) - \hat{T}_Y(u_i))^2 \tag{2.8}$$

3 Experiments

In this section, we perform two experiments to illustrate the utility of the hybrid wasserstein distance [1].

3.1 Synthetic Data

First we apply the H distance to a sample of Gaussian distributions in $\mathbb{R}^2$. 10 means $\mu_1, ..., \mu_{10} \sim$ Unif$[(-10, 10)^2]$ and 100 points $x_{k,1}, ..., x_{k,100} \sim \mathcal{N}(\mu_k, \mathbf{I})$ for each $k = 1, ..., 10$ were generated and are depicted in Figure 2.1a. We then computed the H distance between each pair of empirical distributions, P_k, where $P_k = 1/100 \sum_{i=1}^{100} \delta_{x_{k,i}}$, to obtain a matrix of pairwise H distances. To visualize distances in this embedded space, we apply multi-dimensional scaling to obtain Figure 2.1b. Note that the MDS plot mostly captures the relative centers of each distribution. This is to be expected as the distributions differ only in their means. Furthermore, for Gaussian distributions, H distance reduces to Wasserstein distance which is simply the Euclidean distance between the means for distributions with equal covariance.

3.2 NBA Playoff Data

To investigate the effect of Hybrid Wasserstein distance on real data, we considered the distribution of made shots in the 2018 NBA playoffs [2]. This data contains the details of each shot taken by players in the 2018 NBA playoffs including coordinates of their shot and whether it went in or not. We can think of each player's made shots as comprising a distribution that we hope characterizes some aspect of their shooting style. We can then apply the Hybrid Wasserstein distance to compare players and use MDS to visualize the embedded space as above.

In order to make the data more digestible, we subset to players with more than 150 made shots in this dataset. The Hybrid Wasserstein distance matrix was computed between each player and used as input for MDS as above. The results are included in Figure 3.1.

4 Conclusion and Discussion

We have seen how optimal transport defines a distance between distributions through the expectation of some ground metric. Furthermore, we have shown how to define a more efficient hybrid wasserstein distance. Using this distance, showed how distances between Gaussian data with equal covariances largely depend on the respective means. Finally, we demonstrated how one can measure distances between shot distributions of basketball players and use MDS to visualize the relationships between each.

There are many open questions regarding wasserstein-based clustering methods. For example, what happens to wasserstein-based clustering methods in high dimensions? In the case of the hybrid wasserstein distance, we expect that $\hat{\Sigma}$ will be a poor estimate of Σ in this case. It would be interesting to consider how this may affect clustering methods which use this method.

[1] Code can be found at `https://github.com/oconnor-kevin/clustering-with-ot`

[2] https://www.kaggle.com/boonpalipatana/nba-playoff-shots-2018

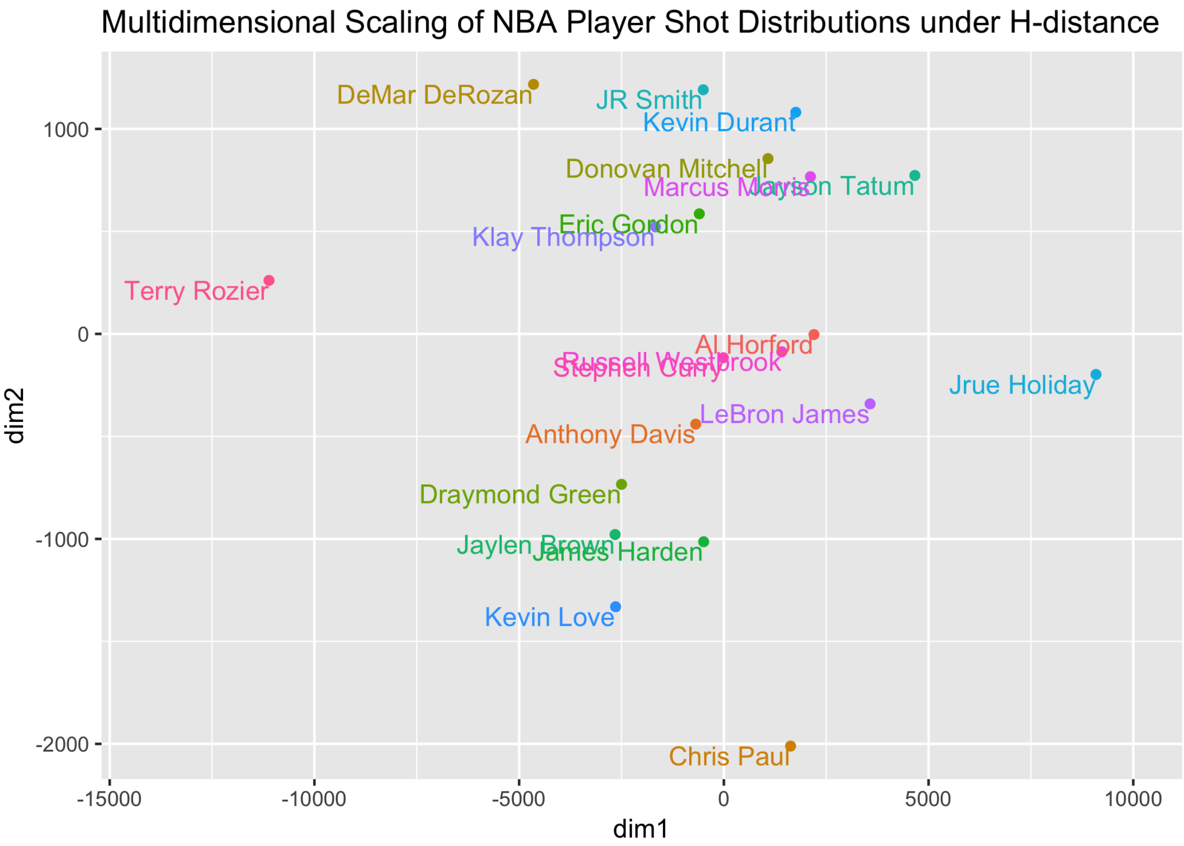


Figure 3.1: Multidimensional scaling plot of NBA player shot distributions.

More generally, one could study what kinds of pathologies for distributions might cause outliers in wasserstein-based distances and clustering. Could one construct more robust distances? For example, one could use medians instead of means. Finally, what are the asymptotic distributional properties of the hybrid wasserstein distance? Few results exist in this area for wasserstein distance without Gaussian assumptions on the two distributions in question. One might try to find special cases where something can be said for the hybrid distance too.

References

[1] Isabella Verdinelli and Larry Wasserman. Hybrid wasserstein distance and fast distribution clustering. *arXiv preprint arXiv:1812.11026*, 2018.

[2] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[3] 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.