# Point clustering with a convex, corrected K-means

Martin Royer

Univ. Paris-Saclay

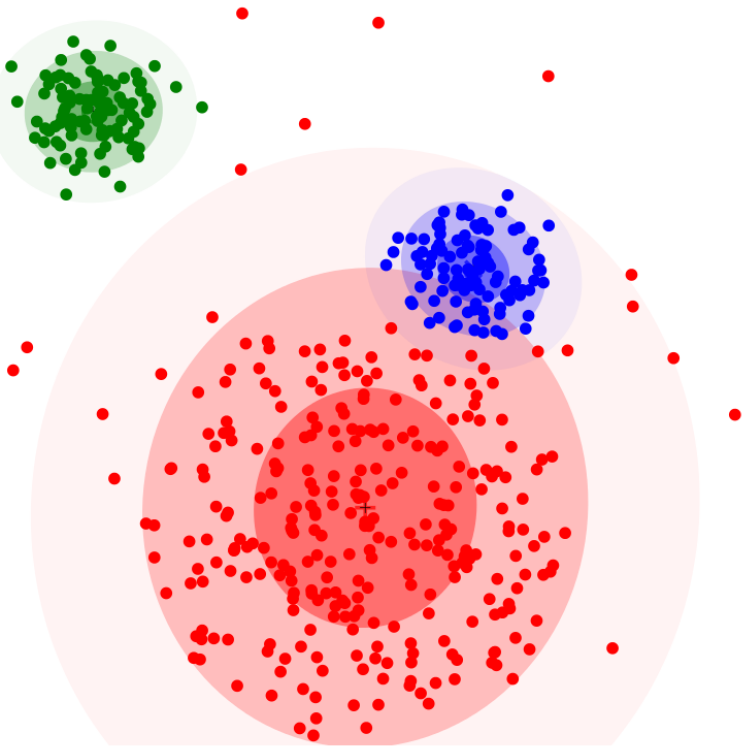
Département de Mathématiques d'Orsay

Let $X_1, \dots, X_n$ be $n$ observations in $\mathbb{R}^p$, $\mathcal{G} = \{G_k\}_{1 \leq k \leq K}$ a partition of interest.

## $(\alpha)$ - Latent model

Assume $\forall k, \forall a \in G_k$:

$$X_a = \mu_k + E_a \quad \text{with} \quad \mathbf{E}[X_a] = \mu_k \quad \text{and} \quad E_a \underset{ind}{\sim} \text{sub-}\mathcal{N}(0, \Sigma_a)$$



**Key quantities to keep in mind:**

- cluster separation $\Delta_{\mathcal{G}}(\mu) := \min_{k \neq l} |\mu_k - \mu_l|_2$
- peak noise $\sigma^2 := \max_{a \in [n]} |\Sigma_a|_{op}$

Solving for MLE on $(\alpha)$ with homoscedastic observations $\Leftrightarrow$ K-means

K-means objective writes:

$$\widehat{\mathcal{G}}_{\text{Kmeans}} \in \underset{\mathcal{G}=\{G_k\}_{1\leq k\leq K}}{\operatorname{argmin}} \ \text{Crit}(\mathcal{G}) := \sum_{k=1}^{K} \sum_{a\in G_k} \|X_a - \bar{X}_{G_k}\|^2 \tag{1}$$

$$= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b\in G_k^2} \|X_a - X_b\|^2 \tag{2}$$

$$= -\sum_{k=1}^{K} \sum_{a,b\in G_k^2} \frac{1}{|G_k|} \langle X_a, X_b \rangle + \sum_{a=1}^{n} \|X_a\|^2 \tag{3}$$

$$= -\langle B_{\mathcal{G}}, XX^T \rangle + \|X\|_F^2 \tag{4}$$

with $X := \begin{bmatrix} X_1^T \\ \cdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n\times p}$, and

$B_{\mathcal{G}} \in \mathbb{R}^{n\times n}$ characteristic of $\mathcal{G}$ s.t. $B_{ab} := \begin{cases} 1/|G_k| & \text{if } a,b \in G_k \\ 0 & \text{otherwise} \end{cases}$

# A pseudo-SDP, Peng and Wei [2007]

## Lemma

We have

$$\widehat{B}_{\text{Kmeans}} \in \underset{B \in \mathcal{D}}{\operatorname{argmax}} \langle XX^T, B \rangle \tag{5}$$

where $\mathcal{D} := \left\{ \text{sym. } B \in \mathbf{R}^{n \times n} : \begin{array}{l} \bullet\ B \geqslant 0 \\ \bullet\ B.1_n = 1_n \\ \bullet\ \mathrm{Tr}(B) = K \\ \bullet\ B^2 = B \end{array} \right.$

Is K-means "optimal"?

## Claim

Suppose $\Sigma_1 = ... = \Sigma_n$, then

$$\underset{B \in \mathcal{D}}{\operatorname{argmax}} \langle \mathbf{E}[XX^T], B \rangle = B_{\mathcal{G}} \tag{6}$$

# Convexifying K-means

K-means as an optimization over set of matrices

$$\mathcal{D} := \left\{ \text{symmetric } B \in \mathbf{R}^{n\times n} : \begin{array}{l} \bullet\ B \geqslant 0 \\ \bullet\ B.1_n = 1_n \\ \bullet\ \mathrm{Tr}(B) = K \\ \bullet\ B^2 = B \end{array} \right.$$

Replace it by an optimization over set of matrices

$$\mathcal{C} := \left\{ \text{symmetric } B \in \mathbf{R}^{n\times n} : \begin{array}{l} \bullet\ B \geqslant 0 \\ \bullet\ B.1_n = 1_n \\ \bullet\ \mathrm{Tr}(B) = K \\ \bullet\ (I \succcurlyeq) B \succcurlyeq 0 \end{array} \right.$$

## Claim

Suppose $\Sigma_1 = ... = \Sigma_n$, then we have

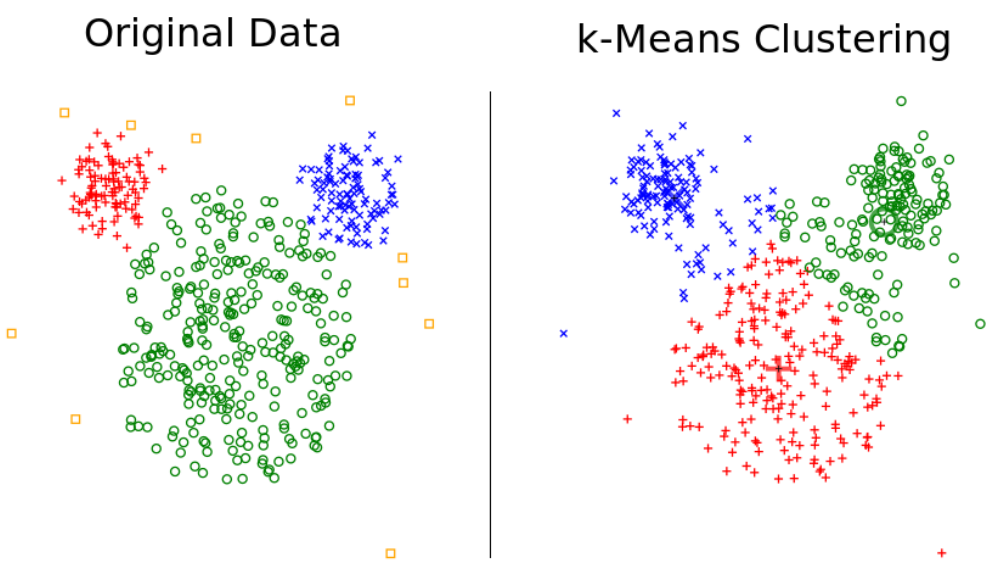
$$\underset{B\in\mathcal{D}}{\mathrm{argmax}} \langle \mathbf{E}[XX^T], B\rangle = \underset{B\in\mathcal{C}}{\mathrm{argmax}} \langle \mathbf{E}[XX^T], B\rangle = B_{\mathcal{G}}$$

# K-means is biased

- define membership matrix of a partition $\mathcal{G}$: $A_{ak} := \mathbf{1}\{a \in G_k\}$, $A \in \mathbb{R}^{n \times K}$.
- define also $\mu := \begin{bmatrix} \mu_1^T \\ \dots \\ \mu_K^T \end{bmatrix} \in \mathbb{R}^{K \times p}$, $E := \begin{bmatrix} E_1^T \\ \dots \\ E_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$.
- then according to $(\alpha)$ we have $X = A\mu + E$ and:

$$XX^T = A(\mu\mu^T)A^T + (A\mu E^T + E\mu^T A^T) + EE^T \tag{7}$$

$$\mathbf{E}[XX^T] = A(\mu\mu^T)A^T + \operatorname{diag}\left(\operatorname{Tr}(\operatorname{Cov}(E_a))\right) = A(\mu\mu^T)A^T + \Gamma \tag{8}$$




## Claim

For a given $\mathcal{G}, \mu$, there exist $\Gamma$ such that
$B_{\mathcal{G}} \notin \operatorname{argmax}_{B \in \mathcal{D}} \langle \mathbf{E}[XX^T], B \rangle$ and
$B_{\mathcal{G}} \notin \operatorname{argmax}_{B \in \mathcal{C}} \langle \mathbf{E}[XX^T], B \rangle$

Adapted from Bunea et al. [2016],

How do we estimate $\Gamma = \mathrm{diag}\left(\mathrm{Tr}(\mathrm{Cov}(E_a))\right)$?
$a \in G_k$, suppose we find neighbours $v_1, v_2$. Then $\mathrm{Tr}(\mathrm{Cov}(E_a))$ estimated by

$$\widehat{\Gamma}_{aa} = \langle X_a - X_{v_1}, X_a - X_{v_2} \rangle = |E_a|_2^2 - \langle E_a, E_{v_1} + E_{v_2} \rangle \tag{9}$$

$v_1, v_2$ unknown!

## Estimator for $\Gamma$

For $(a,b) \in [n]^2$ let $V(a,b) := \max_{(c,d)\in([n]\setminus\{a,b\})^2} \left|\langle X_a - X_b, \frac{X_c - X_d}{|X_c - X_d|_2}\rangle\right|$,
$\widehat{v_1} := \mathrm{argmin}_{b\in[n]\setminus\{a\}} V(a,b)$ and $\widehat{v_2} := \mathrm{argmin}_{b\in[p]\setminus\{a,\widehat{v_1}\}} V(a,b)$. Then

$$\widehat{\Gamma} := \mathrm{diag}\left(\langle X_a - X_{\widehat{v_1}}, X_a - X_{\widehat{v_2}}\rangle_{a\in[n]}\right). \tag{10}$$

Take-away: $\widehat{v_1}, \widehat{v_2}$ chosen so that $\widehat{\Gamma}_{aa}$ is a "good" proxy for $\mathrm{Tr}(\mathrm{Cov}(E_a))$

The following estimator "improves" on K-means:

## Convex, corrected K-means

$$\widehat{B} := \underset{B \in \mathcal{C}}{\operatorname{argmax}} \langle XX^T - \widehat{\Gamma}, B \rangle. \tag{11}$$

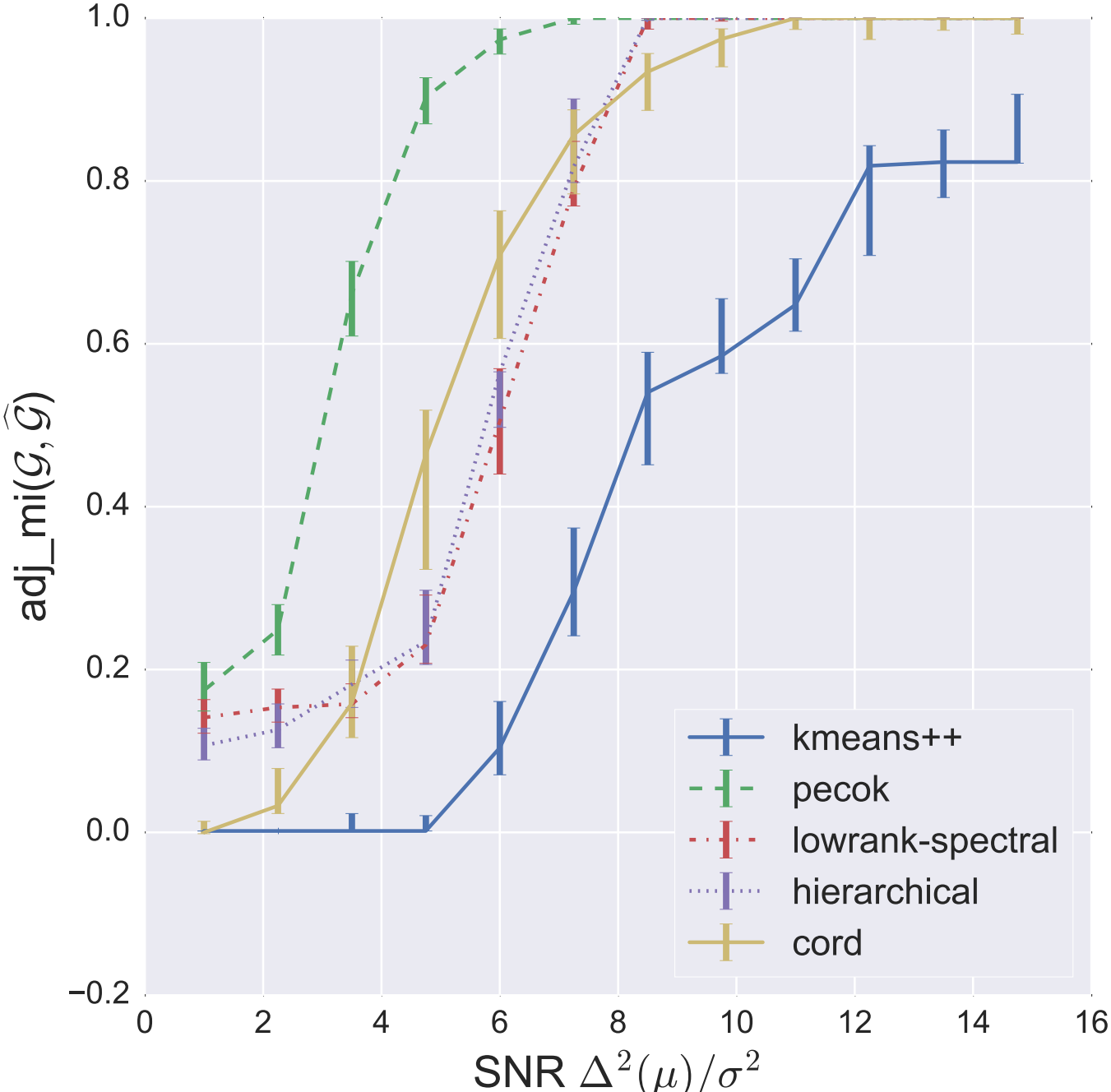
Let effective dimension $r_* := \max_{a \in [n]} \operatorname{Tr}(\Sigma_a) / \max_{a \in [n]} |\Sigma_a|_{op} \leqslant p$,

## Theorem (R. '17)

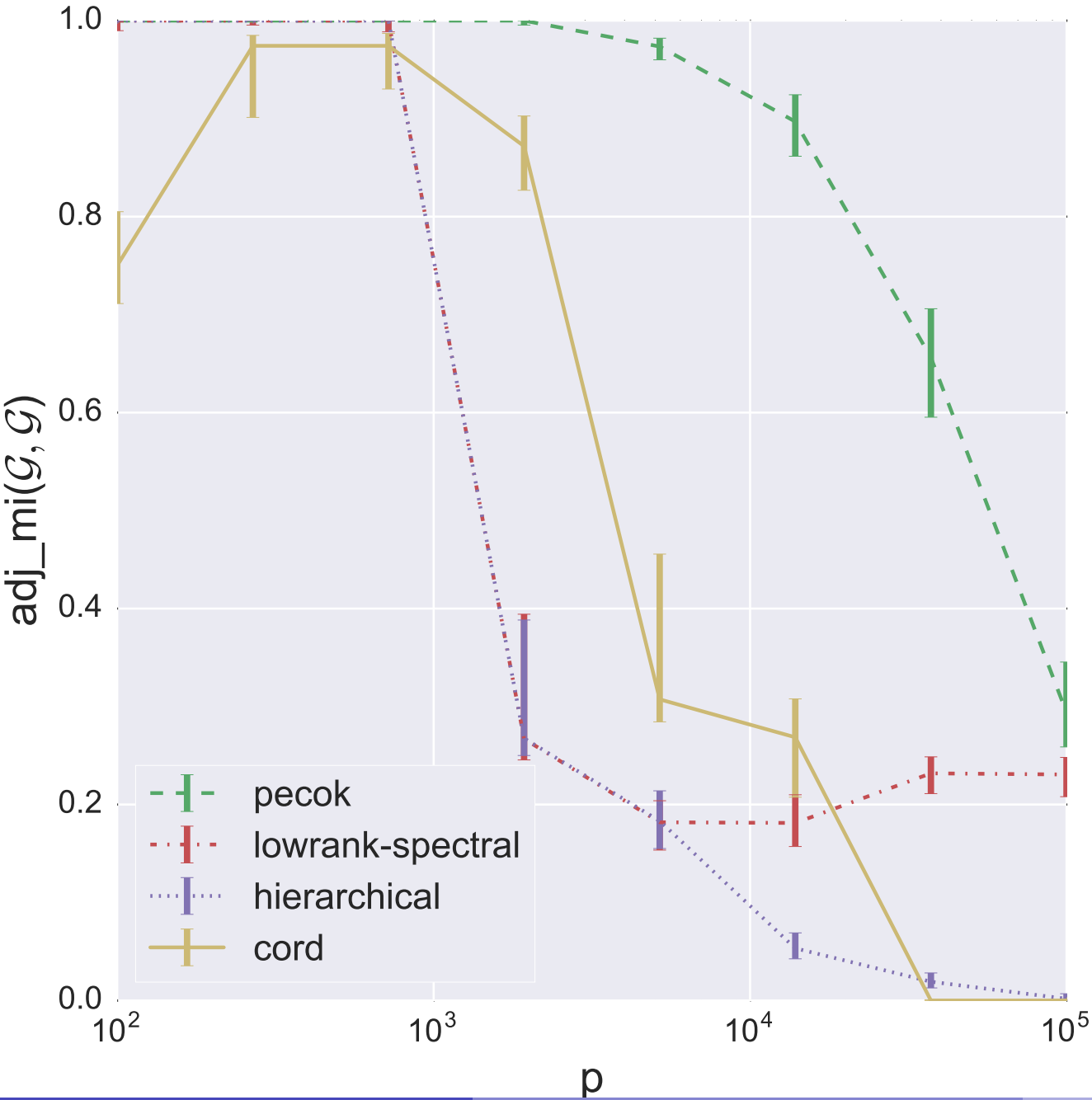
Suppose $m := \min_k |G_k| > 2$. Under latent model $(\alpha)$, if

$$m\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2 \left( n + m \log n + \sqrt{r_*(n + m \log n)} \right) \tag{12}$$

then $\widehat{B} = B_{\mathcal{G}}$ with high probability.


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
2
4
6
8
10
12
14
16
adj_mi($\mathcal{G}, \widehat{\mathcal{G}}$)
SNR $\Delta^2(\mu)/\sigma^2$
kmeans++
pecok
lowrank-spectral
hierarchical
cord

- $n = 100$
- $p = 500$
- $K = 5$


1.0
0.8
0.6
0.4
0.2
0.0
adj_mi($\hat{\mathcal{G}}, \mathcal{G}$)
$10^2$
$10^3$
$10^4$
$10^5$
p
pecok
lowrank-spectral
hierarchical
cord

- $n = 100$
- $K = 5$

# "Low" dimension regimes

## Is the separation condition optimal?

$$m\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\left(n + m\log n + \sqrt{r_*(n + m\log n)}\right)$$

Suppose all groups have equal size $m \approx n/K$

- "Low dimension" $p \lesssim n + m\log n$
- "effective low dimension" $r_* \lesssim n + m\log n \lesssim p$

$$\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\left(K + \log n\right) \tag{13}$$

$\rightarrow$ result by Mixon et al. [2016], recovery in: $\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2 K^2$

# "High" dimension regime

## Is the separation condition optimal?

$$m\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\left(n + m\log n + \sqrt{r_*(n + m\log n)}\right)$$

Suppose all groups have equal size $m \approx n/K$

- "High dimension" $n + m\log n \lesssim r_*$

$$\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\sqrt{r_*\frac{K}{n}\left(K + \log n\right)} \tag{14}$$

$\rightarrow$ result by Banks et al. [2016], lower bound for detecting Gaussian mixture:

$$\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\sqrt{\left(K\log K\right)\frac{p}{n}} \tag{15}$$

## $(\beta)$ - Generalized model

Assume $\exists \delta > 0, \forall k, \forall a \in G_k$:

$$X_a = \nu_a + E_a \quad \text{with} \quad \mathbf{E}[X_a] = \nu_a \in B_f(\mu_k, \delta) \quad \text{and} \quad E_a \underset{ind}{\sim} \text{sub-}\mathcal{N}(0, \Sigma_a)$$

## Theorem (R. '17)

Suppose $m := \min_k |G_k| > 2$. Under model $(\beta)$, if

$$m\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\left(n + m\log n + \sqrt{r_*(n + m\log n)}\right) + \delta\sigma + \delta^2(\sqrt{n} + m) \qquad (16)$$

then $\widehat{B} = B_{\mathcal{G}}$ with high probability

NB: when $\delta$ is of order $\sigma\sqrt{\log n}$, no difference between models $(\alpha)$ and $(\beta)$, i.e. it is the model error one can tolerate (and moral: we couldn't expect more)

# Identifiability given K

## $(\beta)$ - Generalized model

Assume $\exists \delta > 0, \forall k, \forall a \in G_k$:

$$X_a = \nu_a + E_a \qquad \text{with} \qquad \nu_a \in B_f(\mu_k, \delta) \qquad \text{and} \qquad E_a \underset{ind}{\sim} \text{sub-}\mathcal{N}(0, \Sigma_a)$$

**Quantities to consider:**

- discriminating power $\rho(\mathcal{G}, \mu, \delta) := \Delta_{\mathcal{G}}(\mu)/\delta$

## Identifiability

If $\rho(\mathcal{G}, \mu, \delta) > 4$, then $\mathcal{G}$ is the unique maximizer of $\rho$ over partitions of size $|\mathcal{G}|$

# How to account for number of clusters K?

## $K$-adaptive estimator

Let $\widehat{\kappa} \in \mathbb{R}_+$, estimate $B_{\mathcal{G}}$ using the SDP

$$\widehat{B}_{\text{adapt}} = \underset{B \in \underline{\mathcal{C}}}{\operatorname{argmax}} \langle XX^T - \widehat{\Gamma}, B \rangle - \widehat{\kappa} \times tr(B) \tag{17}$$

$$\underline{\mathcal{C}} := \left\{ \text{symmetric } B \in \mathbf{R}^{n \times n} : \begin{array}{l} \bullet\ B \geqslant 0 \\ \bullet\ B.1_n = 1_n \\ \bullet\ (1 \succcurlyeq) B \succcurlyeq 0 \end{array} \right.$$

## Theorem (R. '17)

If

$$\sigma^2 \left( \sqrt{r_* n} + n \right) \gtrsim \widehat{\kappa} \gtrsim m \Delta^2_{\mathcal{G}}(\mu) \tag{18}$$

then we have the same recovery conditions as before, i.e.

$$m \Delta^2_{\mathcal{G}}(\mu) \gtrsim \sigma^2 \left( n + m \log n + \sqrt{r_* (n + m \log n)} \right) \tag{19}$$

# Optimization problem with ADMM

We want to solve the semi-definite program

$$\widehat{B} = \underset{B \in \mathcal{C}}{\operatorname{argmax}} \langle \mathsf{gram}, B \rangle \tag{20}$$

over the set $\mathcal{C} := \{\text{symmetric } B \in \mathbb{R}^{n \times n} : B \succcurlyeq 0, B \geqslant 0, B1 = 1, \operatorname{Tr}(B) = K\}$
We introduce $X, Y, Z$, let $\mathcal{L} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n+1}$ be a linear operator so that $\mathcal{L}(X) = b$ collects the $n + 1$ affine constraints for set $\mathcal{C}$, then problem (20) is exactly equivalent to:

$$\inf_{X \in \mathbb{R}^{n \times n}} \{ \underbrace{-\langle \mathsf{gram}, X \rangle + \delta_{\{X : \mathcal{L}(X) = b\}} + \delta_{\mathcal{S}_n^+}(Y) + \delta_{\mathsf{Pos}}(Z)}_{f(X,Y,Z)} \} \tag{21}$$

$$\text{subject to } X = Y = Z$$

Introduce dual variable $\chi$ from $\mathcal{D} = \{\chi = (x, y, z) \in (\mathbb{R}^{n \times n})^3 | x = y = z\}$:

$$\begin{aligned} &\text{minimize } f(X, Y, Z) + \delta_{\mathcal{D}}(\chi) + (\rho/2) \| (X, Y, Z) - \chi + U \|_2^2 \\ &\text{subject to } (X, Y, Z) - \chi = 0 \end{aligned} \tag{22}$$

# Thank you for your attention


J. Banks, C. Moore, N. Verzelen, R. Vershynin, and J. Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *ArXiv e-prints*, July 2016.

F. Bunea, C. Giraud, M. Royer, and N. Verzelen. PECOK: a convex optimization approach to variable clustering. *ArXiv e-prints*, June 2016.

D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures with k-means. In *2016 IEEE Information Theory Workshop (ITW)*, pages 211–215, Sept 2016. doi: 10.1109/ITW.2016.7606826.

J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. on Optimization*, 18(1):186–205, February 2007. ISSN 1052-6234. doi: 10.1137/050641983.

## Effect of correcting for $\Gamma$ on the separating rate of (12)

Separation rate for un-corrected K-means:

$$m\Delta_{\mathcal{G}}^2(\mu) \gtrsim \sigma^2\left(n + m\log n + \sqrt{r_*(n + m\log n)} + r_*\right) \tag{23}$$

In the "High dimension" $n + m\log n \lesssim r_*$ regime, the leading rate is $\sigma^2 r_*$, whereas with correction $\widehat{\Gamma}$, the leading rate was $\sigma^2\sqrt{r_*(n + m\log n)}$