Lecture Notes VII: Classic and Modern Data Clustering – Part III

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

November, 2020

Issues in parametric clustering

Outliers

Cluster validation

Selecting K for hard clustering

Reading HTF Ch.: , Murphy Ch.:

Issues in parametric clustering

- Selecting K
- Outliers

Clustering with outliers

- What are outliers?
- let $p =$ proportion of outliers (e.g 5%-10%)
- Remedies
  - mixture model: introduce a $K+1$-th cluster with large (fixed) Σ_{K+1}, bound Σ_k away from 0
  - K-means and EM
    - **robust** means and variances
      e.g eliminate smallest and largest $pn_k/2$ samples in mean computation (**trimmed mean**)
    - K-medians [Charikar and Guha, 1999]
    - replace Gaussian with a heavier-tailed distribution (e.g. Laplace)
  - single-linkage: do not count clusters with $< r$ points

Is K meaningful when outliers present?

- alternative: non-parametric clustering

Cluster validation

- External
  - when the true clustering Δ^* is known
  - compares result(s) Δ obtained by algorithm **A** with Δ^*
  - validates algorithms/methods
- Internal - no external reference

External cluster validation

Scenarios

- given data $\mathcal{D}$, truth Δ^*; algorithm **A** produces Δ
  is Δ close to Δ^*?
- given data $\mathcal{D}$, truth Δ^*; algorithm **A** produces Δ, algorithm **A'** produces Δ'
  which of Δ, Δ' is closer to Δ^*?
- multiple datasets, multiple algorithms
  which algorithm is better?

A **distance between clusterings** $d(\Delta, \Delta')$ needed

Requirements for a distance

Depend on the application

- Applies to any two partitions of the same data set
- Makes no assumptions about how the clusterings are obtained
- Values of the distance between two pairs of clusterings comparable under the weakest possible assumptions
- Metric (triangle inequality) desirable
- **understandable, interpretable**

The confusion matrix

- Let $\Delta = \{C_{1:K}\}, \ \Delta' = \{C'_{1:K'}\}$
- Define $n_k = |C_k|, \ n'_{k'} = |C'_{k'}|$
- $m_{kk'} = |C_k \cap C'_{k'}|, \ k = 1:K, k' = 1:K'$
- note: $\sum_k m_{kk'} = n'_{k'}, \ \sum_{k'} m_{kk'} = n_k, \ \sum_{k,k'} m_{kk'} = n$
- The **confusion matrix** $M \in \mathbb{R}^{K \times K'}$ is

$$M = [m_{kk'}]_{k=1:K}^{k'=1:K'}$$

- all distances and comparison criteria are based on M
- the **normalized confusion matrix** $P = M/n$

$$p_{kk'} = \frac{m_{kk'}}{n}$$

- The **normalized cluster sizes** $p_k = n_k/n, \ p'_{k'} = n'_{k'}/n$ are the **marginals** of P

$$p_k = \sum_{k'} p_{kk'} \quad p_{k'} = \sum_{k} p_{kk'}$$

The Misclassification Error (ME) distance

- Define the **Misclassification Error (ME)** distance d_{ME}

$$d_{ME} = 1 - \max_{\pi} \sum_{k=1}^{K} p_{k,\pi(k)} \quad \pi \in \{\text{all } K\text{−permutations}\},\ K \leq K' \text{w.l.o.g}$$

- Interpretation: treat the clusterings as classifications, then minimize the classification error over all possible label matchings
- Or: nd_{ME} is the Hamming distance between the vectors of labels, minimized over all possible label matchings
- can be computed in polynomial time by **Max bipartite matching** algorithm (also known as Hungarian algorithm)
- Is a metric: symmetric, ≥ 0, triangle inequality

$$d_{ME}(\Delta_1, \Delta_2) + d_{ME}(\Delta_1, \Delta_3) \geq d_{ME}(\Delta_2, \Delta_3)$$

- easy to understand (very popular in computer science)
- $d_{ME} \leq 1 - 1/K$
- bad: if clusterings not similar, or K large, d_{ME} is coarse/indiscriminative
- recommended: for small K

The Variation of Information (VI) distance
Clusterings as random variables

- Imagine points in $\mathcal{D}$ are picked randomly, with equal probabilities
- Then $k(i), k'(j)$ are random variables
  with $Pr[k] = p_k$, $Pr[k, k'] = p_{kk'}$

Incursion in information theory

- **Entropy** of a random variable/clustering $H_\Delta = -\sum_k p_k \ln p_k$
- $0 \leq H_\Delta \leq \ln K$
- Measures uncertainty in a distribution (amount of randomness)
- **Joint entropy** of two clusterings

$$H_{\Delta,\Delta'} = -\sum_{k,k'} p_{kk'} \ln p_{kk'}$$

- $H_{\Delta',\Delta} \leq H_\Delta + H_{\Delta'}$ with equality when the two random variables are independent
- **Conditional entropy** of Δ' given Δ

$$H_{\Delta'|\Delta} = -\sum_k p_k \sum_{k'} \frac{p_{kk'}}{p_k} \ln \frac{p_{kk'}}{p_k}$$

- Measures the expected uncertainty about k' when k is known
- $H_{\Delta'|\Delta} \leq H_{\Delta'}$ with equality when the two random variables are independent
- **Mutual information** between two clusterings (or random variables)

$$\begin{aligned} I_{\Delta,\Delta} &= H_\Delta + H_{\Delta'} - H_{\Delta',\Delta} \\ &= H_{\Delta'} - H_{\Delta'|\Delta} \end{aligned}$$

- Measures the amount of information of one r.v. about the other
- $I_{\Delta,\Delta} \geq 0$, symmetric. Equality iff r.v.'s independent

The VI distance

- Define the **Variation of Information (VI)** distance

$$\begin{aligned} d_{VI}(\Delta, \Delta') &= H_\Delta + H_{\Delta'} - 2I_{\Delta',\Delta} \\ &= H_{\Delta|\Delta'} + H_{\Delta'|\Delta} \end{aligned}$$

- Interpretation: d_{VI} is the sum of information gained and information lost when labels are switched from $k()$ to $k'()$
- d_{VI} symmetric, ≥ 0
- d_{VI} obeys triangle inequality (is a metric)

Other properties

- Upper bound
  $d_{VI} \leq 2\ln K_{max}$ if $K, K' \leq K_{max} \leq \sqrt{n}$
  (asymptotically attained)
- $d_{VI} \leq \ln n$ over all partitions (attained)
- Unbounded! and grows fast for small K

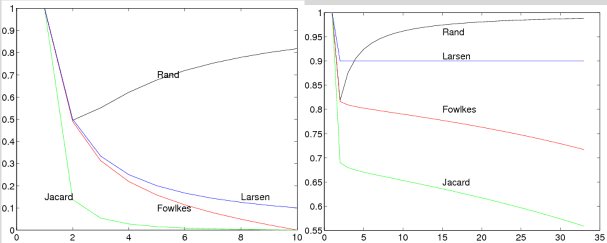
Other criteria and desirable properties

- Comparing clustering by **indices of similarity** $i(\Delta, \Delta')$
  - from statistics (Rand, adjusted Rand, Jaccard, Fowlkes-Mallows ...)
  - range=[0,1], with $i(\Delta, \Delta') = 1$ for $\Delta = \Delta'$
  - the properties of these indices not so good
  - any index can be transformed into a "distance" by $d(\Delta, \Delta') = 1 - i(\Delta, \Delta')$
- Other desirable properties of indices and distances between clusterings
  - n-invariance
  - locality
  - convex additivity

- Define N_{11} = # pairs which are together in both clusterings, N_{12} = # pairs together in Δ, separated in Δ', N_{21} (conversely), N_{22} =#number pairs separated in both clusterings
- Rand index $= \frac{N_{11}+N_{22}}{\#\text{pairs}}$
- Jaccard index $= \frac{N_{11}}{\#\text{pairs}}$
- Fowlkes-Mallows = Precision $\times$ Recall

- all vary strongly with K. Thereforek, **Adjusted** indices used mostly

$$adj(i) = \frac{i - \bar{i}}{\max(i) - \bar{i}}$$



Internal cluster(ing) validation

Why?

- Most algorithms output a clustering even if no clusters in data (parametric algorithms)
  How to decide whether to accept it or not?
- related to selection of K
- Some algorithms are run multiple times (e.g EM)
  How to select the clustering(s) to keep?

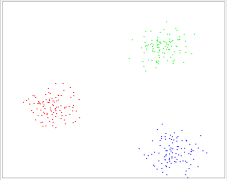
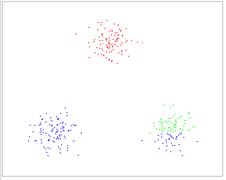
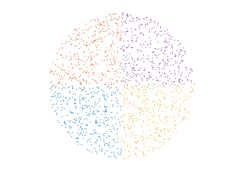
- Validate by the cost $\mathcal{L}$
- Δ is valid if $\mathcal{L}(\Delta)$ is "almost optimal"
  - intractable to know in general (for NP-hard problems)
  - not enough to be "meaningful"

Internal cluster(ing) validation

Why?

- Most algorithms output a clustering even if no clusters in data (parametric algorithms)
  How to decide whether to accept it or not?
- related to selection of K
- Some algorithms are run multiple times (e.g EM)
  How to select the clustering(s) to keep?

- Validate by the cost $\mathcal{L}$
- Δ is valid if $\mathcal{L}(\Delta)$ is "almost optimal"
  - intractable to know in general (for NP-hard problems)
  - not enough to be "meaningful"
- Δ is valid if Δ **stable** and $\mathcal{L}(\Delta)$ is "almost optimal"
  - stable = any other Δ' that is "almost optimal" must be "close" to Δ

Oracle	Yes	No	?
SS method	Yes, OI=$1e^{-4}$	Don't know	Don't know

Heuristics

- **Gap** heuristic
- single linkage:
  - define l_r length of r-th edge added to MST

$$\underbrace{l_1 \leq l_2 \leq \ldots l_{n-K}}_{\text{intracluster}} \leq \underbrace{l_{n-K+1} \leq \ldots}_{\text{deleted}}$$

  - $l_{n-K}/l_{n-K+1} \leq 1$ should be small
- min diameter:

$$\frac{\mathcal{L}(\Delta)}{\max_{i,j\in\mathcal{D}} ||x_i - x_j||}$$

$$\frac{\mathcal{L}(\Delta)}{\min_{k,k'} \text{distance}(C_k, C_{k'})}$$

- etc

Quadratic cost

- $\mathcal{L}(\Delta) = \mathit{const} - \mathrm{trace}\, X^T(\Delta) A X(\Delta)$
- with X = matrix reprentation for Δ
- then, if cost value $\mathcal{L}(\Delta)$ small, we can prove that clustering Δ is almost optimal
- This holds for K-means (weighted, kernelized) and several graph partioning costs (normalized cut, average association, correlation clustering, etc)

Matrix Representations

- matrix reprentations for Δ
  - unnormalized (redundant) representation

$$\tilde{X}_{ik} = \begin{cases} 1 & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

  - normalized (redundant) representation

$$X_{ik} = \begin{cases} 1/\sqrt{|C_k|} & i \in C_k \\ 0 & i \notin C_k \end{cases} \quad \text{for } i = 1:n, k = 1:K$$

    therefore $X_k^T X_{k'} = \delta(k, k')$, X orthogonal matrix
    X_k = column k of X
  - normalized non-redundant reprentation
    - X_K is determined by $X_{1:K-1}$
    - hence we can use $Y \in \mathbb{R}^{n\times(K-1)}$ orthogonal representation
    - intuition: Y represents a subspace (is an orthogonal basis)
    - K centers in $\mathbb{R}^d$, $d \geq K$ determine a $K-1$ dimesional subspace plus a translation

- Example: the K-means cost
  - remember

$$\mathcal{L}(\Delta) = \sum_{k=1}^{K} \sum_{i,j \in C_k} \frac{1}{2|C_k|} ||x_i - x_j||^2 + \text{constant}$$

  - in matrix form

$$\mathcal{L}(\Delta) = -\frac{1}{2} X^T A X + \text{constant}$$

    where

$$A_{ij} = x_i^T x_j$$

    is the Gram matrix of the data
  - if data centered, ie $\sum_i x_i = 0$ and Y rotated appropriately [Meilă, 2006]

$$\mathcal{L}(\Delta) = -\frac{1}{2} Y^T A Y + \text{constant}$$

- Assume k-means cost from now on

A spectral lower bound

- minimizing $\mathcal{L}(\Delta)$ is equivalent to
$$\max Y^T AY$$
over all $Y \in \mathbb{R}^{n\times(K-1)}$ that represent a clustering
- a relaxation
$$\max Y^T AY$$
over all $Y \in \mathbb{R}^{n\times(K-1)}$ orthogonal
- solution to relaxed provlem is
$$Y^* = \text{eigenvectors}_{\ 1:K-1} \text{ of } \ A$$
$$\mathcal{L}^* = \sum_{k=1}^{K-1} \lambda_k(A)$$
- $\mathcal{L}^* = constant - L^* = \operatorname{trace} A - L^*$ is lower bound for $\mathcal{L}$
$$\mathcal{L}^* \leq \mathcal{L}(\Delta) \text{ for all } \Delta$$

A theorem (Meila, 2006)

Theorem

- define

$$\delta = \frac{Y^T A Y - \sum_{k=1}^{K-1} \lambda_k}{\lambda_{K-1} - \lambda_K} \qquad \varepsilon(\delta) = 2\delta[1 - \delta/(K-1)]$$

- define $p_{min}, p_{max} = \frac{\min,\max |C_k|}{n}$
- then, whenever $\varepsilon(\delta) \leq p_{min}$, we have that

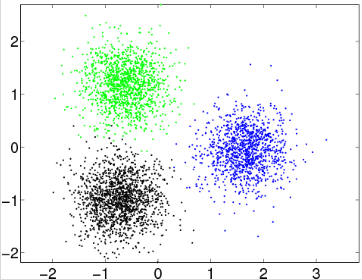
$$d_{ME}(\Delta, \Delta^{opt}) \leq \varepsilon(\delta) p_{max}$$

  where d_{ME} is misclassification error distance

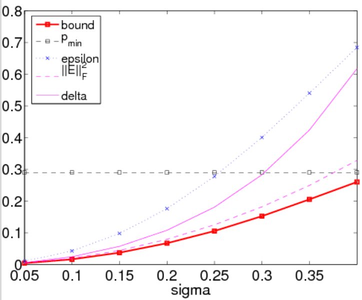
Remarks

- it is a worst-case result
- makes no (implicit) distributional assumptions
- when theorem applies, bound is good $d_{ME}(\Delta, \Delta^{opt}) \leq p_{min}$
- applies only if a good clustering is found (not all data, clusterings)
- intuiton: if data well clustered, $K-1$ principal subspace is aligned with cluster centers

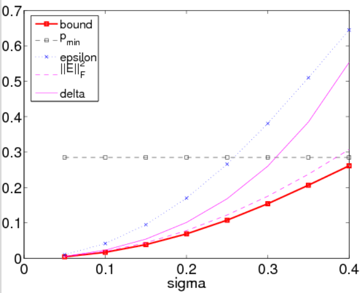
data $d = 35, \ \sigma = 0.4$



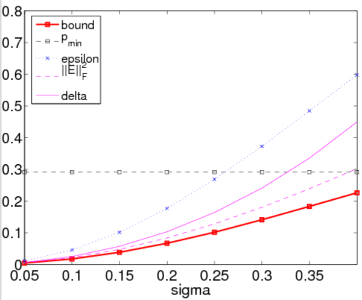
$n = 100$



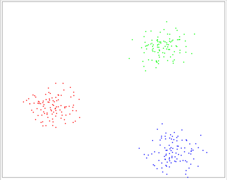
$n = 200$

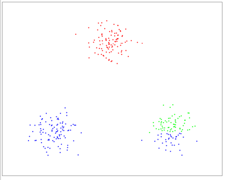


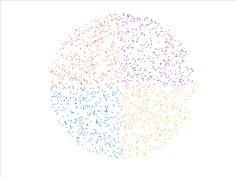
$n = 1000$



Is this clustering approximately correct?

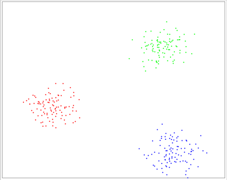
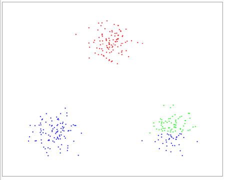
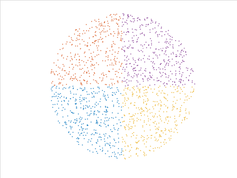






SS method	Yes, OI=$1e^{-4}$	Don't know	Don't know

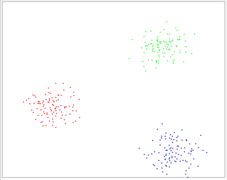
Is this clustering approximately correct?

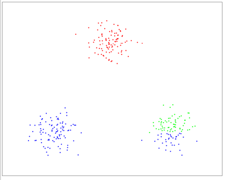




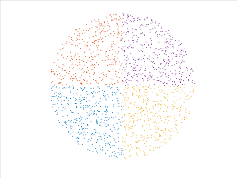
SS method | Yes, OI=$1e^{-4}$ | Don't know | Don't know

- Given data $\mathcal{D}$, clustering Δ
- $\mathcal{L}($ data, clustering$)$ (e.g. K-means)

Is this clustering approximately correct?



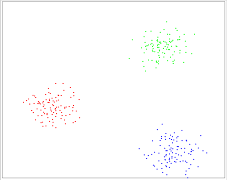
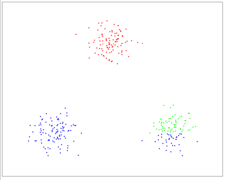
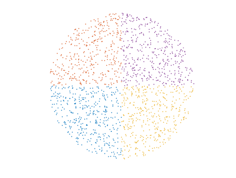




SS method   Yes, OI=$1e^{-4}$ Don't know   Don't know

- Given data $\mathcal{D}$, clustering Δ
- $\mathcal{L}($ data, clustering$)$ (e.g. K-means)
- **"correct"**
  - $=$ the "only" "good" clustering supported by $\mathcal{D}$
  - $\Leftrightarrow$ any other Δ' with smaller $\mathcal{L}$ is ε-close to Δ

Is this clustering approximately correct?

			
SS method	Yes, OI=$1e^{-4}$ good, stable	Don't know bad	Don't know unstable

- Given data $\mathcal{D}$, clustering Δ
- $\mathcal{L}($ data, clustering$)$ (e.g. K-means)
- **"correct"**

  $=$ the "only" "good" clustering supported by $\mathcal{D}$

  $\Leftrightarrow$ any other Δ' with smaller $\mathcal{L}$ is ε-close to Δ

What is an **Optimality Interval (OI)**?

Theorem (Meta-theorem)

If Δ fits the data $\mathcal{D}$ well, then we shall prove that any other clustering Δ' that also fits $\mathcal{D}$ well will be a small perturbation of Δ.

What is an **Optimality Interval (OI)**?

Theorem (Meta-theorem)

If Δ fits the data $\mathcal{D}$ well, then we shall prove that any other clustering Δ' that also fits $\mathcal{D}$ well will be a small perturbation of Δ.

- Δ' is **good** if

$$\mathcal{L}(\Delta') \leq \mathcal{L}(\Delta) + \alpha.$$

- δ is **OI**: for all good Δ',

$$d_{ME}(\Delta', \Delta) \leq \delta$$

  where d_{ME} is the misclassification error/earth mover distance

- if OI exists we say Δ is **stable**

How? 1. Mapping a clustering to a matrix

$n = 5$, $\Delta = (1, 1, 1, 2, 2)$, $X(\Delta) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

1. $X(\Delta)$ is symmetric, positive definite, ≥ 0 elements
2. $X(\Delta)$ has row sums equal to 1
3. $\operatorname{trace} X(\Delta) = K$

$$\|X(\Delta)\|_F^2 = K$$

Let **X** be the space $n \times n$ of matrices with Properties 1, 2, 3 above

- **X** is convex
- $X(C)$ are extreme points of **X**

How? 2. Convex relaxations

Original clustering problem Given data $\mathcal{D}$, K, $\mathcal{L}()$

$$\text{minimize}_{\Delta} \quad \mathcal{L}(\mathcal{D}, \Delta) \quad \text{with solution } \Delta^{\text{opt}}$$

Convex relaxation

- map clustering $\Delta \to$ matrix $X(\Delta) \in \mathbf{X}$
- so that $\mathcal{L}(X)$ convex in X
- Relaxed problem

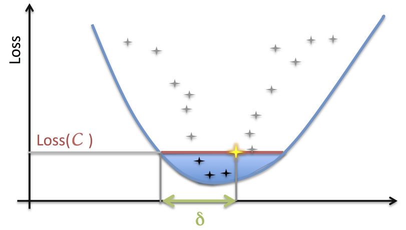
$$L^* = \min_{X \in \mathbf{X}} \mathcal{L}(X), \quad \text{with solution } X^* \tag{1}$$

The Sublevel Set (SS) method

ework Given data, L , convex relaxation

Step 0 Cluster data, obtain a clustering Δ.

$\mathbf{XX}_{\leq l} = \{X \in \mathbf{X},\ \mathcal{L}(X) \leq c\}$ is **sublevel set** of L



The Sublevel Set (SS) method

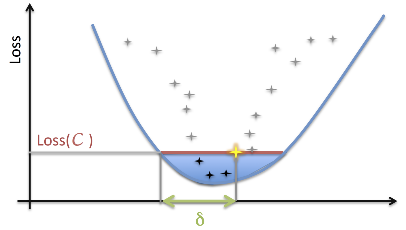
ework Given data, L , convex relaxation

Step 0 Cluster data, obtain a clustering Δ.

Step 1 Use convex relaxation to define new optimization problem

$$\text{SS} \quad \delta \;=\; \max_{X' \in \mathbf{X}} \|X(\Delta) - X'\|_F, \quad \text{s.t. } \mathcal{L}(X') \leq \mathcal{L}(\Delta).$$

$\mathbf{XX}_{\leq l} = \{X \in \mathbf{X},\, \mathcal{L}(X) \leq c\}$ is **sublevel set** of L



The Sublevel Set (SS) method

ework Given data, L , convex relaxation

Step 0 Cluster data, obtain a clustering Δ.

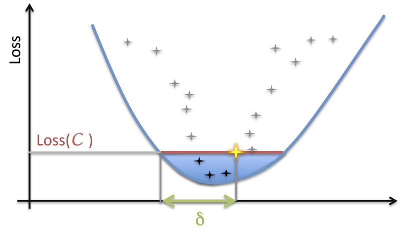
Step 1 Use convex relaxation to define new optimization problem

$$\text{SS} \quad \delta = \max_{X' \in \mathbf{X}} \|X(\Delta) - X'\|_F, \quad \text{s.t. } \mathcal{L}(X') \leq \mathcal{L}(\Delta).$$

Step 2 Prove that $\| \ \|_F \leq \delta \Rightarrow d_{ME}(\ ) \leq \epsilon$ M, MLJ 2012

Done: ϵ is a Optimality Interval (OI) for Δ.

$\mathbf{XX}_{\leq l} = \{X \in \mathbf{X},\ \mathcal{L}(X) \leq c\}$ is **sublevel set** of L



Two technical bits

1. SS is convex only if $||X' - X(\Delta)||$ concave
   - Use $||\ ||_F$ Frobenius norm. $||X(\Delta)||_F^2 = K$ for any clustering.

Two technical bits

1. SS is convex only if $||X' - X(\Delta)||$ concave
   - Use $||\ ||_F$ Frobenius norm. $||X(\Delta)||_F^2 = K$ for any clustering.

2. Relating $\|\ \|_F$ to distance between clusterings.

$$\|X(\Delta) - X(\Delta)'\|_F^2 \leq \delta \quad \Rightarrow \quad d_{ME}(\Delta, \Delta') \leq \epsilon$$

distance between matrices — "misclassification error" metric between clusterings

- Theorem proved in M, Machine Learning Journal, 2012 with $\epsilon = 2\delta p_{\max}$.
- The tightest result known. Upper/lower bounds between d_{ME}, $\|\ \|_F$ and Rand Index
- Proofs use geometry of convex sets + refined analysis for small distances
- Example from Wan,M NIPS16 OI by existing results Rohe et al. 2011 $\sim 10^2$ OI by our method

Relation with other work

- **Previous ideas on OI**
  - Spectral bounds for Spectral Clustering M,Shortreed,Xu AISTATS05
  - Spectral bounds for K-means, NCut and other quadratic costs M, ICML06 and JMVA 2018
  - Spectral bounds for networks model based clustering: Stochastic Block Model and Preference Frame Model Wan,M NIPS2016
- **Previous work we build on**
  - Convex relaxations for clustering MANY! here we use SDP for K-means Peng, Wei 2007
  - Transforming bound on $||X - X'||_F$ into bound on d_{ME} M MLJ 2012
- **Contrast with** work on Clusterability and resilience, e.g. Ben-David, 2015,Bilu,Linial 2009
  - "Their" work: assume $\exists$ stable Δ, prove it can be found efficiently
  - This work: given Δ, prove it is stable

For what clustering paradigms can we obtain OI's?

"All" ways to map Δ to a matrix

space	matrix	definition	size
$\mathcal{X}$	$X(\Delta)$	$X_{ij} = 1/n_k$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\tilde{\mathcal{X}}$	$\tilde{\mathcal{X}}(\Delta)$	$\tilde{\mathcal{X}}_{ij} = 1$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\mathcal{Z}$	$Z(\Delta)$	$Z_{ik} = 1/\sqrt{n_k}$ iff $i \in C_k$	$n \times K$, orthogonal

For what clustering paradigms can we obtain OI's?

"All" ways to map Δ to a matrix

space	matrix	definition	size
$\mathcal{X}$	$X(\Delta)$	$X_{ij} = 1/n_k$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\tilde{\mathcal{X}}$	$\tilde{\mathcal{X}}(\Delta)$	$\tilde{\mathcal{X}}_{ij} = 1$ iff $i, j \in C_k$	$n \times n$, block-diagonal
$\mathcal{Z}$	$Z(\Delta)$	$Z_{ik} = 1/\sqrt{n_k}$ iff $i \in C_k$	$n \times K$, orthogonal

Theorem

M NeurIPS 2018 If L has a convex relaxation involving one of $X, \tilde{\mathcal{X}}, Z$, then

(1) There exists a convex SS problem

$$(SS) \quad \delta = \min_{X' \in \mathbf{XX}_{\leq l}} \langle X(\Delta), X' \rangle \quad \text{(similarly for } \tilde{\mathcal{X}}, Z\text{)}.$$

(2) From optimal δ an OI ε can be obtained, valid when $\varepsilon \leq p_{\min}$.

$$X : X_{ij} = 1/n_k \text{ iff } i, j \in C_k \qquad \varepsilon = (K - \delta) p_{\max}$$

$$\tilde{\mathcal{X}} : \tilde{\mathcal{X}}_{ij} = 1 \text{ iff } i, j \in C_k \qquad \varepsilon = \frac{\sum_{k \in [K]} n_k^2 + (n-K+1)^2 + (K-1) - 2\delta}{2 p_{\min}}$$

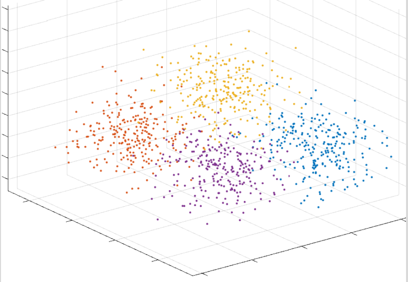
$$Z : Z_{ik} = 1/\sqrt{n_k} \text{ iff } i \in C_k \qquad \varepsilon = (K - \delta^2/2) p_{\max}$$

Existence of guarantee depends only on space of convex relaxation.

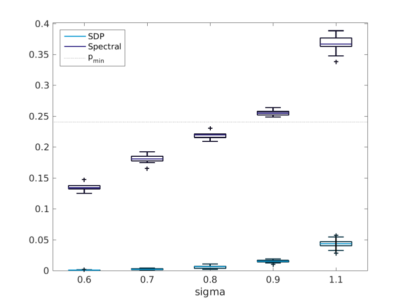
Results for K-means clusterings

$K = 4$ equal Gaussian clusters, $n = 1024$, $||\mu_k - \mu_l|| = 4\sqrt{2} \approx 5.67$

data for $\sigma = 0.9$

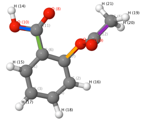


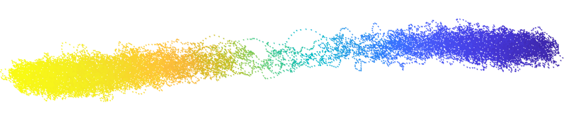
Values of ϵ vs cluster spread σ



Spectral=M ICML06, SDP=M NeurIPS 2018

Aspirin ($C_9 O_4 H_8$) molecular simulation data Chmiela et al. 2017



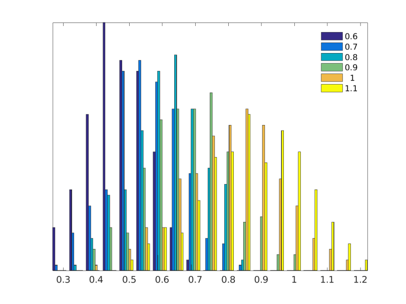


$n = 2118$ $\varepsilon = 0.065$

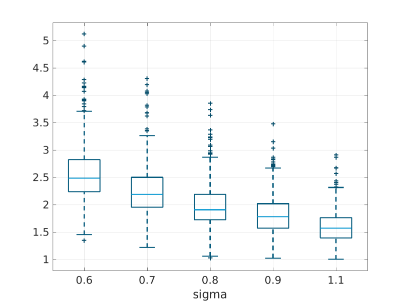
$K = 2$
$p_{\min} = .26$
$p_{\max} = .74$

Separation statistics

distance to own center over min center separation, colored by σ.



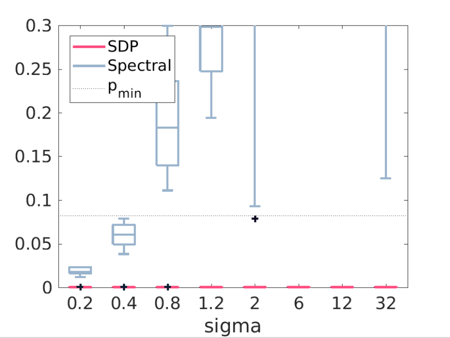
distance to second closest center over distance to own center, versus σ

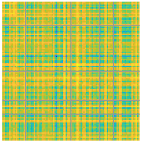


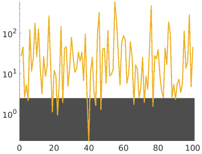
Results for Spectral Clustering by Normalized Cut

Spectral=M AISTATS05, SDP=M NeurIPS 2018

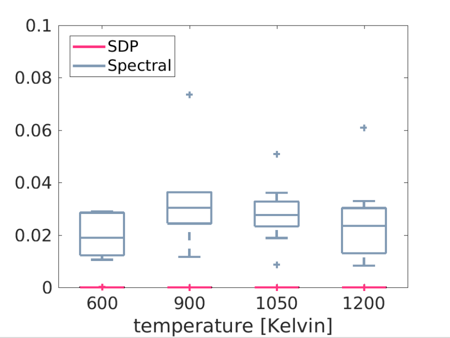
Synthetic S, $n = 100$

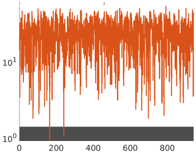


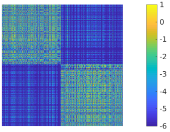




Chemical reaction data, $n \approx 1000$







Stability and the selection of K Cheng,M,Harchaoui (in preparation)

n_200_normal_False_cluster_equal_size_False_full_dimension_True_k_true_8.pdf

Selecting K for hard clusterings

- based on statistical testing: the **gap** statistic (Tibshirani, Walther, Hastie, 2000)
- **X-means** [Pelleg and Moore, 2000] heuristic: splits/merges clusters based on statistical tests of Gaussianity
- Stability methods

The gap statistic

Idea

- for some cost $\mathcal{L}$ compare $\mathcal{L}(\Delta_K)$ with its expected value under a null distribution
  - choose null distribution to have no clusters
    - Gaussian (fit to data)
    - uniform with convex support
    - uniform over K_0 principal components of data
  - null value $= E_{P_0}[\mathcal{L}_{K,n}]$ the expected value of the cost of clustering n points from P_0 into K clusters
- the **gap**

$$g(K) = E_{P_0}[\mathcal{L}_{K,n}] - \mathcal{L}(\Delta_K) = \mathcal{L}_K^0 - \mathcal{L}(\Delta_K)$$

- choose K^* corresponding to the largest gap
- nice: it can also indicate that data has no clusters

Practicalities

- $\mathcal{L}_K^0 = E_{P_0}[\mathcal{L}_{K,n}]$ can rarely be computed in closed form (when P_0 very simple)
- otherwise, estimate $\mathcal{L}_K^0$ be Monte-Carlo sampling i.e generate B samples from P_0 and cluster them
- if sampling, variance s_K^2 of estimate $\hat{\mathcal{L}}_K^0$ must be considered s_K^2 is also estimated from the samples
- selection rule: $K^* =$ smallest K such that $g(K) \geq g(K+1) - s_{K+1}$
- favored $\mathcal{L}^V(\Delta) = \sum_k \frac{1}{|C_k|} \sum_{i \in C_k} ||x_i - \mu_k||^2 \approx$ sum of cluster variances

Stability methods for choosing K

- like bootstrap, or crossvalidation
- **Idea** (implemented by [Ben-Hur et al., 2002])

for each K

1. perturb data $\mathcal{D} \rightarrow \mathcal{D}'$
2. cluster $\mathcal{D}' \rightarrow \Delta'_K$
3. compare Δ_K, Δ'_K. Are they similar?
   If yes, we say Δ_K is **stable to perturbations**

Fundamental assumption If Δ_K is stable to perturbations then K is the correct number of clusters

- these methods are supported by experiments (not extensive)
- **not YET supported by theory** . . . see [von Luxburg, 2009] for a summary of the area

A stability based method for model-based clustering

- **The algorithm of [Lange et al., 2004]**
  1. divide data into 2 halves $\mathcal{D}_1,\ \mathcal{D}_2$ at random
  2. cluster (by EM) $\mathcal{D}_1 \ \rightarrow \Delta_1, \theta_1$
  3. cluster (by EM) $\mathcal{D}_2 \ \rightarrow \Delta_2, \theta_2$
  4. cluster $\mathcal{D}_1$ using $\theta_2 \ \rightarrow \Delta_1'$
  5. compare Δ_1, Δ_1'
  6. repeat B times and average the results
  - repeat for each K
  - select K where Δ_1, Δ_1' are closest on average (or most times)

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002).
A stability based method for discovering structure in clustered data.
In *Pacific Symposium on Biocomputing*, pages 6–17.

Charikar, M. and Guha, S. (1999).
Improved combinatorial algorithms for the facility location and k-median problems.
In *40th Annual Symposium on Foundations of Computer Science*, pages 378–388.

Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M. (2004).
Stability-based validation of clustering solutions.
Neural Comput., 16(6):1299–1323.

Meilă, M. (2006).
The uniqueness of a good optimum for K-means.
In Moore, A. and Cohen, W., editors, *Proceedings of the International Machine Learning Conference (ICML)*, pages 625–632. International Machine Learning Society.

Pelleg, D. and Moore, A. (2000).
X-means: Extending K-means with efficient estimation of the number of clusters.
In Bratko, I. and Džeroski, S., editors, *Proceedings of the 17th International Conference on Machine Learning*, pages 727–734, San Francisco, CA. Morgan Kaufmann.

von Luxburg, U. (2009).
Clustering stability: An overview.
Foundations and Trends in Machine Learning, 2(3):253–274.