Optimal Transport and Wasserstein Distance

The Wasserstein distance — which arises from the idea of *optimal transport* — is being used more and more in Statistics and Machine Learning. In these notes we review some of the basics about this topic. Two good references for this topic are:

Kolouri, Soheil, et al. Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine* 34.4 (2017): 43-59.

Villani, Cedric. *Topics in optimal transportation.* No. 58. American Mathematical Soc., 2003.

As usual, you can find a wealth of information on the web.

1   Introduction

Let $X \sim P$ and $Y \sim Q$ and let the densities be p and q. We assume that $X, Y \in \mathbb{R}^d$. We have already seen that there are many ways to define a distance between P and Q such as:

$$\text{Total Variation} : \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |p - q|$$

$$\text{Hellinger} : \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$$

$$L_2 : \int (p - q)^2$$

$$\chi^2 : \int \frac{(p - q)^2}{q}.$$

These distances are all useful, but they have some drawbacks:

1. We cannot use them to compare P and Q when one is discrete and the other is continuous. For example, suppose that P is uniform on $[0, 1]$ and that Q is uniform on the finite set $\{0, 1/N, 2/N, \ldots, 1\}$. Practically speaking, there is little difference between these distributions. But the total variation distance is 1 (which is the largest the distance can be). The Wasserstein distance is $1/N$ which seems quite reasonable.

2. These distances ignore the underlying geometry of the space. To see this consider Figure 1. In this figure we see three densities p_1, p_2, p_3. It is easy to see that $\int |p_1 - p_2| = \int |p_1 - p_3| = \int |p_2 - p_3|$ and similarly for the other distances. But our intuition tells us that p_1 and p_2 are close together. We shall see that this is captured by Wasserstein distance.
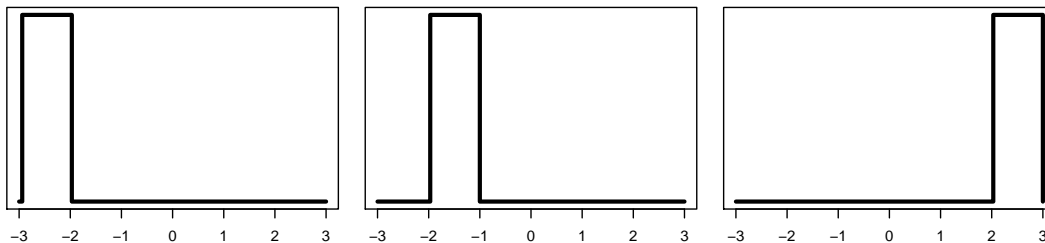
Figure 1: *Three densities p_1, p_2, p_3. Each pair has the same distance in L_1, L_2, Hellinger etc. But in Wasserstein distance, p_1 and p_2 are close.*

3. When we average different objects — such as distributions or images — we would like to make sure that we get back a similar object. The top plot in Figure 2 shows some distributions, each of which is uniform on a circle. The bottom left plot shows the Euclidean average of the distributions which is just a gray mess. The bottom right shows the Wasserstein barycenter (which we will define later) which is a much better summary of the set of images.

4. When we compute the usual distance between two distributions, we get a number but we don't get any qualitative information about why the distributions differ. But with the Wasserstein distance we also get a map that shows us how we have to move the mass of P to morph it into Q.

5. Suppose we want to create a path of distributions (a geodesic) P_t that interpolates between two distributions P_0 and P_1. We would like the distributions P_t to preserve the basic structure of the distributions. Figure 5 shows an example. The top row shows the path between P_0 and P_1 using Wasserstein distance. The bottom row shows the path using L_2 distance. We see that the Wasserstein path does a better job of preserving the structure.

6. Some of these distances are sensitive to small wiggles in the distribution. But we shall see that the Wasserstein distance is insensitive to small wiggles. For example if P is uniform on $[0, 1]$ and Q has density $1 + \sin(2\pi k x)$ on $[0, 1]$ then the Wasserstein distance is $O(1/k)$.

2   Optimal Transport

If $T : \mathbb{R}^d \to \mathbb{R}^d$ then the the distribution of $T(X)$ is called the push-forward of P, denoted by $T_\# P$. In other words,

$$T_\# P(A) = P\Big(\{x : \ T(x) \in A \} \Big) = P(T^{-1}(A)).$$
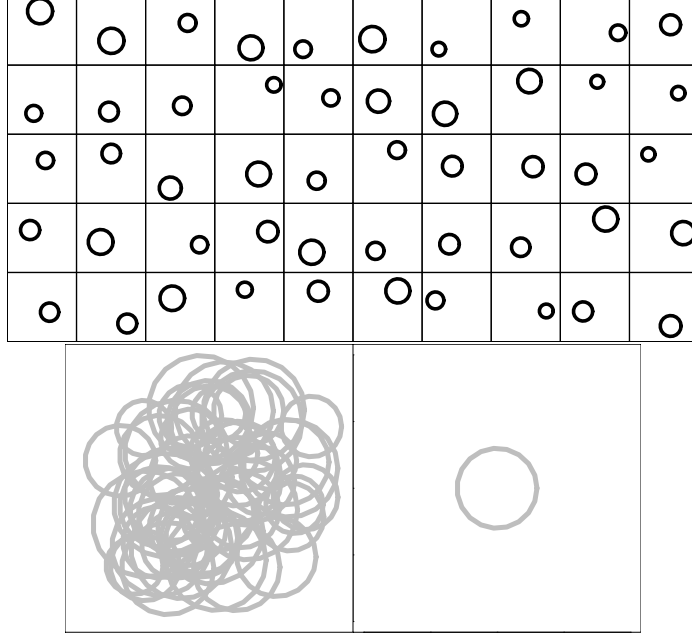
Figure 2: *Top: Some random circles. Bottom left: Euclidean average of the circles. Bottom right: Wasserstein barycenter.*

The *Monge* version of the optimal transport distance is

$$\inf_{T} \int ||x - T(x)||^p dP(x)$$

where the infimum is over all T such that $T_{\#}P = Q$. Intuitively, this measures how far you have to move the mass of P to turn it into Q. A minimizer T^*, if one exists, is called the *optimal transport map.*

If P and Q both have densities than T^* exists. The map $T_t(x) = (1-t)x + tT^*(x)$ gives the path of a particle of mass at x. Also, $P_t = T_{t\#}P$ is the geodesic connecting P to Q.

But, the minimizer might not exist. Consider $P = \delta_0$ and $Q = (1/2)\delta_{-1} + (1/2)\delta_1$ where δ_a. In this case, there is no map T such that $T_{\#}P = Q$. This leads us to the Kantorovich formulation where we allow the mass at x to be split and move to more than one location.

Let $\mathcal{J}(P, Q)$ denote all joint distributions J for (X, Y) that have marginals P and Q. In other words, $T_{X\#}J = P$ and $T_{Y\#}J = Q$ where $T_X(x, y) = x$ and $T_Y(x, y) = y$. Figure 4 shows an example of a joint distribution with two given marginal distributions. Then the Kantorovich, or Wasserstein, distance is

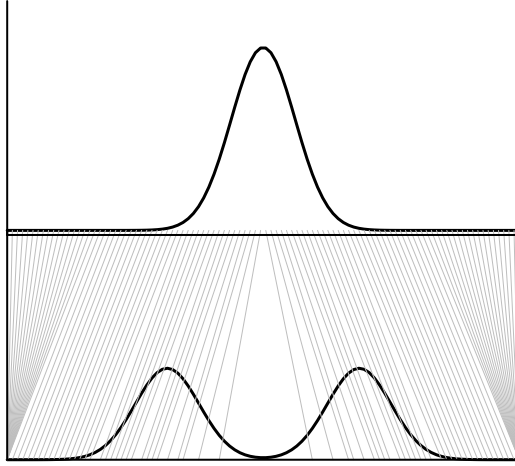$$W_p(P, Q) = \left(\inf_{J \in \mathcal{J}(P,Q)} \int ||x - y||^p dJ(x, y) \right)^{1/p}$$

3

Figure 3: *Two densities p and q and the optimal transport map to that morphs p into q.*

where $p \geq 1$. When $p = 1$ this is also called the *Earth Mover distance*. The minimizer J^* (which does exist) is called the *optimal transport plan* or *the optimal coupling*. In case there is an optimal transport map T then J is a singular measure with all its mass on the set $\{(x, T(x))\}$.

It can be shown that

$$W_p^p(P, Q) = \sup_{\psi, \phi} \int \psi(y) dQ(y) - \int \phi(x) dP(x)$$

where $\psi(y) - \phi(x) \leq ||x - y||^p$. This is called the dual formulation. In special case where $p = 1$ we have the very simple representation

$$W_1(P, Q) = \sup \left\{ \int f(x) dP(x) - \int f(x) dQ(x) : \; f \in \mathcal{F} \right\}$$

where $\mathcal{F}$ denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(y) - f(x)| \leq ||x - y||$ for all x, y.

When $d = 1$, the distance has a closed form:

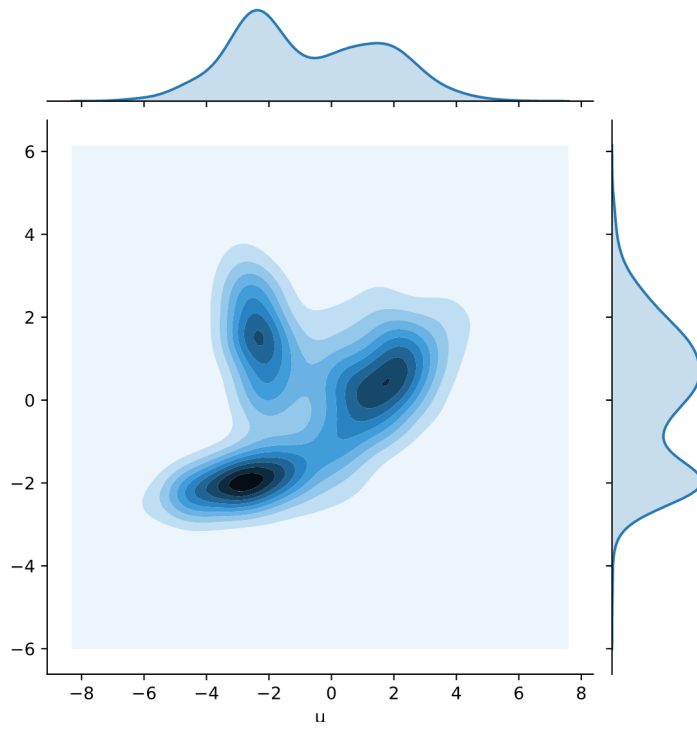$$W_p(P, Q) = \left(\int_0^1 |F^{-1}(z) - G^{-1}(z)|^p \right)^{1/p}$$

4

Figure 4: *This plot shows one joint distribution J with a given X marginal and a given Y marginal. Generally, there are many such joint distributions. Image credit: Wikipedia.*

and F and G are the cdf's of P and Q. If P is the empirical distribution of a dataset $X_1, \ldots, X_n$ and Q is the empirical distribution of another dataset $Y_1, \ldots, Y_n$ of the same size, then the distance takes a very simple function of the order statistics:

$$W_p(P,Q) = \left(\sum_{i=1}^{n} ||X_{(i)} - Y_{(i)}||^p \right)^{1/p}.$$

Because the one-dimensional case is so easy, some authors use this as the basis for defining a modified distance, called the *sliced Wasserstein distance* defined by

$$S(P,Q) = \left(\int W_p^p(P_\theta, Q_\theta) d\theta \right)^{1/p}$$

where the integral is over the unit sphere and P_θ denotes the distribution of $X^T\theta$ and Q_θ denotes the distribution of $Y^T\theta$.

3   Geodesics

Let P_0 and P_1 be two distributions. Consider a map c taking $[0,1]$ to the set of distributions, such that $c(0) = P_0$ and $c(1) = P_1$. Thus $(P_t : 0 \le t \le 1)$ is a path connecting P_0
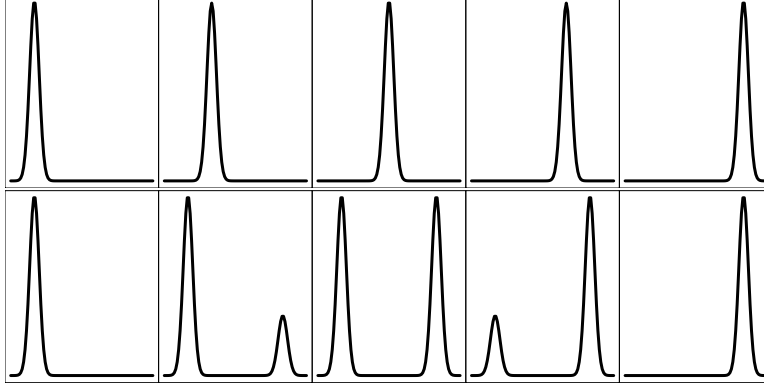
Figure 5: *Top row: Geodesic path from P_0 to P_1. Bottom row: Euclidean path from P_0 to P_1.*

Figure 6: *Morphing one image into another using the Wasserstein geodesic. Image credit: Bauer, Joshi and Modin 2015.*

and P_1, where $P_t = c(t)$. The length of c — denoted by $L(c)$ — is the supremum of $\sum_{i=1}^{m} W_p(c(t_{i-1}), c(t_i))$ over all m and all $0 = t_1 < \cdots < t_m = 1$. There exists such a path c such that $L(c) = W(P_0, P_1)$. In other words, $(P_t : 0 \leq t \leq 1)$ is the geodesic connecting P_0 and P_1. It can be shown that

$$P_t = F_{t\#}J$$

where J is the optimal coupling and $F_t(x, y) = (1-t)x + ty$. Examples are shown in Figures 5 and 6.

4   Barycenters and PCA

Suppose we have a set of distributions $P_1, \ldots, P_N$. How do we summarize these distributions with one "typical" distribution? We could take the average $\frac{1}{N} \sum_{j=1}^{n} P_j$. But the resulting average won't look like any of the P_j's. See Figure 7.

Instead we can use the Wasserstein barycenter which is the distribution P that minimizes

$$\sum_{j=1}^{N} W(P, P_j).$$

The bottom right plot of Figure 7 shows an example. You can see that this does a much better job.

We can do the same thing for data sets. See Figure 8. Here we simple regard a dataset as an empirical distribution. The average (red dots) $N^{-1} \sum_j \widehat{P}_j$ of these empirical distributions $\widehat{P}_j$ is useless. But the Wasserstein barycenter (blue dots) gives us a better sense of what a typical dataset looks like.

Let's pursue this last example a bit more since it will give us some intuition. Suppose we have N datasets $\mathcal{X}_1, \ldots, \mathcal{X}_N$ where $\mathcal{X}_j = \{X_{j1}, \ldots, X_{jn}\}$. For simplicity, suppose that each is of the same size n. In this case, we can describe the Wasserstein barycenter in a simple way. First we find the order statistics for each data set:

$$X_{(j1)} \leq X_{(j2)} \leq \cdots \leq X_{(jn)}.$$

Now for each $1 \leq r \leq n$, we find the average r^{th} average order statistic:

$$Y_{(r)} = \frac{1}{N} \sum_{j=1}^{N} X_{(jr)}.$$

Then $\mathcal{Y} = \{Y_{(1)}, \ldots, Y_{(n)}\}$ is the Wasserstein barycenter. In a sense, all we are really doing is converting to quantiles and averaging.

Now that we have a notion of average, it is possible to define a Wasserstein version of PCA. There are several approaches; see, for example Seguy and Cuturi (2015), Boissard et al (2013), Bigot (2014), Wang, Wei and Slepcev (2013). The idea, as with the barycenters, is to find orthogonal directions of variation in the space of measures (or images).

In the usual Euclidean version of PCA we first find the mean μ. Then we find the linear subspace centered at μ with the largest variance, or, equivalently, we find a unit vector v_1 to minimize $\mathbb{E}[||X - (\mu + tv_1)||^2]$. This is the first PC. Then we find v_2 to minimize the same quantity subject to being orthogonal to v_1 and so on. For Wasserstein distance we instead use Wasserstein geodesics. Alternatively, one finds a template image Q and transport maps g_j taking Q to P_j. Then PCA is performed on the maps $g_1, \ldots, g_n$. The resulting principal components can then be used for dimension reduction. The details are somewhat complicated.
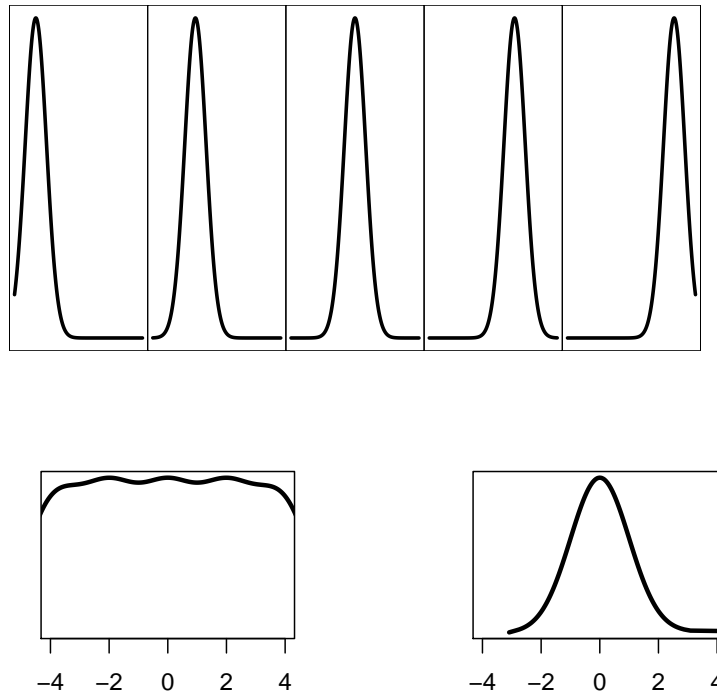
7

Figure 7: *Top: Five distributions. Bottom left: Euclidean average of the distributions. Bottom right: Wasserstein barycenter.*
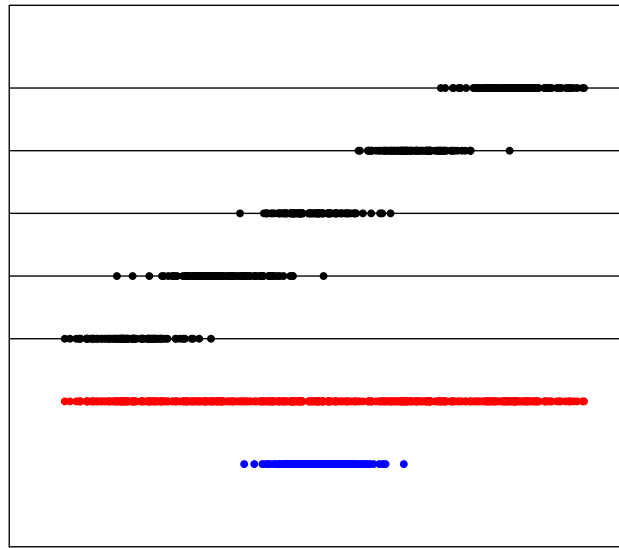

Figure 8: *The top five lines show five, one-dimensional datasets. The red points the what happens if we simple average the give empirical distributions. The blue dots show the Wasserstein barycenter which, in this case, can be obtained simply by averaging the order statistics.*

5 Minimax Rates

All the minimax proofs for Wasserstein distance make use of the following bound. Suppose we observe $X_1, \ldots, X_n \sim P$ supported on $[0, \Delta]^d$. Divide the space into cubes with side length $\Delta/2^i$. Let $p^{(i)}$ denote the multinomial obtained by coarsening P over the partition. The, for any P and Q,

$$W_1(P, Q) \le d \left(2 \sum_{i=1}^{m} \frac{\Delta}{2^i} ||p^{(i)} - q^{(i)}||_1 \right) + \frac{\epsilon}{2}$$

where $m = \log(2\Delta d/\epsilon)$. Also,

$$W_1(P, Q) \ge \max_i \frac{\Delta ||p^{(i)} - q^{(i)}||_1}{2^{i+1}}.$$

More generally, as discussed in Weed and Bach (2017), for any sequence of dyadic partitions $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m$ we have

$$W_p^p(P, Q) \le \delta^{mp} + \sum_{j=1}^{m} \delta^{(j-1)p} \sum_{A \in \mathcal{A}_j} |P(A) - Q(A)|$$

where $\text{diam}(A) \le \delta^j$ for every $A \in \mathcal{A}_j$. These bounds allows us to make use of minimax bounds for L_1 distance. They also reveal that, in some sense, that Wasserstein distance behaves like a multiscale L_1 distance.

Weed and Bach show that

$$\mathbb{E}[W_p(P, P_n)] \preceq \left(\frac{1}{n} \right)^{1/d}$$

where P_n is the empirical distribution. Singh and Poczos (2018) showed that this is minimax. Actually, I have simplified the results a bit. The bounds in Weed and Bach (2017) and Singh and Poczos (2018) are expressed in terms of certain covering numbers. When P is smooth, the rate is $n^{-1/d}$. But when P is concentrated on a lower dimensional set, the rate is faster.

Now suppose we observe $X_1, \ldots, X_n \sim P$ supported on $[0, \Delta]^d$. We want to test $H_0 : P = P_0$ versus $H_1 : W_1(P, P_0) > \epsilon$. Ba et al (2013) and Deng, Li and Wu (2017) showed that the minimax separation rate is (ignoring some log terms)

$$\epsilon_n \asymp \frac{2\Delta d}{n^{\frac{3}{2d}}}.$$

In the special case were P and P_0 are concentrated in k small clusters, the rate becomes

$$\epsilon_n \asymp d\Delta \left(\frac{k}{n} \right)^{1/4}.$$

6   Confidence Intervals

How do we get hypothesis tests and confidence intervals for the Wasserstein distance? Usually, we would use some sort of central limit theorem. Such results are available when $d = 1$ but are elusive in general.

del Barrio and Loubes (2017) show that

$$\sqrt{n}(W_2^2(P, P_n) - \mathbb{E}[W_2^2(P, P_n)]) \rightsquigarrow N(0, \sigma^2(P))$$

for some $\sigma^2(P)$. And, in the two sample case

$$\sqrt{\frac{nm}{n+m}}\left(W_2^2(P_n, Q_m) - \mathbb{E}[W_2^2(P_n, Q_m)]\right) \rightsquigarrow N(0, \sigma^2(P, Q))$$

for some $\sigma^2(P, Q)$. Unfortunately, these results do not give a confidence interval for $W(P, Q)$ since the limit is centered around $\mathbb{E}[W_2^2(P_n, Q_m)]$ instead of $W_2^2(P, Q)$.

The only case, so far, where some sort of confidence interval is available is when the support $\mathcal{X} = \{x_1, \ldots, x_k\}$ is a finite metric space. In this case, Sommerfeld and Munk (2017) obtained some precise results. First, they showed that

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} W_p(P_n, Q_m) \rightsquigarrow \left(\max_u \langle G, u \rangle\right)^{1/p}$$

G is a mean 0 Gaussian random vector and u varies over a convex set. By itself, this does not yield a confidence set. But they showed that the distribution can be approximated by subsampling, where the subsamples of size m with $m \to \infty$ and $m = o(n)$.

You might wonder why the usual bootstrap does not work. The reason is that the map $(P, Q) \mapsto W_p^p(P, Q)$ is not Hadamard differentiable. This means that the map does not have smooth derivatives.

In general, the problem of constructing confidence intervals for Wasserstein distance is unsolved.

7   Robustness

One problem with the Wasserstein distance is that it is not robust. To see this, note that $W(P, (1 - \epsilon)P + \epsilon\delta_x) \to \infty$ as $x \to \infty$.

However, a partial solution to the robustness problem is available due to Alvarez-Esteban, del Barrio, Cuesta Albertos and Matran (2008). They define the α-trimmed Wasserstein

distance

$$\tau(P,Q) = \inf_A W_2(P_A, Q_A)$$

where $P_A(\cdot) = P(A \bigcap \cdot)/P(A)$, $Q_A(\cdot) = Q(A \bigcap \cdot)/Q(A)$ and A varies over all sets such that $P(A) \geq 1 - \alpha$ and $Q(A) \geq 1 - \alpha$. When $d = 1$, they show that

$$\tau(P,Q) = \inf_A \left(\frac{1}{1-\alpha} \int_A (F^{-1}(t) - G^{-1}(t))^2 dt \right)^{1/2}$$

where A varies over all sets with Lebesgue measure $1 - \alpha$.

8   Inference From Simulations

Suppose we have a parametric model $(P_\theta : \theta \in \Theta)$. We can estimate θ using the likelihood function $\prod_i p_\theta(X_i)$. But in some cases we cannot actually evaluate p_θ. Instead, we can simulate from P_θ. This happens quite often, for example, in astronomy and climate science. Berntom et al (2017) suggest replacing maximum likelihood with minimum Wasserstein distance. That is, given data $X_1, \ldots, X_n$ we use

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \, W(P_\theta, P_n)$$

where P_n is the ampirical measure. We estimate $W(P_\theta, P_n)$ by $W(Q_N, P_n)$ where Q_N is the empirical measure based on a sample $Z_1, \ldots, Z_N \sim P_\theta$.

9   Computing the Distance

We saw that, when $d = 1$,

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

and F and G are the cdf's of P and Q. If P is the empirical distribution of a dataset $X_1, \ldots, X_n$ and Q is the empirical distribution of another dataset $Y_1, \ldots, Y_n$ of the same size, then the distance takes a very simple function of the order statistics:

$$W_p(P,Q) = \left(\sum_{i=1}^n ||X_{(i)} - Y_{(i)}||^p \right)^{1/p}.$$

The one dimensional case is, perhaps, the only case where computing W is easy.

For any d, if P and Q are empirical distributions — each based on n observations — then

$$W_p(P, Q) = \inf_\pi \left(\sum_i ||X_i - Y_{\pi(i)}||^p \right)^{1/p}$$

where the infimum is over all permutations π. This is not an easy minimization but there are greedy heuristics. For example, start with an arbitrary permutation. Now check each pair i and j. If $||X_i - Y_{\pi(i)}||^p + ||X_j - Y_{\pi(j)}||^p > ||X_i - Y_{\pi(j)}||^p + ||X_j - Y_{\pi(i)}||^p$ then swap $\pi(i)$ and $\pi(j)$. Repeat until there is no improvement. This is simple but not guaranteed to lead to a minimum.

Suppose that P has density p and that $Q = \sum_{j=1}^m q_j \delta_{y_j}$ is discrete. Given weights $w = (w_1, \ldots, w_m)$ define the power diagram $V_1, \ldots, V_m$ where $y \in V_j$ if y is closer to the ball $B(y_j, w_j)$ and any other ball $B(y_s, w_s)$. Define the map $T(x) = y_j$ when $x \in V_j$. According to a result known as Bernier's theorem, if have that $P(V_j) = q_j$ then

$$W_2(P, Q) = \left(\sum_j \int_{V_j} ||x - y_j||^2 dP(x) \right)^{1/2}.$$

The problem is: how do we choose w is that we end up with $P(V_j) = q_j$? It was shown by Aurenhammer, Hoffmann, Aronov (1998) that this corresponds to minimizing

$$F(w) = \sum_j \left(q_j w_j - \int_{V_j} [||x - y_j||^2 - w_j] dP(x) \right).$$

Merigot (2011) gives a multiscale method to minimize $F(w)$.

There are a few papers (Merigot 2011 and Gerber and Maggioni 2017) use multiscale methods for computing the distance. These approaches make use of decompositions like those used for the minimax theory.

Cuturi (2013) showed that if we replace $\inf \mathbb{E}||x - y||^p dJ(x, y)$ with the regularized version $\inf \mathbb{E}||x - y||^p dJ(x, y) + \int j(x, y) \log j(x, y)$ then a minimizer can be found using a fast, iterative algorithm called the Sinkhorn algorithm. However, this requires discretizing the space and it changes the metric.

10   Applications

The Wasserstein distance is now being used for many tasks in statistical machine learning including:

- Two-sample testing

- goodness-of-fit

- analysis of mixture models

- image processing

- dimension reduction

- generative adversarial networks

- domain adaptation

- signal processing

The domain adaptation application is very intriguing. Suppose we have two data sets $\mathcal{D}_1 = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\mathcal{D}_2 = \{(X_1', Y_1'), \ldots, (X_N', Y_N')\}$ from two related problems. We want to construct a predictor for the first problem. We could use just $\mathcal{D}_1$. But if we can find a transport map T that makes $\mathcal{D}_2$ similar to $\mathcal{D}_1$, then we can apply the map to $\mathcal{D}_2$ and effectively increase the sample size for problem 1. This kind of reasoning can be used for many statistical tasks.

11   Summary

Wasserstein distance has many nice properties and has become popular in statistics and machine learning. Recently, for example, it has been used for Generative Adversarial Networks (GANs).

But the distance does have problems. First, it is hard to compute. Second, as we have seen, we do not have a way to do inference for the distance. This reflects the fact that the distance is not a smooth functional which is, itself not a good thing. We have also seen that the distance is not robust although, the trimmed version may fix this.