Introduction to Optimal Transport With Application to Estimating Background Distributions in Particle Physics

Tudor Manole, Patrick Bryant
John Alison, Mikael Kuusela
Larry Wasserman

Motivating Example

Two similar distributions: P_{3b} and P_{4b}.

Motivating Example

Two similar distributions: P_{3b} and P_{4b}.

Sample space $\mathcal{X} = C \bigcup S$.

Motivating Example

Two similar distributions: P_{3b} and P_{4b}.

Sample space $\mathcal{X} = C \bigcup S$.

Given: a sample

$$X_1, \ldots, X_n \sim P_{3b}$$

and a sample

$$Y_1, \ldots, Y_m \sim P_{4b}(\cdot|C)$$

estimate $P_{4b}(\cdot|S)$.

Motivating Example

Two similar distributions: P_{3b} and P_{4b}.

Sample space $\mathcal{X} = C \bigcup S$.

Given: a sample

$$X_1, \ldots, X_n \sim P_{3b}$$

and a sample

$$Y_1, \ldots, Y_m \sim P_{4b}(\cdot|C)$$

estimate $P_{4b}(\cdot|S)$.

$C =$ control region, $S =$ signal region

Motivating Example

Two similar distributions: P_{3b} and P_{4b}.

Sample space $\mathcal{X} = C \bigcup S$.

Given: a sample

$$X_1, \ldots, X_n \sim P_{3b}$$

and a sample

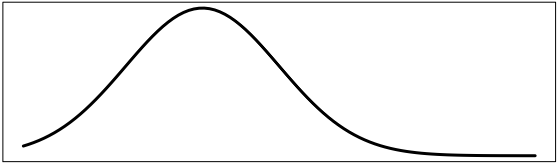
$$Y_1, \ldots, Y_m \sim P_{4b}(\cdot|C)$$

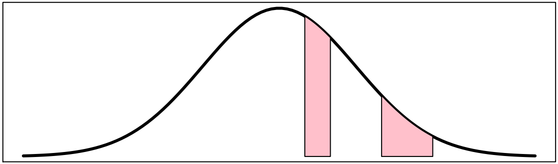
estimate $P_{4b}(\cdot|S)$.

C = control region, S = signal region

The problem is ill-posed; we will have to make (reasonable)

Motivating Example





Motivating Example

- The data are collider events.

Motivating Example

- The data are collider events.
- The data are in 16 dimensions.

Motivating Example

- The data are collider events.
- The data are in 16 dimensions.
- The set S can be a very complex set.

Motivating Example

- The data are collider events.
- The data are in 16 dimensions.
- The set S can be a very complex set.
- The metric on $\mathcal{X}$ is non-standard.

Motivating Example

- The data are collider events.

- The data are in 16 dimensions.

- The set S can be a very complex set.

- The metric on $\mathcal{X}$ is non-standard.

plus a bunch of other complications.

Three Methods

1. Density ratio: estimate $\frac{p_{3b}(x)}{p_{4b}(x)}$ over C and extend to S.

Three Methods

1. Density ratio: estimate $\frac{p_{3b}(x)}{p_{4b}(x)}$ over C and extend to S.

2. Optimal transport: Use P_{3b} to find a map T that transports mass from C to S. Apply the map to P_{4b}.

Three Methods

1. Density ratio: estimate $\frac{p_{3b}(x)}{p_{4b}(x)}$ over C and extend to S.

2. Optimal transport: Use P_{3b} to find a map T that transports mass from C to S. Apply the map to P_{4b}.

3. Combination: Transport S to C then apply ratio.

Optimal Transport

Introduction: What is Optimal Transport?

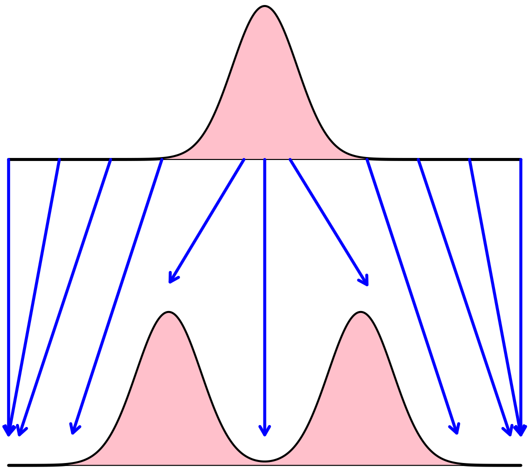
We have two distributions P_0 and P_1.

Introduction: What is Optimal Transport?

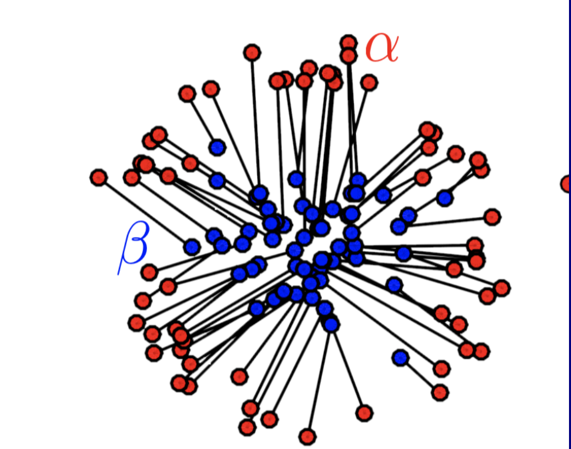
We have two distributions P_0 and P_1.
Goals:

- Define an "optimal map" that transforms P_0 into P_1.
- Define a distance based on transport (Wasserstein distance)
- Define a path (geodesic) between P_1 and P_2 (morphing) in the space of distributions.
- Define a shape-preserving notion of" averages" of distributions.

Optimal Transport (Monge 1781)



Point Cloud Example (from Peyre, Cuturi 2019)



Optimal Transport: Monge Version

Let $X \sim P_0$.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p \right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance. We will use a metric betwen collider events (which is itself a type of transport).

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p \right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance. We will use a metric betwen collider events (which is itself a type of transport).

For now, assume that the minimizer exists. The the minimizer is called the optimal transport map.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance. We will use a metric betwen collider events (which is itself a type of transport).

For now, assume that the minimizer exists. The the minimizer is called the optimal transport map.

Common choices: $p = 2$ or $p = 1$.

Wasserstein (transport) distance

$$W_p(X,Y) \equiv W_p(P_0,P_1) = \left(\int ||x - T^*(x)||^p dP_0(x) \right)^{1/p}$$

where T^* is the optimal transport map.

Defines a metric on the space of (nearly) all distributions.

W_1 is called the Earth Mover Distance

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$

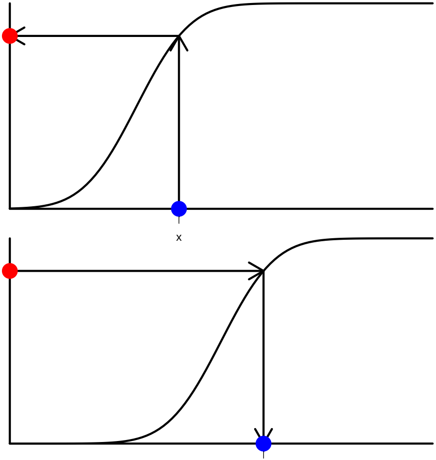
Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$
- $W_p(P_0, P_1) = \big(\int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds\big)^{1/p}$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \le s)$
- $F_1(s) = P_1(Y \le s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$
- $W_p(P_0, P_1) = \big(\int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds\big)^{1/p}$
- The morphing — geodesic linking F_0 and F_1 — is

$$F_s = \big[(1-s)F_0^{-1} + sF_1^{-1}\big]^{-1}$$


x

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

$$\widehat{F}_0(s) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le s)$$

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

$$\widehat{F}_0(s) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq s)$$

$$\widehat{F}_1(s) = \frac{1}{m}\sum_{i=1}^{m} I(Y_i \leq s)$$

Finding the Transport Map: Gaussian Case

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Then:

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

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Then:

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

$$W_2^2(P_0, P_1) = ||\mu_0 - \mu_1||^2 + B(\Sigma_0, \Sigma_1)$$

where

$$B(\Sigma_0, \Sigma_1) = \mathrm{trace}(\Sigma_0) + \mathrm{trace}(\Sigma_1) - 2\mathrm{trace}\Big[(\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2}\Big].$$

Finding the Transport Map: Two Point Clouds

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$
- $\mathcal{Y} = \{Y_1, \ldots, Y_n\} \quad Y_i \in \mathbb{R}^d$

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$
- $\mathcal{Y} = \{Y_1, \ldots, Y_n\} \quad Y_i \in \mathbb{R}^d$
- $T : X_i \to Y_{\pi(i)}$ where π minimizes
$$\sum_i ||X_i - Y_{\pi(i)}||^2$$
  over all permutations π.
- Hungarian algorithm $O(n^3)$ time.

How Accurate is This?

$X_1, \ldots, X_n \sim P$

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

$\widehat{T}$ is estimated from data (and extended by one-nearest-neighbor):

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

$\widehat{T}$ is estimated from data (and extended by one-nearest-neighbor):

under conditions (Manole, Balakrishnan and Wasserman, in progress):

$$\mathbb{E}||\widehat{T}(X) - T(X)||^2 = O(n^{-2/d})$$

and this is optimal without further conditions.

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.
estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.
estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.
estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

This is suboptimal but easy.

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.
estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

This is suboptimal but easy.
It does estimate the smoothed transport $T_h : p \star K_h \to q \star K_h$ optimally. The rate is $n^{-1/2}$ independent of dimension.

Smooth Transport

With smoothness assumptions (on P and Q or T) we can estimate T at a faster rate (Hutter and Rigollet 2019). But the method is impractical. (Requires wavelet estimator with difficult constraints.)

Instead we can:

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.
estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

This is suboptimal but easy.
It does estimate the smoothed transport $T_h : p \star K_h \to q \star K_h$ optimally. The rate is $n^{-1/2}$ independent of dimension.
We are currently trying to show that the bootstrap gives valid confidence intervals for $T_h(x)$. (And bias correction.)

Other Computing Methods

Other Computing Methods

- Sinkhorn (Cuturi 2013)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)
- Subsampling (Sommerfeld, Schrieber, Zemel and Munk, 2018)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)
- Subsampling (Sommerfeld, Schrieber, Zemel and Munk, 2018)
- Hubs (Forrow et al 2018)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)
- Subsampling (Sommerfeld, Schrieber, Zemel and Munk, 2018)
- Hubs (Forrow et al 2018)
- See: POT (Python Optimal Transport) `https://pot.readthedocs.io/en/stable/`

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a shortest path (geodesic) between them.

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a shortest path (geodesic) between them.
- $T_s(x) = (1 - s)x + sT(x)$

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a shortest path (geodesic) between them.

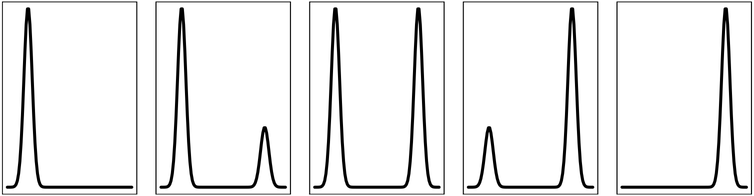
- $T_s(x) = (1 - s)x + sT(x)$
- $P_s = T_{s\#}P$.
  In other words, P_s is the distribution of the random variable $(1 - s)X + sT(X)$ where $X \sim P_0$.

Geodesics (Morphing)

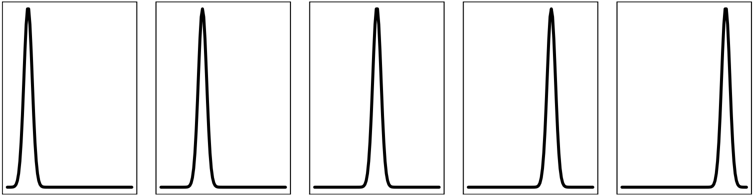
- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a shortest path (geodesic) between them.

- $T_s(x) = (1 - s)x + sT(x)$
- $P_s = T_{s\#}P$.
  In other words, P_s is the distribution of the random variable $(1 - s)X + sT(X)$ where $X \sim P_0$.
- Then $(P_s : 0 \leq t \leq 1)$ is the geodesic.
  Length of the path $= W(P_0, P_1)$.

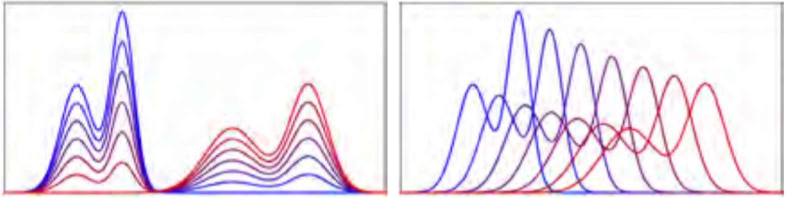
Euclidean Path between Two Gaussians



Geodesic Path between Two Gaussians



Geodesic Path between Two Mixtures: Bonneel, Peyre, Cuturi 2016

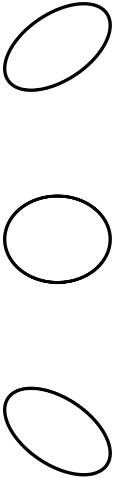


ℓ_2 interpolation   Wasserstein interpolation

Geodesic Path Between Two Images

Image credit: Bauer, Joshi and Modin 2015.

Bivariate Gaussian



Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N}\sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Wasserstein barycenter: find P to minimize:

$$\sum_j W^2(P, P_j).$$

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N}\sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Wasserstein barycenter: find P to minimize:

$$\sum_j W^2(P, P_j).$$

This is the barycenter and it is shape preserving.

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N}\sum_j P_j$$

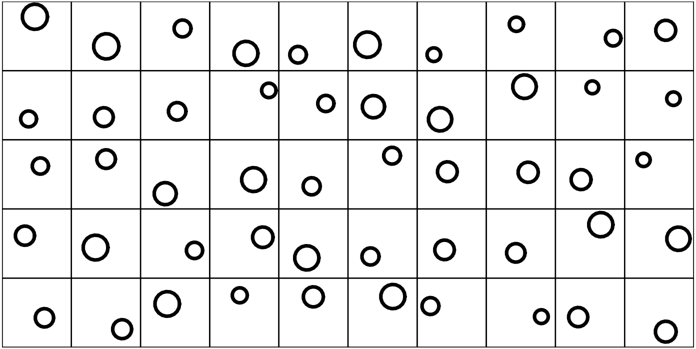
Same problem as before: this does not look like any of the P_j's.

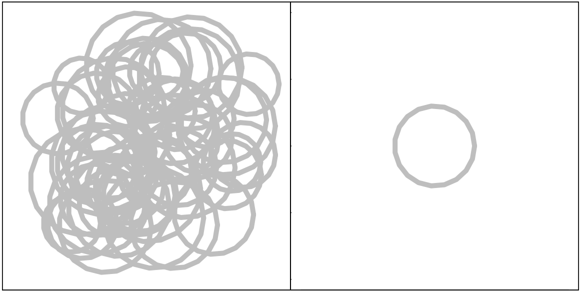
Wasserstein barycenter: find P to minimize:

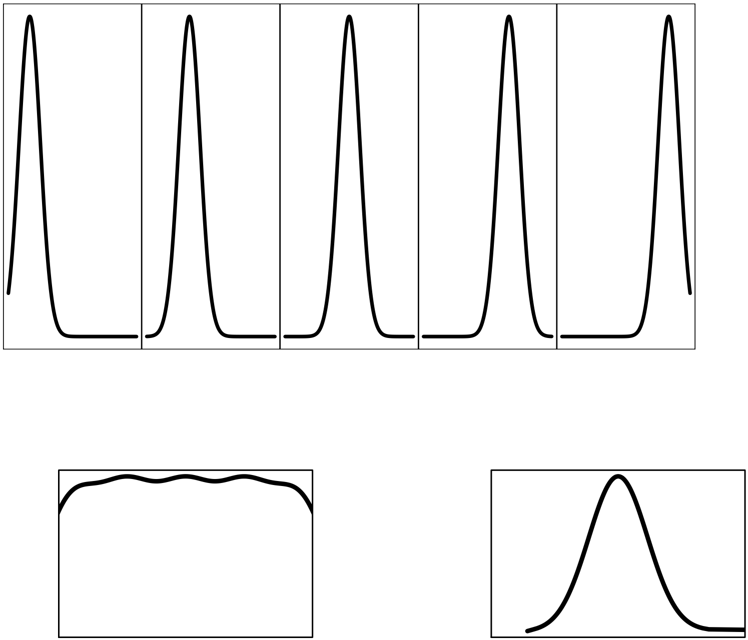
$$\sum_j W^2(P, P_j).$$

This is the barycenter and it is shape preserving.

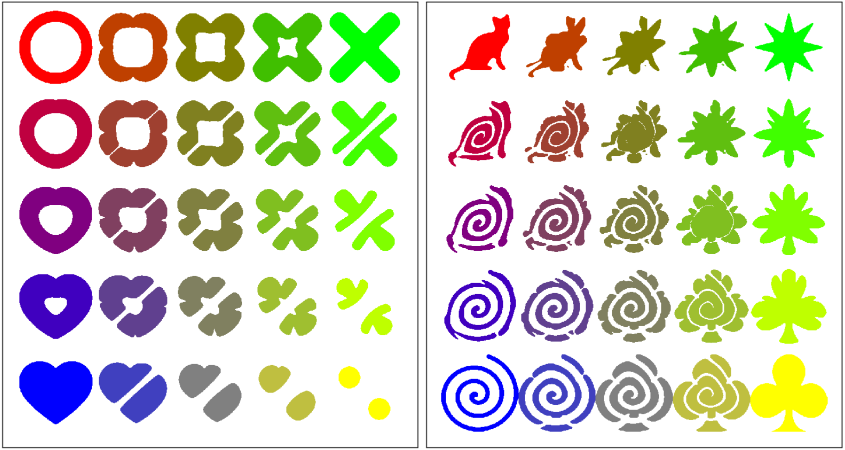
We can then define morphings from the barycenter to each of the P_j.







Example from Peyre and Cuturi 2019



How to Compute the Barycenter?

Active research area.

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

$\overline{F} = Q^{-1}$ where

$$Q(u) = \frac{1}{N}\sum_j F_j^{-1}(u)$$

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

$\overline{F} = Q^{-1}$ where

$$Q(u) = \frac{1}{N} \sum_j F_j^{-1}(u)$$

See Claici, Chien, Solomon (arXiv:1802.05757) and references therein.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Solution: Kantoravich relaxation:

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Solution: Kantoravich relaxation:

Take mass at x, and split it into many small pieces.

Optimal Transport (Kantorovich Version)

Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.

Optimal Transport (Kantorovich Version)

Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.
Find J (optimal transport plan) to minimize

$$\mathbb{E}_J[||X - Y||] = \left(\int ||x - y||^p \; dJ(x, y) \right)^{1/p}.$$

Optimal Transport (Kantorovich Version)

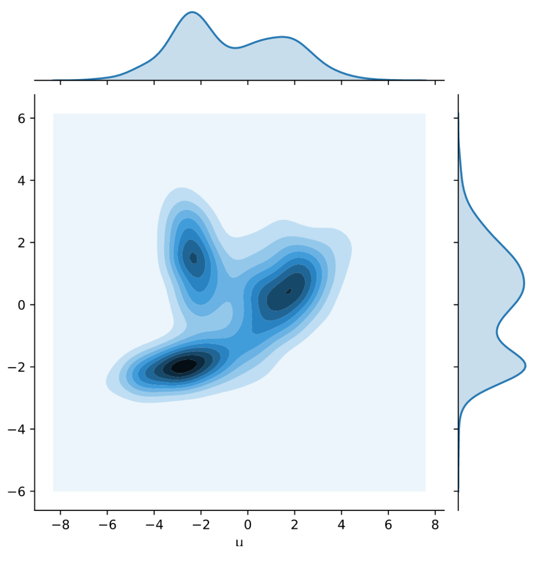
Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.
Find J (optimal transport plan) to minimize

$$\mathbb{E}_J[||X - Y||] = \left(\int ||x - y||^p \; dJ(x, y)\right)^{1/p}.$$

Again, this defines a distance

$$W(P, Q) = W(X, Y) = \left(\inf_J \int (||x - y||^2 dJ(x, y)\right)^{1/2}$$

called the Wasserstein distance.



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

Morphing

In this case, the morphing (geodesic) can be described as follows.

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

that is,

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

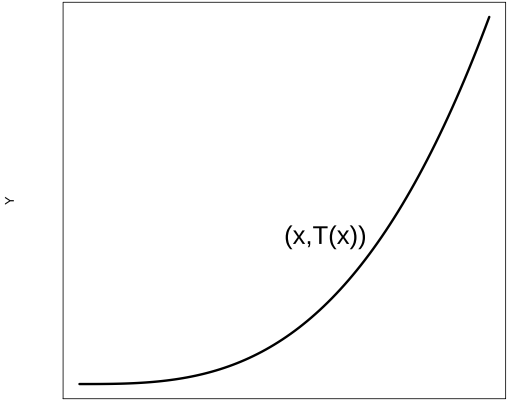
that is,

$P_s = F_{s\#}J$.

If a Transport Map Exists

If an optimal transport map T exists the the optimal coupling J is degenerate and is supported on the curve

$$\mathcal{S} = \{(x, T(x))\}$$



Regularized Optimal Transport

Find J (optimal transport plan) to minimize

$$\left(\int ||x - y||^p dJ(x,y) \right)^{1/p} + \lambda f(J)$$

for some f.

Regularized Optimal Transport

Find J (optimal transport plan) to minimize

$$\left(\int ||x - y||^p dJ(x, y) \right)^{1/p} + \lambda f(J)$$

for some f.

For example, Cuturi (2013) uses the entropy:

$$f(J) = - \int j(x, y) \log j(x, y)$$

Regularized Optimal Transport

Advantages:

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Disadvantages:

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Disadvantages:

(i) How to choose λ?

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

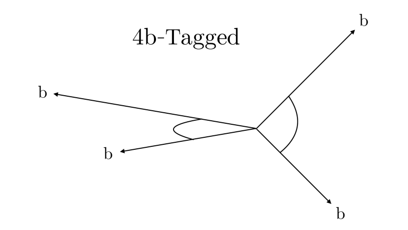
Disadvantages:

(i) How to choose λ?

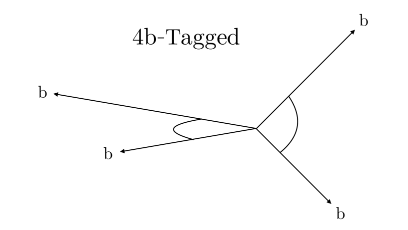
(ii) Effect of regularization is not clear.

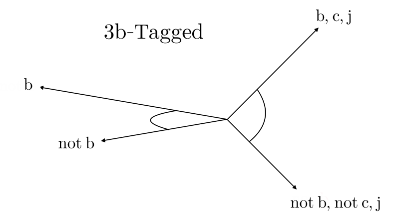
Background Modelling for Double Higgs Boson Production

Background Modelling via 3b Events



Background Modelling via 3b Events





The Metric Space of Collider Events

A jet is (p, η, ϕ, m) where p = momentum, m = mass, ϕ and η are angles.

The Metric Space of Collider Events

A jet is (p, η, ϕ, m) where p = momentum, m = mass, ϕ and η are angles.

An event is 4 jets. We treat it as a measure:

$$\mathcal{E} = \sum_{i=1}^{4} p_i \delta_i$$

where δ_i is a point mass at (η_i, ϕ_i, m_i).

The Metric Space of Collider Events

A jet is (p, η, ϕ, m) where p = momentum, m = mass, ϕ and η are angles.

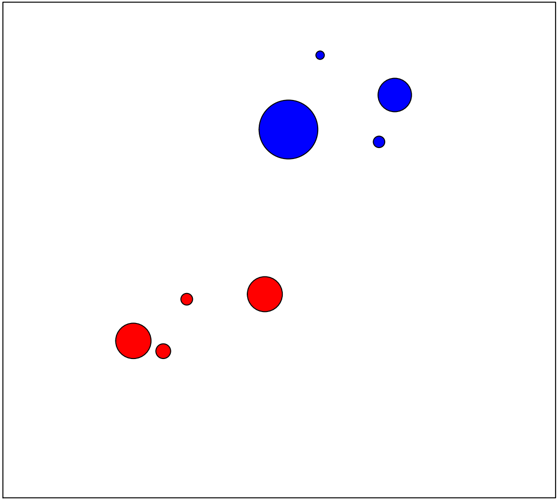
An event is 4 jets. We treat it as a measure:

$$\mathcal{E} = \sum_{i=1}^{4} p_i \delta_i$$

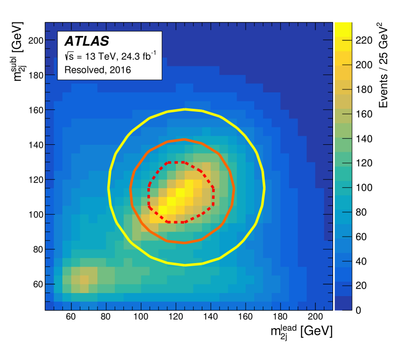
where δ_i is a point mass at (η_i, ϕ_i, m_i).

The metric between two events g_1 and g_2 is the (modified) Wasserstein distance, a metric between measures.
or: $\mathcal{E}$ is a vector in $\mathbb{R}^{16}$ with a weird geometry.
see Komiske, Metodiev and Thaler (2019).

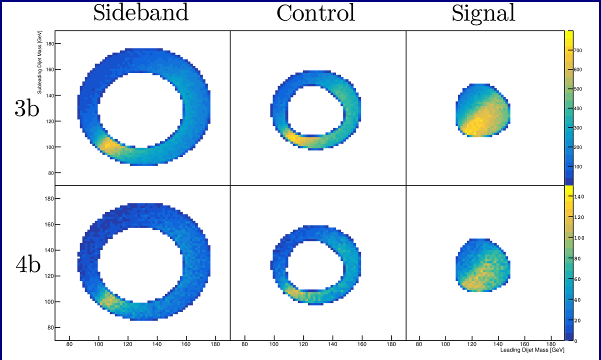
Events as Measures



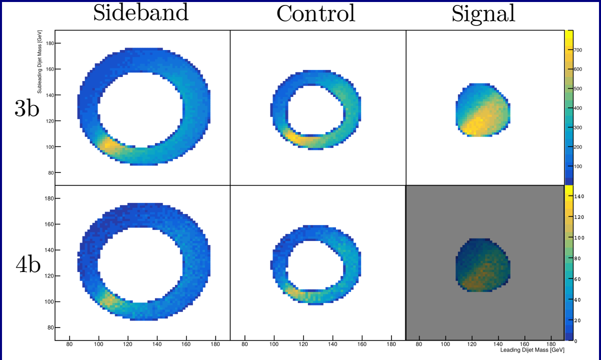
Sideband, Control and Signal Regions



Sideband, Control and Signal Regions



Sideband, Control and Signal Regions



Density Ratios and Classifiers

In general, given two densities p and q and samples

$$X_1, \ldots, X_n \sim p$$

$$Y_1, \ldots, Y_n \sim q$$

Density Ratios and Classifiers

In general, given two densities p and q and samples

$$X_1, \ldots, X_n \sim p$$

$$Y_1, \ldots, Y_n \sim q$$

$$\begin{array}{c|cccccc} & X_1 & \ldots & X_n & Y_1 & \ldots & Y_n \\ Z & 1 & \ldots & 1 & 0 & \ldots & 0 \end{array}$$

Density Ratios and Classifiers

In general, given two densities p and q and samples

$$X_1, \ldots, X_n \sim p$$

$$Y_1, \ldots, Y_n \sim q$$

$$\begin{array}{c|cccccc} & X_1 & \ldots & X_n & Y_1 & \ldots & Y_n \\ Z & 1 & \ldots & 1 & 0 & \ldots & 0 \end{array}$$

Classifier ψ:

$$\psi(u) = P(Z = 1|u) = \frac{p}{p+q}$$

and so

$$\frac{p}{q} = \frac{\psi}{1-\psi}.$$

Density Ratios and Classifiers

Modern classifiers (neural nets, random forests) are very accurate so we use classifiers to estimate the density ratios. No one really knows why.

Density Ratios and Classifiers

Modern classifiers (neural nets, random forests) are very accurate so we use classifiers to estimate the density ratios. No one really knows why.

We will assume in what follows that the ratio can be estimated well.

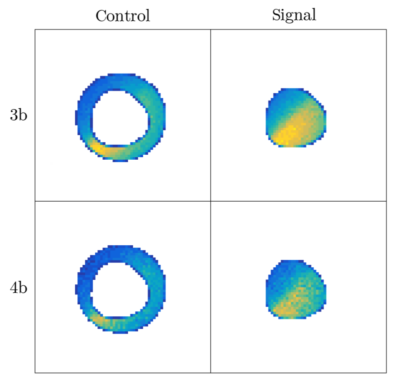
Density Ratios and Classifiers

Modern classifiers (neural nets, random forests) are very accurate so we use classifiers to estimate the density ratios. No one really knows why.

We will assume in what follows that the ratio can be estimated well.

We use a specially designed neural net built by Patrick. The model uses knowledge of the structure of the data. (Respects certain symmetries.)

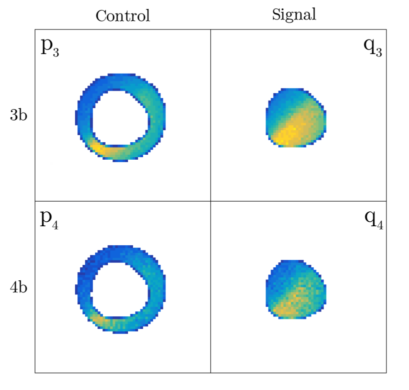
Extrapolating Density Ratios



Extrapolating Density Ratios

At the population level,

- Let $\psi(x) = \mathbb{P}(X \text{ is in 4b} | X = x)$.

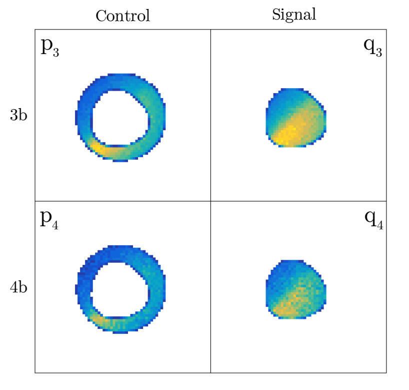


Extrapolating Density Ratios

At the population level,

- Let $\psi(x) = \mathbb{P}(X \text{ is in 4b} | X = x)$.
- Then,

$$p_4(x) \propto \frac{\psi(x)}{1 - \psi(x)} p_3(x).$$



Extrapolating Density Ratios

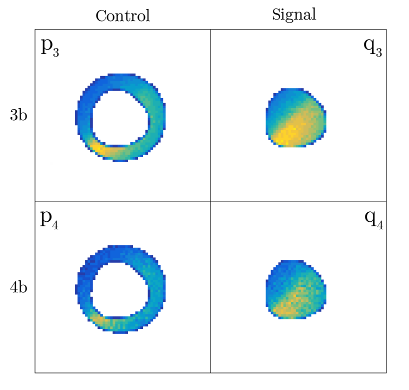
At the population level,

- Let $\psi(x) = \mathbb{P}(X \text{ is in 4b} | X = x)$.
- Then,

$$p_4(x) \propto \frac{\psi(x)}{1-\psi(x)} p_3(x).$$

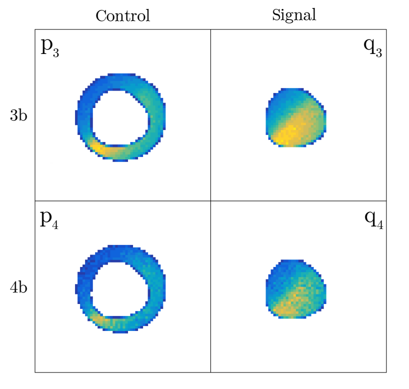
- Similarly,

$$q_4(x) \propto \frac{\psi(x)}{1-\psi(x)} q_3(x).$$



Extrapolating Probability Ratios

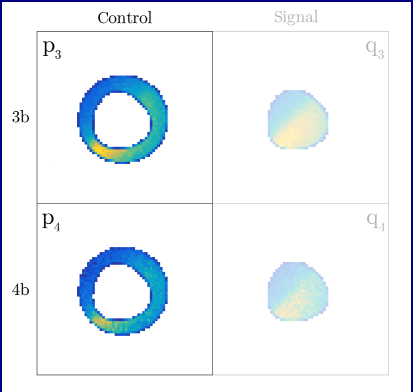
In practice,



Extrapolating Probability Ratios

In practice,

- Train a classifier $\widehat{h}$ on the 3b and 4b control regions.

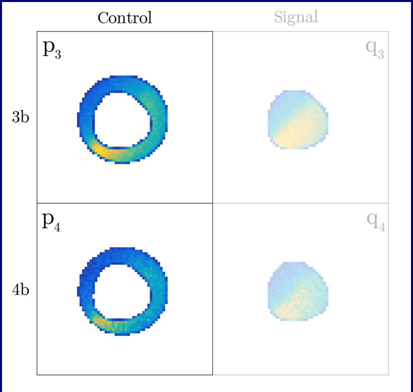


Extrapolating Probability Ratios

In practice,

- Train a classifier $\widehat{h}$ on the 3b and 4b control regions.
- For all x in the control region,

$$p_4(x) \approx \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} p_3(x).$$



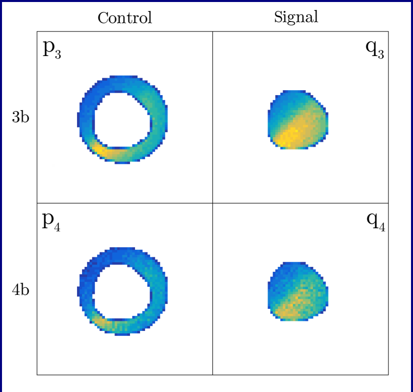
Extrapolating Probability Ratios

In practice,

- Train a classifier $\widehat{h}$ on the 3b and 4b control regions.
- For all x in the control region,

$$p_4(x) \approx \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} p_3(x).$$

- Estimate a histogram $\widehat{q}_3$ of q_3.



Extrapolating Probability Ratios

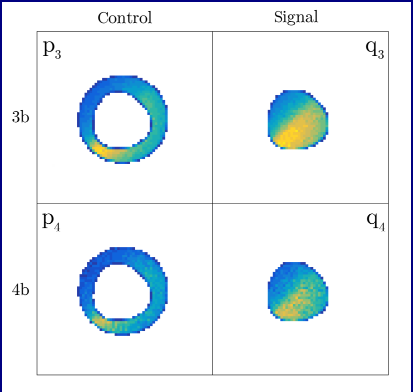
In practice,

- Train a classifier $\widehat{h}$ on the 3b and 4b control regions.
- For all x in the control region,

$$p_4(x) \approx \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} p_3(x).$$

- Estimate a histogram $\widehat{q}_3$ of q_3.
- Final estimate:

$$\boxed{\widehat{q}_4(x) := \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} \widehat{q}_3(x)}$$



Extrapolating Probability Ratios

In practice,

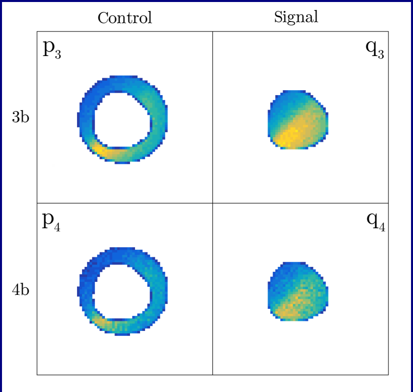
- Train a classifier $\widehat{h}$ on the 3b and 4b control regions.
- For all x in the control region,

$$p_4(x) \approx \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} p_3(x).$$

- Estimate a histogram $\widehat{q}_3$ of q_3.
- Final estimate:

$$\boxed{\widehat{q}_4(x) := \frac{\widehat{\psi}(x)}{1-\widehat{\psi}(x)} \widehat{q}_3(x)}$$

- Assumption: Transfer learning to a phase space with different support.



Optimal Transport

Let $X \sim P_{3b}(\cdot|C)$ and $Y \sim P_{3b}(\cdot|S)$. Find $T : C \to S$ to minimize

$$\int ||T(x) - x||^2 dP_{3b}(\cdot|C)$$

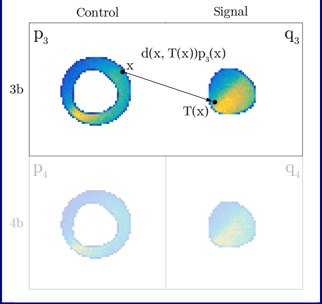
subject to $T(X) \sim P_{3b}(\cdot|S)$. (Monge map).

Optimal Transport

Let $X \sim P_{3b}(\cdot|C)$ and $Y \sim P_{3b}(\cdot|S)$. Find $T : C \to S$ to minimize

$$\int ||T(x) - x||^2 dP_{3b}(\cdot|C)$$

subject to $T(X) \sim P_{3b}(\cdot|S)$. (Monge map).



Double Optimal Transport

Double Optimal Transport

If P_n and Q_n are two empirical measures with the same sample size then:

$$T(X_i) = Y_{\pi(i)}$$

where π minimizes

$$\sum_i d(X_i, Y_{\pi(i)}).$$

Double Optimal Transport

If P_n and Q_n are two empirical measures with the same sample size then:

$$T(X_i) = Y_{\pi(i)}$$

where π minimizes

$$\sum_i d(X_i, Y_{\pi(i)}).$$

Note that computing d is itself an optimal transport problem!

Double Optimal Transport

If P_n and Q_n are two empirical measures with the same sample size then:

$$T(X_i) = Y_{\pi(i)}$$

where π minimizes

$$\sum_i d(X_i, Y_{\pi(i)}).$$

Note that computing d is itself an optimal transport problem!

T can be found in $O(n^3)$ time.

Unequal Sample Sizes

Unequal Sample Sizes

When the sample sizes are unequal, we instead use the Kantorovich Relaxation (allow mass to go to more than one point).

Unequal Sample Sizes

When the sample sizes are unequal, we instead use the Kantorovich Relaxation (allow mass to go to more than one point).

Find a coupling h to minimize

$$\int \int d(x, y) h(x, y) dx dy$$

over all h such that

$$\int h(x, y) dx = q_{3b}(y|C), \quad \int h(x, y) dy = q_{3b}(x|S).$$

Unequal Sample Sizes

When the sample sizes are unequal, we instead use the Kantorovich Relaxation (allow mass to go to more than one point).

Find a coupling h to minimize

$$\int \int d(x,y)h(x,y)dxdy$$

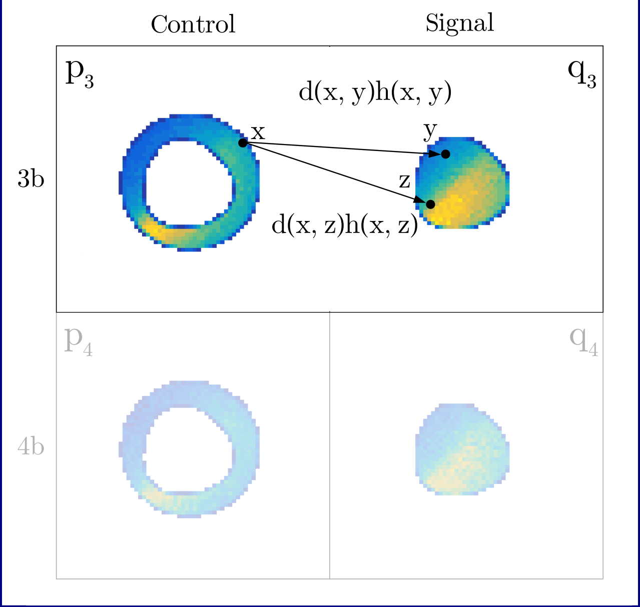
over all h such that

$$\int h(x,y)dx = q_{3b}(y|C), \qquad \int h(x,y)dy = q_{3b}(x|S).$$

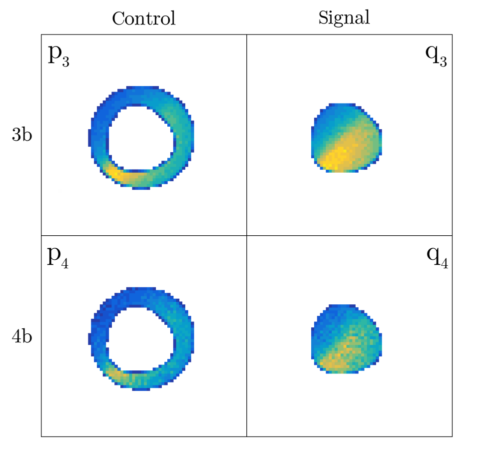
For empirical measures $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, $Q_m = \frac{1}{m}\sum_{j=1}^{m} \delta_{Y_j}$,

$$\underset{\substack{H=(h_{ij})\in\mathbb{R}_+^{n\times m} \\ \sum_{i=1}^{n} h_{ij}=1/m \\ \sum_{j=1}^{m} h_{ij}=1/n}}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij} d(X_i, Y_i)$$

The Kantorovich Relaxation



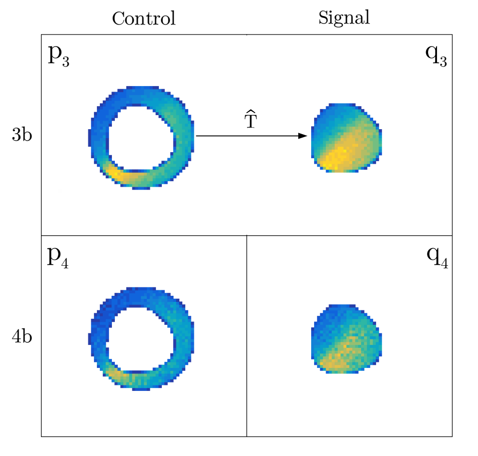
Estimating q_4 using Optimal Transport



Estimating q_4 using Optimal Transport

Procedure:

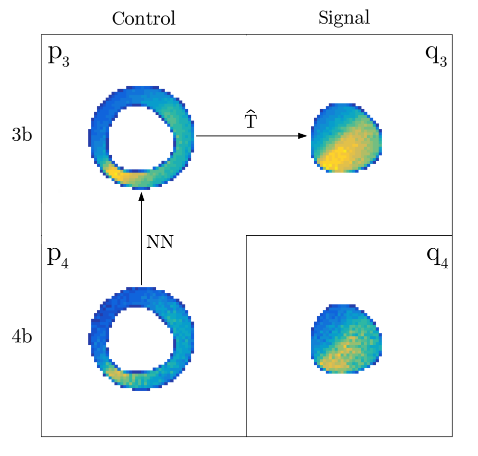
1. Compute CR→SR coupling in 3b.



Estimating q_4 using Optimal Transport

Procedure:

1. Compute CR→SR coupling in 3b.
2. Find 4b→3b nearest neighbor in CR.

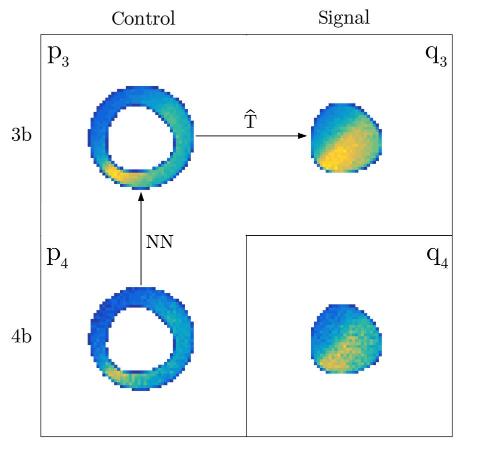


Estimating q_4 using Optimal Transport

Procedure:

1. Compute CR→SR coupling in 3b.
2. Find 4b→3b nearest neighbor in CR.
3. Form a histogram $\widehat{q}_4$ of the resulting point cloud. Loosely,

$$\widehat{q}_4(x) \propto \widehat{p}_4(\widehat{T}(x))$$

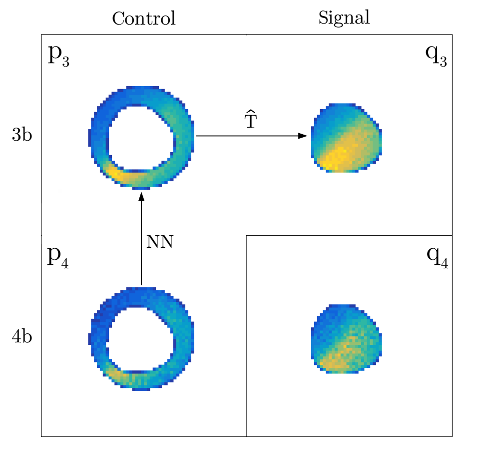


Estimating q_4 using Optimal Transport

Procedure:

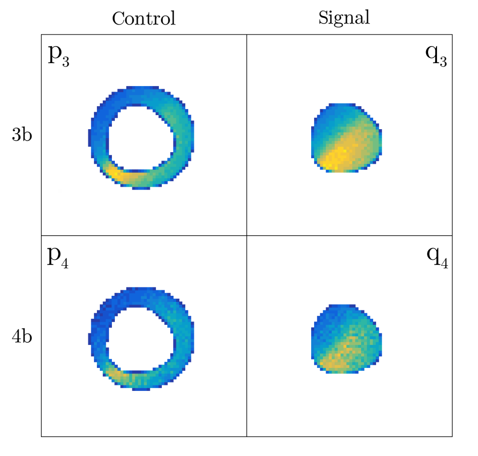
1. Compute CR→SR coupling in 3b.
2. Find 4b→3b nearest neighbor in CR.
3. Form a histogram $\widehat{q}_4$ of the resulting point cloud. Loosely,

$$\widehat{q}_4(x) \propto \widehat{p}_4(\widehat{T}(x))$$



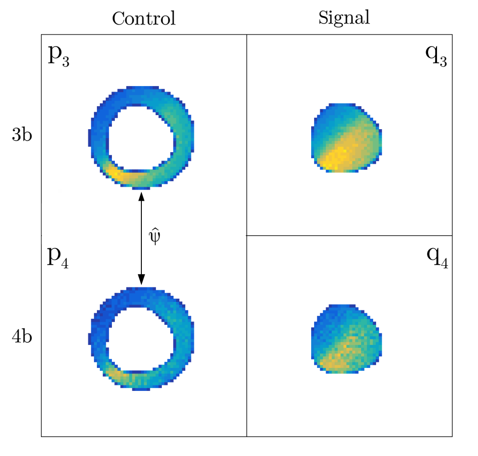
Modelling Assumption: The optimal transport map T^* between p_3 and p_4 maps q_3 to q_4.

Combining Optimal Transport with the Classifier



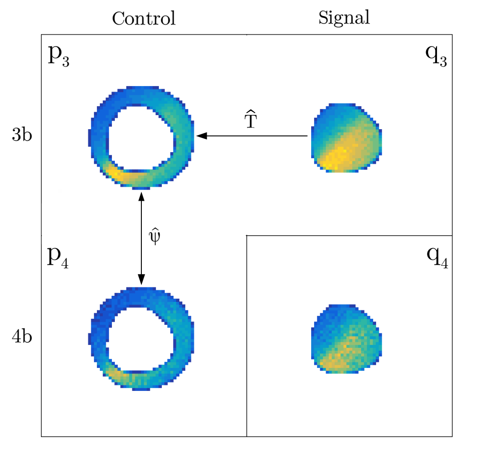
Combining Optimal Transport with the Classifier

1. Train 3b→4b classifier $\widehat{\psi}$ in CR.



Combining Optimal Transport with the Classifier

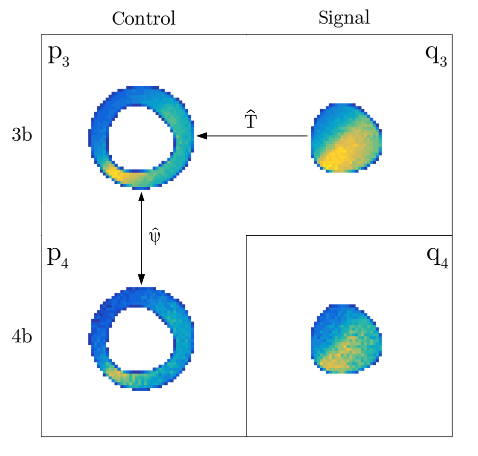
1. Train 3b→4b classifier $\widehat{\psi}$ in CR.
2. Estimate SR→CR transport map $\widehat{T}$.



Combining Optimal Transport with the Classifier

1. Train 3b→4b classifier $\widehat{\psi}$ in CR.
2. Estimate SR→CR transport map $\widehat{T}$.
3. Estimate q_4 by

$$\widehat{q}_4(x) = \frac{\widehat{\psi}(\widehat{T}(x))}{1 - \widehat{\psi}(\widehat{T}(x))}\widehat{q}_3(x).$$



Results

Results

We will compare the methods by using simulated data and comparing one-dimensional histograms.

Results

We will compare the methods by using simulated data and comparing one-dimensional histograms.

In practice, we can use all the methods. They provide a check on each other.

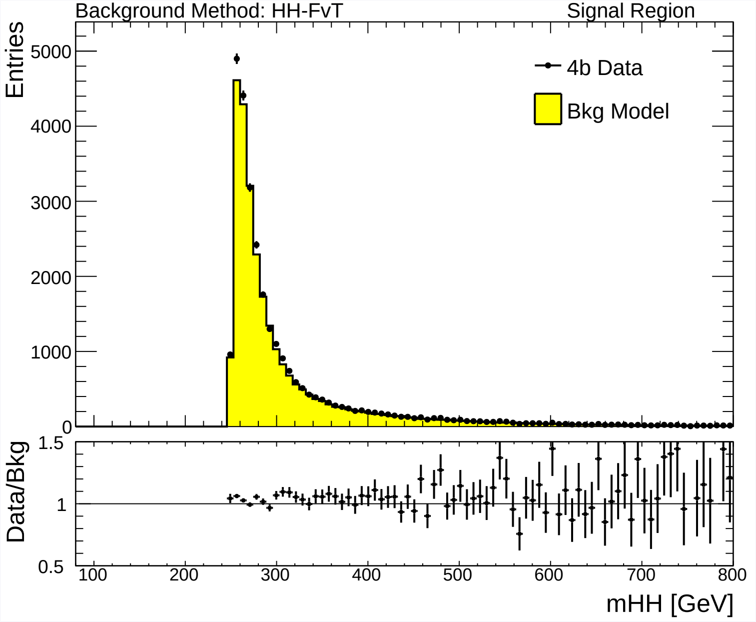
Results

We will compare the methods by using simulated data and comparing one-dimensional histograms.

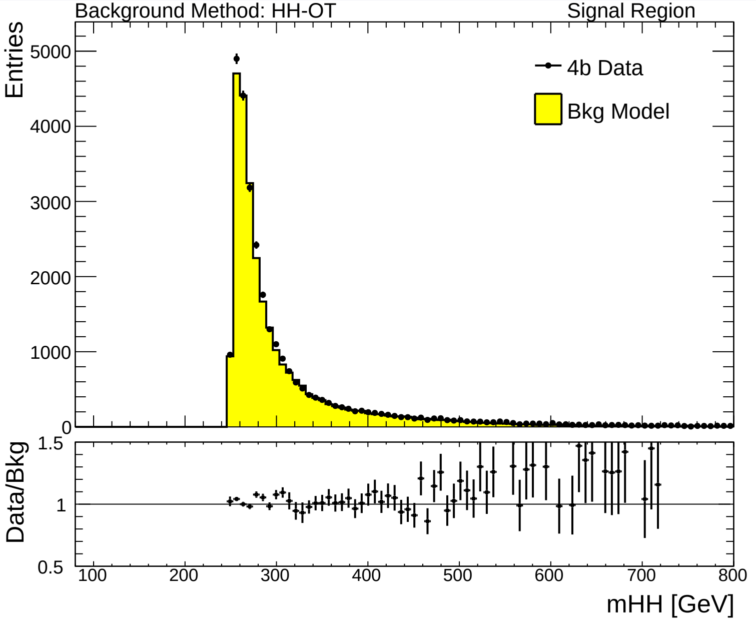
In practice, we can use all the methods. They provide a check on each other.

Lots of computational details to produce what follows.

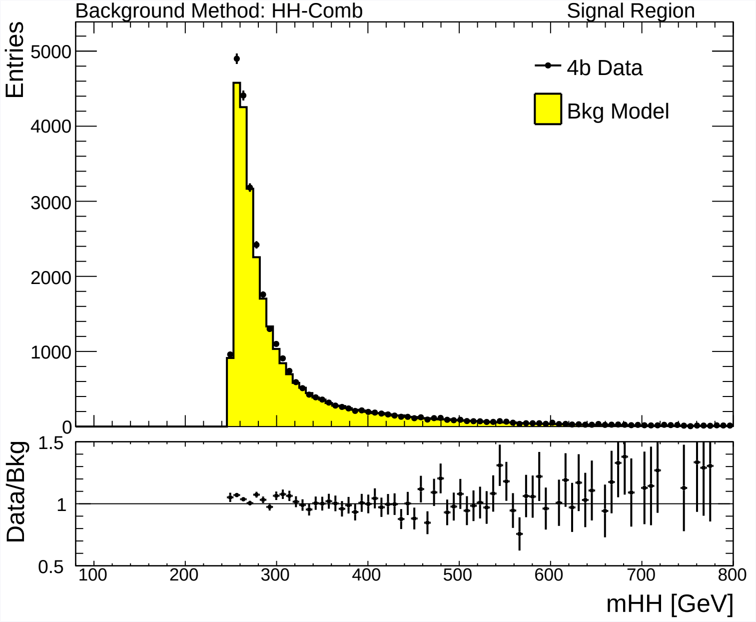
Results: Density Ratio Method



Results: Transport



Results: Combination



Conclusions

Conclusions

Other topics in Optimal Transport

1. Clustering distributions (see Verdinelli, Wasserman 2020)
2. Domain adaptation
3. Hypothesis testing
4. Finding anomalous data sets
5. PCA in Waserstein space
6. Image processing

Background modeling

1. Still tweaking
2. working on inference (confidence sets)

Conclusions

Other topics in Optimal Transport

1. Clustering distributions (see Verdinelli, Wasserman 2020)
2. Domain adaptation
3. Hypothesis testing
4. Finding anomalous data sets
5. PCA in Waserstein space
6. Image processing

Background modeling

1. Still tweaking
2. working on inference (confidence sets)

THE END