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 $\mathcal{X} = C \cup S$. 

Motivating Example

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

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

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

The problem is ill-posed; we will have to make (reasonable) assumptions.
Motivating Example

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

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

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

$C = \text{control region}, \ S = \text{signal region}$
Motivating Example

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

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

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

$C = \text{control region}, \ S = \text{signal region}$

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 $X$ 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 $P_1$. 

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.
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)
Let $X \sim P_0$. 

Optimal Transport: Monge Version
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 between 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 $p = 1$. 
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 between 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 between 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 between collider events (which is itself a type of transport).

For now, assume that the minimizer exists. 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: $T(s) = F_1^{-1}(F_0(s))$
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: \( T(s) = F_1^{-1}(F_0(s)) \)
- \( W_p(P_0, P_1) = \left( \int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds \right)^{1/p} \)
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: \( T(s) = F_1^{-1}(F_0(s)) \)
- \( W_p(P_0, P_1) = \left( \int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds \right)^{1/p} \)
- The morphing — geodesic linking \( F_0 \) and \( F_1 \) — is
  \[
  F_s = \left[ (1 - s)F_0^{-1} + sF_1^{-1} \right]^{-1}
  \]
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:

\[ \hat{F}_0(s) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq s) \]
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:

\[ \hat{F}_0(s) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq s) \]

\[ \hat{F}_1(s) = \frac{1}{m} \sum_{i=1}^{m} I(Y_i \leq s) \]
Finding the Transport Map: Gaussian Case

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

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

Then:

\[
T(X) = \mu_1 + \frac{1}{2} \Sigma_1 \left( X - \mu_0 \right)
\]

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

where

\[
B(\Sigma_0, \Sigma_1) = \text{trace}(\Sigma_0) + \text{trace}(\Sigma_1) - 2 \text{trace}\left( \frac{1}{2} \Sigma_0 \Sigma_1 \right)
\]
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

\[
B(\Sigma_0, \Sigma_1) = \text{trace}(\Sigma_0) + \text{trace}(\Sigma_1) - 2\text{trace}[(\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2}].
\]
Finding the Transport Map: Two Point Clouds

\[ \mathbf{X} = \{ \mathbf{X}_1, \ldots, \mathbf{X}_n \} \quad \mathbf{X}_i \in \mathbb{R}^d \]

\[ \mathbf{Y} = \{ \mathbf{Y}_1, \ldots, \mathbf{Y}_n \} \quad \mathbf{Y}_i \in \mathbb{R}^d \]

\[ T : \mathbf{X}_i \rightarrow \mathbf{Y}_{\pi(i)} \]

where \( \pi \) minimizes

\[ \sum_i \| \mathbf{X}_i - \mathbf{Y}_{\pi(i)} \|^2 \]

over all permutations \( \pi \).

Hungarian algorithm \( O(n^3) \) time.
Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\}$  $X_i \in \mathbb{R}^d$
Finding the Transport Map: Two Point Clouds

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

$T: X_i \rightarrow Y_{\pi(i)}$ where $\pi$ minimizes $\sum_i ||X_i - Y_{\pi(i)}||^2$ over all permutations $\pi$.

Hungarian algorithm $\mathcal{O}(n^3)$ time.
Finding the Transport Map: Two Point Clouds

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

- \( \mathcal{Y} = \{Y_1, \ldots, Y_n\} \quad Y_i \in \mathbb{R}^d \)

- \( T : X_i \rightarrow Y_{\pi(i)} \) where \( \pi \) minimizes

\[
\sum_i ||X_i - Y_{\pi(i)}||^2
\]

over all permutations \( \pi \).

- Hungarian algorithm \( O(n^3) \) time.
$X_1, \ldots, X_n \sim P$
$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$
How Accurate is This?

\[ X_1, \ldots, X_n \sim P \]
\[ Y_1, \ldots, Y_n \sim Q \]
\[ T \text{ is true map from } P \text{ to } Q. \]
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 \( Q \).

\( \hat{T} \) 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 \( Q \).

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

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

\[
\mathbb{E} \| \hat{T}(X) - T(X) \|^2 = O(n^{-2/d})
\]

and this is optimal without further conditions.
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:

1. estimate $p$ with kernel estimator $\hat{p}_h$ using bandwidth $h$.
2. estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.
3. Sample from $\hat{p}_h$ and $\hat{q}_h$ 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.)
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:

1. Estimate $p$ with kernel estimator $\hat{p}_h$ using bandwidth $h$.
2. Estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.
3. Sample from $\hat{p}_h$ and $\hat{q}_h$ and apply Hungarian algorithm.

This is suboptimal but easy. It does estimate the smoothed transport $T_h$: $p_\star K_h \rightarrow q_\star K_h$. 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.)
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 $\hat{p}_h$ using bandwidth $h$.
- estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.

Sample from $\hat{p}_h$ and $\hat{q}_h$ and apply Hungarian algorithm. This is suboptimal but easy. It does estimate the smoothed transport $T_h$: $p \ast K_h \rightarrow q \ast K_h$ 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.)
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 $\hat{p}_h$ using bandwidth $h$.
- estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.

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

This is suboptimal but easy. It does estimate the smoothed transport $T_h$: $p^\star K_h \rightarrow q^\star K_h$ 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.)
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 $\hat{p}_h$ using bandwidth $h$.
estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.

Sample from $\hat{p}_h$ and $\hat{q}_h$ 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 $\hat{p}_h$ using bandwidth $h$.

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

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

This is suboptimal but easy.

It does estimate the smoothed transport $T_h : p \ast K_h \rightarrow q \ast K_h$ 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 $\hat{p}_h$ using bandwidth $h$.
- estimate $q$ with kernel estimator $\hat{q}_h$ using bandwidth $h$.

Sample from $\hat{p}_h$ and $\hat{q}_h$ 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

• 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)
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)

See: POT (Python Optimal Transport)
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)

See: POT (Python Optimal Transport)
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)
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.

$T_s(x) = (1-s)x + sT_s(x)$

$P_s = T_s # P_0$. In other words, $P_s$ is the distribution of the random variable $(1-s)X + sT_s(X)$ where $X \sim P_0$.

Then $(P_s: 0 \leq t \leq 1)$ is the geodesic. Length of the path = $W(P_0, P_1)$. 
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 $P$ 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 $X \sim P_0$. 

Length of the path = $W(P_0, P_1)$. 


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

\[ \ell_2 \text{ interpolation} \quad \text{Wasserstein interpolation} \]
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?
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:

\[ F = Q - 1 \]

where \( Q(u) = \frac{1}{N} \sum_{j} F_{j}(u) \)

See Claici, Chien, Solomon (arXiv:1802.05757) and references therein.
How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

\[ \bar{F} = Q^{-1} \]

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.
An important technical detail that we have ignored:
An important technical detail that we have ignored:
There may not be a map that takes $P$ to $Q$. 
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.
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: Kantorovich relaxation:
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: Kantorovich relaxation:

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

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

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.
In this case, the morphing (geodesic) can be described as follows. Let $J$ be the optimal transport plan for $P_0$ and $P_1$. 
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 $(X, Y) \sim J$. 
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,
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

$$S = \{(x, T(x))\}$$
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 $f$. 

For example, Cuturi (2013) uses the entropy:

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

(i) fast algorithms (Sinkhorn-Knopp algorithm)

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

Disadvantages:

(i) How to choose $\lambda$?

(ii) Effect of regularization is not clear.
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 $\lambda$?
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 $\lambda$?

(ii) Effect of regularization is not clear.
Background Modelling for Double Higgs Boson Production
Background Modelling via 3b Events
Background Modelling via 3b Events

4b-Tagged

3b-Tagged

b

b

b, c, j

not b, not c, j
A jet is \((p, \eta, \phi, m)\) where \(p\) = momentum, \(m\) = mass, \(\phi\) and \(\eta\) are angles.

\[
E = 4 \sum_{i=1}^{4} p_i \delta_i
\]

where \(\delta_i\) is a point mass at \((\eta_i, \phi_i, m_i)\).

The metric between two events \(g_1\) and \(g_2\) is the (modified) Wasserstein distance, a metric between measures.

or:

\(E\) is a vector in \(\mathbb{R}^{16}\) with a weird geometry.

see Komiske, Metodiev and Thaler (2019).
A jet is \((p, \eta, \phi, m)\) where \(p\) = momentum, \(m\) = mass, \(\phi\) and \(\eta\) are angles.

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

\[
E = \sum_{i=1}^{4} p_i \delta_i
\]

where \(\delta_i\) is a point mass at \((\eta_i, \phi_i, m_i)\).
The Metric Space of Collider Events

A jet is \((p, \eta, \phi, m)\) where \(p\) = momentum, \(m\) = mass, \(\phi\) and \(\eta\) are angles.

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

\[
E = \sum_{i=1}^{4} p_i \delta_i
\]

where \(\delta_i\) is a point mass at \((\eta_i, \phi_i, m_i)\).

The metric between two events \(g_1\) and \(g_2\) is the (modified) Wasserstein distance, a metric between measures.

or: \(E\) 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

**Image Description**

A graph from the ATLAS experiment showing the distribution of events as a function of $m_{\ell^+\ell^-}$ in the CMS frame. The graph indicates the observed events with a yellow contour, and the expected signal with a red dashed contour. The inset in the top left corner indicates the CMS center-of-mass energy ($\sqrt{s}$) as 13 TeV, and the integrated luminosity as 24.3 fb$^{-1}$ for the year 2016. The color scale on the right represents the number of events per 25 GeV$^2$.
Sideband, Control and Signal Regions

3b

4b
Sideband, Control and Signal Regions

3b

4b
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

\[ X_1, \ldots, X_n \sim p \]
\[ Y_1, \ldots, Y_n \sim q \]

\[
\begin{array}{c|cccc}
Z & X_1 & \ldots & X_n & Y_1 & \ldots & Y_n \\
1 & 1 & \ldots & 1 & 0 & \ldots & 0 \\
\end{array}
\]
Density Ratios and Classifiers

In general, given two densities $p$ and $q$ and samples

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

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

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$\ldots$</th>
<th>$X_n$</th>
<th>$Y_1$</th>
<th>$\ldots$</th>
<th>$Y_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
<td>0</td>
<td>$\ldots$</td>
<td>0</td>
</tr>
</tbody>
</table>

Classifier $\psi$:

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

and so

$$\frac{p}{q} = \frac{\psi}{1 - \psi}.$$
Modern classifiers (neural nets, random forests) are very accurate so we use classifiers to estimate the density ratios. No one really knows why.
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.
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

At the population level,

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

- Then, \( p_4(x) \propto \psi(x)^{1 - \psi(x)} p_3(x) \).

- Similarly, \( q_4(x) \propto \psi(x)^{1 - \psi(x)} q_3(x) \).

![Signal Control](image)
Extrapolating Density Ratios

At the population level,

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

\[ p_4(x) \propto \psi(x)^{1 - \psi(x)} p_3(x). \]

- Similarly,

\[ q_4(x) \propto \psi(x)^{1 - \psi(x)} q_3(x). \]
At the population level,

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

Signal Control

$3b$  $3b$
$p_3$  $q_3$

$4b$  $4b$
$p_4$  $q_4$
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,
  $$q_4(x) \propto \frac{\psi(x)}{1 - \psi(x)} q_3(x).$$
In practice,

- Train a classifier \( \hat{h} \) on the 3b and 4b control regions.
- For all \( x \) in the control region, \( p_{4}(x) \approx \hat{\psi}(x) - \hat{\psi}(x) p_{3}(x) \).
- Estimate a histogram \( \hat{q}_{3} \) of \( q_{3} \).
- Final estimate: \( \hat{q}_{4}(x) := \hat{\psi}(x) - \hat{\psi}(x) \hat{q}_{3}(x) \).

Assumption: Transfer learning to a phase space with different support.
Extrapolating Probability Ratios

In practice,

- Train a classifier \( \hat{h} \) on the 3b and 4b control regions.

\[
\begin{align*}
\text{Control} & \quad \text{Signal} \\
\hat{p}_3 & \quad \hat{q}_3 \\
3b & \quad \text{3b} \\
\hat{p}_4 & \quad \hat{q}_4 \\
4b & \quad \text{4b}
\end{align*}
\]
Extrapolating Probability Ratios

In practice,

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

\[
p_4(x) \approx \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} p_3(x).
\]

\[\begin{array}{cccc}
\text{Control} & \text{Signal} \\
\hline
p_3 & q_3 \\
3b & \text{ } & \text{ } \\
p_4 & q_4 \\
4b & \text{ } & \text{ }
\end{array}\]
Extrapolating Probability Ratios

In practice,

- Train a classifier \( \hat{h} \) on the 3b and 4b control regions.
- For all \( x \) in the control region,
  \[
  p_4(x) \approx \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} p_3(x).
  \]
- Estimate a histogram \( \hat{q}_3 \) of \( q_3 \).

\[\begin{array}{|c|c|}
\hline
\text{Control} & \text{Signal} \\
\hline
p_3 & q_3 \\
\hline
3b & \text{ } \\
\hline
p_4 & q_4 \\
\hline
4b & \text{ } \\
\hline
\end{array}\]
Extrapolating Probability Ratios

In practice,

- Train a classifier $\hat{h}$ on the 3b and 4b control regions.
- For all $x$ in the control region,
  
  \[ p_4(x) \approx \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} p_3(x). \]

- Estimate a histogram $\hat{q}_3$ of $q_3$.
- Final estimate:
  
  \[ \hat{q}_4(x) := \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} \hat{q}_3(x) \]
Extrapolating Probability Ratios

In practice,

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

• For all $x$ in the control region,
  
  $$p_4(x) \approx \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} p_3(x).$$

• Estimate a histogram $\hat{q}_3$ of $q_3$.

• Final estimate:
  
  $$\hat{q}_4(x) := \frac{\hat{\psi}(x)}{1 - \hat{\psi}(x)} \hat{q}_3(x)$$

• 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 $T : C \rightarrow S$ 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

If \( P \) and \( Q \) are two empirical measures with the same sample size, then:

\[
T(X_i) = Y_{\pi(i)}
\]

where \( \pi \) 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.
If \( P_n \) and \( Q_n \) are two empirical measures with the same sample size then:

\[
T(X_i) = Y_{\pi(i)}
\]

where \( \pi \) 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 $\pi$ 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 $\pi$ 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

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

$$\int h(x, y) \, dx = q_3 b(y|C), \quad \int h(x, y) \, dy = q_3 b(x|S).$$

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

$$\arg\min_{H = (h_{ij})} \in \mathbb{R}^{n \times m} + \sum_{i=1}^{n} h_{ij} = \frac{1}{m} \sum_{j=1}^{m} h_{ij} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{m} h_{ij} d(X_i, Y_i).$$
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

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

over all $h$ such that

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

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

$$\arg\min_{H=(h_{ij}) \in \mathbb{R}^{n \times m}_{+}} \sum_{i=1}^{n} \sum_{j=1}^{m} h_{ij} d(X_i, Y_i)$$

subject to

$$\sum_{i=1}^{n} h_{ij} = 1/m, \quad \sum_{j=1}^{m} h_{ij} = 1/n.$$
The Kantorovich Relaxation

\[
\begin{align*}
\text{Control} & \quad \text{Signal} \\
\begin{array}{c}
p_3 \\
3b
\end{array} & \quad \begin{array}{c}
d(x, y)h(x, y) \\
x
y
z
\end{array} & \quad \begin{array}{c}
q_3 \\
4b
\end{array} \\
\begin{array}{c}
p_4 \\
4b
\end{array} & \quad \begin{array}{c}
d(x, z)h(x, z) \\
\end{array} & \quad \begin{array}{c}
q_4
\end{array}
\end{align*}
\]
Estimating $q_4$ using Optimal Transport

1. Compute $\text{CR} \rightarrow \text{SR}$ coupling in 3b.
2. Find $4b \rightarrow 3b$ nearest neighbor in CR.
3. Form a histogram $\hat{q}_4$ of the resulting point cloud. Loosely, $\hat{q}_4(x) \propto \hat{p}_4(\hat{T}(x))$.

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

[Diagram showing distributions in Control and Signal for $p_3$, $q_3$, $p_4$, and $q_4$.]
Estimating $q_4$ using Optimal Transport

Procedure:
1. Compute $\text{CR} \rightarrow \text{SR}$ coupling in 3b.

$\hat{q}_4(x) \propto \hat{p}_4(\hat{T}(x))$
Estimating $q_4$ using Optimal Transport

Procedure:
1. Compute $CR \rightarrow SR$ coupling in 3b.
2. Find $4b \rightarrow 3b$ nearest neighbor in CR.

Modelling Assumption: The optimal transport map $T^*$ between $p_3$ and $p_4$ maps $q_3$ to $q_4$. 
Estimating $q_4$ using Optimal Transport

Procedure:

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

$$\hat{q}_4(x) \propto \hat{p}_4(\hat{T}(x))$$
Estimating $q_4$ using Optimal Transport

**Procedure:**

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

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

**Modelling Assumption:** The optimal transport map $\mathcal{T}^*$ between $p_3$ and $p_4$ maps $q_3$ to $q_4$. 

![Diagram of signal control and optimal transport](attachment:image.png)
Combining Optimal Transport with the Classifier

1. Train $\hat{\psi}$ classifier in CR.
2. Estimate $\hat{T}$ transport map.
3. Estimate $q_4$ by $\hat{q}_4(x) = \hat{\psi}(\hat{T}(x)) - \hat{\psi}(\hat{T}(x)) \hat{q}_3(x)$.
1. Train $3b \rightarrow 4b$ classifier $\hat{\psi}$ in CR.

<table>
<thead>
<tr>
<th>Control</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_3$</td>
<td>$q_3$</td>
</tr>
<tr>
<td>$3b$</td>
<td></td>
</tr>
<tr>
<td>$p_4$</td>
<td>$q_4$</td>
</tr>
<tr>
<td>$4b$</td>
<td></td>
</tr>
</tbody>
</table>
Combining Optimal Transport with the Classifier

1. Train $3b \rightarrow 4b$ classifier $\hat{\psi}$ in CR.
2. Estimate $SR \rightarrow CR$ transport map $\hat{T}$.
Combining Optimal Transport with the Classifier

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

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

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

Background Method: HH-FvT

Signal Region

- 4b Data
- Bkg Model

Entries vs. mHH [GeV]

Data/Bkg vs. mHH [GeV]
Results: Transport

Background Method: HH-OT

Signal Region

- 4b Data
- Bkg Model

Entries

Data/Bkg

mHH [GeV]
Results: Combination

Background Method: HH-Comb

Signal Region

- 4b Data
- Bkg Model

Entries

Data/Bkg

mHH [GeV]
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 Wasserstein space

Background modeling

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

THE END
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