VARIABILITY REDUCTION USING OPTIMAL TRANSPORT

Kai Hung, Andrew Lipnick, Ryan Shìjié Dù, Nina Mortensen, Esteban G. Tabak

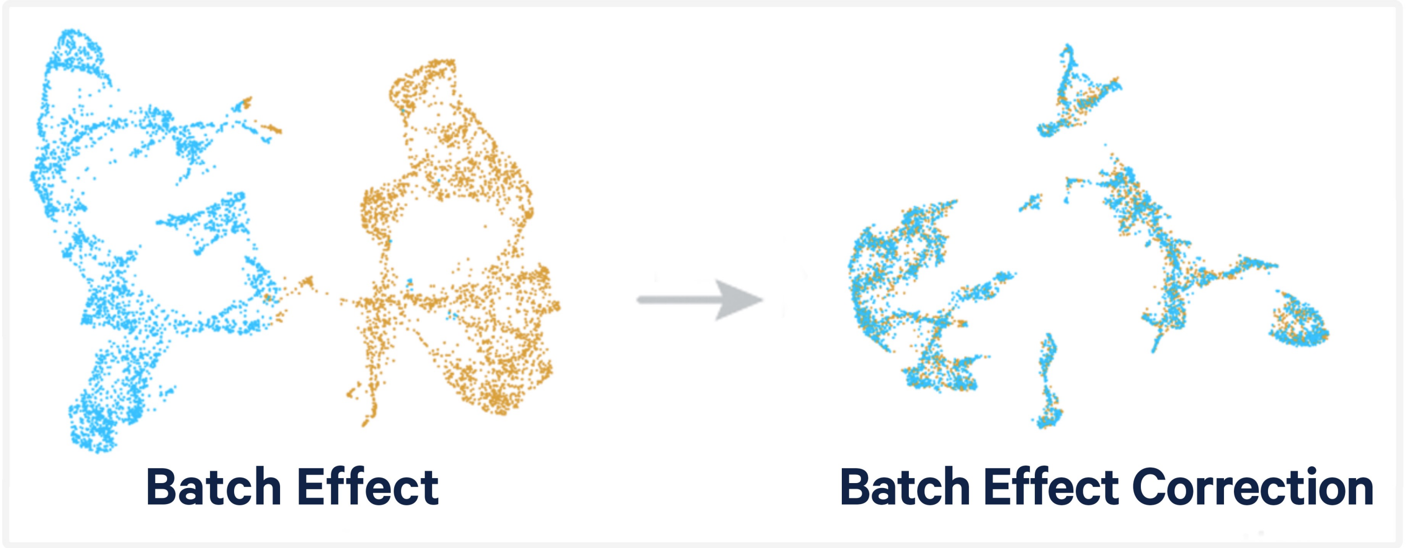
AGENDA

- Problem Formulation
- Motivation
- Methods
  - Data-driven optimal transport barycenter
- Simulation Results
- Future Works

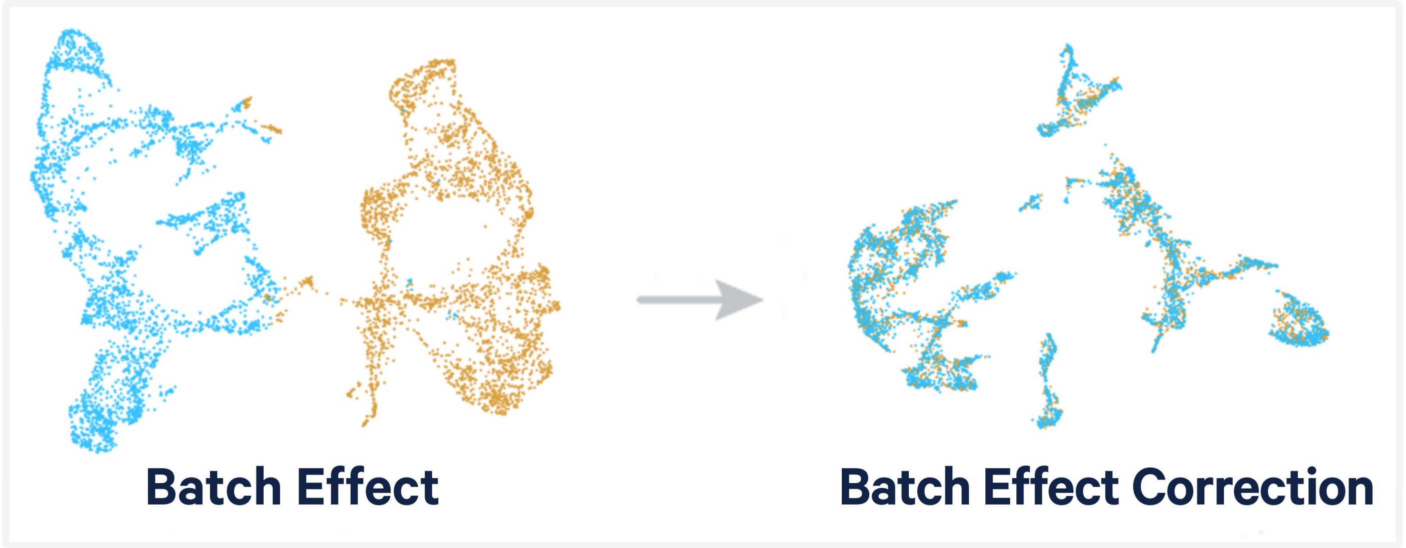
Often, we would like a method to filter out the effect of unwanted variables from a data distribution while preserving the original data as best as possible.

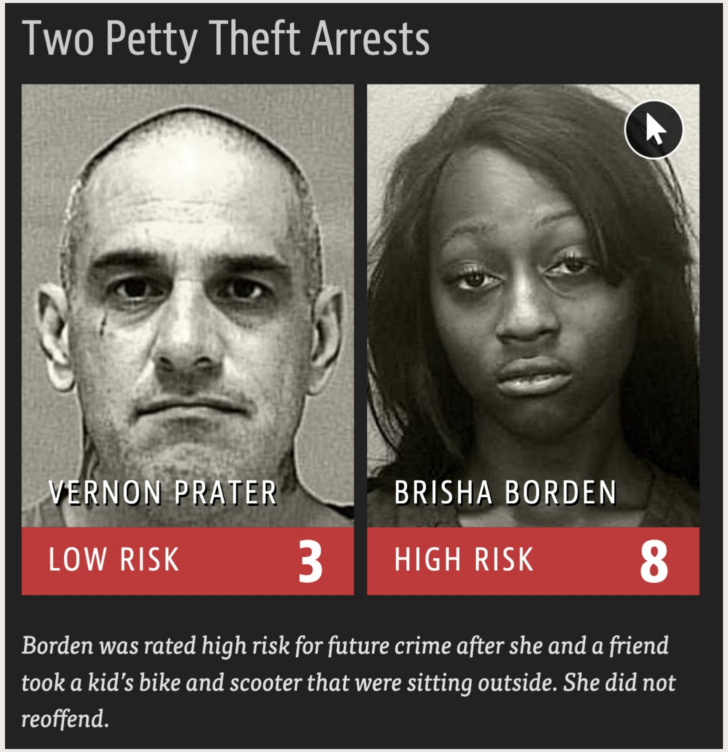
Data variability induced by unwanted variables is ubiquitous.

Data variability induced by unwanted variables is ubiquitous.

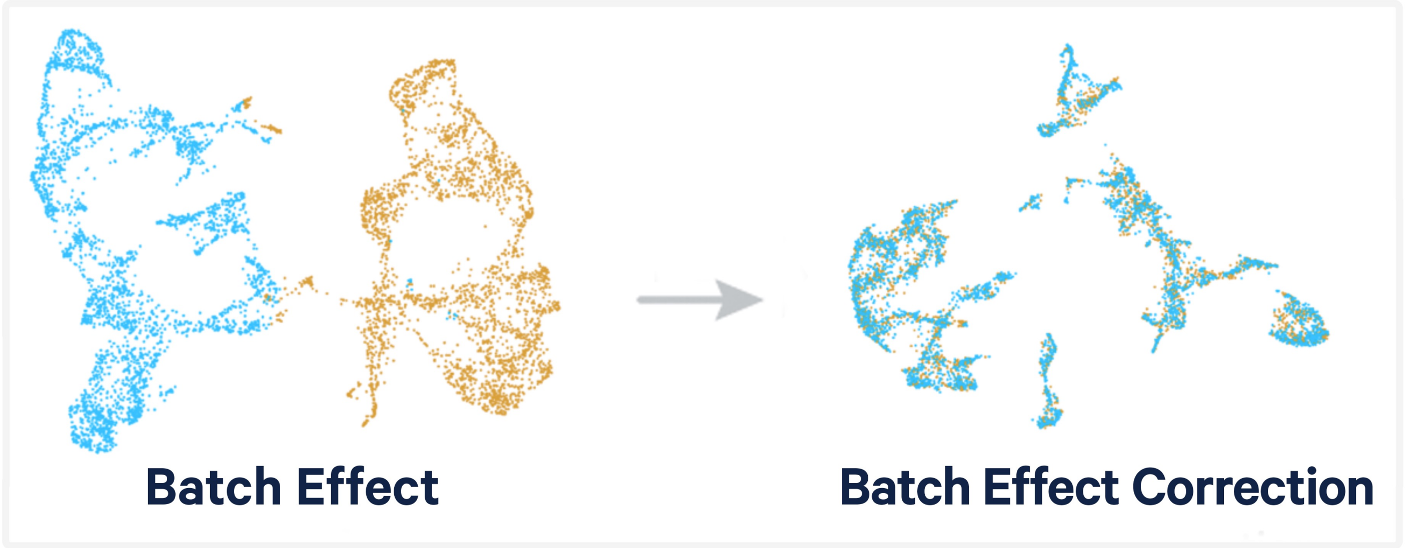


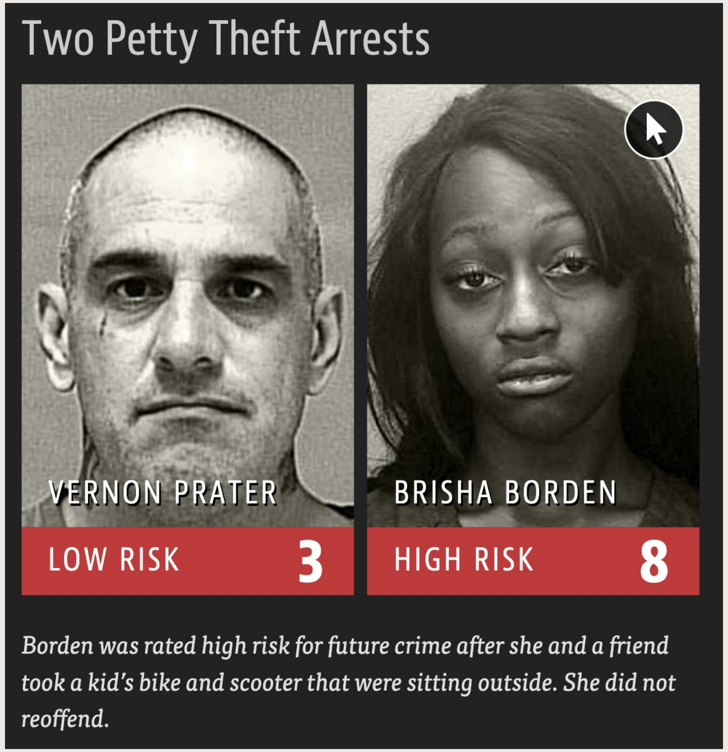
Data variability induced by unwanted variables is ubiquitous.





Data variability induced by unwanted variables is ubiquitous.





For more on batch effect correction, check out the resources curated by 10X Genomics.
Also, please refer to this article on machine learning bias in criminal sentencing.

Let X be the original data, Z be the "unwanted" factor, and Y be the "filtered" version of X.

$$\min_{Y} \max_{\lambda} \text{data\_deformation}(X, Y) - \lambda \cdot \text{independence}(Y, Z).$$

Let X be the original data, Z be the "unwanted" factor, and Y be the "filtered" version of X.

$$\min_{Y} \max_{\lambda} \text{data\_deformation}(X, Y) - \lambda \cdot \text{independence}(Y, Z).$$

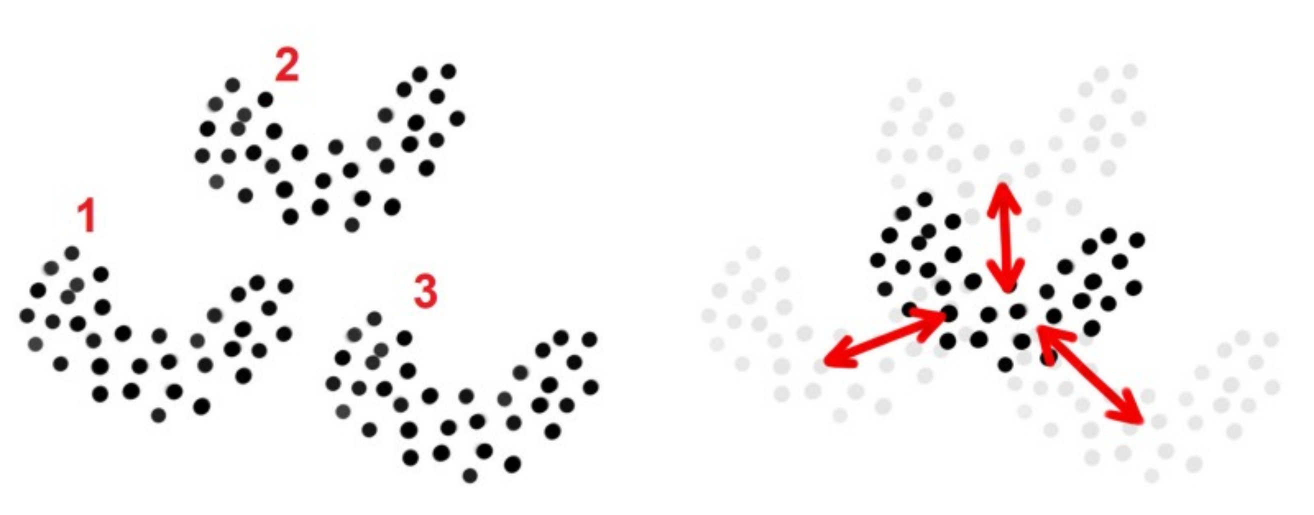


Figure 1: Each data point x is characterized by hidden factor $z \in \{1, 2, 3\}$. The filtration process is akin to computing the optimal transport barycenter of the conditional distributions $\rho(x \mid z)$ for all z.

Source: H. Yang, E. Tabak. *Conditional Density Estimation, Latent Variable Discovery, and Optimal Transport*. 2019.

Formally, we can frame this problem in terms of **optimal transport.** Specifically, we quantify the data deformation using an optimal transport distance.

$$\min_{y=T(x,z)} \int c(x,y)\rho(x|z)\gamma(z)dxdz$$

Here, $c(x, y)$ is the cost function between x and y, $\rho(x|z)$ is the conditional probability distribution of x given z, and $\gamma(z)$ is the probability distribution of z.

Formally, we can frame this problem in terms of **optimal transport.** Specifically, we quantify the data deformation using an optimal transport distance.

$$\min_{y=T(x,z)} \int c(x,y)\rho(x|z)\gamma(z)dxdz$$

And we measure the independence between y and z using their mutual information, which is the KL divergence between the joint and the product of their distributions.

$$D_{KL}(\pi(y,z)||\mu(y)\gamma(z))$$

Altogether, we have the following objective.

$$\min_{y=T(x,z)} \max_{\lambda} \int c(x,y)\rho(x|z)\gamma(z)dxdz + \lambda \cdot D_{KL}(\pi(y,z)\|\mu(y)\gamma(z))$$

Altogether, we have the following objective.

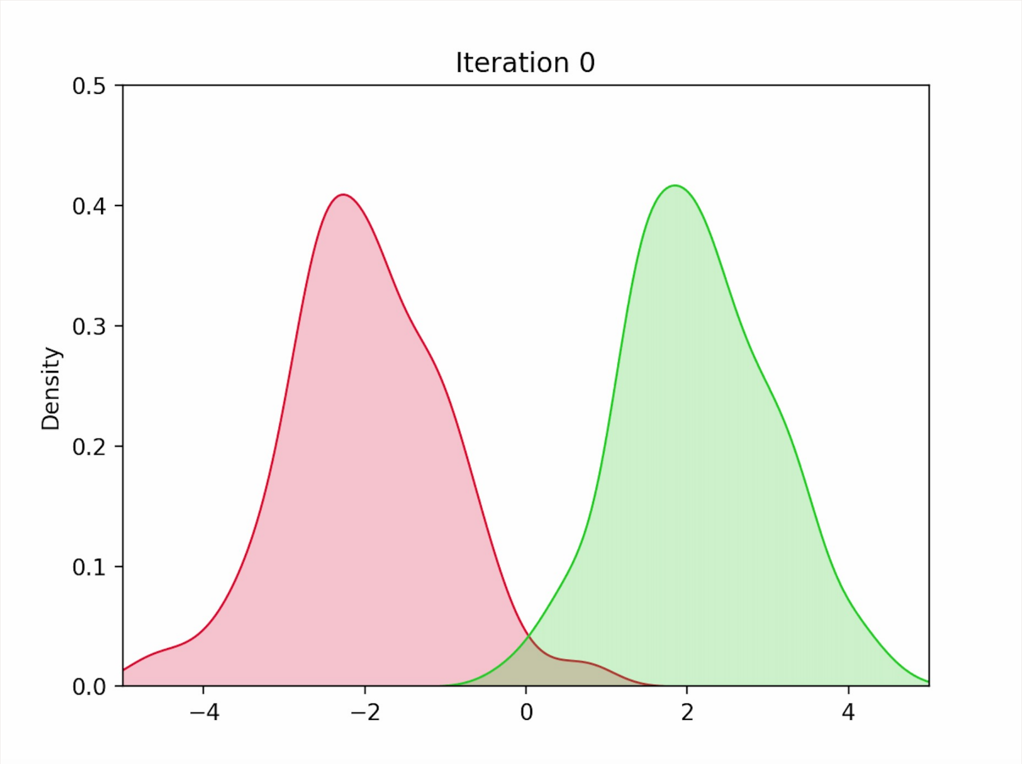
$$\min_{y=T(x,z)} \max_{\lambda} \int c(x,y)\rho(x|z)\gamma(z)dxdz + \lambda \cdot D_{KL}(\pi(y,z)\|\mu(y)\gamma(z))$$

The data-based formulation using samples $\{x_i, z_i\}_i$ yield us

$$\min_{y_i} \max_{\lambda} \sum_i c(x_i, y_i) + \lambda \cdot \sum_i \log\left(\frac{\pi(y_i, z_i)}{\mu(y_i)\gamma(z_i)}\right)$$

where $\pi(y_i, z_i)$, $\mu(y_i)$, and $\gamma(z_i)$ are approximated using KDE.

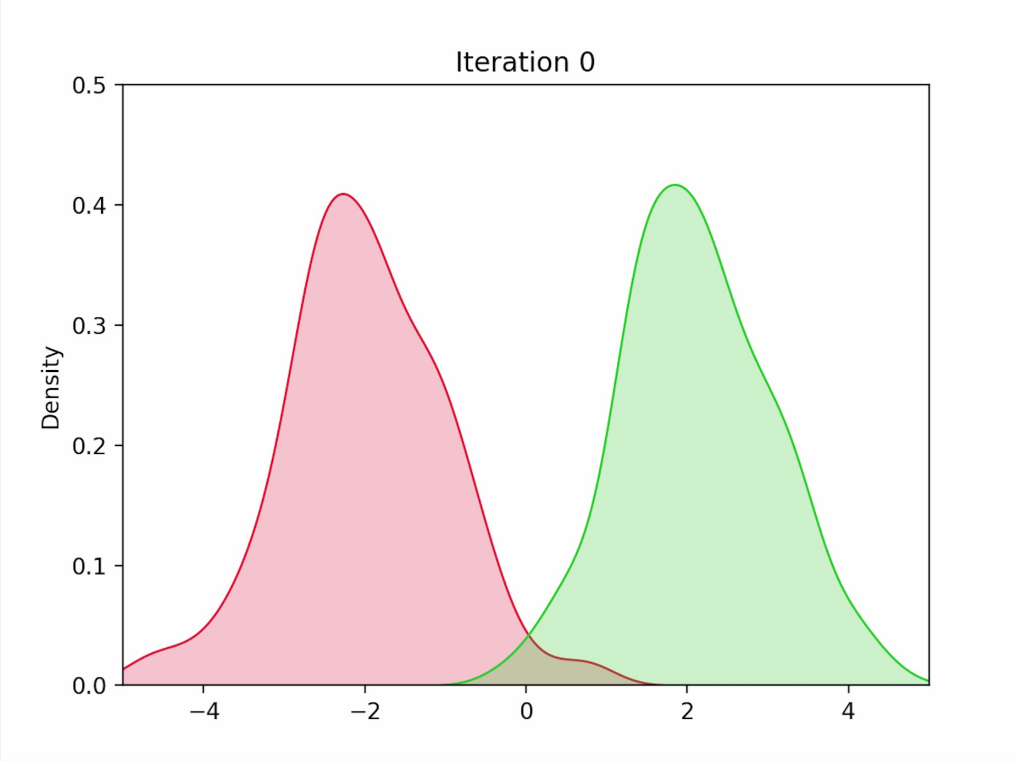
Simulation Results: Gaussian Experiment

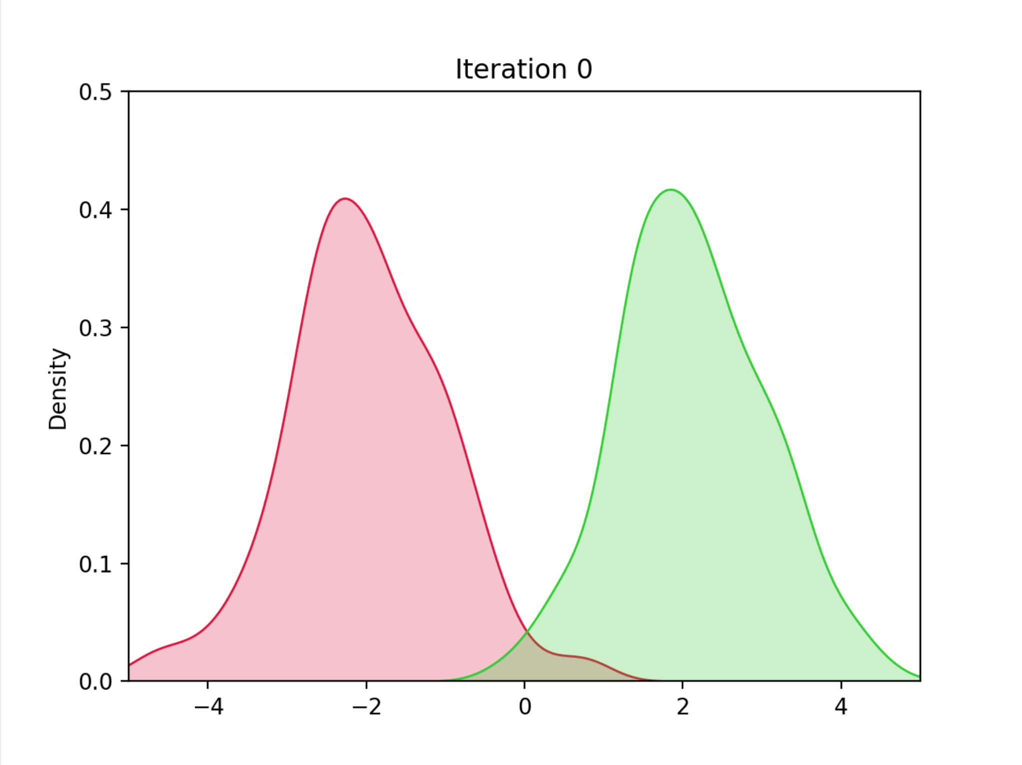


Let $\{x_i, z_i\}_i$ be a sample of points such that x_i is sampled from $\mathcal{N}(-2, 1)$ if z_i = 0 and $\mathcal{N}(2, 1)$ otherwise.

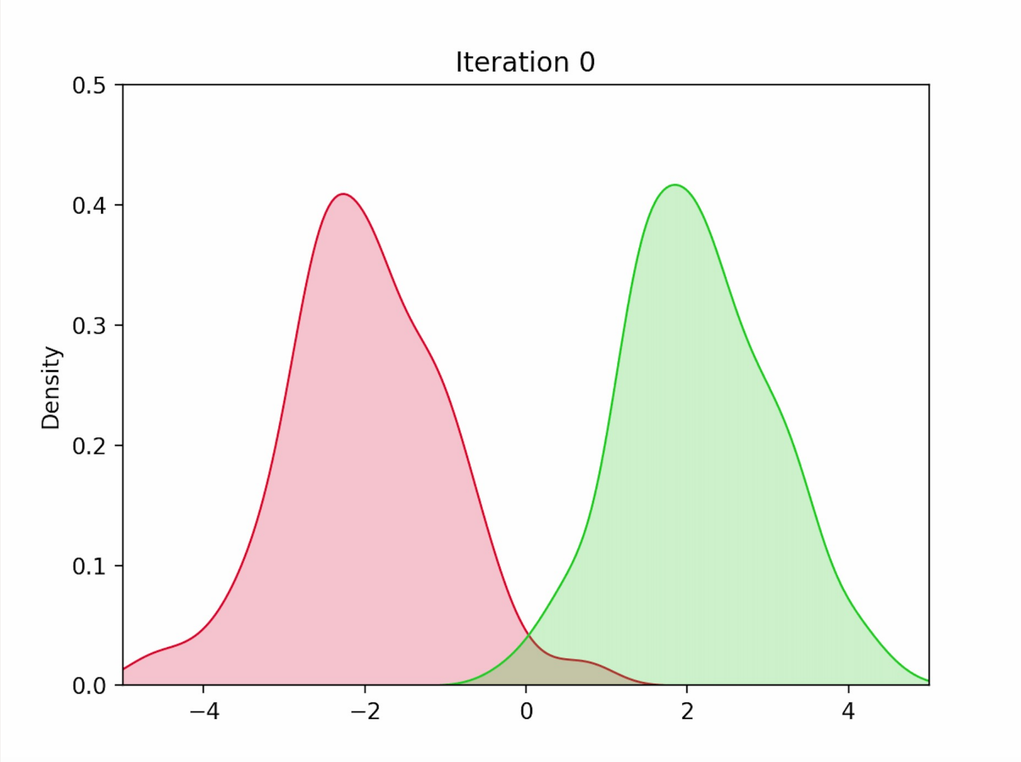
We wish to find the distribution $\mu(y)$.

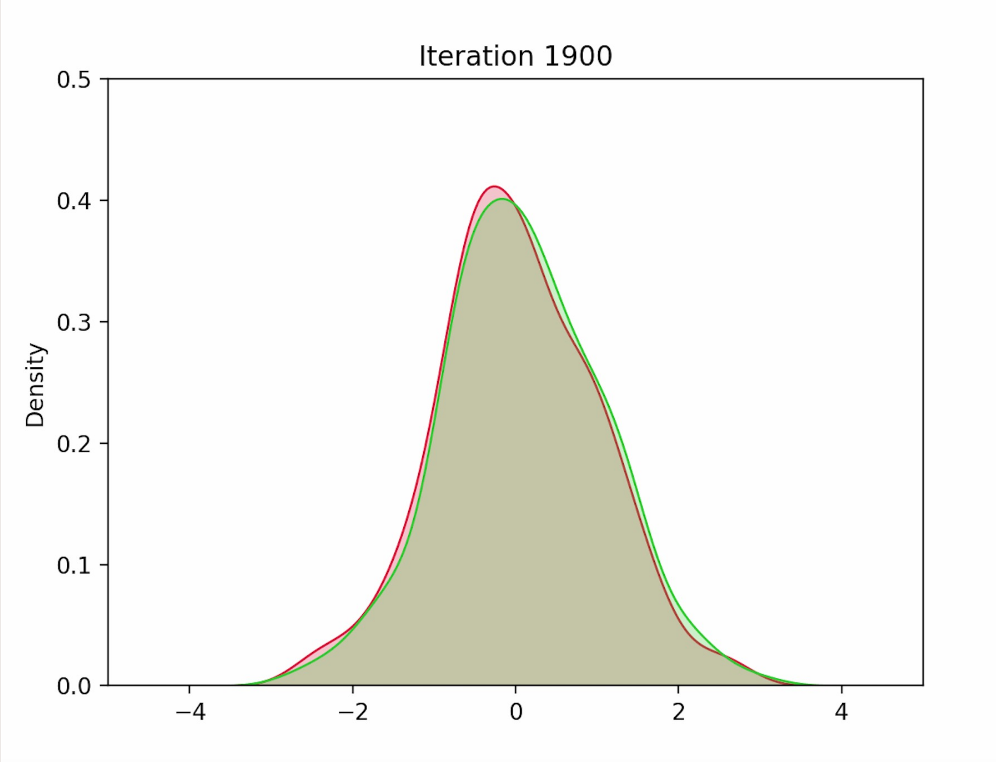
Simulation Results: Gaussian Experiment



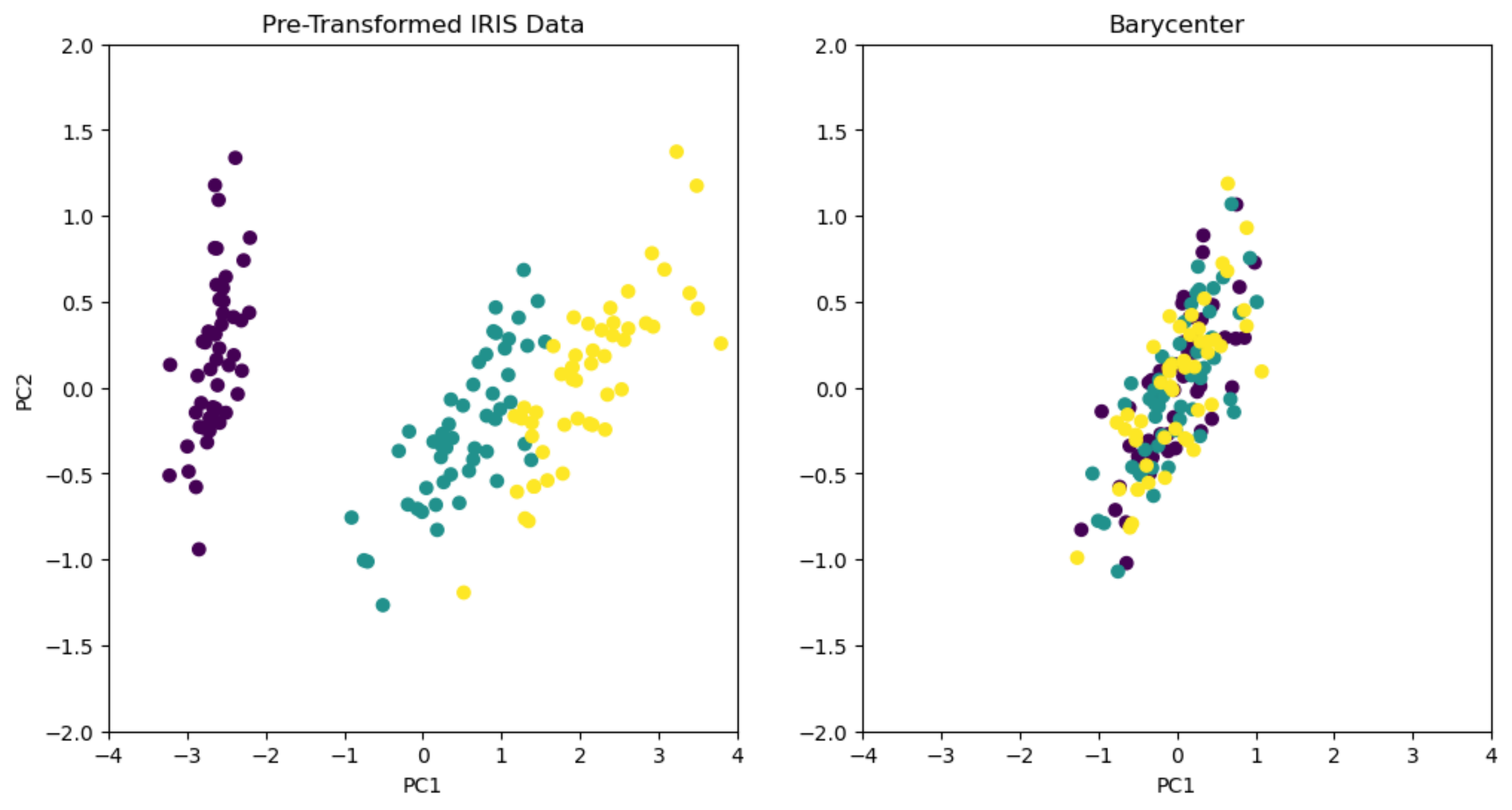


Simulation Results: Gaussian Experiment





Simulation Results: Iris Dataset



The variability reduction problem paves the way to...

The variability reduction problem paves the way to…

- **Factor discovery** through identifying the factor z that *maximizes the reduction in variability* among all factors.

The variability reduction problem paves the way to…

- **Factor discovery** through identifying the factor z that *maximizes the reduction in variability* among all factors.
- **Data augmentation** through reversal of the transformed data points to a specific factor z^* that may have little samples otherwise.

The variability reduction problem paves the way to…

- **Factor discovery** through identifying the factor z that *maximizes the reduction in variability* among all factors.
- **Data augmentation** through reversal of the transformed data points to a specific factor z^* that may have little samples otherwise.
- **Semi-supervised classification** by solving the barycenter problem with observations with unknown factors.

That’s it! Questions?