Recovery of a mixture of Gaussians by sum-of-norms clustering

Tao Jiang[1] Stephen Vavasis[1] Chen Wen Zhai [1] [2]

[1]Department of Combinatorics & Optimization
University of Waterloo

[2]Department of Statistics & Actuarial Science
University of Waterloo

The Sixth International Conference on Continuous Optimization

Outline

1. Sum-of-norms clustering
2. Recovery result of a mixture of Gaussians
3. Cluster characterization theorem
4. Recovery theorem of a mixture of Gaussians
5. Discussion

Clustering

Given n points $a_1, a_2, ..., a_n$ lying in $\mathbb{R}^d$, one seeks to partition $\{1, \ldots, n\}$ into K sets $C_1, \ldots, C_K$ such that the a_i's for $i \in C_m$ are closer to each other than to the a_i's for $i \in C_{m'}$, $m' \neq m$.

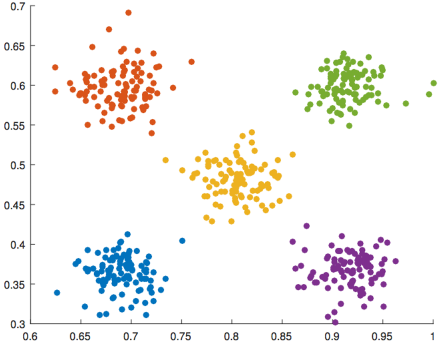


Figure: Visualization of a possible clustering

Traditional clustering models

Here is a hierarchical clustering model

$$\begin{aligned} \min_{x_1,\ldots,x_n\in\mathbb{R}^d} \quad & \frac{1}{2}\sum_{i=1}^{n}\|x_i - a_i\|^2 \\ \text{subject to} \quad & \sum_{i<j} 1_{x_i\neq x_j} \leq t \end{aligned} \tag{1}$$

Let $x_1^*, x_2^*, ..., x_n^*$ be the optimizer of (1). For any distinct pair $i, j \in \{1, 2, ..., n\}$,

- if $x_i^* = x_j^*$, points i, j are assigned to the same cluster;
- Otherwise, points i, j are assigned to different clusters;

Traditional clustering models

Here is a hierarchical clustering model

$$\begin{aligned} \min_{x_1,\ldots,x_n \in \mathrm{R}^d} \quad & \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 \\ \text{subject to} \quad & \sum_{i<j} 1_{x_i \neq x_j} \leq t \end{aligned}$$

Remark

- If $t \geq \frac{n(n-1)}{2}$, (1) is unconstrained and $x_i^* = a_i$ for all $i \in \{1, 2, \ldots, n\}$;
- If $t = \frac{n(n-1)}{2} - 1$, one distinct pair $i, j \in \{1, 2, \ldots, n\}$ is forced to fuse;
- If $t = 0$, $x_i^* = \sum_{i=1}^{n} \frac{a_i}{n}$.

Traditional clustering models

Problems of traditional clustering models:

- Most are hard combinatorial optimization problems;
- Prior knowledge about the number of clusters is often required;
- Initialization affects the clustering assignment.

A convex clustering model

Hocking et al. (2011) proposed the following convex relaxation of the Hierarchical clustering model.

$$\begin{aligned} \min_{x_1,\ldots,x_n \in \mathbb{R}^d} \quad & \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 \\ \text{subject to} \quad & \sum_{i<j} \|x_i - x_j\| \le t \end{aligned} \tag{2}$$

Here is the Lagrangian formulation of (2).

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \lambda \sum_{i<j} \|x_i - x_j\| .$$

A convex clustering model

$$\min_{x_1,\dots,x_n\in\mathbb{R}^d} \frac{1}{2}\sum_{i=1}^{n}\|x_i - a_i\|^2 + \lambda\sum_{i<j}\|x_i - x_j\|. \tag{3}$$

The formulation (3) is known as sum-of-norms clustering, convex clustering, or clusterpath clustering.

Remark

The formulation (3) is strongly convex.

Let $x_1^*, x_2^*, \dots, x_n^*$ be the optimizer of (3). For any distinct pair $i, j \in \{1, 2, \dots, n\}$,

- if $x_i^* = x_j^*$, points i, j are assigned to the same cluster;
- Otherwise, points i, j are assigned to different clusters;

A mixture of Gaussians

- Setup: Given K Gaussians with means $\mu_1, \ldots, \mu_K \in \mathrm{R}^d$, variances $\sigma_1^2, \ldots, \sigma_K^2$, and probabilities $w_1, \ldots, w_K$, positive and summing to 1.
- Generative model: One draws n i.i.d. samples from K Gaussians.
  - An index $m \in \{1, \ldots, K\}$ is selected at random according to probabilities $w_1, \ldots, w_K$,
  - A point a is chosen according to the spherical Gaussian distribution $N(\mu_m, \sigma_m^2 I)$.

Recovery condition by Panahi et al. (2017)

Recovery condition (Panahi et al., 2017)

For the appropriate choice of λ, sum-of-norms clustering formulation (3) exactly recovers a mixture of Gaussians provided that for all m, m', $1 \leq m < m' \leq K$,

$$\|\mu_m - \mu_{m'}\| \geq \frac{CK\sigma_{\max}}{w_{\min}} \mathrm{polylog}(n). \tag{4}$$

- C: some constant
- K: the number of Gaussians
- polylog(n): a polynomial function with respect to log(n)
- σ_{max}: $max\{\sigma_1, \sigma_2, ..., \sigma_K\}$
- w_{min}: $min\{w_1, w_2, ..., w_K\}$

Recovery condition by Panahi et al. (2017)

$$\|\mu_m - \mu_{m'}\| \geq \frac{CK\sigma_{\max}}{w_{\min}} \mathrm{polylog}(n).$$

Remark

As the number of samples n tends to infinity, the bound implies that distinguishing the clusters becomes increasingly difficult

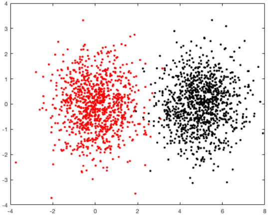


Figure: 2D Gaussians with 1000 samples

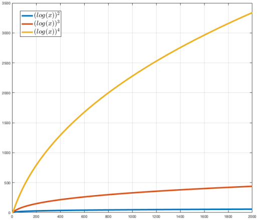
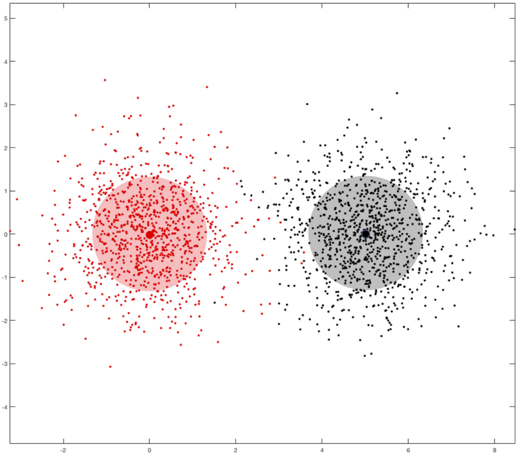


Figure: $polylog(n)$ VS n

Main contributions

We prove (3) can correctly cluster the points lying within some fixed number (θ) of standard-deviations for each mean even as $n \to \infty$.



Our recovery condition

Define $V_m = \{a_i : \|a_i - \mu_m\| \leq \theta\sigma_m\}, \quad m = 1, \ldots, K.$

Recovery condition

There is a λ such that with probability tending to 1 exponentially fast in n, the points in V_m are in the same cluster for any $m = 1, \ldots, K$, and these clusters are distinct, provided that

$$\min_{1 \leq m < m' \leq K} \|\mu_m - \mu_{m'}\| > \frac{4\theta\sigma_{\max}}{F(\theta, d)w_{\min} - \epsilon}. \tag{5}$$

* d: the dimension of the data space
* θ: the number of standard-deviations from the mean
* $\epsilon > 0$: an arbitrary number
* $F(\theta, d)$ denotes the cumulative density function of the chi distribution with d degrees of freedom

Our recovery condition

$$\min_{1 \le m < m' \le K} \|\mu_m - \mu_{m'}\| > \frac{4\theta\sigma_{\max}}{F(\theta, d) w_{\min} - \epsilon}.$$

Remark

The dependence of the right-hand side on n as well as the factor of K has been removed.

Cluster characterization theorem by Chiquet et al. (2017)

Let $x_1^*, \ldots, x_n^*$ denote the optimizer of (3). Let $x^* := \begin{bmatrix} x_1^* \\ x_2^* \\ \\ x_n^* \end{bmatrix} \in \mathbb{R}^{nd}$.

Suppose $\emptyset \neq C \subseteq \{1, \ldots, n\}$.

(a) Necessary condition

If for some $\hat{x} \in \mathrm{R}^d$, $x_i^* = \hat{x}$ for $i \in C$ and $x_i^* \neq \hat{x}$ for $i \notin C$, then there exist z_{ij}^* for $i, j \in C$, $i \neq j$, which solve

$$\begin{aligned} a_i - \frac{1}{|C|} \sum_{l \in C} a_l &= \lambda \sum_{j \in C - \{i\}} z_{ij}^* && \forall i \in C, \\ \left\| z_{ij}^* \right\| &\leq 1 && \forall i, j \in C, i \neq j, \\ z_{ij}^* &= -z_{ji}^* && \forall i, j \in C, i \neq j. \end{aligned} \tag{6}$$

Cluster characterization theorem by Chiquet et al. (2017)

Suppose $\emptyset \neq C \subseteq \{1, \ldots, n\}$.

(b) Sufficient condition

Suppose there exists a solution z_{ij}^* for $j \in C - \{i\}$, $i \in C$ to the following conditions.

$$\begin{aligned} a_i - \frac{1}{|C|}\sum_{l \in C} a_l &= \lambda \sum_{j \in C - \{i\}} z_{ij}^* && \forall i \in C, \\ \left\| z_{ij}^* \right\| &\leq 1 && \forall i, j \in C, i \neq j, \\ z_{ij}^* &= -z_{ji}^* && \forall i, j \in C, i \neq j. \end{aligned}$$

Then there exists an $\hat{x} \in \mathrm{R}^d$ such that the minimizer x^* of (3) satisfies $x_i^* = \hat{x}$ for $i \in C$.

Cluster characterization theorem by Chiquet et al. (2017)

With the cluster chracterization theorem,

- one can chacterize the cluster assignment without the information of other points;
- one can prove the agglomeration property of sum-of-norms clustering with unitary weight (conjectured by Hocking et al. (2011)).

Consider a $\bar{\lambda} \geq \lambda$ and its corresponding sum-of-norms cluster model:

$$\min_{x_1,\ldots,x_n} \frac{1}{2}\sum_{i=1}^{n} \|x_i - a_i\|^2 + \bar{\lambda}\sum_{i<j} \|x_i - x_j\| . \tag{7}$$

Corollary (Chiquet et al., 2017)

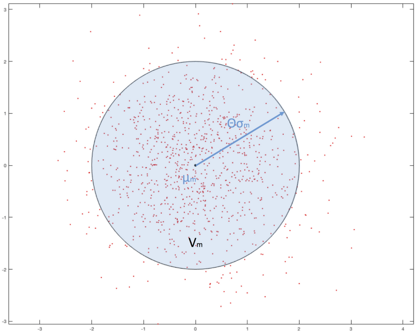
If there is a C such that minimizer x^* of (3) satisfies $x_i^* = \hat{x}$ for $i \in C$, $x_i^* \neq \hat{x}$ for $i \notin C$ for some $\hat{x} \in \mathrm{R}^d$, then there exists an $\hat{x}' \in \mathrm{R}^d$ such that the minimizer of (7), $\bar{x}^*$, satisfies $\bar{x}_i^* = \hat{x}'$ for $i \in C$.

Recovery of a mixture of Gaussians theorem

Let the vertices $a_1, \ldots, a_n \in \mathrm{R}^d$ be generated from a mixture of K Gaussian distributions with parameters $\mu_1, \ldots, \mu_K$, $\sigma_1^2, \ldots, \sigma_K^2$, and $w_1, \ldots, w_K$. Let $\theta > 0$ be given, and let

$$V_m = \{a_i : \|a_i - \mu_m\| \leq \theta\sigma_m\}, \quad m = 1, \ldots, K.$$

Let $\epsilon > 0$ be arbitrary.



Recovery of a mixture of Gaussians theorem

Theorem (Lower Bound)

For any $m = 1, \ldots, K$, with probability exponentially close to 1 (and depending on ϵ) as $n \to \infty$, for the solution x^* computed by (3), the points in V_m are in the same cluster provided

$$\lambda \geq \frac{2\theta\sigma_m}{(F(\theta, d)w_m - \epsilon)n}. \tag{8}$$

* $F(\theta, d)$: the cumulative density function of the chi distribution with d degrees of freedom.

Recovery of a mixture of Gaussians theorem

Theorem (Upper Bound)

Furthermore, the cluster associated with V_m is distinct from the cluster associated with $V_{m'}$, $1 \leq m < m' < k$, provided that

$$\lambda < \frac{\|\mu_m - \mu_{m'}\|}{2(n-1)}. \tag{9}$$

Recovery of a mixture of Gaussians theorem

Provided that

$$\frac{2\theta\sigma_m}{(F(\theta,d)w_m-\epsilon)n} < \frac{\|\mu_m-\mu_{m'}\|}{2(n-1)},$$

there exists a λ so that the solution to (3) can simultaneously place all points in V_m into the same cluster for each $m = 1, \ldots, K$ while distinguishing the clusters.

Discussion

- The key technique is the cluster characterization theorem, which decouples the clusters from each other so that each can be analyzed in isolation.
- The analysis can be extended to Gaussians with a more general covariance matrix, uniform distributions and many kinds of deterministic distributions.
- The cluster characterization theorem does not apply to most other clustering algorithms, or even to sum-of-norm clustering in the case of unequal weights.

Reference I

Emmanuel J. Cand'es and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, Apr 2009

E. Chi and S. Steinerberger. Recovering trees with convex clustering. https://arxiv.org/abs/1806.11096, 2018

J. Chiquet, P. Gutierrez, and G. Rigaill. Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26:205216, 2017.

Jean-Baptiste Hiriart-Urruty and Claude Lemar echal. Fundamentals of convex analysis.Springer, 2012.

T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: An algorithm for clustering using convex fusion penalties. In *International Conference on Machine Learning*, 2011

Reference II

F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *IEEE Statistical Signal Pro-cessing Workshop (SSP)*, 2011.

A. Panahi, D. Dubhashi, F. Johansson, and C. Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. *Journal of Machine Learning Research*, 70, 2017.

Peter Radchenko and Gourab Mukherjee. Convex clustering via l1 fusion penalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):15271546, 2017.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Reference III

D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: model, theoretical guarantees and efficient algorithm. https://arxiv.org/abs/1810.02677, 2018.

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Convex cluster shrinkage. Available online at ftp://ftp.esat.kuleuven.ac.be/sista/kpelckma/ccs$_pelckmans2005$ .pdf, 2005.