Sparse Convex Clustering

Binhuan Wang, Yilong Zhang, Will Wei Sun, Yixin Fang

Reporter : Yanhang Zhang
2021.11.10

Outline

Cluster Analysis

Cluster analysis aims to assign observations into a number of clusters such that observations in the same group are similar to each other.

Traditional clustering methods:

- K-means.
- Hierarchical clustering.
- Gaussian mixture models.

However, these methods suffer from instabilities due to their non-convex optimization formulations.

Convex Clustering

Convex Clustering [Lindsten et al., 2011, Hocking et al., 2011]:

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times p}} \frac{1}{2}\sum_{i=1}^{n}\|X_i - A_{i.}\|_2^2 + \gamma \sum_{i_1<i_2} \left\|A_{i_1.} - A_{i_2.}\right\|_q$$

where $X \in \mathbb{R}^{n\times p}$, $A_{i.}$ is the i-th row of $\mathbf{A}$ and $\|\cdot\|_q$ is the L_q-norm of a vector with $q \in \{1,2,\infty\}$.

- K-means clustering and hierarchical clustering consider L_0-norm in the second term, which leads to a non-convex optimization problem.
- Small γ (e.g. $\gamma = 0$) makes each observation by itself is a cluster.
- Large γ (e.g. $\gamma = \infty$) makes all the row of $\hat{A}$ be identical.

Convex Clustering

- In recent years, much effort has been spent on developing algorithms and theory for convex clustering [Chi and Lange, 2015, Tan and Witten, 2015].
- When the number of features becomes large, many of them may contain no information. Thus the performance of these methods can be severely deteriorated.
- To overcome this problem, an algorithm that can simultaneously perform cluster analysis and select informative variables is in demand.

Sparse Convex Clustering

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times p}} \frac{1}{2}\sum_{i=1}^{n} \|X_{i.} - A_{i.}\|_2^2 + \gamma \sum_{i_1<i_2} w_{i_1,i_2} \left\|A_{i_1.} - A_{i_2.}\right\|_q \tag{1}$$

where the weight $w_{i_1,i_2} \geqslant 0$.

- [Hocking et al., 2011] considered a pairwise affinity weight $w_{i_1,i_2} = \exp\left(-\phi \left\|X_{i_1.} - X_{i_2.}\right\|_2^2\right)$.
- [Chi and Lange, 2015] suggested $w_{i_1,i_2} = \iota^m_{i_1,i_2} \exp\left(-\phi \left\|X_{i_1.} - X_{i_2}\right\|_2^2\right)$ where $\iota^m_{i_1,i_2}$ is 1 if observation i_2 is among i_1 's m nearest neighbors or vice verse, and 0 otherwise.

Sparse Convex Clustering

Write the data matrix X in feature-level as column vector $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$, where $\mathbf{x}_j = \left(X_{1j}, \cdots, X_{nj}\right)^{\mathrm{T}}, j = 1, \ldots, p$ and denote $\mathbf{A}$ in feature level as column vector $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_p)$. Simple algebra implies that (1) can be reformulated as

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times p}} \frac{1}{2}\sum_{j=1}^{p}\left\|\mathbf{x}_j - \mathbf{a}_j\right\|_2^2 + \gamma\sum_{l\in\mathscr{E}} w_l \left\|A_{i_1\cdot} - A_{i_2\cdot}\right\|_q \tag{2}$$

where $\mathscr{E} = \{l = (i_1, i_2) : 1 \leqslant i_1 < i_2 \leqslant n\}$.

- Without loss of generality, we assume the feature vectors are centered, i.e., $\sum_{i=1}^{n} X_{ij} = 0$ for each $j = 1, \ldots, p$.
- When $\hat{\mathbf{a}}_j$ are identical, when the corresponding feature j is not informative for clustering, i.e., $\|\hat{\mathbf{a}}_j\|_2^2 = 0$.

Sparse Convex Clustering

Sparse convex clustering solves

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times p}} \frac{1}{2}\sum_{j=1}^{p}\left\|\mathbf{x}_j-\mathbf{a}_j\right\|_2^2+\gamma_1\sum_{l\in\mathscr{E}} w_l\left\|A_{i_1}\cdot-A_{i_2}\cdot\right\|_q+\gamma_2\sum_{j=1}^{p}u_j\left\|\mathbf{a}_j\right\|_2 \tag{3}$$

where tuning parameter γ_1 controls the cluster size and tuning parameter γ_2 controls the number of informative features.

- In the group-lasso penalty, the weight u_j plays an important role to adaptively penalize the features.

Algorithm

Two optimization approaches similar to [Chi and Lange, 2015].

- Sparse alternating direction method of multipliers (S-ADMM).
- Sparse alternating minimization algorithm (S-AMA).

Equivalent Form

This is equivalent to minimize the augmented Lagrangian function,

$$L_\nu(\mathbf{A}, \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{2}\sum_{j=1}^{p}\left\|\mathbf{x}_j - \mathbf{a}_j\right\|_2^2 + \gamma_1 \sum_{l\in\mathscr{E}} w_l \left\|\mathbf{v}_l\right\|_q + \gamma_2 \sum_{j=1}^{p} u_i \left\|\mathbf{a}_j\right\|_2$$
$$+ \sum_{l\in\mathscr{E}} \left\langle \lambda_l, \mathbf{v}_l - A_{i_1} \cdot + A_{i_2} \right\rangle + \frac{\nu}{2}\sum_{l\in\mathscr{E}} \left\|\mathbf{v}_l - A_{i_1} \cdot + A_{i_2}\right\|_2^2$$

where ν is a small constant, $\mathbf{V} = \left(\mathbf{v}_1, \ldots, \mathbf{v}_{|\mathscr{E}|}\right)$, and $\mathbf{\Lambda} = \left(\lambda_1, \ldots, \lambda_{|\mathscr{E}|}\right)$.

S-ADMM

S-ADMM solves

$$\mathbf{A}^{m+1} = \underset{\mathbf{A}}{\operatorname{argmin}} L_\nu \left(\mathbf{A}, \mathbf{V}^m, \mathbf{\Lambda}^m\right),$$
$$\mathbf{V}^{m+1} = \underset{\mathbf{V}}{\operatorname{argmin}} L_\nu \left(\mathbf{A}^{m+1}, \mathbf{V}, \mathbf{\Lambda}^m\right),$$
$$\boldsymbol{\lambda}_l^{m+1} = \boldsymbol{\lambda}_l^m + \nu\left(\mathbf{v}_l^{m+1} - A_{i_1\cdot}^{m+1} + A_{i_2\cdot}^{m+1}\right), l \in \mathscr{E}.$$

Step 1 : Update A Denote $\widetilde{\mathbf{v}}_l = \mathbf{v}_l + \frac{1}{\nu}\boldsymbol{\lambda}_l$. Updating A is equivalent to minimizing

$$f(\mathbf{A}) = \frac{1}{2}\sum_{j=1}^{p}\left\|\mathbf{x}_j - \mathbf{a}_j\right\|_2^2 + \frac{\nu}{2}\sum_{l\in\mathscr{E}}\left\|\widetilde{\mathbf{v}}_l - A_{i_1\cdot} + A_{i_2\cdot}\right\|_2^2 + \gamma_2\sum_{j=1}^{p} u_j\left\|\mathbf{a}_j\right\|_2 \quad (4)$$

This optimization problem is challenging because the objective function involves both rows and columns of the matrix A.

Lemma 1

Let $\mathbf{I}_n$ be an $n \times n$ identity matrix, $\mathbf{1}_n \in \mathbb{R}^n$ be a vector with each component being 1 , and $\mathbf{e}_i$ be a vector with each component being 0 but its i-th component being 1. Define $\mathbf{N}^{-1} = (1+n\nu)^{-1/2}\left[\mathbf{I}_n + n^{-1}(\sqrt{1+n\nu}-1)\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\right]$ and denote $\mathbf{y}_j = \mathbf{N}^{-1}\left[\mathbf{x}_j + \ \nu\sum_{l\in\mathscr{E}}\tilde{v}_{jl}\left(\mathbf{e}_{i_1} - \mathbf{e}_{i_2}\right)\right]$ with $\widetilde{v}_{jl}$ the j-th element of $\tilde{\mathbf{v}}_l$. Then, minimizing (4) is equivalent to

$$\min_{\mathbf{a}_j} \frac{1}{2}\left\|\mathbf{y}_j - \mathbf{N}\mathbf{a}_j\right\|_2^2 + \gamma_2 u_j \left\|\mathbf{a}_j\right\|_2, \text{ for each } j = 1,\ldots,p$$

remark: Based on this property, we are able to solve the minimization of $f(A)$ by p separate sub-optimization problems.

S-ADMM

Step 2 : Update V For any $\sigma > 0$ and norm $\Omega(\cdot)$, we define a proximal map,

$$\text{prox}_{\sigma\Omega}(\mathbf{u}) = \underset{\mathbf{v}}{\text{argmin}} \left[\sigma\Omega(\mathbf{v}) + \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 \right]$$

In S-ADMM, $\Omega(\cdot)$ is a q-norm $\|\cdot\|_q$ with $q = 1, 2$, or ∞, and $\sigma = \gamma_1 w_l / \nu$. Because vectors $\mathbf{v}_l$ are separable, they can be solved via proximal maps, that is

$$\begin{aligned} \mathbf{v}_l &= \underset{\mathbf{v}_l}{\text{argmin}} \frac{1}{2} \left\| \mathbf{v}_l - \left(A_{i_1 \cdot} - A_{i_2 \cdot} - \nu^{-1} \boldsymbol{\lambda}_l \right) \right\|_2^2 + \frac{\gamma_1 w_l}{\nu} \|\mathbf{v}_l\|_q \\ &= \text{prox}_{\sigma_l \|\cdot\|_q} \left(A_{i_1 \cdot} - A_{i_2 \cdot} - \nu^{-1} \boldsymbol{\lambda}_l \right) \end{aligned}$$

Step 2 : Update $\boldsymbol{\Lambda}$ $\boldsymbol{\lambda}_l$ can be updated by $\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_l + \nu \left(\mathbf{v}_l - A_{i_1 \cdot} + A_{i_2 \cdot} \right)$.

S-ADMM

1. Initialize $\mathbf{V}^0$ and $\boldsymbol{\Lambda}^0$. For $m = 1, 2, \ldots$
2. For $j = 1, \ldots, p$, do

$$\widetilde{\mathbf{v}}_l^{m-1} = \mathbf{v}_l^{m-1} + \frac{1}{\nu}\boldsymbol{\lambda}_l^{m-1}, l \in \mathscr{E}$$

$$\mathbf{y}_j^{m-1} = \mathbf{N}^{-1}\left(\mathbf{x}_j + \nu \sum_{l \in \mathscr{E}} \widetilde{v}_{lj}^{m-1}\left(\mathbf{e}_{i_1} - \mathbf{e}_{i_2}\right)\right)$$

$$\mathbf{a}_j^m = \underset{\mathbf{a}_j}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{y}_j^{m-1} - \mathbf{N}\mathbf{a}_j\right\|_2^2 + \gamma_2 u_j \left\|\mathbf{a}_j\right\|_2$$

$$\mathbf{a}_j^m = \mathbf{a}_j^m - \overline{\mathbf{a}}_j^m \mathbf{1}_n, \text{ where } \overline{\mathbf{a}}_j^m = \mathbf{1}_n^T \mathbf{a}_j^m / n$$

3. For $l \in \mathscr{E}$, do

$$\mathbf{v}_l^m = \operatorname{prox}_{\sigma_l \|\cdot\|_q}\left(A_{i_1\cdot}^m - A_{i_2\cdot}^m - \nu^{-1}\boldsymbol{\lambda}_l^{m-1}\right)$$

4. For $l \in \mathscr{E}$, do

$$\boldsymbol{\lambda}_l^m = \boldsymbol{\lambda}_l^{m-1} + \nu\left(\mathbf{v}_l^m - A_{i_1\cdot}^m + A_{i_2\cdot}^m\right)$$

5. Repeat Steps 2-4 until convergence.

S-AMA

S-AMA aims to increase the computational efficiency of S-ADMM.

- S-AMA solves A by treating $v = 0$, i.e., $\mathbf{A}^{m+1} = \text{argmin}_{\mathbf{A}} L_0(\mathbf{A}, \mathbf{V}^m, \mathbf{\Lambda}^m)$. When $v = 0$, we have $N = \mathbf{I_n}$ and $y_j = x_j$. According to Lemma 1 , updating A requires to solve p group-lasso problems:

$$\min_{\mathbf{a}_j} \frac{1}{2} \left\| \mathbf{x}_j - \mathbf{a}_j \right\|_2^2 + \gamma_2 u_j \left\| \mathbf{a}_j \right\|_2, j = 1, \ldots, p \tag{5}$$

  By Karush-Kuhn-Tucker (KKT) conditions of the group lasso problem [Yuan and Lin, 2006], the solution to (5) has a closed form as

$$\widehat{\mathbf{a}}_j = \left(1 - \frac{\gamma_2 u_j}{\left\| \mathbf{z}_j \right\|_2} \right)_+ \mathbf{z}_j$$

  where $\mathbf{z}_j = \mathbf{x}_j + \sum_{l \in \mathscr{E}} \lambda_{jl} (\mathbf{e}_{i_1} - \mathbf{e}_{i_2})$ and $(z)_+ = \max\{0, z\}$.

- S-AMA does not need to update V.

S-AMA

1. Initialize $\boldsymbol{\Lambda}^0$. For $m = 1, 2, \ldots$
2. For $j = 1, \ldots, p$, do
$$\mathbf{z}_j^m = \mathbf{x}_j + \sum_{l \in \mathscr{E}} \lambda_{lj}^{m-1} (\mathbf{e}_{i_1} - \mathbf{e}_{i_2})$$
$$\mathbf{a}_j^m = \left(1 - \frac{\gamma_2 u_i}{\left\|\mathbf{z}_i^m\right\|_2}\right)_+ \mathbf{z}_j^m$$
$$\mathbf{a}_j^m = \mathbf{a}_j^m - \bar{\mathbf{a}}_j^m \mathbf{1}_n, \text{ where } \bar{\mathbf{a}}_j^m = \mathbf{1}_n^T \mathbf{a}_j^m / n$$
3. For $l \in \mathscr{E}$, do
$$\boldsymbol{\lambda}_l^m = P_{C_l}\left[\boldsymbol{\lambda}_l^{m-1} - \nu\left(A_{i_1}^m - A_{i_2}^m\right)\right]$$
where $C_l = \{\boldsymbol{\lambda}_l : \|\boldsymbol{\lambda}_l\|_\dagger \leqslant \gamma_1 w_l\}$
4. Repeat Steps 2-3 until convergence.

remark: $P_{C_l}(\cdot)$ denotes projection onto C_l, and $\|\cdot\|_\dagger$ denotes the dual norm.

Some Notations

- Assume $\mathbf{x} = \mathbf{a}_0 + \varepsilon$, where $\varepsilon \in \mathbb{R}^{np}$ is a vector of independent sub-Gaussian noise terms with mean zero and variance σ^2, and $\mathbf{a}_0 = \left(\mathbf{a}_{01}^{\mathrm{T}}, \ldots, \mathbf{a}_{0p}^{\mathrm{T}}\right)^{\mathrm{T}}$ is a np-dimensional mean vector.
- Assume that only the first $p_0 < p$ features are informative, i.e., $\left\|\mathbf{a}_{0j}\right\|_2 \neq 0$ for $j \leqslant p_0$ and $\left\|\mathbf{a}_{0j}\right\|_2 = 0$ for $j > p_0$. The informative feature set is denoted as $A = \{1, \ldots, p_0\}$ and the noninformative feature set is $A^c = \{p_0 + 1, \ldots, p\}$. For simplicity, we consider the case with $w_l = 1$.
- Sparse convex clustering in (3) can be reformulated as the following problem:

$$\widehat{\mathbf{a}} = \underset{\mathbf{a} \in \mathbb{R}^{np}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{x} - \mathbf{a}\|_2^2 + \gamma_1 \sum_{l \in \mathscr{E}} \left\|\mathbf{C}_l \mathbf{a}\right\|_q + \gamma_2 \sum_{j=1}^{p} u_j \left\|\mathbf{a}_j\right\|_2 \tag{6}$$

where $\mathbf{C}_l = \mathbf{I}_p \otimes (\mathbf{e}_{i_1} - \mathbf{e}_{i_2})^{\mathrm{T}}$ and hence $\mathbf{C}_l \mathbf{a} = A_{i_1} - A_{i_2}$.. Define $\mathbf{C} = \left(\mathbf{C}_1^{\mathrm{T}}, \ldots, \mathbf{C}_{|\mathscr{E}|}^{\mathrm{T}}\right)^{\mathrm{T}}$ and denote $\mathbf{u} = (u_1, \ldots, u_p)^{\mathrm{T}}$.

Prediction error for $q = 1$

Theorem 1

Let $\hat{\mathbf{a}}$ *be the estimate of* (6) *with* $q = 1$. *If* $\gamma_1 > 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n}}$, *then*

$$\frac{1-\gamma_2}{2np}\|\widehat{\mathbf{a}}-\mathbf{a}_0\|_2^2 \leqslant \frac{3\gamma_1}{2np}\|\mathbf{C}\mathbf{a}_0\|_1 + \frac{\gamma_2\|\mathbf{u}\|_2^2}{2np} + \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2 p}}\right] + \frac{1}{np}$$

holds with probability at least $1 - c_3$, *where*

$$c_3 = \frac{2}{p\cdot\binom{n}{2}} + \exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\} + 2\exp\left(-\frac{np}{(2\sigma^2\gamma_2^2\|\mathbf{u}\|_1^2)}\right)$$

for some positive constants c_1 *and* c_2 *defined in Lemma* S.1

Prediction error for $q = 2$

Theorem 2

Let $\widehat{\mathbf{a}}$ *be the estimate of* (6) *with* $q = 2$. *If* $\gamma_1 > 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n}}$, *then*

$$\frac{1-\gamma_2}{2np}\|\widehat{\mathbf{a}}-\mathbf{a}_0\|_2^2 \leqslant \frac{3\gamma_1}{2np}\sum_{l\in\mathscr{E}}\|\mathbf{C}_l\mathbf{a}_0\|_2 + \frac{\gamma_2\|\mathbf{u}\|_2^2}{2np} + \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right] + \frac{1}{np}$$

holds with probability at least $1 - c_3$, *where* c_3 *is defined in Theorem 1 .*

Theorem 3

Theorem 3

If $\gamma_1 > 4\sigma\sqrt{\log\left(p\cdot\left(\begin{array}{c} n \\ 2 \end{array}\right)\right)/n}, \gamma_1 \|\mathrm{Ca}_0\|_1/(2np) = o(1), \gamma_2 \to 0$ *and* $\gamma_2\|\mathbf{u}\|_1^2/(np) \to 0$ *as* $n, p \to \infty$*, then* $P\left(\left\|\widehat{\mathbf{a}}_j\right\|_2 = 0\right) \to 1$ *for any* $j \in A^c$*, with the solution* $\hat{\mathbf{a}}$ *to* (6) *with either* $q = 1$ *or* $q = 2$*.*

Remark: $\gamma_2\|\mathbf{u}\|_1^2/(np) \to 0$ generally implies that the adaptive weights cannot be too large. For example, uniform weights satisfy this condition.

Selection of weights

- Following [Chi and Lange, 2015], we choose weights by incorporating the m-nearest-neighbors methods with Gaussian kernel. In specific, the weight between the pair (i_1, i_2) is

$$w_{i_1,i_2} = \iota^m_{i_1,i_2} \exp\left(-\phi \left\|X_{i_1}\cdot - X_{i_2}\cdot\right\|_2^2\right),$$

  where $\iota^m_{i_1,i_2}$ equals 1 if observation i_2 is among observation i_1's m nearest neighbors. In application, we set $m = 5$ and $\phi = 0.5$.

- μ_j can be chosen as $1/\|\hat{\mathbf{a}}_j^0\|_2$, where $\|\hat{\mathbf{a}}_j^0\|_2$ is the estimate of $\mathbf{a}_j$ in (3) with $\gamma_2 = 0$.

Selection of Tuning Parameters

- γ_1 controls the number of estimated clusters.
- γ_2 controls the number of selected informative features.
- Use stability selection to tune both γ_1 and γ_2:
  - For any given γ_1 and γ_2, based on two sets of bootstrapped samples, two clustering results can be produced by (3).
  - Compute the stability measurement [Fang and Wang, 2012] to measure the agreement between the two clustering result.
  - Repeat this procedure 50 times and then compute the averaged stability selection method.
- To speed up tunning process, stability path can be computed over of a coarse grid of γ_1 and a fine grid of γ_2.

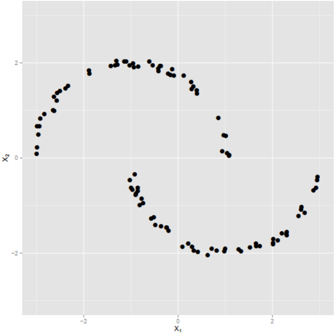
Case One

- Sample size $n = 60$ with the number of clusters either $K = 2$ or 4.
- The number of features either $p = 150$ or 500.
- For each $i = 1, \ldots, n$, cluster label Z_i is uniformly sampled from $\{1, \ldots, K\}$.
- The first 20 informative features are generated from $MVN_p(\mu_K(Z_i), \mathbf{I}_{20})$, where $\mu_K(Z_i)$ is defined as:
  - If $K = 2, \boldsymbol{\mu}_2(Z_i) = \mu \mathbf{1}_{20} I(Z_i = 1) - \mu \mathbf{1}_{20} I(Z_i = 2)$.
  - If $K = 4, \boldsymbol{\mu}_4(Z_i) = \left(\mu \mathbf{1}_{10}^{\mathrm{T}}, -\mu \mathbf{1}_{10}^{\mathrm{T}}\right)^{\mathrm{T}} I(Z_i = 1) + \left(-\mu \mathbf{1}_{10}^{\mathrm{T}}, -\mu \mathbf{1}_{10}^{\mathrm{T}}\right)^{\mathrm{T}} I(Z_i = 2) + \left(-\mu \mathbf{1}_{10}^{\mathrm{T}}, \mu \mathbf{1}_{10}^{\mathrm{T}}\right)^{\mathrm{T}} I(Z_i = 3) + \left(\mu \mathbf{1}_{10}^{\mathrm{T}}, \mu \mathbf{1}_{10}^{\mathrm{T}}\right)^{\mathrm{T}} I(Z_i = 4)$.
- The rest $p - 20$ noise features are generated from $N(0, 1)$.

Remark : μ controls the distance between cluster centers. A large μ indicates that clusters are well-separated, whereas a small μ indicates that clusters are overlapped.

Case Two

- $n = 100$, $K = 2$ and $p = 40$, where the first two features are informative, and the rest 38 noisy features are generated from $N(0, 0.5)$.
- The plot of the first two features for one example of two interlocking half moons.



Five Settings for Simulation

- Spherical settings
  - Setting 1 : $K = 2, n = 60, p = 150, \mu = 0.6$.
  - Setting 2 : $K = 2, n = 60, p = 500, \mu = 0.7$.
  - Setting 3 : $K = 4, n = 60, p = 150, \mu = 0.9$.
  - Setting 4 : $K = 4, n = 60, p = 500, \mu = 1.2$.
- Non-spherical settings
  - Setting 5 : $K = 2, n = 100, p = 40$.

Evaluation Criteria

- **RAND index** : RAND index ranges between 0 and 1, and a higher value indicates better performance.
- **False Negative Ratio (FNR).**
- **False Positive Ratio (FPR).**

Due to the high computational burden for S-ADMM in high-dimensional settings, S-ADMM is not evaluated for $p = 500$. Additionally, we run 200 repetitions for each setting.

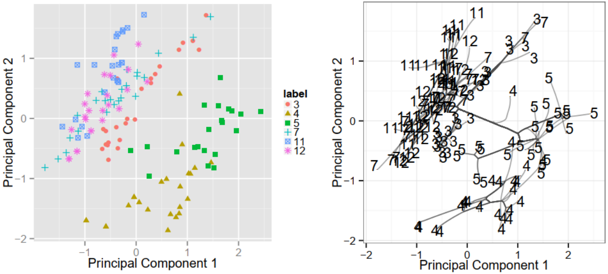
Results

		RAND		FNR		FPR	
Setting 1	k-means	0.95	0.06	0.00	0.00	1.00	0.00
	ADMM	0.53	0.39	0.00	0.00	1.00	0.00
	AMA	0.66	0.40	0.00	0.00	1.00	0.00
	S-ADMM	0.82	0.24	0.04	0.05	0.25	0.16
	S-AMA	**0.96**	0.06	0.03	0.07	0.30	0.21
Setting 2	k-means	0.95	0.11	0.00	0.00	1.00	0.00
	ADMM	0.14	0.20	0.00	0.00	1.00	0.00
	AMA	0.08	0.21	0.00	0.00	1.00	0.00
	S-AMA	**0.97**	0.07	0.07	0.09	0.11	0.10
Setting 3	k-means	0.83	0.15	0.00	0.00	1.00	0.00
	ADMM	0.56	0.22	0.00	0.00	1.00	0.00
	AMA	0.47	0.21	0.00	0.00	1.00	0.00
	S-ADMM	0.82	0.14	0.04	0.06	0.25	0.24
	S-AMA	**0.84**	0.13	0.02	0.04	0.11	0.18
Setting 4	k-means	0.89	0.14	0.00	0.00	1.00	0.00
	ADMM	0.31	0.23	0.00	0.00	1.00	0.00
	AMA	0.31	0.20	0.00	0.00	1.00	0.00
	S-AMA	**0.94**	0.09	0.01	0.02	0.01	0.03
Setting 5	k-means	0.51	0.07	0.00	0.00	1.00	0.00
	ADMM	0.54	0.08	0.00	0.00	1.00	0.00
	AMA	0.53	0.09	0.00	0.00	1.00	0.00
	S-AMA	**0.57**	0.07	0.00	0.00	0.34	0.27
	SPECC	0.52	0.08	0.00	0.00	1.00	0.00

- Convex clustering does not perform well when feature dimension is high.
- Sparse convex clustering selects informative features with great clustering accuracy.

Application : hand movement clustering

- Dataset contains 15 classes with each class referring to a hand movement type.
- Each class contains 24 observations and each observation has 90 features.



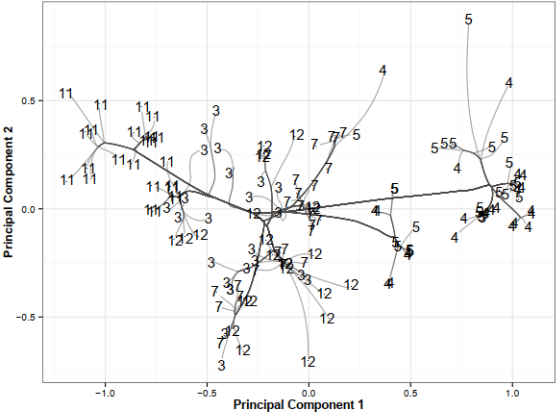
Convex clustering is only able to distinguish clusters 4 and 5 and treat the rest clusters as one class.

Results

Algorithm	# of clusters	# of features	RAND index
k-means	2	90	0.06
AMA	3	90	0.31
S-AMA	3	13	0.45

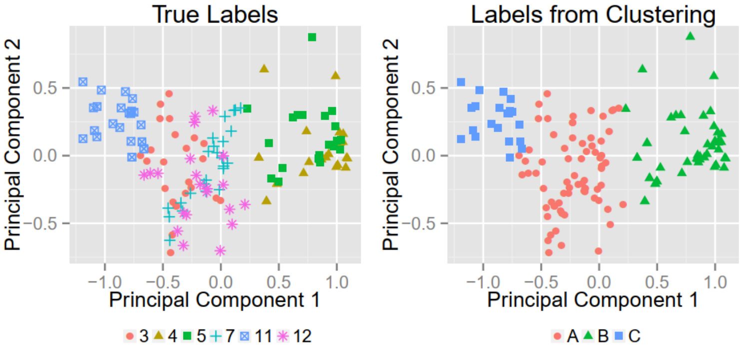
- Both convex clustering (AMA) and sparse convex clustering (S-AMA) perform better than k-means, which indicates that the performance of convex clustering or sparse convex clustering is less sensitive to the assumption of spherical clustering centers.
- By using only 13 informative features, our S-AMA is able to improve the clustering accuracy of convex clustering (AMA) by 45%. This indicates the importance of variable selection in high-dimensional clustering.

Clustering path of S-AMA



As tuning parameter γ_1 increases, the clustering path of S-AMA tends to merge clusters 3, 7 and 12 into one big cluster, merge cluster 4 and 5 into another big cluster, and identify cluster 11 as the third cluster.

Results


True Labels
Principal Component 2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Principal Component 1
3 4 5 7 11 12
Labels from Clustering
Principal Component 2
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
Principal Component 1
A B C

Conclusion and Future Work

- An extension of convex clustering, sparse convex clustering, is proposed to simultaneously cluster observations and conduct feature selection.
- The numerical results show that S-AMA is computationally faster and delivers better performance than S-ADMM.
- The numerical results show that the selection of tuning parameters in sparse convex clustering is important and the tuning method based on clustering stability performs well.
- Future work:
    - Extend convex bi-clustering [Chi et al., 2017] to sparse bi-clustering.
    - Use group L_0 penalty [Zhang et al., 2021] to replace the group lasso penalty applying on the feature level.

Any questions or comments?

References I

Chi, E. C., Allen, G. I., and Baraniuk, R. G. (2017).
Convex biclustering.
Biometrics, 73(1):10–19.

Chi, E. C. and Lange, K. (2015).
Splitting methods for convex clustering.
Journal of Computational and Graphical Statistics, 24(4):994–1013.

Fang, Y. and Wang, J. (2012).
Selection of the number of clusters via the bootstrap method.
Computational Statistics & Data Analysis, 56(3):468–477.

Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. (2011).
Clusterpath an algorithm for clustering using convex fusion penalties.
In *28th international conference on machine learning*, page 1.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011).
Just relax and come clustering!: A convexification of k-means clustering.
Linköping University Electronic Press.

References II

Tan, K. M. and Witten, D. (2015).
Statistical properties of convex clustering.
Electronic journal of statistics, 9(2):2324.

Yuan, M. and Lin, Y. (2006).
Model selection and estimation in regression with grouped variables.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

Zhang, Y., Zhu, J., Zhu, J., and Wang, X. (2021).
Certifiably polynomial algorithm for best group subset selection.
arXiv preprint arXiv:2104.12576.