Sample title

Anonymous

Overleaf

2024

SDP relaxation of K-means

From

$$\text{minimize}_{G_1,\ldots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i,j \in G_k} \|\boldsymbol{X}_i - \boldsymbol{X}_j\|_2^2, \tag{1}$$

through several steps, relax the problem into

$$\begin{aligned} &\text{minimize}_{\boldsymbol{Z} \in \mathcal{S}_+^n} \ \langle A, \boldsymbol{Z} \rangle \\ &\qquad s.t.\ \text{tr}(\boldsymbol{Z}) = K, \boldsymbol{Z}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{Z} \geq 0, \end{aligned} \tag{2}$$

where $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{X}^\top$ (Chen and Yang, 2021).

Statistical optimality

Chen and Yang (2021) uses the following model to investigate the statistical optimality:

- Balanced Gaussian mixture model with K components of equal size $|G_1^*| = \ldots |G_K^*|$:

$$\boldsymbol{X}_i = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p) \tag{3}$$

- Minimal separation: $\Theta^2_{\min} = \min_{1 \le k \ne l \le K} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2^2$

Statistical optimality

- Suppose $K \leq \log n/(\log\log n)$. Then
  - [minimax upper bound, through an algorithm] If $\Theta^2_{\min} \geq (1+\alpha)\bar{\Theta}^2$, where

$$\Theta^2 = 4\sigma^2\left(1+\sqrt{1+\frac{Kp}{n\log n}\log n}\right)$$

    the **SDP relaxation** (2) yields the true cluster partitions, with probability tending to one. **This is a sharp threshold; not up to a constant**.
- [minimax lower bound, information-theoretic] If $\Theta^2_{\min} \leq (1-\alpha)\bar{\Theta}^2$, then probability of exact recovery of any estimator vanishes to zero.

Our goal

For simplicity, let $K = 2$ and assume symmetric mean: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_0$

1. Assume hard sparsity: only s entries of $\boldsymbol{\mu}_0$ are non-zero, and their positions are unknown
2. Assume sparse precision matrix

Then we aim to derive an analog of the theorems of Chen and Yang (2021):

1. an algorithm-specific bound
2. an information theoretic bound

that match, **up to a constant**.
If things work well, we can try a sharp threshold next.

A conjecture on the bound

Dependence on p inside the square root improves from p to $s(\log p)^\gamma$:

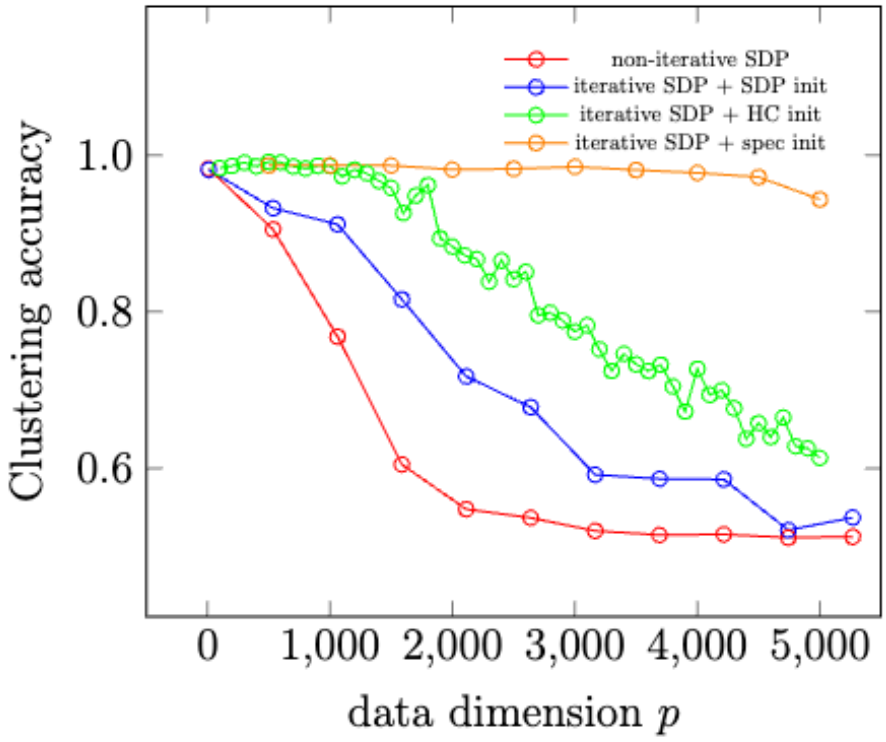
$$4\left(1+\sqrt{1+\frac{Ks(\log p)^\gamma}{n\log n}}\right)\log n, \tag{4}$$

i.e. dependence on p becomes a log factor.

1.1. Algorithm-specific bound: Adapting to sparsity through iterative thresholding

1. The signal (differences in non-zero entries) would be larger than the noise (differences of zero-mean Gaussian noises)
2. Thus given current clustering result, we can estimate the support by entrywise thresholding. threshold $= \sqrt{2 \log p}$.
3. Using this estimated support, we do the clustering again, and using this clustering result, we estimate the support again. This iteration goes on up to some reasonable number of iterations.
4. If we have initialization e.g. by spectral clustering that is not too bad, estimation accuracy of the support will gradually improve over iterations

1. Algorithm-specific bound: Adapting to sparsity through iterative thresholding



(c) $\Delta = 5$

$2\|\boldsymbol{\mu}_0\|_2^2 = 5, n = 200, s = 10, p = 100 - 5000$, number of iterations fixed at 10.

1.2. Algorithm-specific bound: Generalization from isotropic to sparse covariance matrix

Assume that $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ is sparse

- Change the data sparsity assumption: $\boldsymbol{\Omega}\boldsymbol{\mu}_0$ is s-sparse, instead of $\boldsymbol{\mu}_0$. The data-generating process is the same.
- Use the same iterative algorithm, but now the estimation of Ω kicks in and deteriorates the performance
- After some explorations, found out that the full estimation of $\boldsymbol{\Omega}$ is unnecessary, and what we really need to estimate are
  1. the diagonals of $\boldsymbol{\Omega}$
  2. the transformed data vectors $\boldsymbol{\Omega}\boldsymbol{X}_1, \ldots, \boldsymbol{\Omega}\boldsymbol{X}_n$,

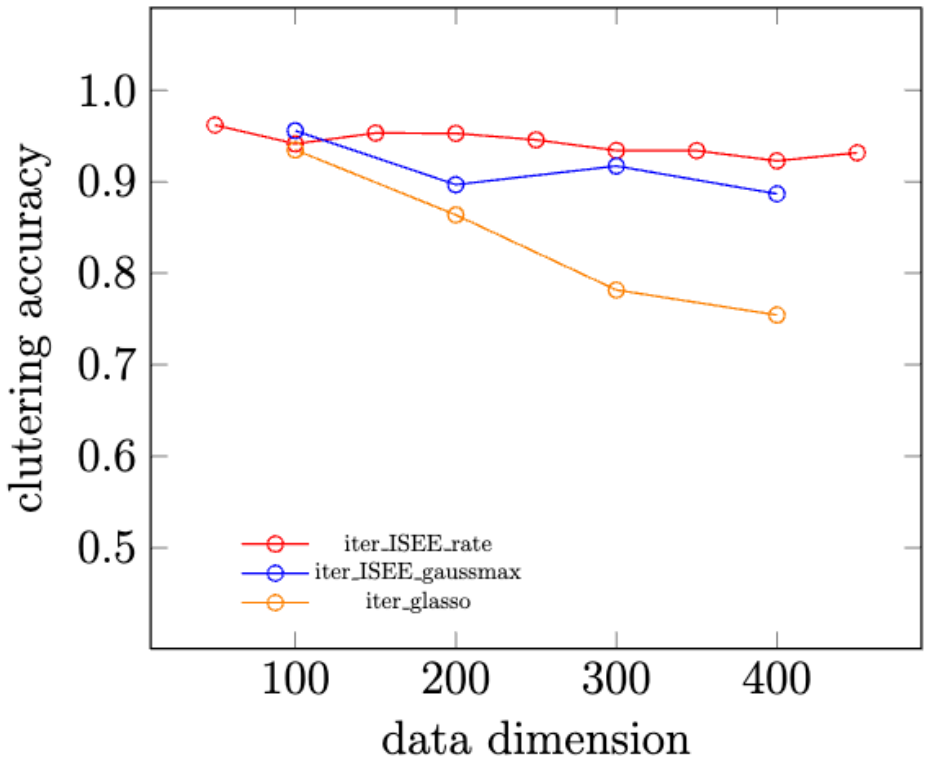
  which can be effectively estimated, in parallel, with ISEE algorithm Fan and Lv (2016) (runs LASSO for small blocks of $\boldsymbol{\Omega}\boldsymbol{X}_i$, requires Gaussian noise assumption)

1.2. Algorithm-specific bound: Generalization from isotropic to sparse covariance matrix

- The thresholding now exploits the convergence rate result of ISEE (Fan and Lv, 2016).
- Due to the regression nature, ISEE estimates the mean and the residual separately.
- The formal method compares mean+noise vs expected noise level ($\sqrt{2 \log p}$.
- If we use ISEE we can just compare mean vs noise.

Using ISEE and thresholding through the convergence rate is shows the best performance so far.

1.2. Algorithm-specific bound: Generalization from isotropic to sparse covariance matrix

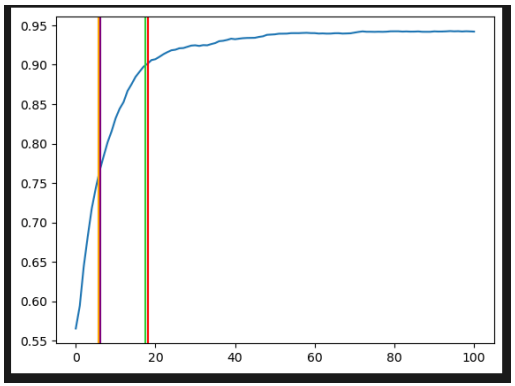


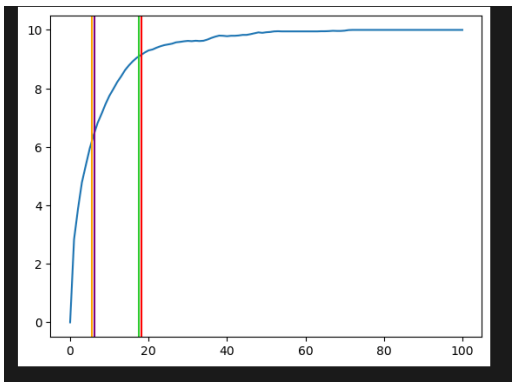
$\mathbf{\Omega} = AR(1), \|\mathbf{\Sigma}2\boldsymbol{\mu}_0\|_2^2 = 4, n = 500, s = 10, p = 100 - 400,$ off-diagonal of $\mathbf{\Omega} = 0.45$, number of iterations fixed at 10.

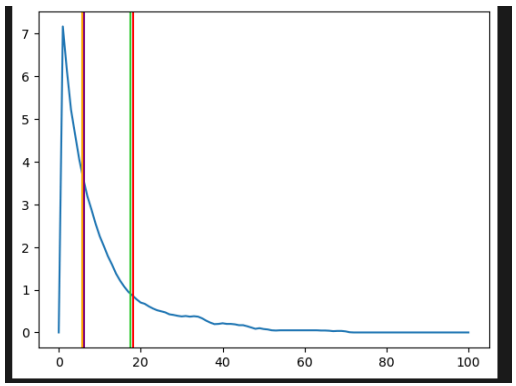
1.3. Algorithm-specific bound: when to stop the iteration

1. percent change criterion: SDP objective or original K-means objective
2. Early stopping with patience parameter: keep track of the minimum objective value so far. If the minimum does not change for {patience} steps, stop the iteration and use the result of {current iteration - patience}

1.3. Algorithm-specific bound: when to stop the iteration







$\mathbf{\Omega} = AR(1), \|\mathbf{\Sigma} 2\boldsymbol{\mu}_0\|_2^2 = 4, n = 500, s = 10, p = 400$, off-diagonal of $\mathbf{\Omega} = 0.45$,

1. purple: early stopping using the k-means objective
2. orange: early stopping using the SDP objective
3. limegreen: percent change using the k-means objective
4. red: percent change using the SDP objective

1.4. Algorithm-specific bound: deriving the bound

Proof techniques of Chen and Yang (2021): ... conditions on the dual variable λ (trace(Z)=K constraint) are met with high probability... requires degenerate U-process...

Chen, X. and Yang, Y. (2021). Cutoff for Exact Recovery of Gaussian Mixture Models. *IEEE transactions on information theory*, 67(6):4223–4238. Publisher: IEEE.

Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics*, 44(5):2098–2126. Publisher: Institute of Mathematical Statistics.