Optimal Transport Wasserstein distance and Hypothesis Testing

Larry Wasserman
larry@cmu.edu

Two Sample Testing

$$X_1, \ldots, X_n \sim P$$

$$Y_1, \ldots, Y_m \sim Q$$

Two Sample Testing

$$X_1, \ldots, X_n \sim P$$
$$Y_1, \ldots, Y_m \sim Q$$

$$H_0 : P = Q$$

Two Sample Testing

$$X_1, \ldots, X_n \sim P$$
$$Y_1, \ldots, Y_m \sim Q$$

$$H_0 : P = Q$$

$$H_0 : W(P, Q) = 0$$

where $W(P, Q)$ is the Wasserstein distance.

Goodness of Fit

$$X_1, \ldots, X_n \sim P$$

Goodness of Fit

$$X_1, \ldots, X_n \sim P$$

$$H_0 : P = P_0$$

Goodness of Fit

$$X_1, \ldots, X_n \sim P$$

$$H_0 : P = P_0$$

$$H_0 : W(P, P_0) = 0$$

where $W(P, Q)$ is the Wasserstein distance.

Questions

Questions

1. What is Wasserstein distance?

Questions

1. What is Wasserstein distance?
2. Should we use it for testing?

The Wasserstein Distance

Assume that P has a density (but not really required). The distance $W(P, Q)$ is defined by

$$W^2(P, Q) = \mathbb{E}\left[||T(X) - X||^2 \right]$$

where T is the optimal transport map.

The Wasserstein Distance

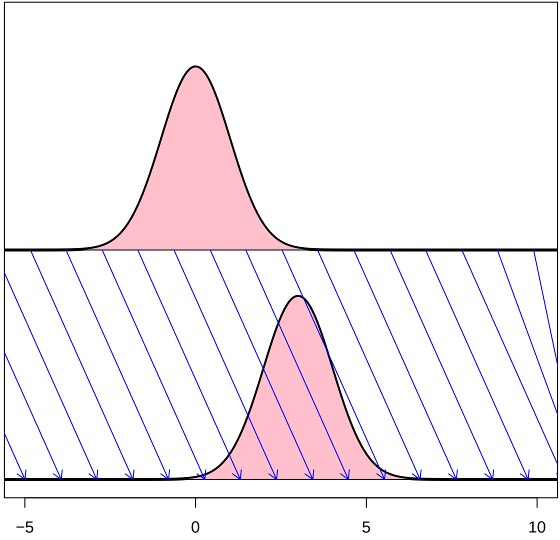
Assume that P has a density (but not really required). The distance $W(P, Q)$ is defined by

$$W^2(P, Q) = \mathbb{E}\left[||T(X) - X||^2 \right]$$

where T is the optimal transport map.

But what is the optimal transport map?

Optimal Transport



Formal Definition

Given two distributions: P_0 and P_1.

Formal Definition

Given two distributions: P_0 and P_1.

$X_0 \sim P_0$ (X_0 is a draw from P_0.)

Formal Definition

Given two distributions: P_0 and P_1.

$X_0 \sim P_0$ (X_0 is a draw from P_0.)

Find a map T that minimizes

$$\mathbb{E}\Big[||X_0 - T(X_0)||^2\Big] = \int ||x - T(x)||^2 dP_0(x)$$

subject to: $T(X_0) \sim P_1$.

Formal Definition

Given two distributions: P_0 and P_1.

$X_0 \sim P_0 \quad (X_0$ is a draw from P_0.)

Find a map T that minimizes

$$\mathbb{E}\Big[||X_0 - T(X_0)||^2\Big] = \int ||x - T(x)||^2 dP_0(x)$$

subject to: $T(X_0) \sim P_1$.

Can replace $(\cdots)^2$ with any cost.

What is T? Four special cases

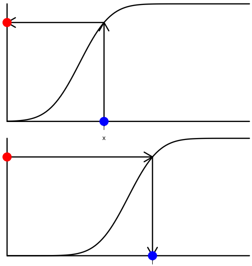
What is T? Four special cases

1. One dimension.

$$T(x) = F_1^{-1}(F_0(x))$$

where

$$F_0(t) = P_0(X \le t) \;\; \text{and} \;\; F_1(t) = P_1(X \le t).$$



What T?

2. If $P_0 = N(\mu_0, \Sigma_0)$ and $P_1 = N(\mu_1, \Sigma_1)$ then:

$$T(x) = \mu_1 + \Sigma_1^{1/2} \Sigma_0^{-1/2} (x - \mu_0).$$

What T?

What T?

3. Data clouds: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Then $T(X_i) = Y_{\pi(i)}$ where π is the permutation that minimizes

$$\sum_i ||X_i - Y_{\pi(i)}||^2.$$

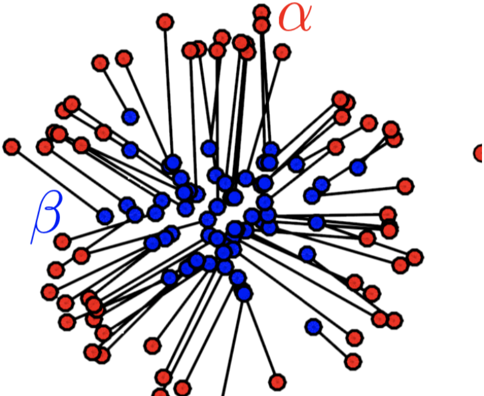
Hungarian algorithm: $O(n^3)$.

What T?

3. Data clouds: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Then $T(X_i) = Y_{\pi(i)}$ where π is the permutation that minimizes

$$\sum_i ||X_i - Y_{\pi(i)}||^2.$$

Hungarian algorithm: $O(n^3)$.



What T?

4. Convex optimization.

What T?

4. Convex optimization.

Brenier's theorem: $T = \nabla\phi$ where ϕ is the convex function that maximizes

$$\int \phi(x) dP_0(x) + \int \phi^*(x) dP_1(x)$$

where $\phi^*(x) = \sup_u\{\langle x, u\rangle - \phi(u)\}$.

What T?

4. Convex optimization.

Brenier's theorem: $T = \nabla\phi$ where ϕ is the convex function that maximizes

$$\int \phi(x)dP_0(x) + \int \phi^*(x)dP_1(x)$$

where $\phi^*(x) = \sup_u\{\langle x, u\rangle - \phi(u)\}$.

Now parameterize ϕ_θ using a (convex) neural net.

What if there is no such T?: More General Definition

The distance $W(P, Q)$ is defined by

$$W^2(P, Q) = \inf_{\pi} \mathbb{E}_{\pi} ||X - Y||^2$$

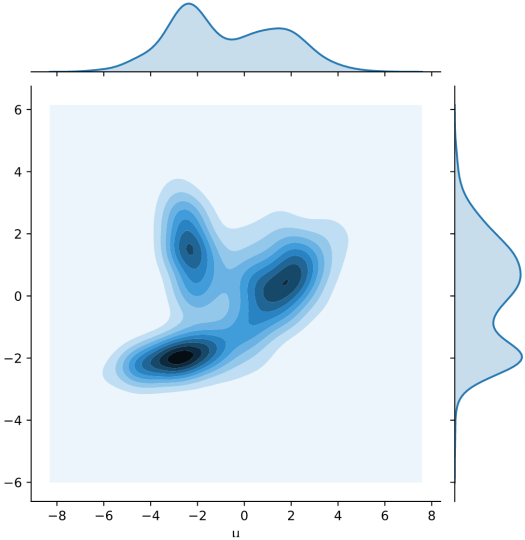
where

$$X \sim P$$

$$Y \sim Q$$

and the infimum is over all joint distributions π with marginals P and Q.

Optimal Transport



Joint distribution π with a given X marginal and a given Y marginal. Image credit: Wikipedia

Wasserstein Distance

$$P = Q \quad \text{iff} \quad W^2(P, Q) = 0 \quad \text{iff} \quad T(x) = x$$

Wasserstein Distance

$$P = Q \quad \text{iff} \quad W^2(P, Q) = 0 \quad \text{iff} \quad T(x) = x$$

We can test $H_0 : P = Q$ using an estimate of $W(P, Q)$. i.e. reject H_0 if

$$\widehat{W} > t$$

for some t.

Wasserstein Distance

$$P = Q \quad \text{iff} \quad W^2(P, Q) = 0 \quad \text{iff} \quad T(x) = x$$

We can test $H_0 : P = Q$ using an estimate of $W(P, Q)$. i.e. reject H_0 if

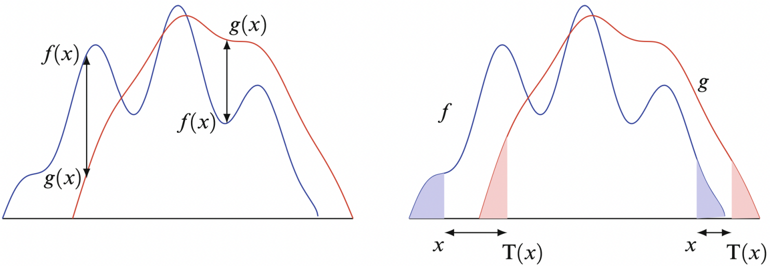
$$\widehat{W} > t$$

for some t.

Why use Wasserstein?

It has nice properties $\cdots$

Wasserstein versus $\int |p - q|^2$ (image from Santambrogio)



Wasserstein Distance has Nice Properties

Wasserstein distance is geometry sensitive.

Wasserstein Distance has Nice Properties

Wasserstein distance is geometry sensitive.

Let P be a point mass x. Let Q be a point mass y.

Wasserstein Distance has Nice Properties

Wasserstein distance is geometry sensitive.

Let P be a point mass x. Let Q be a point mass y.

$KS(P, Q) = 1$.

Wasserstein Distance has Nice Properties

Wasserstein distance is geometry sensitive.

Let P be a point mass x. Let Q be a point mass y.

$KS(P, Q) = 1$.

$W(P, Q) = |x - y|$

Wasserstein Distance has Nice Properties

Wasserstein distance is geometry sensitive.

Let P be a point mass x. Let Q be a point mass y.

$KS(P, Q) = 1$.

$W(P, Q) = |x - y|$

Suggests that this may have more power for certain deviations from the null.

Wasserstein Distance has Nice Properties

Wasserstein Distance has Nice Properties

What is the average of $N(-3, 1)$ and $N(3, 1)$?

Wasserstein Distance has Nice Properties

What is the average of $N(-3,1)$ and $N(3,1)$?

Euclidean average is

$$\frac{1}{2}N(-3,1)+\frac{1}{2}N(3,1)$$

Wasserstein Distance has Nice Properties

What is the average of $N(-3, 1)$ and $N(3, 1)$?

Euclidean average is

$$\frac{1}{2}N(-3,1) + \frac{1}{2}N(3,1)$$

The Wasserstein average (barycenter): B minimizes

$$W^2(P_1, B) + W^2(P_2, B)$$

Wasserstein Distance has Nice Properties

What is the average of $N(-3,1)$ and $N(3,1)$?

Euclidean average is

$$\frac{1}{2}N(-3,1) + \frac{1}{2}N(3,1)$$

The Wasserstein average (barycenter): B minimizes

$$W^2(P_1, B) + W^2(P_2, B)$$

The solution is

$$B = N(0,1)$$

Connection to Fluid Dynamics

$$W^2(P, Q) = \min_v \int_0^1 \int ||v(x, t)||^2 \rho_t(x) dx\, dt$$

where

$$\rho_0 = P_0$$

$$\rho_1 = P_1$$

and

$$\partial_t \rho_t + \nabla(\rho_t v_t) = 0$$

Negative Sobolov Norm

$$c||p - q||_{\dot{H}^{-1}} \leq W(P, Q) \leq C||p - q||_{\dot{H}^{-1}}$$

where

$$||f||_{\dot{H}^{-1}} = \sup\left\{ \int gf : \int |\nabla g|^2 \leq 1 \right\}$$

How Do We Estimate $W(P, Q)$?

Plugin estimator: Estimate $W(P, Q)$ with

$$\widehat{W} = W(P_n, Q_n)$$

where P_n is the empirical distribution of the data that puts mass $1/n$ at each X_i. Q_n is the empirical distribution of the data that puts mass $1/n$ at each Y_i.

How Do We Estimate $W(P, Q)$?

Plugin estimator: Estimate $W(P, Q)$ with

$$\widehat{W} = W(P_n, Q_n)$$

where P_n is the empirical distribution of the data that puts mass $1/n$ at each X_i. Q_n is the empirical distribution of the data that puts mass $1/n$ at each Y_i.

Then use the Hungarian algorithm $O(n^3)$.

How Do We Estimate $W(P, Q)$?

Plugin estimator: Estimate $W(P, Q)$ with

$$\widehat{W} = W(P_n, Q_n)$$

where P_n is the empirical distribution of the data that puts mass $1/n$ at each X_i. Q_n is the empirical distribution of the data that puts mass $1/n$ at each Y_i.

Then use the Hungarian algorithm $O(n^3)$.

Two problems:

How Do We Estimate $W(P, Q)$?

Plugin estimator: Estimate $W(P, Q)$ with

$$\widehat{W} = W(P_n, Q_n)$$

where P_n is the empirical distribution of the data that puts mass $1/n$ at each X_i. Q_n is the empirical distribution of the data that puts mass $1/n$ at each Y_i.

Then use the Hungarian algorithm $O(n^3)$.

Two problems:

1. This is slow.

How Do We Estimate $W(P, Q)$?

Plugin estimator: Estimate $W(P, Q)$ with

$$\widehat{W} = W(P_n, Q_n)$$

where P_n is the empirical distribution of the data that puts mass $1/n$ at each X_i. Q_n is the empirical distribution of the data that puts mass $1/n$ at each Y_i.

Then use the Hungarian algorithm $O(n^3)$.

Two problems:

1. This is slow.
2. $\widehat{W}$ is a poor estimate of W:

$$\widehat{W} - W = O(n^{-1/d})$$

where $d =$ the dimension of X.

Better estimator

Use $W(\widehat{p}, \widehat{q})$
where $\widehat{p}$ is a smooth estimate of the density of P and $\widehat{q}$ is a smooth estimate of the density of Q.

Better estimator

Use $W(\widehat{p}, \widehat{q})$
where $\widehat{p}$ is a smooth estimate of the density of P and $\widehat{q}$ is a smooth estimate of the density of Q.

Then

$$\widehat{W} - W \approx \left(\frac{1}{n}\right)^{\frac{2\alpha}{2(\alpha-1)+d}}$$

where $T \in \mathrm{Holder}(\alpha)$.

Better estimator

Use $W(\widehat{p}, \widehat{q})$
where $\widehat{p}$ is a smooth estimate of the density of P and $\widehat{q}$ is a smooth estimate of the density of Q.

Then

$$\widehat{W} - W \approx \left(\frac{1}{n}\right)^{\frac{2\alpha}{2(\alpha-1)+d}}$$

where $T \in \mathrm{Holder}(\alpha)$.

If $\alpha + 1 > d/2$ then

$$\sqrt{n}(\widehat{W}^2 - W^2) \rightsquigarrow N(0, \sigma^2)$$

which can simplify inference.

(Manole et al arXiv:2107.12364)

Computation

Speeding up computations is a very active area.

Computation

Speeding up computations is a very active area.

1. Regularized (entropic) transport can be computed in $O(n^2)$.

Computation

Speeding up computations is a very active area.

1. Regularized (entropic) transport can be computed in $O(n^2)$.
2. Can use neural net, convex optimization?

Computation

Speeding up computations is a very active area.

1. Regularized (entropic) transport can be computed in $O(n^2)$.
2. Can use neural net, convex optimization?
3. Mini-batch. Take subsamples of size k and average.

Computation

Speeding up computations is a very active area.

1. Regularized (entropic) transport can be computed in $O(n^2)$.
2. Can use neural net, convex optimization?
3. Mini-batch. Take subsamples of size k and average.

Many others ...

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.
3. Might have good power?

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.
3. Might have good power?

Con's:

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.
3. Might have good power?

Con's:

1. Expensive.

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.
3. Might have good power?

Con's:

1. Expensive.
2. Getting the rejection threshold is not easy.

Is This Useful?

Pro's:

1. Get a transport map $\widehat{T}$.
2. Wasserstein distance is a meaningful distance.
3. Might have good power?

Con's:

1. Expensive.
2. Getting the rejection threshold is not easy.

THE END