[Oral] Statistically Optimal K-means Clustering via Nonnegative Low-rank Semidefinite Programming

Yubo Zhuang[1], Xiaohui Chen[2], Yun Yang[1], Richard Y. Zhang[3]. [1]UIUC Statistics, [2]USC Mathematics, [3]UIUC ECE

yubo2@illinois.edu, xiaohuic@usc.edu, yy84@lllinois.edu, ryz@lllinois.edu

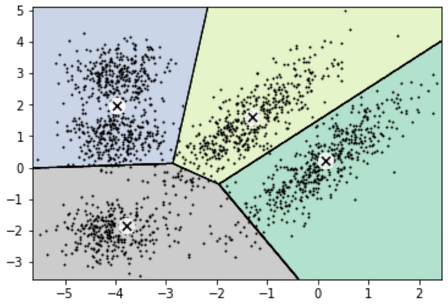
Summary & Highlight

- Clustering is a hard problem: computationally & statistically.
- Various approximations and relaxations: Lloyd, spectral, nonnegative matrix factorization (NMF), semidefinite programming (SDP).
- SDP achieves sharp information-theoretical threshold for exact recovery.
- **Goal: computational scalability and strong theoretical guarantee.**
- **This paper: an algorithm simultaneously achieving O(n) per iteration complexity + local linear convergence + same SDP recovery guarantee.**
- Future work: Partial recovery? Optimization landscape?

Clustering analysis: divide data $x_1, \dots, x_n \in \mathbb{R}^p$ into K groups $G_1^*, \dots, G_K^*$ based on their similarity.

K-means clustering (NP-hard):

$$\max_{G_1,\dots,G_K} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{i\in G_k} \sum_{j\in G_k} x_i^T x_j \quad \text{subject to} \quad \bigsqcup_{k=1}^{K} G_k = [n].$$

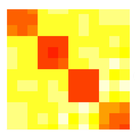


	Scalability	Stability	Optimality
NMF	+	-	-
Lloyd	+	-	-
Spectral	+	+	-
SDP	-	+	+
Ours	+	+	+

Relaxed formulations

SDP (Peng & Wei 2007; Giraud & Verzelen 2018)

$$\max_{Z\in\mathbb{R}^{n\times n}} \langle XX^T, Z\rangle \quad \text{s.t. } Z\geq 0,\ Z1_n = 1_n,\ \mathrm{tr}(Z) = K,\ Z\succeq 0.$$



$Z^* = U^*(U^*)^T$

Clustering membership matrix

Low-rank solution

NMF (He et al. 2011; Kuang et al. 2015)

$$\min_{U\in\mathbb{R}^{n\times r}} \|XX^T - UU^T\|_F \quad \text{s.t. } U\geq 0.$$

Exact recovery sharp information-threshold (Chen & Yang 2021)

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

Gaussian mixture model (GMM): $i \in G_k^* \Rightarrow x_i = \mu_k + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2 I_p)$.

Minimum centroid separation (statistical hardness): $\Theta_{\min} := \min_{1\leq j\neq k\leq K} \|\mu_j - \mu_k\|_2$

(-) $\Theta_{\min} \leq (1-\delta)\overline{\Theta}$ → No algorithm (regardless complexity) can exactly recover $G_1^*, \dots, G_K^*$

(+) $\Theta_{\min} \geq (1+\delta)\overline{\Theta}$ → Unique SDP solution that perfectly recovers $G_1^*, \dots, G_K^*$

Nonnegative low-rank (NLR) factorization formulation

$$\max_{U\in\mathbb{R}^{n\times r}} \left\{ \langle XX^T, UU^T\rangle :\ UU^T 1_n = 1_n,\ \|U\|_F^2 = K,\ U\geq 0 \right\}.$$

Algorithm design

$$\max_{U\in\mathbb{R}^{n\times r}} \left\{ \langle XX^T, UU^T\rangle :\ \underbrace{UU^T 1_n = 1_n}_{\text{dual}},\ \underbrace{\|U\|_F^2 = K, U\geq 0}_{\text{primal} =: \Omega} \right\}.$$

Key insight: primal constraints can be analytically solved by projection onto Ω.

Augmented Lagrangian method (ALM)

$$\mathscr{L}_\beta(U,y) = \langle L\cdot \mathrm{Id}_n - XX^T, UU^T\rangle + \langle y, UU^T 1_n - 1_n\rangle + \frac{\beta}{2}\|UU^T 1_n - 1_n\|_2^2.$$

iterate between:

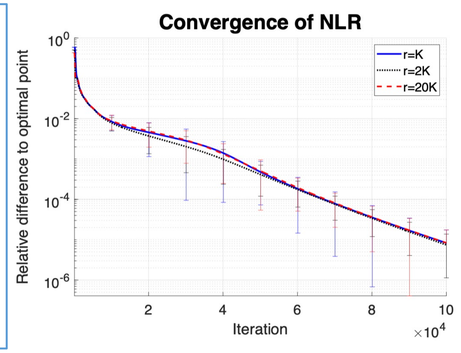
- Primal update: $U_{\text{new}} := U_{\text{new},y} = \arg\min_{U\in\Omega} \{f(U) := \mathscr{L}_\beta(U,y)\}$ via iterative projected gradient descent $U^{t+1} = \Pi_\Omega(U^t - \alpha\nabla f(U^t))$.
- Dual update (one-step): $y_{\text{new}} = y + \beta(U_{\text{new}}U_{\text{new}}^T 1_n - 1_n)$.

Theorem: local linear convergence

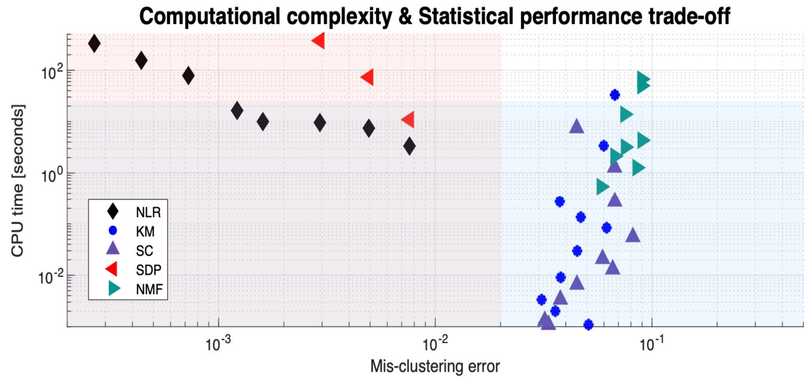
Initialize U^0 within an **O(1) neighborhood** of the optimal solution U^*. Under GMM, if $\Theta_{\min} \geq (1+\delta)\overline{\Theta}$, then for all y close to the optimal dual y^* and $t > t_0$, we have with high prob

$$\|U^{t+1} - U_y^*\|_F \leq \gamma\|U^t - U_y^*\|_F.$$

Here, $\gamma\in(0,1)$, U^t is the t-th iterate, and t_0 is a constant.



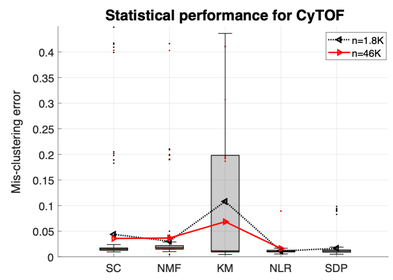
Overall time complexity of NLR: $O(nrK^6)$

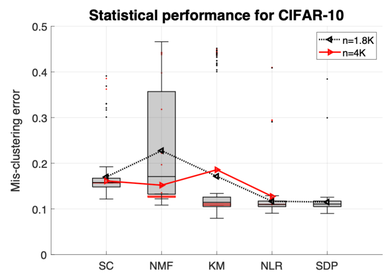


• Varying sample size n. • Centroid separation $O(\log n)$

Real-data performance

- Mass cytometry (CyTOF) dataset
- Sample size n = 1,800, 46,258.
- CIFAR-10 dataset (colored images of size $32\times 32\times 3$)
- Sample size n = 1,800, 4,000.

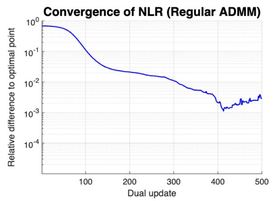




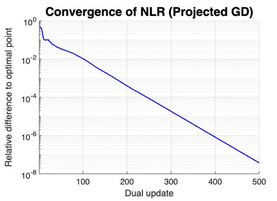
Technical highlight: two-phase convergence

Phase 1: PGD becomes block diagonal after O(K³) iterates.

Phase 2: PGD iterates remain block diagonal and attain linear convergence rate since ALM objective function is (locally) restricted strongly convex.



Lagrangian function contains all constraints.



NLR via primal-dual projected gradient descent.

arXiv: 2305.18436