ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Nonparametric consistency for maximum likelihood of mixtures of elliptically symmetric distributions (ESD)

Pietro Coretto and Christian Hennig

1. Introduction

Interested in estimating finite mixture models of the type

$$\psi(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

K fixed, where f elliptically symmetric:

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} g\left((\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

This includes Gaussian mixtures where

$$g(r) = c\exp(-\frac{r^2}{2}),\ \ f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \varphi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

The meaning of model assumptions?

Parametric method;
"We have to believe that data were iid generated by $\psi(\bullet; \boldsymbol{\theta})$."

"*K*-means is a nonparametric method; this is better
if we don't know that above assumption is fulfilled."

???

In fact, K-means…

$$\begin{aligned} T_n(\tilde{\boldsymbol{X}}_n) &= (\boldsymbol{m}_{1n}, \ldots, \boldsymbol{m}_{Kn}, g_{in}, \ldots, g_{nn}) \\ &= \underset{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_K, g_1,\ldots,g_n}{\arg\min} \sum_{i=1}^{n} \|\boldsymbol{X}_{in} - \boldsymbol{m}_{g_i}\|^2 \end{aligned}$$

…is ML for "fixed partition model":

$$\mathcal{L}(\boldsymbol{X}_i) = \mathcal{N}_p\left(\boldsymbol{\mu}_{\gamma_i}, \sigma^2 \boldsymbol{I}_p\right),\ \gamma_i \in \{1, \ldots, K\},\ K > 1,\ \sigma^2 \geq 0.$$

Who calls K-means "nonparametric" either doesn't know this, or argues that originally it was defined nonparametrically, without reference to the model. Or…

...or makes reference to the following:
Pollard (1981) showed that under nonparametric P,
K-means is a consistent estimator for
its own canonical functional ($T_n(\tilde{\boldsymbol{X}}_n) = C(\hat{P}_n)$)

$$(\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_K^*) = \underset{(\boldsymbol{m}_1,\ldots,\boldsymbol{m}_K)\in(\mathbb{R}^p)^k}{\arg\min} \int \min_{\boldsymbol{m}\in\{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_k\}} \|\boldsymbol{x} - \boldsymbol{m}\|^2 dP(\boldsymbol{x}).$$

Interestingly (Bryant 1991), it's *not* consistent for $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ in

$$\mathcal{L}(\boldsymbol{X}_i) = \mathcal{N}_p\left(\boldsymbol{\mu}_{\gamma_i}, \sigma^2 \boldsymbol{I}_p\right),\ \gamma_i \in \{1, \ldots, K\},\ K > 1,\ \sigma^2 \geq 0.$$

May wonder whether $(\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_K^*)$ is really of interest!
(Depends on application; Voronoi tesselation)

The meaning of “model assumptions” is not usually well communicated!

Model assumptions do *not* have to be fulfilled in practice. (They never are!)

“Method X assumes Y” means that there’s a theorem that states that under Y, X has certain “good” properties.

The K-means example shows that a property may look good under one assumption but not so good under another.
(One could claim K-means assumes
a fixed partition spherical Gaussian model,
or a nonparametric P, i.i.d.)

Model can be assumed in order to derive/develop a method
that works well under that assumption
- the model is an *inspiration* -
but in reality it is always applied to data
that don't obey the assumption.

Need then new theory or simulations to find out what happens
if method assuming Y is applied in situation Z$\neq$Y.

(Obviously, *Z is not the reality either,* but
gives broader understanding of characteristics of method X.)

2. The ESD mixture setup

"Assuming" mixture

$$\psi(\boldsymbol{x};\theta) := \sum_{k=1}^{K} \pi_k f(\boldsymbol{x};\boldsymbol{\mu}_k,\Sigma_k)$$

to derive ML-estimator,
what happens if data comes from nonparametric P?

- Consistency for canonical functional
  (Gaussian mixture done by Garcia-Escudero et al., 2015),
- result on value of canonical functional in case of
  well separated nonparametric mixture components.

$$
\begin{aligned}
\ell_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\log(\psi(\boldsymbol{x}_i;\boldsymbol{\theta})),\\
\boldsymbol{\theta}_n &\in \arg\max_{\boldsymbol{\theta}\in\tilde{\Theta}_K} \ell_n(\boldsymbol{\theta}),
\end{aligned}
$$

Can show that

$$
\lambda^*_{\min}(\boldsymbol{\Sigma}) \searrow 0 \Rightarrow f(\boldsymbol{\mu};\boldsymbol{\mu},\boldsymbol{\Sigma}) \longrightarrow +\infty.
$$

Degeneration of likelihood!

In order to avoid degeneration, require

$$\boldsymbol{\theta} \in \tilde{\Theta}_K = \left\{ \boldsymbol{\theta} : \ \ \pi_k \geq 0 \ \forall k \geq 1, \ \sum_{k=1}^{K} \pi_k = 1; \ \frac{\lambda_{\max}(\boldsymbol{\theta})}{\lambda_{\min}(\boldsymbol{\theta})} \leq \gamma \right\}.$$

(Garcia-Escudero et al. 2014 etc.)

$$\begin{aligned} L(\boldsymbol{\theta}, P) &= \int \log \psi(\boldsymbol{x}, \boldsymbol{\theta}) dP(\boldsymbol{x}), \\ L_K(P) &= \sup_{\boldsymbol{\theta} \in \tilde{\Theta}_K} L(\boldsymbol{\theta}, P), \\ \boldsymbol{\theta}^\star(P) &\in \arg\max_{\boldsymbol{\theta} \in \tilde{\Theta}} L(\boldsymbol{\theta}, P). \end{aligned}$$

3. Existence and consistency

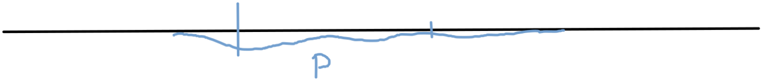
Assumptions:

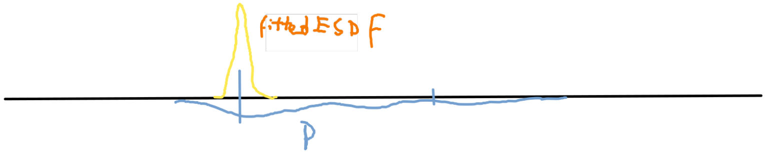
- A1 For every $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_K\} \subset \mathbb{R}^p$: $P(S) < 1$.
- A2 With $h_g(y) = \mathrm{E}_P\left[\log(g(y^{-1}\|X - \boldsymbol{\mu}\|^2))\right]$, for all $\boldsymbol{\mu},\ y \searrow 0:\ \log(y^{-1}) \in o(h_g(y))$.
- A3 $L_{K-1}(P) < L_K(P)$.

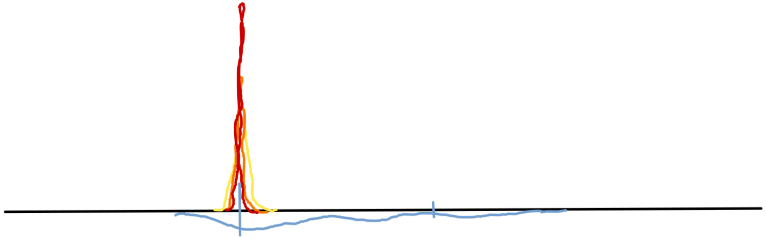
Without A1, degeneration cannot be avoided.
A2 states that if $\lambda_{\min}(\boldsymbol{\theta}) \searrow 0$, then for all k,

$$\mathrm{E}_P[\log(f(X; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))] \longrightarrow -\infty.$$

This regards the combination (P, g) and should be rather mild, (for f Gaussian with $\mathrm{E}_P\left[(\|\boldsymbol{x}\|^2)\right] < \infty$ it holds).
A3 is required to avoid parameter identification issues.


P


fitted ESD F
P



Theorem 1 (existence of the ML functional).
Under A1-A3,

$$\exists \text{ compact } T \subset \tilde{\Theta}_K : \ \exists \boldsymbol{\theta} \in T : -\infty < L(\boldsymbol{\theta}, P) < +\infty,$$

$$\boldsymbol{\theta} \notin T \Rightarrow \exists c : \ L(\boldsymbol{\theta}, P) < c < L_K(P).$$

...but the maximiser is not unique
(label switching and potentially other issues).

$$S(\dot{\boldsymbol{\theta}}) = \left\{ \boldsymbol{\theta} \in \tilde{\Theta}_K(P) : \; L(\boldsymbol{\theta}, P) = L(\dot{\boldsymbol{\theta}}, P) \right\},$$

$$\mathcal{T}(\dot{\boldsymbol{\theta}}, \varepsilon) = \left\{ \boldsymbol{\theta} \in \tilde{\Theta}_K(P) : \|\boldsymbol{\theta} - \ddot{\boldsymbol{\theta}}\| < \varepsilon \; \forall \; \ddot{\theta} \in S(\dot{\boldsymbol{\theta}}) \right\}$$

Theorem 2 (consistency).
Under A1-A3,
$\forall \varepsilon > 0$ and every sequence of maximizers $\boldsymbol{\theta}_n$ of $\ell_n(\cdot)$:

$$\lim_{n \to \infty} \Pr[\boldsymbol{\theta}_n \in \mathcal{T}(\boldsymbol{\theta}^\star(P), \varepsilon)] = 1.$$

(For Gaussian f, assumptions are almost same as for nonparametric K-means consistency!)

4. The mixture ML functional for nonparametric mixtures

Given distributions $Q_1, \ldots, Q_K$ "centered" at zero,
$\xi_1, \ldots, \xi_K > 0$ mixture proportions with $\sum_{k=1}^K \xi_k = 1$,
For $m \in \mathbb{N},\ k \in \{1, \ldots, K\}$, $\boldsymbol{\rho}_{mk} \in \mathbb{R}^p$ so that

$$\lim_{m\to\infty} \min_{k_1 \neq k_2 \in \{1,\ldots,K\}} \|\boldsymbol{\rho}_{mk_1} - \boldsymbol{\rho}_{mk_2}\| = \infty.$$

Define sequence of nonparametric mixture distributions

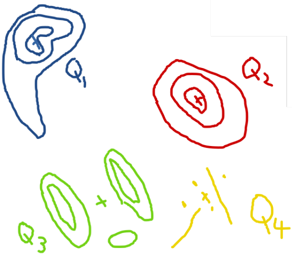
$$P_m(\boldsymbol{x}) = \sum_{k=1}^{K} \xi_k Q_{mk}(\boldsymbol{x}),\ Q_{mk}(\boldsymbol{x}) = Q_k(\boldsymbol{x} - \boldsymbol{\rho}_{mk}).$$

"Central set"

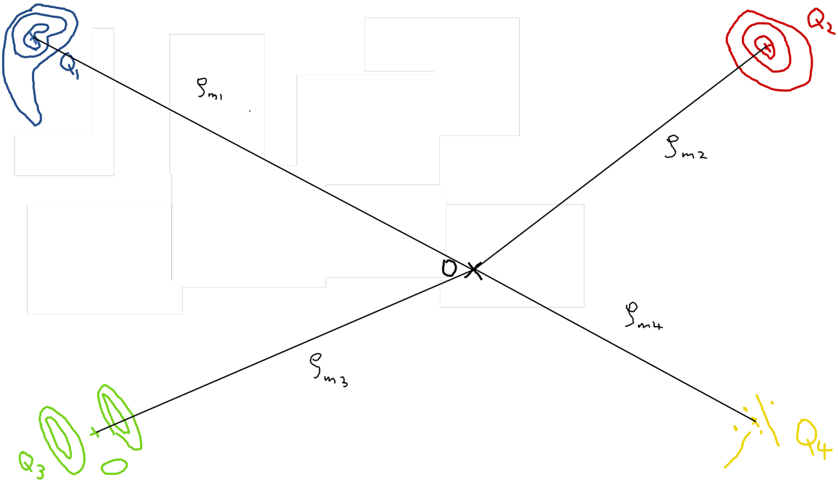
$$B_\epsilon(\boldsymbol{\rho}_{mk}) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{\rho}_{mk}\| < \epsilon\}$$

ϵ large enough: for arbitrarily small $\eta > 0$:

$$\forall m, k : Q_{mk}(B_\epsilon(\boldsymbol{\rho}_{mk})) \geq 1 - \eta.$$


Q1
Q2
Q3
Q4


ε
Q1
Q2
Q3
Q4


Q_1
Q_2
Q_3
Q_4
ρ_{m1}
ρ_{m2}
ρ_{m3}
ρ_{m4}
0

Clustering assuming $P = \sum_{k=1}^{K} \pi_k F_k$,
F_k with density $f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$:
Model for $(\boldsymbol{x}, Z_1, \ldots, Z_K),\ Z_k \in \{0, 1\}$ unobserved,
$\sum_{k=1}^{K} Z_k = 1$.

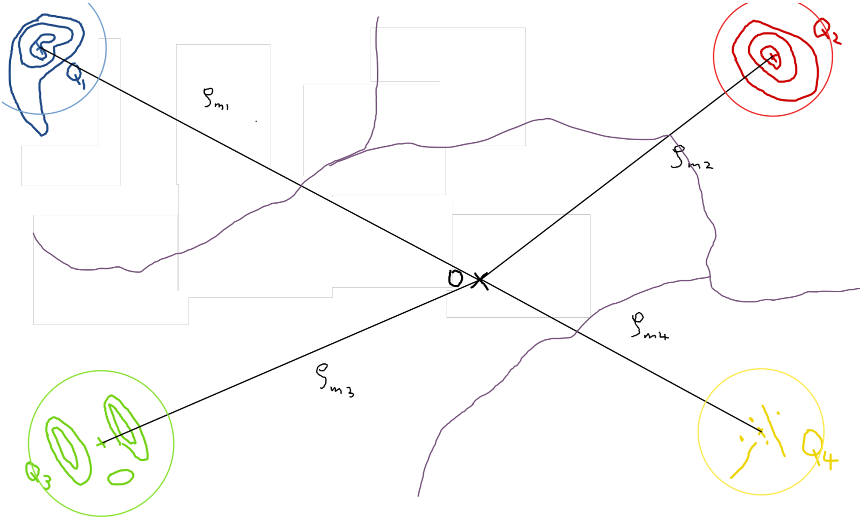
$$\begin{aligned} P\{Z_k = 1\} &= \pi_k, \\ p(\boldsymbol{x}|Z_k = 1) &= f_k(\boldsymbol{x}) \Rightarrow \\ \Pr[Z_k = 1 \mid \boldsymbol{x}] &= \tau_k(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\pi_k f(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\psi(\boldsymbol{x}; \boldsymbol{\theta})}, \\ \mathrm{cl}(\boldsymbol{x}) &= \underset{1 \le k \le K}{\arg\max}\ \tau_k(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned}$$

Assumption:

A4 $\exists c_0 < \infty : \ \forall k \in \{1, \ldots, K\} \ :$
$\int \log g(\|\boldsymbol{x}\|) dQ_k(\boldsymbol{x}) \leq c_0.$

Theorem 3 (functional components correspond to Q_k)**.**
Under A2 and A4, for large enough m,
components of $\theta^\star(P_m)$ can be numbered so that $\forall k$:

$$B_\epsilon(\boldsymbol{\rho}_{mk}) \subseteq C_{mk} = \{\boldsymbol{x} : \ \mathrm{cl}(\boldsymbol{x}, \theta^\star(P_m)) = k\}.$$


Q1
Q2
Q3
Q4
O
ρm1
ρm2
ρm3
ρm4

- For separation between $Q_{mk} \to \infty$, this may not seem surprising.
- Can prove similar theorem for K-means (requires second moments).
- Q_{mk} may still overlap (nonzero density).
- Results about functional values for nonparametric P_m hardly exist.
- Does not hold for all clustering methods:
  - Single linkage, Q_{mk} Gaussian, will for any m, large enough n, produce one-point cluster.
  - Same average linkage (conjecture).
  - α-trimmed clustering can trim complete central set of Q_{mk} if $\xi_k \leq \alpha$.

With growing separation, also parameter estimators converge.

$$\tilde{\boldsymbol{\kappa}} = (\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k) = \arg\max_{\boldsymbol{\kappa}} \tilde{L}(\boldsymbol{\kappa}, Q_k),\ \tilde{L}(\boldsymbol{\kappa}, Q) = \int \log f(\boldsymbol{x}; \boldsymbol{\kappa}) dQ(\boldsymbol{x}).$$

Corresponding functionals for $Q_{mk} = Q_k(\bullet - \boldsymbol{\rho}_{mk})$ are

$$\tilde{\boldsymbol{\mu}}_{mk} = \tilde{\boldsymbol{\mu}}_k + \boldsymbol{\rho}_{mk},\ \tilde{\boldsymbol{\Sigma}}_{mk} = \tilde{\boldsymbol{\Sigma}}_k.$$

Assumption A5 For given Q_k,

$$\forall \varepsilon > 0\ \exists \beta > 0 :\ \|\boldsymbol{\kappa} - \tilde{\boldsymbol{\kappa}}_k\| > \varepsilon \Rightarrow L(\tilde{\boldsymbol{\kappa}}_k, Q_k) - L(\boldsymbol{\kappa}, Q_k) > \beta.$$

"Distinguished" maximum exists for Q_k - holds e.g. if f Gaussian.

Theorem 4 (functional parameters correspond to Q_k).
Under A2 and A4, for large enough m, components of $\theta^\star(P_m)$ can be numbered so that

$$\lim_{m\to\infty} \|\pi^\star_{mk} - \xi_k\| = 0,$$

and for Q_k fulfilling A5,

$$\lim_{m\to\infty} \|\boldsymbol{\kappa}^\star_{mk} - \tilde{\boldsymbol{\kappa}}_{mk}\| = 0.$$

Corollary. With $f(\bullet;\mu,\Sigma)$ p-variate Gaussian, under A2 and A4,

$$\lim_{m\to\infty}\left\|\boldsymbol{\mu}^{\star}_{mk}-\int \boldsymbol{x}\,dQ_k(\boldsymbol{x})-\boldsymbol{\rho}_{mk}\right\| = 0,$$

$$\lim_{m\to\infty}\left\|\Sigma^{\star}_{mk}-\int(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_k)(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_k)^{\mathsf{T}}dQ_k(\boldsymbol{x})\right\| = 0.$$

5. Conclusion

- ML estimators based on parametric ESD mixtures are consistent on nonparametric distributions.
- For well separated nonparametric mixtures, nonparametric mixture components will eventually be found.
- Such parametric mixture ML-estimators are at least as “nonparametric” as K-means; the parametric mixture assumption does *not* need to hold.
- Still work to do: Better characterisation of assumptions!
- Estimating number of components?

References

Bryant, P. G. (1991) Large-sample results for optimization-based clustering meth- ods. *Journal of Classification* 8, 31–44.

Coretto, P. and C. Hennig (2017) Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research* 18 (142), 1–39.

Garcia-Escudero, L. A., A. Gordaliza, C. Matran and A. Mayo-Iscar (2015) Avoiding spurious local maximizers in mixture modeling. *Statistics and Computing* 25, 1–15.

García-Escudero, L. A., A. Gordaliza and A. Mayo-Iscar (2014) A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification* 8, 27–43.

Pollard, D. (1981) Strong consistency of k-means clustering. *Annals of Statistics* 9, 135–140.