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

\[ \{ \mathbf{x}_i \}_{i=1}^{n} \sim \frac{1}{2} N(\mu, I_d) + \frac{1}{2} N(-\mu, I_d) \]
Spherical Clusters

\[ \{x_i\}_{i=1}^n \sim \frac{1}{2} N(\mu, I_d) + \frac{1}{2} N(-\mu, I_d) \]

- PCA: \( \max_{\beta \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i=1}^{n} (\beta^\top x_i)^2 \)
- k-means: \( \min_{\mu_1, \mu_2, y} \frac{1}{n} \sum_{i=1}^{n} \| x_i - \mu_{y_i} \|_2^2 \)
- SDP relaxations of k-means, etc
- Density-based methods require large samples
Finding a Needle in a Haystack

They are powerful but not omnipotent.

\[ \frac{1}{2} N(\mu, \Sigma) + \frac{1}{2} N(-\mu, \Sigma): \text{covariance } \mu \mu^\top + \Sigma \]

- Max variance \( \neq \) useful
- PCA: \( \| \mu \|_2^2 / \| \Sigma \|_2 \gg 1 \) or \( \Sigma \approx I \)

Reduction to the spherical case?
- Estimation of \( \Sigma \) 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 Uncoupled Regression

- The CURE methodology
- Theoretical guarantees
Vanilla CURE

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

$$\beta^\top x_i \approx y_i, \quad i \in [n].$$
Vanilla CURE

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

$$\beta^\top x_i \approx y_i, \quad i \in [n].$$

Cluster via Uncoupled REgression:

$$\frac{1}{n} \sum_{i=1}^n \delta \beta^\top x_i \approx \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1.$$
Vanilla CURE

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

\[
\beta^\top \mathbf{x}_i \approx y_i, \quad i \in [n].
\]

Clustering via Uncoupled REgression:

\[
\frac{1}{n} \sum_{i=1}^n \delta \beta^\top \mathbf{x}_i \approx \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1.
\]

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

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

\[ \frac{1}{n} \sum_{i=1}^{n} f(\beta^\top x_i) \] 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 \( \{x_i\}_{i=1}^n \subseteq \mathbb{R}^d \), find \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R}^d \) s.t.

\[
\frac{1}{n} \sum_{i=1}^n \delta_{\alpha + \beta^\top x_i} \approx \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1.
\]

The naïve extension

\[
\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f(\alpha + \beta^\top x_i)
\]

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

It only forces \(|\alpha + \beta^\top x_i| \approx 1\) rather than

\[
\#\{i : \alpha + \beta^\top x_i \approx 1\} \approx \frac{n}{2}.
\]
Vanilla CURE with Intercept

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

\[
\frac{1}{n} \sum_{i=1}^{n} \delta_{\alpha + \beta^\top \mathbf{x}_i} \approx \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_{1}.
\]

**CURE:**

\[
\min_{\alpha \in \mathbb{R}, \, \beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(\alpha + \beta^\top \mathbf{x}_i) + \frac{1}{2} (\alpha + \beta^\top \bar{\mathbf{x}})^2 \right\}.
\]
Vanilla CURE with Intercept

Given \( \{ \mathbf{x}_i \}_{i=1}^n \subseteq \mathbb{R}^d \), find \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R}^d \) s.t.
\[
\frac{1}{n} \sum_{i=1}^{n} \delta_{\alpha + \beta^\top \mathbf{x}_i} \approx \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_1.
\]

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

- \( \frac{1}{n} \sum_{i=1}^{n} f(\alpha + \beta^\top \mathbf{x}_i) \): \( |\alpha + \beta^\top \mathbf{x}_i| \approx 1 \);
- \( (\alpha + \beta^\top \bar{\mathbf{x}})^2 \): \( \#\{i : \alpha + \beta^\top \mathbf{x}_i \approx 1\} \approx n/2 \).
  
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 \( \{x_i\}_{i=1}^n \subseteq \mathcal{X} \), find \( f : \mathcal{X} \rightarrow \mathcal{Y} \) in \( \mathcal{F} \) s.t.

\[
\frac{1}{n} \sum_{i=1}^{n} \delta_{f(x_i)} \approx \sum_{j=1}^{K} \pi_j \delta_{y_j}.
\]
General CURE

Given \( \{x_i\}_{i=1}^n \subseteq \mathcal{X} \), find \( f : \mathcal{X} \rightarrow \mathcal{Y} \) in \( \mathcal{F} \) s.t.

\[
\frac{1}{n} \sum_{i=1}^n \delta_{f(x_i)} \approx \sum_{j=1}^K \pi_j \delta_{y_j}.
\]

**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;
Clustering Algorithms

- **Generative:** \((X, Y) \rightarrow (Y \mid X)\)
  - Distribution learning (EM, DBSCAN)
  - \(~\) Linear discriminant analysis

- **Discriminative:** \((Y \mid X) \rightarrow \text{CURE}\) 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: \( \{x_i\}_{i=1}^{n} \sim \frac{1}{2}N(\mu, I_d) + \frac{1}{2}N(-\mu, I_d) \) with \( d \gg n \)

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

Never ask for more than you need!
Clustering via Uncoupled REgression

- The CURE methodology
- Theoretical guarantees
Elliptical Mixture Model

### Main Assumptions

\[ x_i \sim \begin{cases} 
(\mu_1, \Sigma), & \text{if } y_i = 1 \\
(\mu_{-1}, \Sigma), & \text{if } y_i = -1.
\end{cases} \]

- \( \mathbb{P}(y_i = 1) = \mathbb{P}(y_i = -1) = 1/2, \ x_i = \mu_{y_i} + \Sigma^{1/2} z_i; \)
- \( z_i \) spherically symmetric, leptokurtic, sub-Gaussian.

**CURE:**

\[
\min_{\alpha \in \mathbb{R}, \ \beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(\alpha + \beta^T x_i) + \frac{1}{2} (\alpha + \beta^T \bar{x})^2 \right\}.
\]
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

$$\tilde{O}\left(\frac{n}{d} \sqrt{\frac{d^2}{n}}\right)$$

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 \( \mathbf{0} \) achieves **stat. precision** within

\[
\tilde{O}\left( \frac{n}{d} \sqrt{\frac{d^2}{n}} \right)
\]

iterations (hiding polylog factors).

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

Consider the centered case $x_i \sim (\pm \mu, \Sigma)$:

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

Theorem (population landscape)

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

- Two minima $\pm \beta^*$, where $\beta^* \propto \Sigma^{-1} \mu$, locally strongly cvx;
- Local maximum 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**.

![Graph of the function](image)
Proof Sketch: Finite Samples

**Theorem (empirical landscape)**

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

- Approx. second-order stationary points are good:
  
  - $\nabla \hat{L}$ is $\tilde{O}(1)$-Lipschitz, $\nabla^2 \hat{L}$ is $\tilde{O}(1 \lor \frac{d}{\sqrt{n}})$-Lipschitz.

Nice landscape ensures efficiency and accuracy of optimization.
Proof Sketch: Finite Samples

**Theorem (empirical landscape)**

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

- Approx. second-order stationary points are good:
  
  If \( \|\nabla \hat{L}(\beta)\|_2 \leq \delta, \quad \lambda_{\min}[\nabla^2 \hat{L}(\beta)] \geq -\delta \), then

  \[
  \|\beta - \beta^*\|_2 \lesssim \|\nabla \hat{L}(\beta)\|_2 + \sqrt{\frac{d}{n} \log \left( \frac{n}{d} \right)}; \\
  \]

- \( \nabla \hat{L} \) is \( \tilde{O}(1) \)-Lipschitz, \( \nabla^2 \hat{L} \) is \( \tilde{O}(1 \lor \frac{d}{\sqrt{n}}) \)-Lipschitz.

Nice landscape ensures efficiency and accuracy of optimization.
Summary

A general CURE for clustering problems.


- **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!