Clustering via Uncoupled REgression (CURE)

Kaizheng Wang

Department of ORFE
Princeton University

May 8th 2020

Collaborators

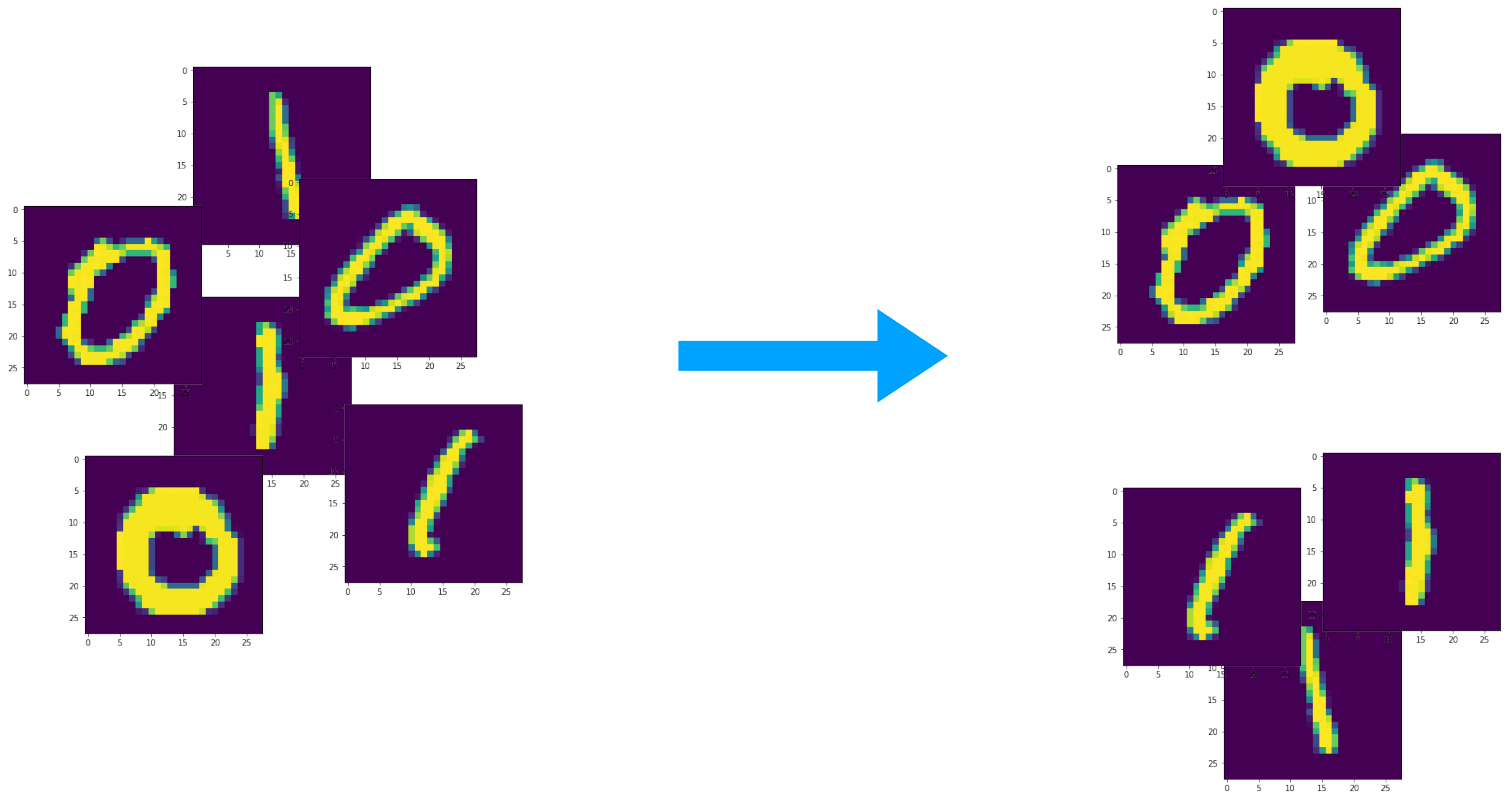
Yuling Yan

Princeton ORFE

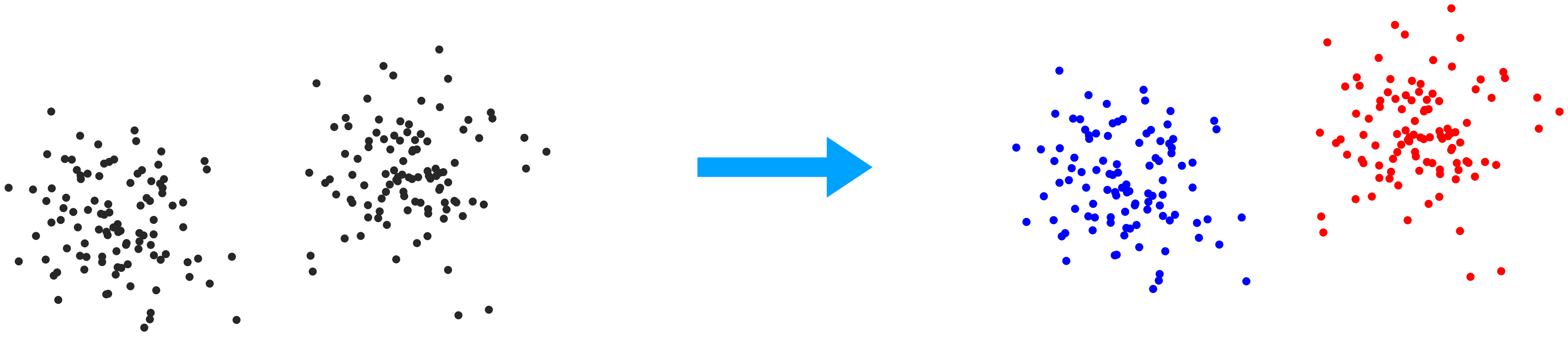
Mateo Díaz

Cornell CAM

Clustering

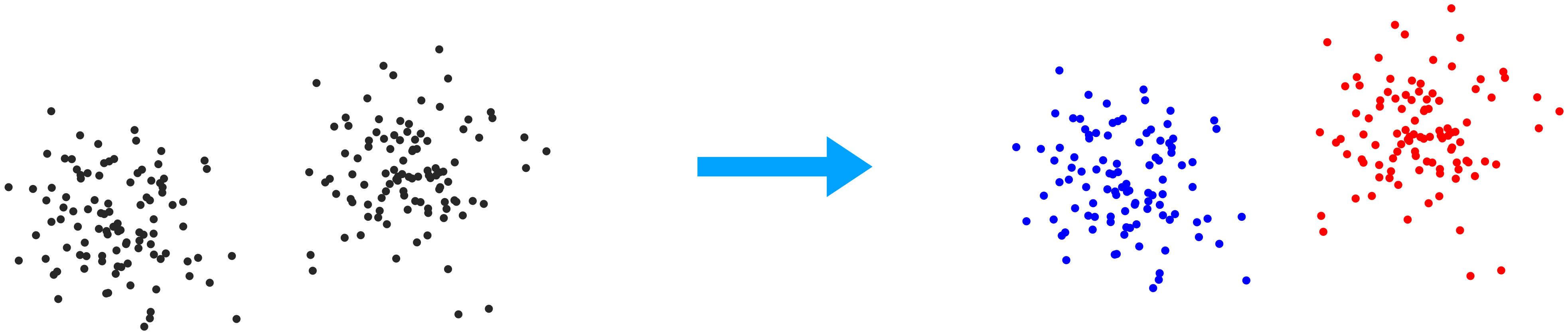


Spherical Clusters



$$\{\boldsymbol{x}_i\}_{i=1}^n \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$$

Spherical Clusters



$$\{\boldsymbol{x}_i\}_{i=1}^n \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$$

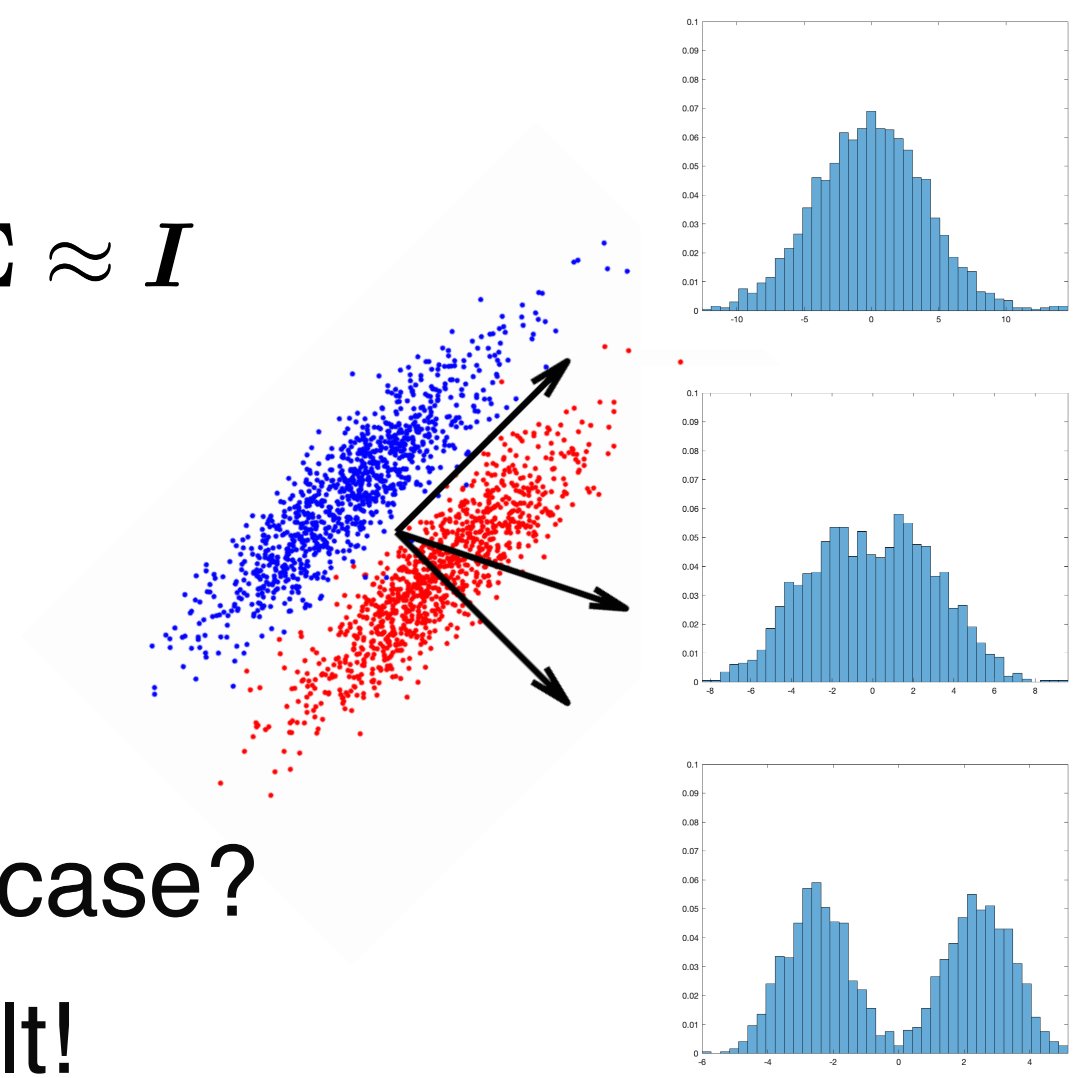
- PCA: $\max_{\boldsymbol{\beta} \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i=1}^n (\boldsymbol{\beta}^\top \boldsymbol{x}_i)^2$
- k-means: $\min_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{y}} \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\mu}_{y_i}\|_2^2$
- SDP relaxations of k-means, etc
- Density-based methods require large samples

Finding a Needle in a Haystack

They are **powerful** but **not omnipotent**.

$\frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$: covariance $\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}$

- Max variance $\neq$ useful
- PCA: $\|\boldsymbol{\mu}\|_2^2 / \|\boldsymbol{\Sigma}\|_2 \gg 1$ or $\boldsymbol{\Sigma} \approx \boldsymbol{I}$



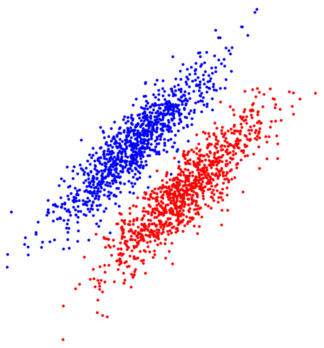
Reduction to the spherical case?

- Estimation of $\boldsymbol{\Sigma}$ is difficult!

Headaches

- PCA and many: **nice shapes** & large separations.
- Learning with non-convex losses:
  1. Initialization (e.g. **spectral methods**);
  2. Refinement (e.g. gradient descent).

Stretched mixtures can be **catastrophic**.



Commonly-used: isotropic, Gaussian, uniform, etc.

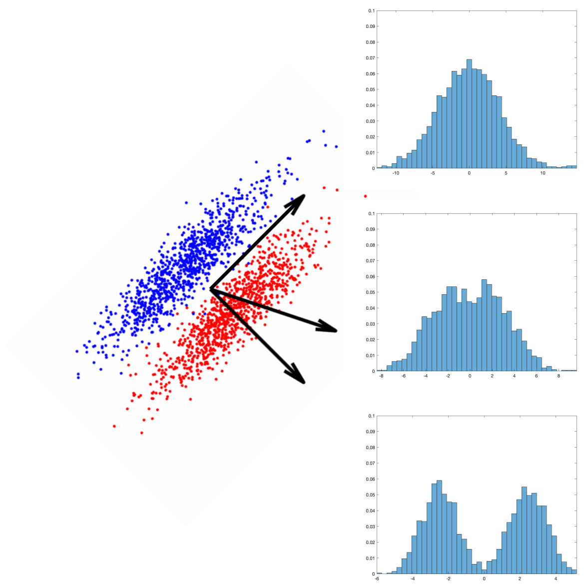
Clustering via **U**ncoupled **RE**gression

- **The CURE methodology**
- Theoretical guarantees

Vanilla CURE

Given centered $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, want $\boldsymbol{\beta} \in \mathbb{R}^d$ such that

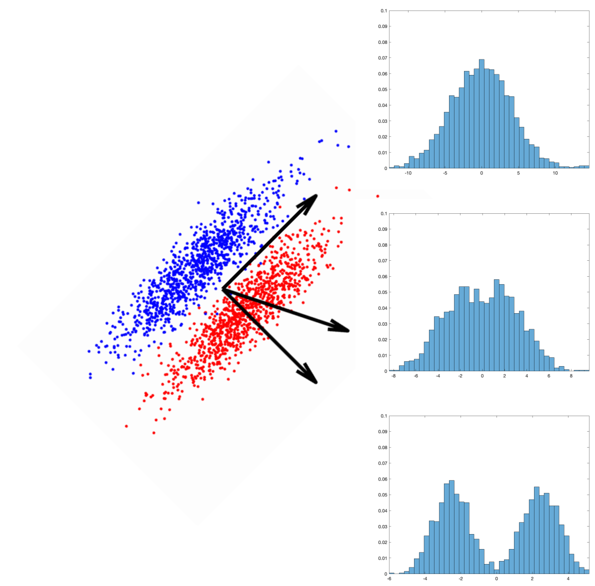
$$\boldsymbol{\beta}^\top \boldsymbol{x}_i \approx y_i, \qquad i \in [n].$$



Vanilla CURE

Given centered $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, want $\boldsymbol{\beta} \in \mathbb{R}^d$ such that

$$\boldsymbol{\beta}^\top \boldsymbol{x}_i \approx y_i, \qquad i \in [n].$$



<u>C</u>lustering via **<u>U</u>**ncoupled **<u>RE</u>**gression:

$$\frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\beta}^\top \boldsymbol{x}_i} \approx \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1.$$

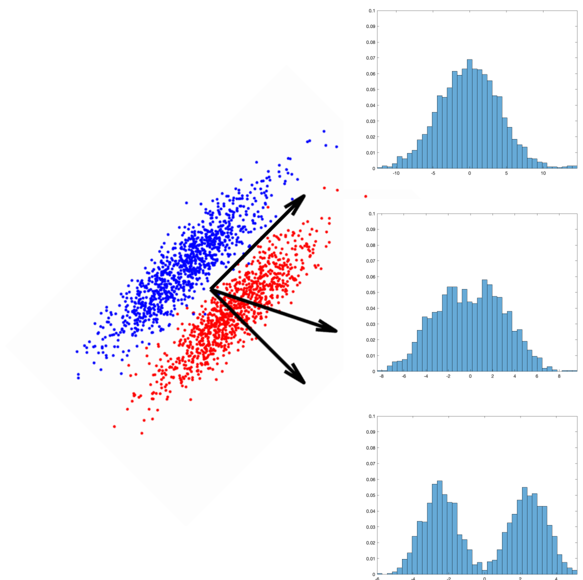
Vanilla CURE

Given centered $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, want $\boldsymbol{\beta} \in \mathbb{R}^d$ such that

$$\boldsymbol{\beta}^\top \boldsymbol{x}_i \approx y_i, \qquad i \in [n].$$

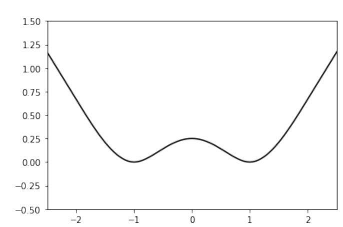
<u>**C**</u>lustering via <u>**U**</u>ncoupled <u>**RE**</u>gression:

$$\frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\beta}^\top \boldsymbol{x}_i} \approx \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{1}.$$



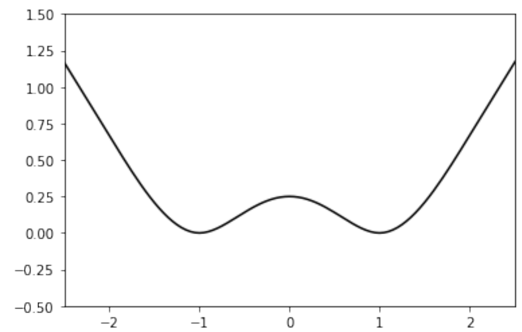
CURE: take f with valleys at ± 1, e.g. $f(x) = (x^2-1)^2$;

solve $\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$; return $\hat{y}_i = \mathrm{sgn}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_i)$.



Vanilla CURE

$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\beta}^{\top}\boldsymbol{x}_i)$ is **non-convex** by nature.



- **Projection pursuit** (Friedman and Tukey, 1974),

  ICA (Hyvärinen and Oja, 2000)
  - Maximize deviation from the null (Gaussian);
  - Limited algorithmic guarantees.

- **Phase retrieval** (Candès et al. 2011)
  - Isotropic measurements, spectral initialization.

Vanilla CURE with Intercept

Given $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, find $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ s.t.

$$\frac{1}{n}\sum_{i=1}^n \delta_{\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i} \approx \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1.$$

The naïve extension

$$\min_{\alpha\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n f(\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i)$$

yields **trivial** solutions $(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = (\pm 1, \mathbf{0})$.

It only forces $|\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i| \approx 1$ rather than

$$\#\{i:\ \alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i \approx 1\} \approx n/2.$$

Vanilla CURE with Intercept

Given $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, find $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ s.t.

$$\frac{1}{n}\sum_{i=1}^n \delta_{\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i} \approx \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1.$$

CURE: $$\min_{\alpha\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n f(\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i) + \frac{1}{2}(\alpha+\boldsymbol{\beta}^\top \bar{\boldsymbol{x}})^2 \right\}.$$

Vanilla CURE with Intercept

Given $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, find $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ s.t.

$$\frac{1}{n}\sum_{i=1}^n \delta_{\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i} \approx \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1.$$

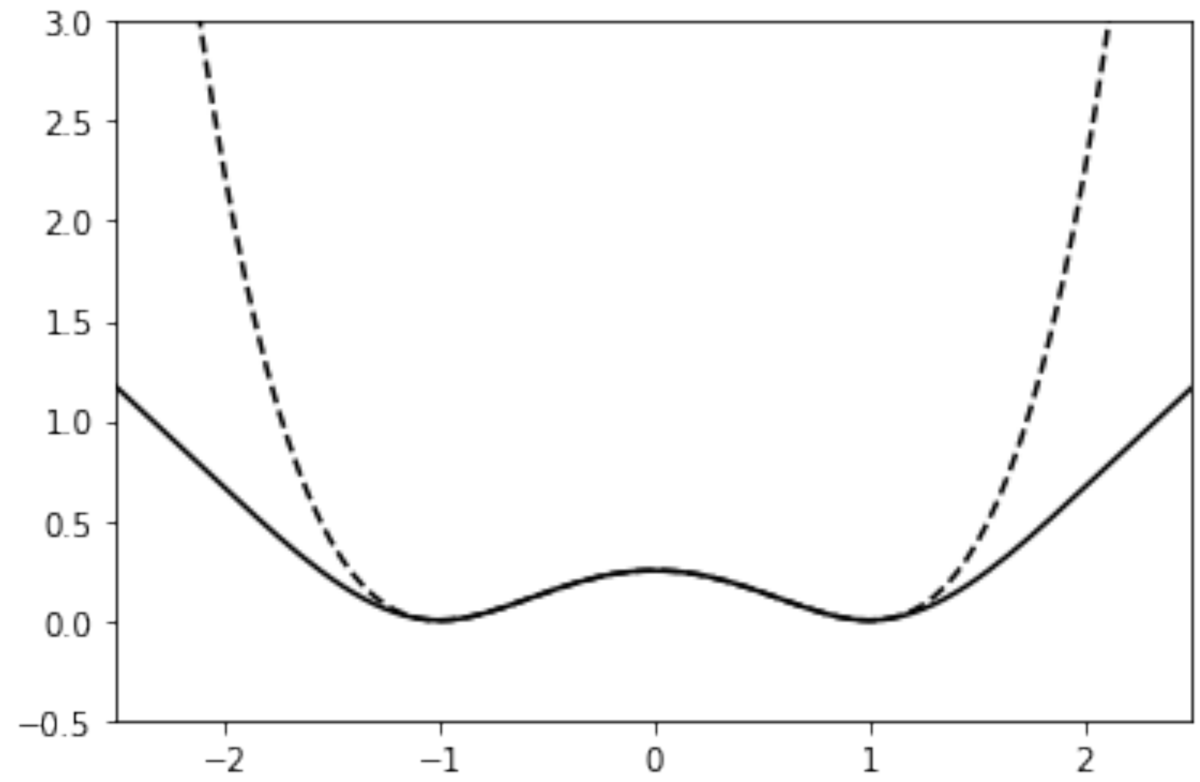
CURE: $$\min_{\alpha\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n f(\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i) + \frac{1}{2}(\alpha+\boldsymbol{\beta}^\top \bar{\boldsymbol{x}})^2 \right\}.$$

- $\frac{1}{n}\sum_{i=1}^n f(\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i)$: $|\alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i| \approx 1$;
- $(\alpha+\boldsymbol{\beta}^\top \bar{\boldsymbol{x}})^2$: $\#\{i:\ \alpha+\boldsymbol{\beta}^\top \boldsymbol{x}_i \approx 1\} \approx n/2$.
  - **Moment matching**. Extension: imbalanced cases.

Loss Function

Clip $(x^2 - 1)^2/4$ to improve

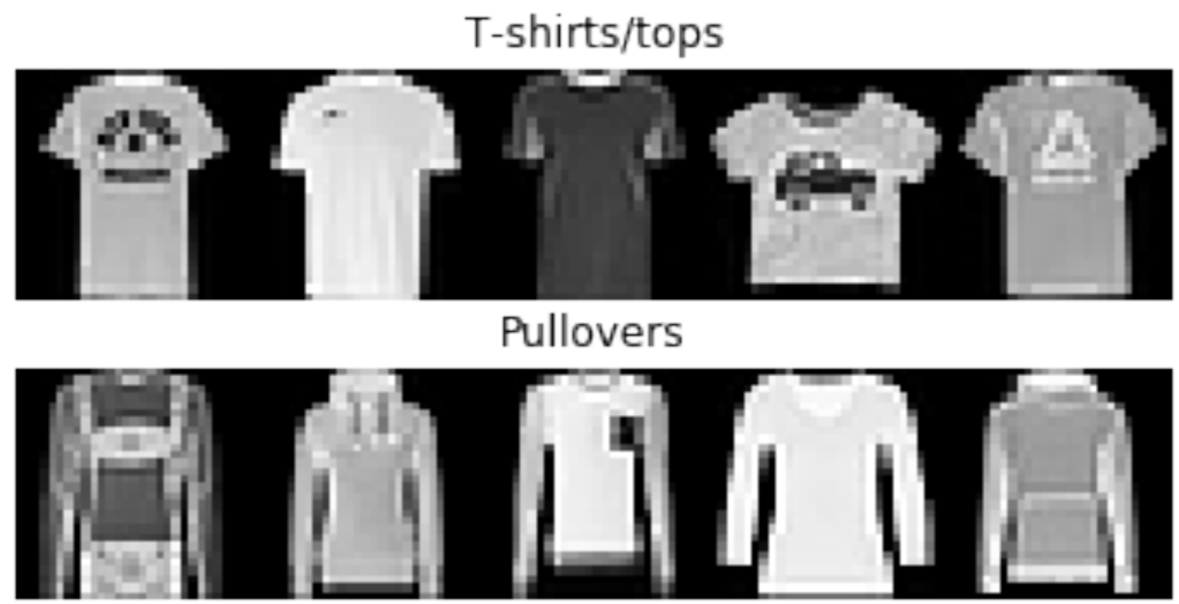
- concentration and robustness for **statistics**;
- growth condition and smoothness for **optimization**.



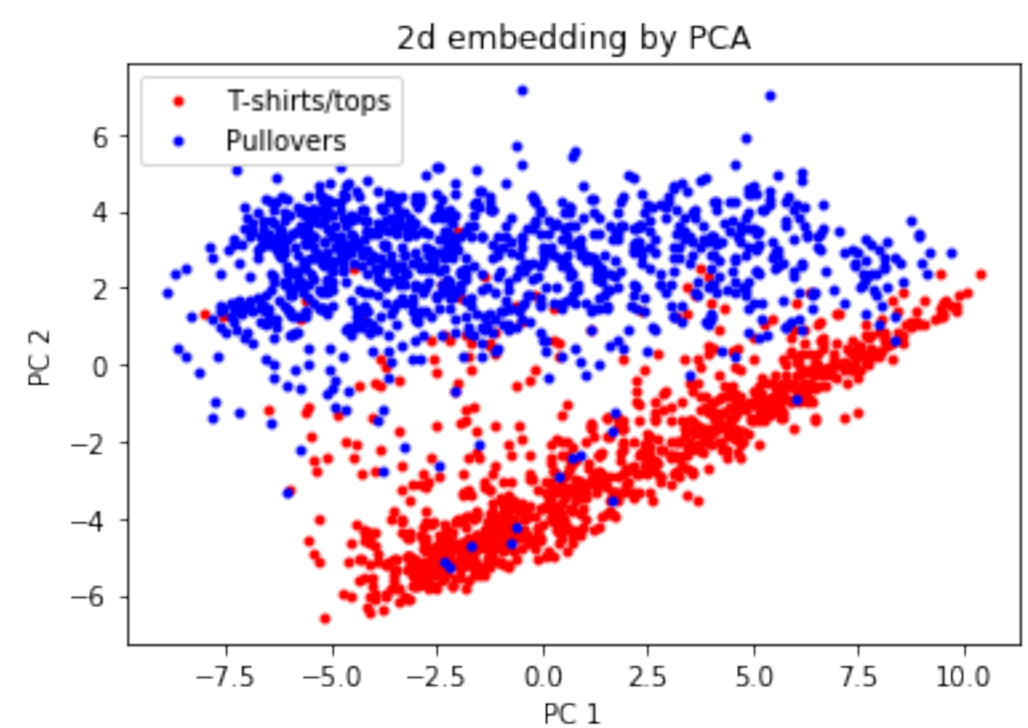
Example: Fashion-MNIST

70000 fashion products, 10 categories (Xiao et al. 2017).

- T-shirts/tops
- Pullovers



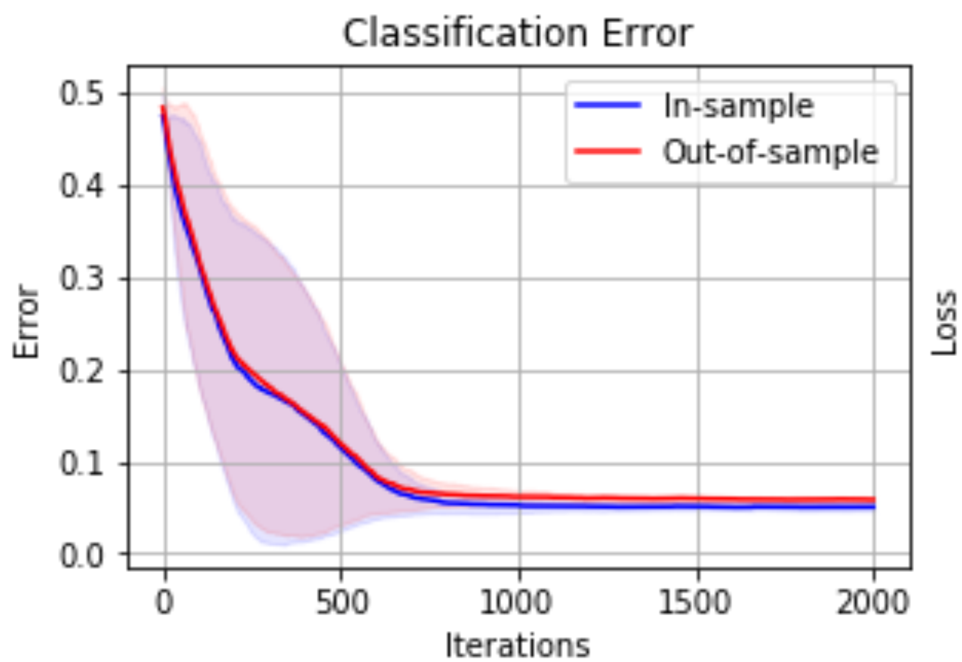
Visualization by PCA

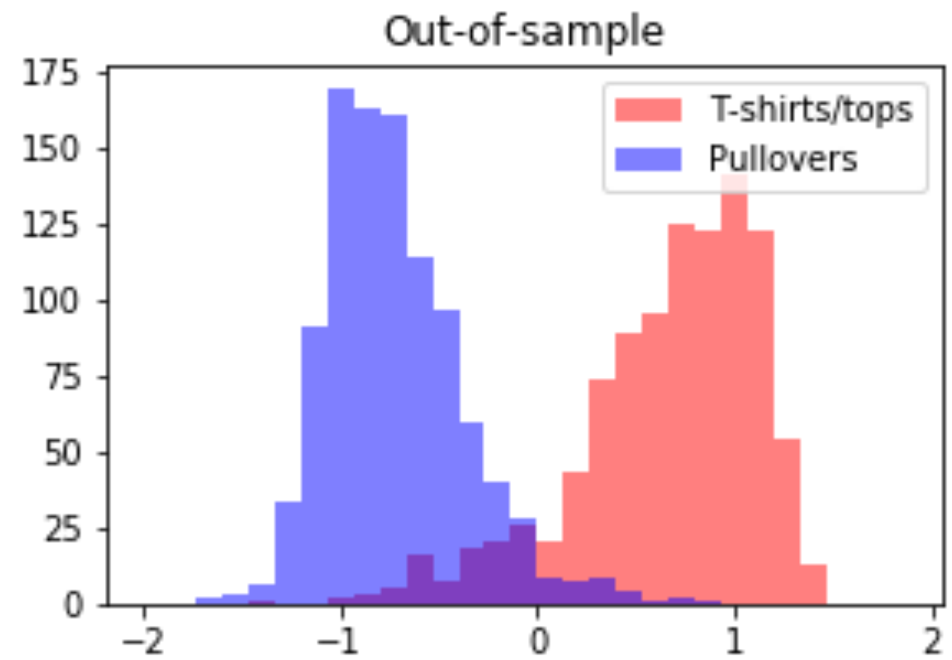


Example: Fashion-MNIST

Goal: cluster 1000 T-shirts/tops and 1000 Pullovers.

Alg.: gradient descent, random initialization from unit sphere.





Err.: **CURE 5.2%**, kmeans 44.3%, spectral (vanilla) 41.9%;

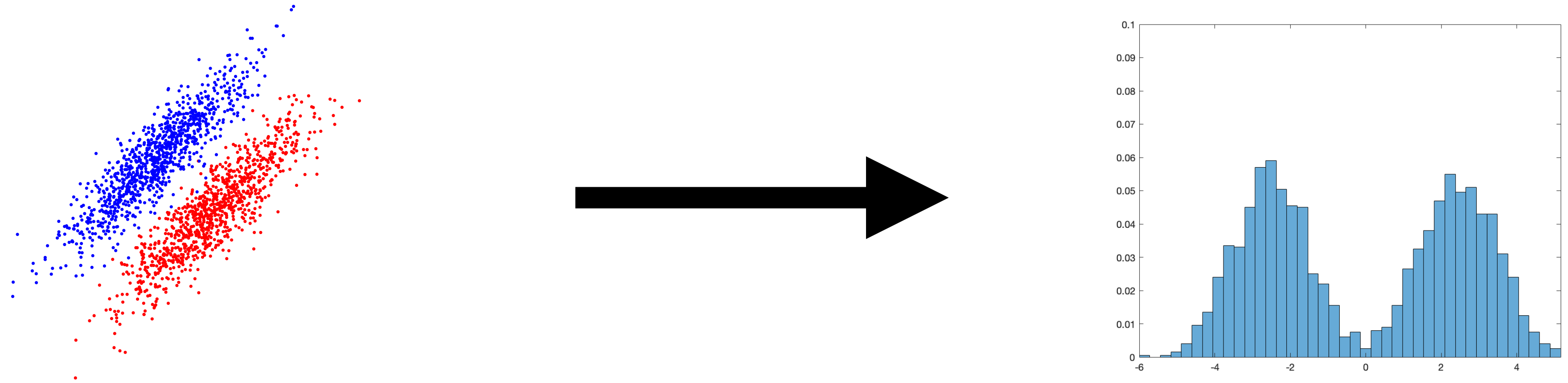
spectral (Gaussian kernel) 10.5%.

Also works when the classes are **imbalanced**.

General CURE

Given $\{\boldsymbol{x}_i\}_{i=1}^{n} \subseteq \mathcal{X}$, find $f : \mathcal{X} \to \mathcal{Y}$ in $\mathcal{F}$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{f(\boldsymbol{x}_i)} \approx \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{y}_j}.$$



General CURE

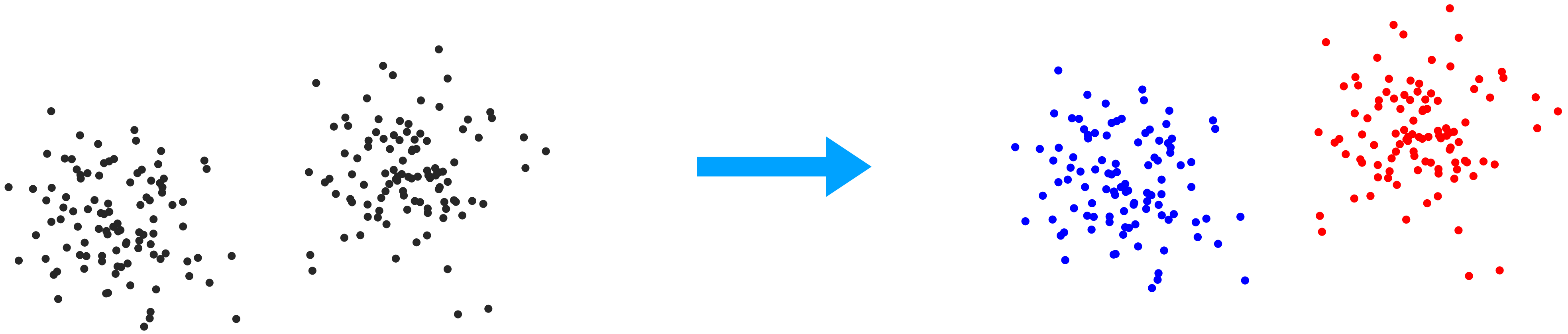
Given $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{X}$, find $f: \mathcal{X} \to \mathcal{Y}$ in $\mathcal{F}$ s.t.

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{f(\boldsymbol{x}_i)} \approx \sum_{j=1}^{K} \pi_j \delta_{\boldsymbol{y}_j}.$$

CURE: $$\min_{f \in \mathcal{F}} D(f_{\#}\hat{\rho}_n,\ \nu).$$

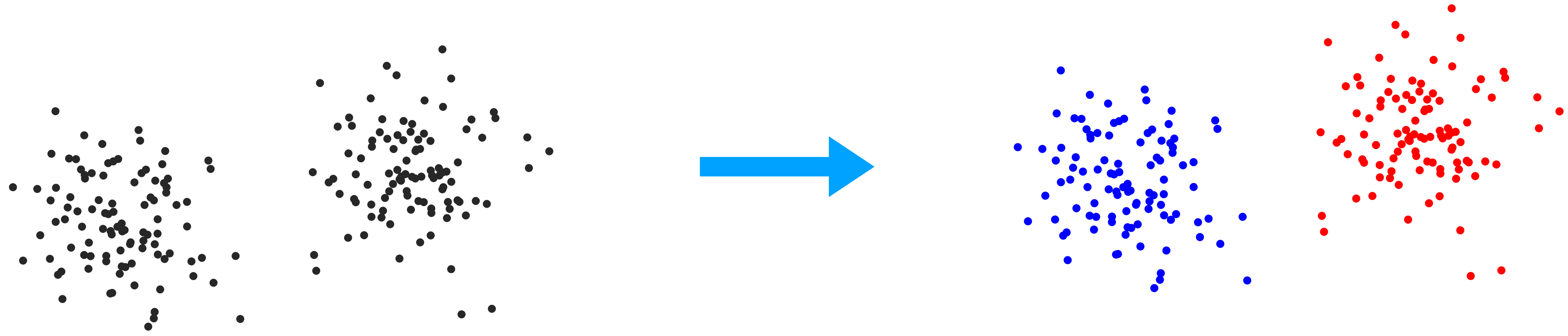
- Discrepancy measure: divergence; MMD; W_p;
- Fashion (**10** classes), CNN + W_1: state-of-the-art;
- Bridle et al. (1992), Krause et al. (2010), Springenberg (2015), Xie et al. (2016), Yang et al. (2017), Hu et al. (2017), Shaham et al. (2018).

Clustering Algorithms



- **Generative: (X, Y) -> (Y | X)**
  - Distribution learning (EM, DBSCAN)
  - ~ Linear discriminant analysis
- **Discriminative: (Y | X)** — **CURE** belongs to this.
  - Criterion opt. (projection pursuit, Transductive SVM)
  - ~ Logistic regression

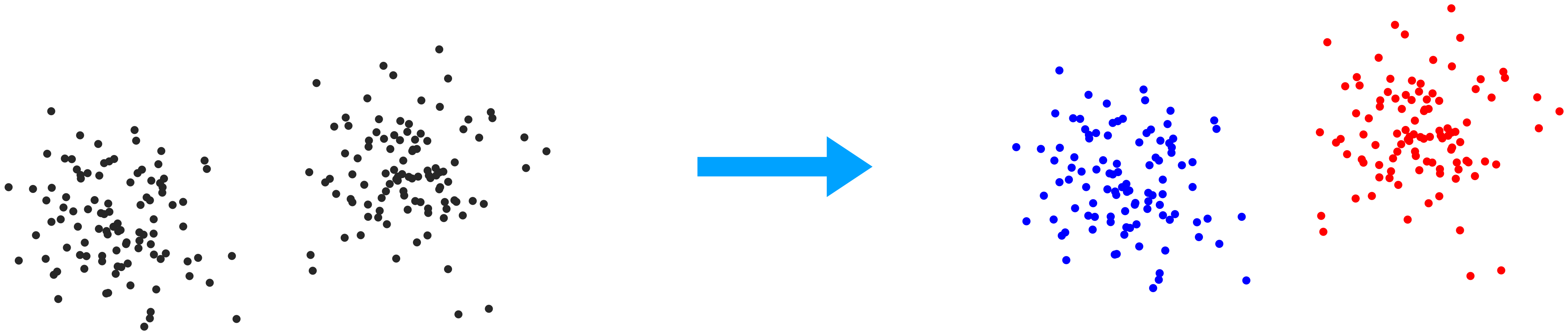
Clustering Algorithms



Drawbacks of generative approaches

- Model dependency
- Unnecessary parameters
- Computational challenges
- Strong conditions

Clustering Algorithms



Example: $\{\boldsymbol{x}_i\}_{i=1}^n \sim \frac{1}{2}N(\boldsymbol{\mu}, \boldsymbol{I}_d) + \frac{1}{2}N(-\boldsymbol{\mu}, \boldsymbol{I}_d)$ with $d \gg n$

- Parameter estimation: $\|\boldsymbol{\mu}\|_2 \gg \sqrt{d/n}$
- Clustering: $\|\boldsymbol{\mu}\|_2 \gg (d/n)^{1/4}$

Never ask for more than you need!

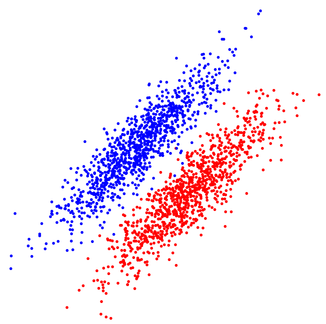
Clustering via **U**ncoupled **RE**gression

- The CURE methodology
- **Theoretical guarantees**

Elliptical Mixture Model

Main Assumptions

$$\boldsymbol{x}_i \sim \begin{cases} (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), & \text{if } y_i = 1 \\ (\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}), & \text{if } y_i = -1 \end{cases}.$$



- $\mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2$, $\boldsymbol{x}_i = \boldsymbol{\mu}_{y_i} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$;
- $\boldsymbol{z}_i$ spherically symmetric, leptokurtic, sub-Gaussian.

CURE: $$\min_{\alpha\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} f(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i) + \frac{1}{2}(\alpha + \boldsymbol{\beta}^\top \bar{\boldsymbol{x}})^2 \right\}.$$

Theoretical Guarantees

Theorem (WYD'20)

Suppose n/d is large. The **perturbed gradient descent** alg. (Jin et al. 2017) starting from **0** achieves **stat. precision** within

$$\widetilde{O}\left(\frac{n}{d} \vee \frac{d^2}{n}\right)$$

iterations (hiding polylog factors).

Theoretical Guarantees

Theorem (WYD'20)

Suppose n/d is large. The **perturbed gradient descent** alg. (Jin et al. 2017) starting from **0** achieves **stat. precision** within

$$\widetilde{O}\left(\frac{n}{d} \vee \frac{d^2}{n}\right)$$

iterations (hiding polylog factors).

- **Efficient** clustering for **stretched** mixtures **without** warm start;
- Two terms: prices for accuracy **(stat.)** and smoothness **(opt.)**;
- Angular error: $\widetilde{O}(\sqrt{d/n})$; excess risk: $\widetilde{O}(d/n)$.

Proof Sketch: Population

Consider the centered case $\boldsymbol{x}_i \sim (\pm\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\beta}^\top \boldsymbol{x}_i).$$

Theorem (population landscape)

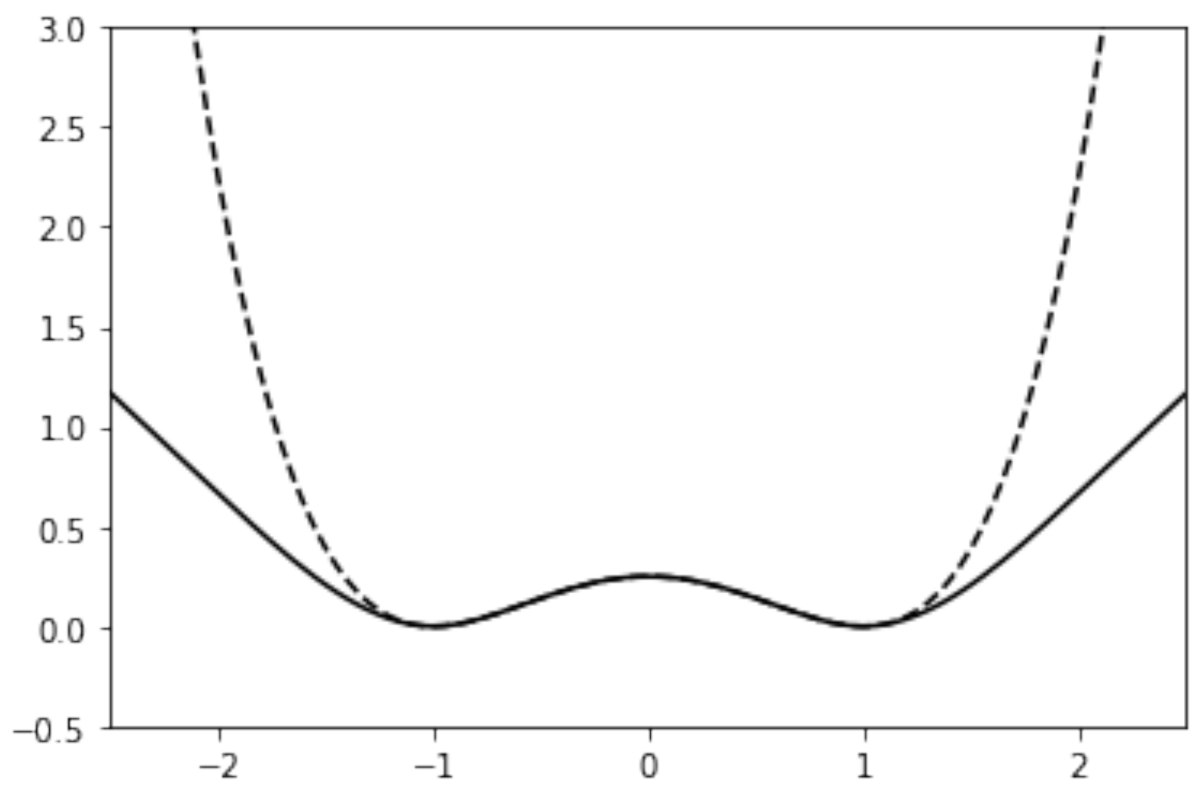
Let $f(x) = (x^2-1)^2/4$. For the infinite-sample loss:

- Two minima $\pm\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^* \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, locally strongly cvx;
- Local maximum $\mathbf{0}$; all saddles are strict.

Loss Function

Clip $(x^2 - 1)^2/4$ to improve

- concentration and robustness for **statistics**;
- growth condition and smoothness for **optimization**.



Proof Sketch: Finite Samples

Theorem (empirical landscape)

Suppose n/d is large and let $\widehat{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$. W.h.p.,

- Approx. second-order stationary points are good:

- $\nabla \widehat{L}$ is $\widetilde{O}(1)$-Lipschitz, $\nabla^2 \widehat{L}$ is $\widetilde{O}(1 \vee \frac{d}{\sqrt{n}})$-Lipschitz.

Nice landscape ensures efficiency and accuracy of optimization.

Proof Sketch: Finite Samples

Theorem (empirical landscape)

Suppose n/d is large and let $\widehat{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$. W.h.p.,

- Approx. second-order stationary points are good:
  if $\|\nabla \widehat{L}(\boldsymbol{\beta})\|_2 \le \delta, \quad \lambda_{\min}[\nabla^2 \widehat{L}(\boldsymbol{\beta})] \ge -\delta,$ then

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \lesssim \underbrace{\|\nabla \widehat{L}(\boldsymbol{\beta})\|_2}_{\text{opt err.}} + \underbrace{\sqrt{\frac{d}{n}\log\left(\frac{n}{d}\right)}}_{\text{stat err.}};$$

- $\nabla \widehat{L}$ is $\widetilde{O}(1)$-Lipschitz, $\nabla^2 \widehat{L}$ is $\widetilde{O}(1 \vee \frac{d}{\sqrt{n}})$-Lipschitz.

Nice landscape ensures efficiency and accuracy of optimization.

Summary

A general **CURE** for clustering **problems**.

Wang, Yan and Díaz. Efficient clustering for stretched mixtures: landscape and optimality. Submitted.

- **Clustering** -> **classification**;
- Flexible choices of transforms, OOS-extensions;
- **Stat.** and **comp.** guarantees under mixture models.

Extensions

- High dim., significance testing, model selection;
- Representation learning, semi-supervised version.

Q & A

Thank you!