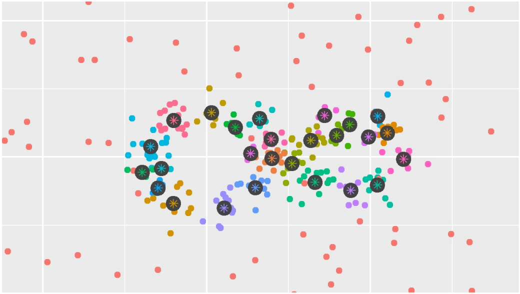
Robust approximation of compact sets with unions of ellipsoïds. Application to data clustering.

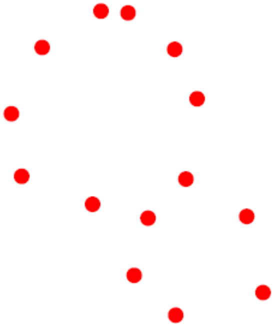
Claire Brécheteau

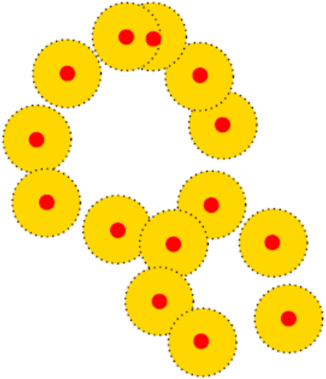
Université Rennes 2, IRMAR - with Clément Levrard, Université de Paris, LPSM

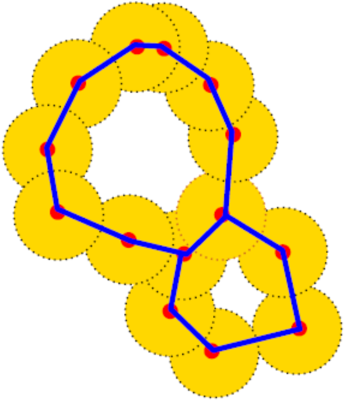
January, 29th 2021

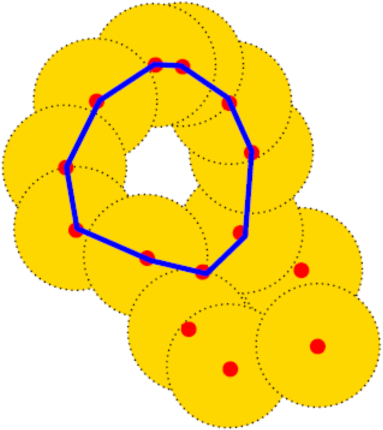
How to approximate a manifold with a set of k points, from a noisy sample ?

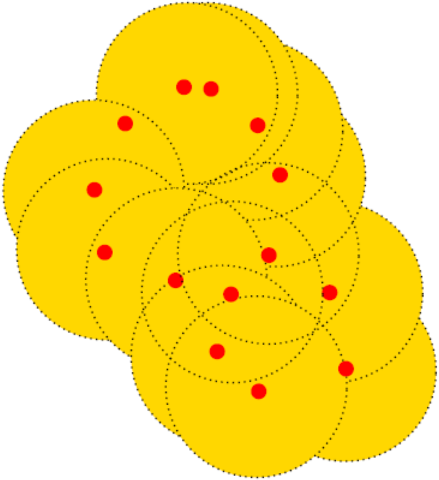












The k-means method

P distribution on $\mathbb{R}^d$
$\mathbf{c} = (c_1, c_2, \ldots, c_k) \in (\mathbb{R}^d)^k$ codebook

Definition

The optimal codebook $\mathbf{c}^*$ minimizes the k-means loss function

$$R : \mathbf{c} \mapsto P \min_{i=1..k} \| \cdot - c_i \|^2.$$

FIGURE – Lloyd's algorithm method

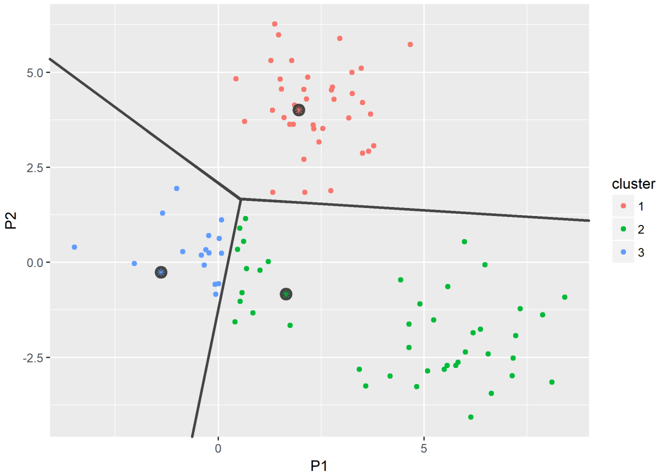


FIGURE – Lloyd's algorithm method

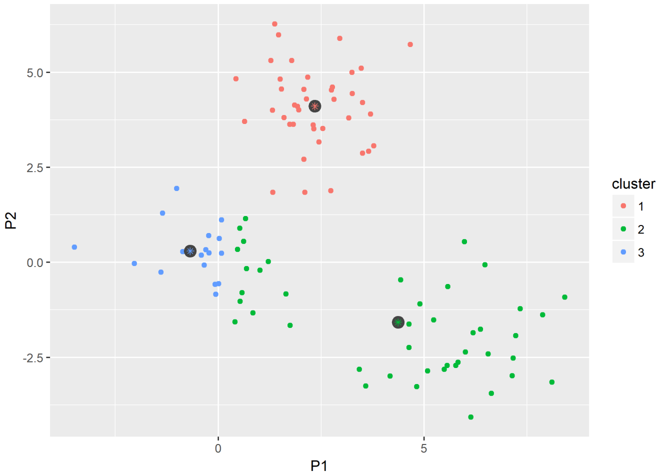


FIGURE – Lloyd's algorithm method

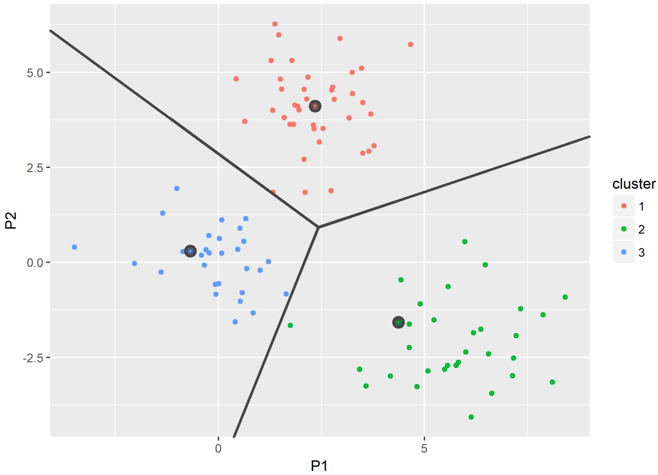


Figure – Lloyd's algorithm method

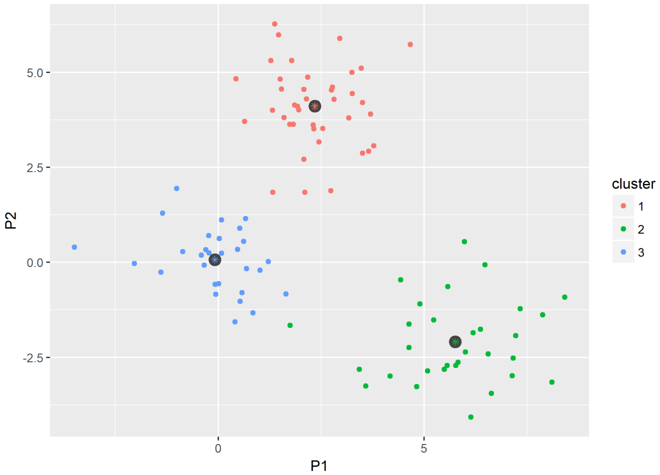
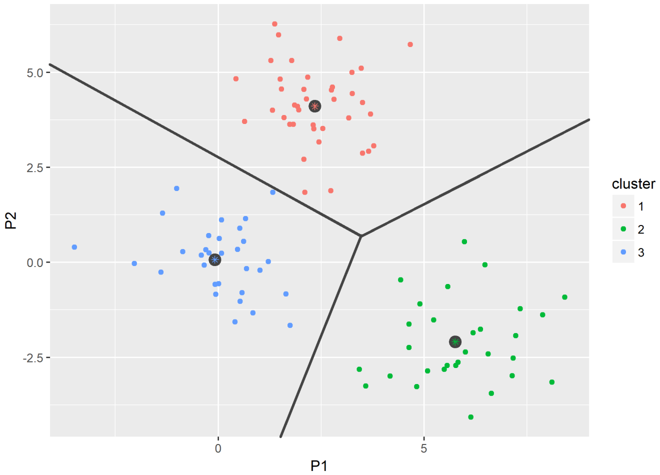


FIGURE – Lloyd's algorithm method

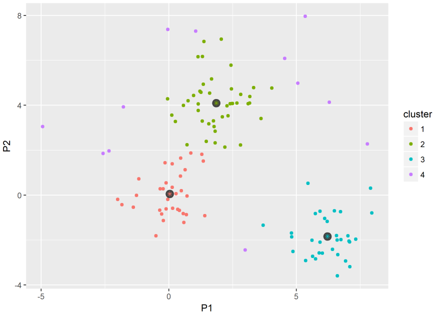


Definition

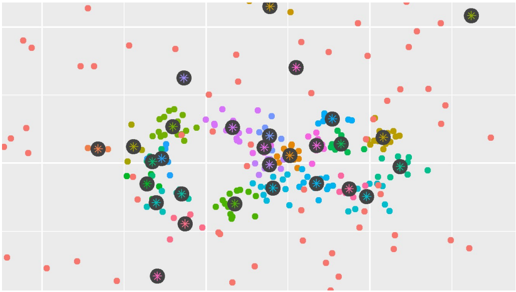
The optimal trimmed codebook $\mathbf{c}_h^*$ minimizes the trimmed k-means loss function

$$R_h : \mathbf{c} \mapsto \inf_{\tilde{P},\, h\tilde{P} \leq P} \tilde{P} \min_{i=1..k} \| \cdot - c_i \|^2 .$$

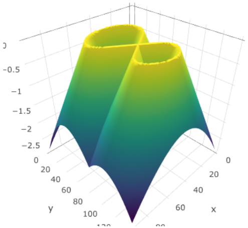
FIGURE – Trimmed k-means



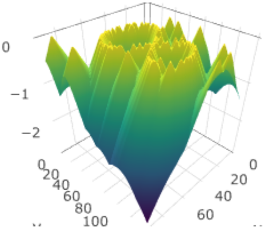
Trimmed k-means :



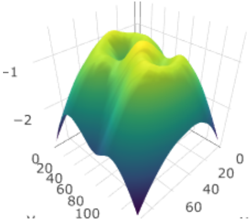
$\Rightarrow$ Not working...



Distance to the compact

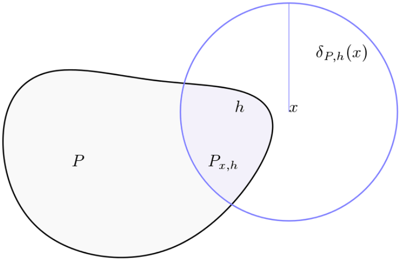


Distance to the sample



Distance to the empirical measure

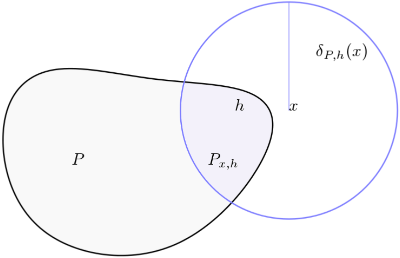
Distance-to-measure (DTM)



$$
\begin{aligned}
\mathrm{d}^2_{P,h}(x) &= P_{x,h}\|\cdot - x\|^2 \\
&= \inf_{t\in\mathbb{R}^d} P_{t,h}\|\cdot - x\|^2 \\
&= \|m(P_{x,h}) - x\|^2 + v(P_{x,h}) \\
&= \inf_{t\in\mathbb{R}^d} \|m(P_{t,h}) - x\|^2 + v(P_{t,h})
\end{aligned}
$$

Notation : Mean $m(P)$, Variance $v(P)$.

Distance-to-measure (DTM)



$$
\begin{aligned}
\mathrm{d}^2_{P,h}(x) &= P_{x,h}\|\cdot - x\|^2\\
&= \inf_{t\in\mathbb{R}^d} P_{t,h}\|\cdot - x\|^2\\
&= \|m(P_{x,h}) - x\|^2 + v(P_{x,h})\\
&= \inf_{t\in\mathbb{R}^d} \|m(P_{t,h}) - x\|^2 + v(P_{t,h})
\end{aligned}
$$

Notation : Mean $m(P)$, Variance $v(P)$.

$\rightsquigarrow$ Sublevel sets of the DTM : union of balls.

The k-PDTM

$$\mathbf{t}^* \in \arg\min_{\mathbf{t}} P \min_{j=1..k} \|\cdot - m(P_{t_j,h})\|^2 + v(P_{t_j,h}).$$

Definition

The k-power distance-to-measure (k-PDTM) $\mathrm{d}_{P,h,k}$ is defined for $x \in \mathbb{R}^d$ by :

$$\mathrm{d}^2_{P,h,k}(x) = \min_{j=1..k} \|x - m(P_{t_j^*,h})\|^2 + v(P_{t_j^*,h})$$

Approximation of the DTM with the empirical k-PDTM

$$P\left|\mathrm{d}^2_{Q_n,h,k}(\cdot)-\mathrm{d}^2_{P,h}(\cdot)\right|\leq P\left|\mathrm{d}^2_{Q_n,h,k}(\cdot)-\mathrm{d}^2_{Q,h,k}(\cdot)\right|+P\left|\mathrm{d}^2_{Q,h,k}(\cdot)-\mathrm{d}^2_{P,k}(\cdot)\right|$$

Proposition

If $\mathrm{Supp}(P) \subset \mathrm{B}(0,K)$, *and* $Q\|.\| < \infty$,
then $P\left|\mathrm{d}^2_{Q,h,k}(.) - \mathrm{d}^2_{P,h}(.)\right|$ *is bounded from above by*

$$3\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,\mathrm{B}(0,K)} + P\left(\mathrm{d}^2_{P,h,k}(.) - \mathrm{d}^2_{P,h}(.)\right) + 4W_1(P,Q)\sup_{s\in\mathbb{R}^d}\|m(P_{s,h})\|$$

with $P\left(\mathrm{d}^2_{P,h,k}(.) - \mathrm{d}^2_{P,h}(.)\right)$ of order $k^{-\frac{2}{d'}}$ for a "d'-dimensional distribution".

Approximation of the k-PDTM from point clouds

$\mathrm{Supp}(P) = \mathscr{X} \subset \mathrm{B}(0,K)$
$X_i = Y_i + Z_i$, Y_i and Z_i all independent, $Y_i \sim P$, Z_i sub-Gaussian with variance $\sigma^2 \leq K^2$
$Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Theorem (B. - Levrard 2020)

For every $p > 0$, with probability larger than $1 - 10n^{-p}$, we have

$$\left| P\mathrm{d}^2_{Q_n,h,k}(.) - \mathrm{d}^2_{Q,h,k}(.) \right| \leq C\sqrt{k\log(k)d}\,\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}.$$

Approximation of the k-PDTM from point clouds

$\mathrm{Supp}(P) = \mathscr{X} \subset \mathrm{B}(0, K)$
$X_i = Y_i + Z_i$, Y_i and Z_i all independent, $Y_i \sim P$, Z_i sub-Gaussian with variance $\sigma^2 \leq K^2$
$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.

Theorem (B. - Levrard 2020)

For every $p > 0$, with probability larger than $1 - 10n^{-p}$, we have

$$\left|P\mathrm{d}^2_{Q_n,h,k}(.) - \mathrm{d}^2_{Q,h,k}(.)\right| \leq C\sqrt{k\log(k)d}\,\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}.$$

$\rightsquigarrow$ optimize in k the quantity

$$\frac{C\sqrt{k\log(k)}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_{P,h}k^{-\frac{2}{d'}}.$$

Approximation of the k-PDTM from point clouds

$\mathrm{Supp}(P) = \mathscr{X} \subset \mathrm{B}(0, K)$
$X_i = Y_i + Z_i$, Y_i and Z_i all independent, $Y_i \sim P$, Z_i sub-Gaussian with variance $\sigma^2 \leq K^2$
$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.

Theorem (B. - Levrard 2020)

For every $p > 0$, with probability larger than $1 - 10n^{-p}$, we have

$$\left| P\mathrm{d}^2_{Q_n,h,k}(.) - \mathrm{d}^2_{Q,h,k}(.) \right| \leq C\sqrt{k\log(k)d}\,\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}.$$

$\rightsquigarrow$ optimize in k the quantity

$$\frac{C\sqrt{k\log(k)}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_{P,h}k^{-\frac{2}{d'}}.$$

Optimal choice $k \sim n^{\frac{d'}{d'+4}}$.

An algorithm

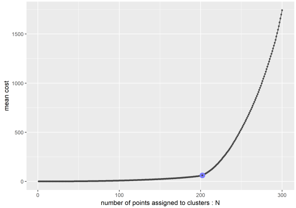
Algorithm Approximation of the k-PDTM centers

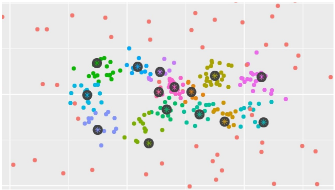
Input $\mathbb{X}_n$ a n-sample from P, q and k
Sample t_1, t_2,... t_k from $\mathbb{X}_n$ without replacement.
while the t_i's vary **do**
  for j in $1..n$ **do**
    Add X_j to some $\mathscr{C}(t_i)$ satisfying
    $\|X_j - m(t_i)\|^2 + v(t_i) \leq \|X_j - m(t_l)\|^2 + v(t_l)\,\forall l \neq i$
  end for
  for i in $1..k$ **do**
    $t_i = \frac{1}{|\mathscr{C}(t_i)|}\sum_{X\in\mathscr{C}(t_i)} X$
  end for
end while
Output $(t_1, t_2, \ldots, t_k)$.

Let $h = \frac{q}{n}$, $q \in \mathbb{N}^*$. For $t \in \mathbb{R}^d$, $m(t) = \frac{1}{q}\sum_{i=1}^{q} X_i(t)$, $v(t) = \frac{1}{q}\sum_{i=1}^{q}\left(X_i(t) - m(t)\right)^2$ with $X_i(t)$ an i-th nearest neighbor of t and $\mathscr{C}(t)$ the weighted Voronoï cell of t.

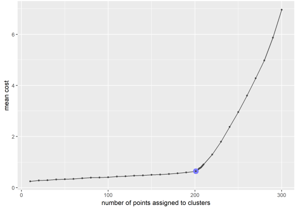
B. - Levrard, 2020

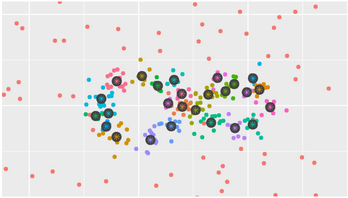
Convergence to a local minimum of $\mathbf{t} \mapsto P_n \min_{i=1..k}\|\cdot - m(P_{n\,t_i,h})\|^2 + v(P_{n\,t_i,h})$.





k-PDTM





Trimmed k-PDTM

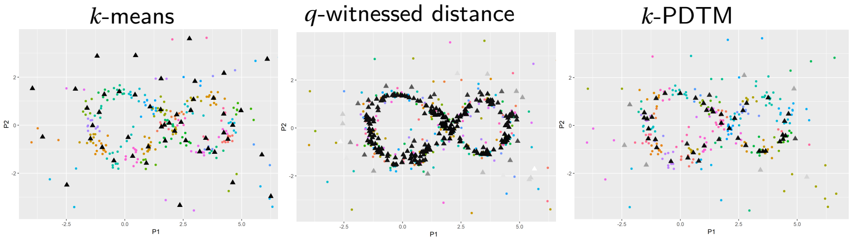


FIGURE – Comparison of the basic methods

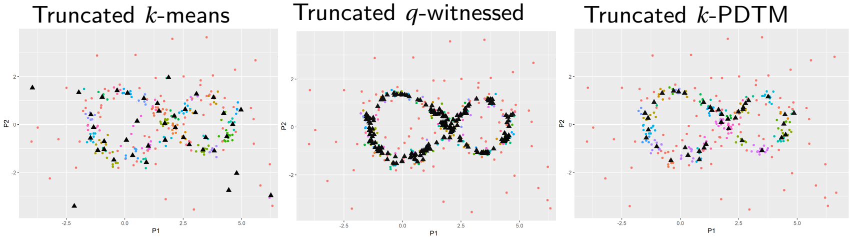
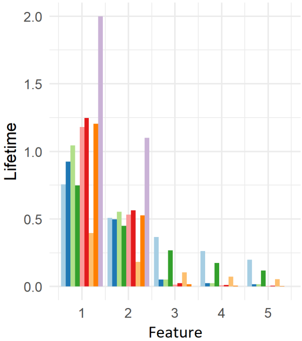
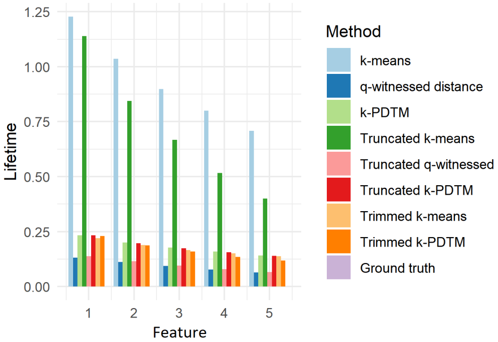


FIGURE – Comparison of the methods after thresholding

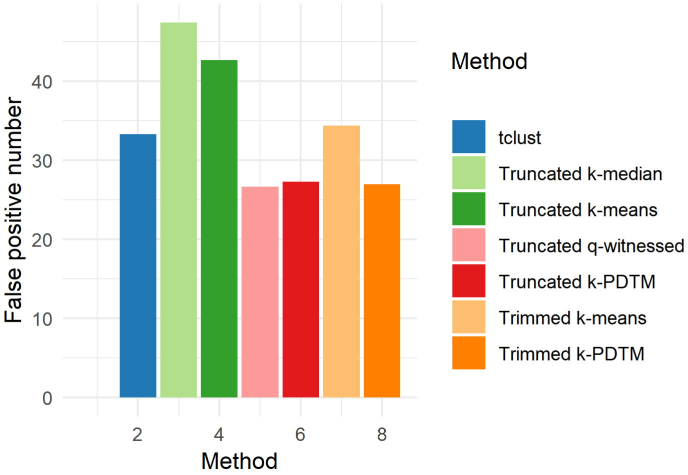
Features lifetimes comparison



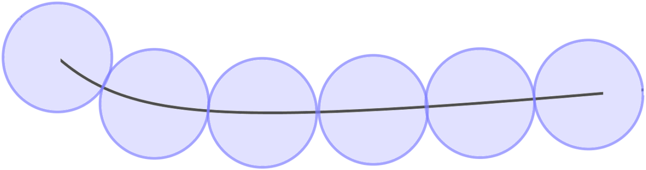
Holes

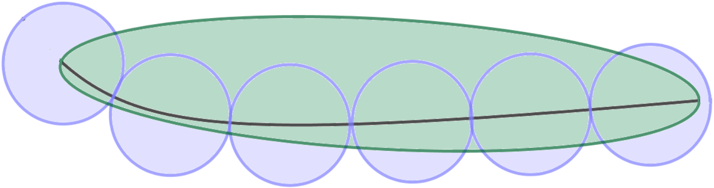


Connected components



False positive number





How to approximate a manifold with a set of k ellipsoids, from a noisy sample ?

How to approximate a manifold with a set of k ellipsoids, from a noisy sample ?

By modifying the criterion, with Mahalanobis norms...

Squared Σ-Mahalanobis norm : $\|y\|^2_{\Sigma^{-1}} = y^T \Sigma^{-1} y$.

Definition

The optimal codebook $\boldsymbol{\theta}^* = (\theta_i^*)_{i=1..k} = ((t_i^*, \Sigma_i^*))_{i=1..k}$: a minimizer of

$$\boldsymbol{\theta} \mapsto P\left(\min_{i=1..k} \|\cdot - m_{\theta_i,h}\|^2_{\Sigma_i^{-1}} + v^{\Sigma_i}_{\theta_i,h} + \log(\det(\Sigma_i))\right),$$

with $m_{\theta_i,h}$ the expectation of $P_{\theta_i,h}$ and $v^{\Sigma_i}_{\theta_i,h}$ its variance (for $\|\cdot\|_{\Sigma_i^{-1}}$).

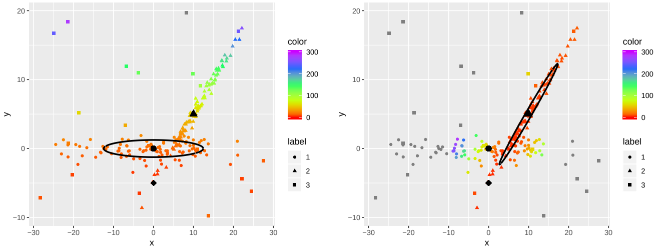


FIGURE – Illustration of the assignment phase

An algorithm

Algorithm Approximation of the k-PLM centers

Input $\mathbb{X}_n$ a n-sample from P, $q = hn$ and k

Sample t_1, t_2,... t_k from $\mathbb{X}_n$ without replacement. Set $\Sigma_i = I_d$ for i in $1..k$.

while the $\theta_i = (t_i, \Sigma_i)$s vary **do**

    Set $\mathscr{C}(\theta_i) = \{\}$ for i in $1..k$.

    for j in $1..n$ **do**

        Add X_j to some $\mathscr{C}(\theta_i)$ satisfying

        $\|X_j - m(\theta_i)\|^2_{\Sigma_i^{-1}} + w(\theta_i) \leq \|X_j - m(\theta_l)\|^2_{\Sigma_l^{-1}} + w(\theta_l)\,\forall l \neq i$

    end for

    for i in $1..k$ **do**

        $t_i = \frac{1}{|\mathscr{C}(\theta_i)|}\sum_{X \in \mathscr{C}(\theta_i)} X$; $\Sigma_i = \Sigma(t_i, \Sigma_i, \mathscr{C}(\theta_i))$

        $\theta_i = (t_i, \Sigma_i)$

    end for

end while

Output $(\theta_1, \theta_2, \ldots, \theta_k)$.

For $\theta = (t, \Sigma)$, $m(\theta) = \frac{1}{q}\sum_{i=1}^q X_i(\theta)$, $w(\theta) = \frac{1}{q}\sum_{i=1}^q \|X_i(t) - m(\theta)\|^2_{\Sigma^{-1}} + \log(\det(\Sigma))$ with $X_i(\theta)$ an i-th $\|\cdot\|_\Sigma$-nearest neighbor of t and $\mathscr{C}(\theta)$ the cell of θ. Also, for $l, m = 1..d$, $\Sigma(t, \Sigma, \mathscr{C})_{l,m} = \frac{1}{|\mathscr{C}|}\sum_{X \in \mathscr{C}} P_{n,(t,\Sigma),h}(X_{(l)} - \cdot_{(l)})(X_{(m)} - \cdot_{(m)})$.

Optimum when $k=1$:

f : density on $\mathbb{R}$
Σ : scatter matrix
$\mu \in \mathbb{R}^d$: location parameter
P on $\mathbb{R}^d$, with density :

$$f_{\mu,\Sigma,f} : x \mapsto \frac{C_{d,f}}{\sqrt{\det(\Sigma)}} f(\|x-\mu\|_{\Sigma^{-1}}).$$

Theorem (B. - Levrard - Michel, 2020)

If f is non-increasing, then, $t^* = \mu$ *and* $\Sigma^* = \left(1 + \frac{1}{h}\frac{M_{d+1}^{r_h}(f)}{M_{d+1}(f)}\right)\mathrm{Cov}(P)$.

$M_{d+1}^{r}(f) = \int_{u=0}^{r} u^{d+1} f(u)\mathrm{d}u$
$r_h \in \mathbb{R}_+$ such that $h = \frac{M_{d-1}^{r_h}(f)}{M_{d-1}(f)}$.

$$\mathscr{S}^*_{d,d',\sigma^2_{\min}} = \{\Sigma = PDP^T \mid PP^T = I_d,$$
$$D = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d'}, \sigma^2 \ldots \sigma^2),$$
$$\lambda_1 \geq \ldots \geq \lambda_{d'} \geq \sigma^2 \geq \sigma^2_{\min}\}.$$

P sub-Gaussian vith variance V^2 : $\left(\forall r > V, P(\mathrm{B}(0,r)^c) \leq \exp\left(-\frac{r^2}{2V^2}\right)\right)$

$$R_h(\theta) = P\|\cdot - m_{\theta,h}\|^2_{\Sigma^{-1}} + v^{\Sigma}_{\theta,h} + \log(\det(\Sigma))$$

$\theta^*_{d'}$ (resp. $\hat{\theta}_{d'}$) an R_h-minimizer (resp. $\hat{R}_h$) in $\mathbb{R}^d \times \mathscr{S}^*_{d,d',\sigma^2_{\min}}$.

Theorem (B. - Levrard - Michel, 2020)

$$\mathbb{E}\left[R_h(\hat{\theta}_{d'}) - R_h(\theta^*_{d'})\right] \leq \frac{CV^2}{h\sigma^2_{\min}}\sqrt{\mathscr{D}_{d'}}\frac{\log n}{\sqrt{n}}$$

for some absolute constant $C > 0$, with

$$\mathscr{D}_{d'} = \max\left(d'\left(d - \frac{d'+1}{2}\right)\log(d'), (d'+1)\left(d + 1 - \frac{d'}{2}\right)\right).$$

$$\mathscr{C}^*_{d,d'} = \{\{x \mid \|x-c\|^2_{\Sigma^{-1}} \leq r\}, \Sigma \in \mathscr{S}^*_{d,d',0}, c \in \mathbb{R}^d, r > 0\}.$$

Lemma (B. - Levrard - Michel, 2020)

The Vapnik-Chervonenkis dimension of $\mathscr{C}^*_{d,d'}$ is bounded above by

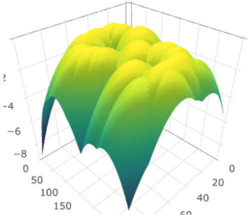
$$(12.416...)\left(d+1-\frac{d'}{2}\right)(d'+1).$$

(Dudley, 1979)

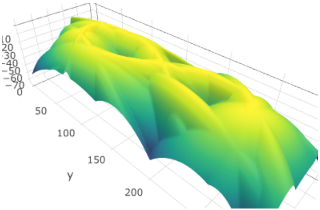
$$VC(\mathscr{C}^*_{d,0}) = d+2$$

(Akamaa, Irie, 2011)

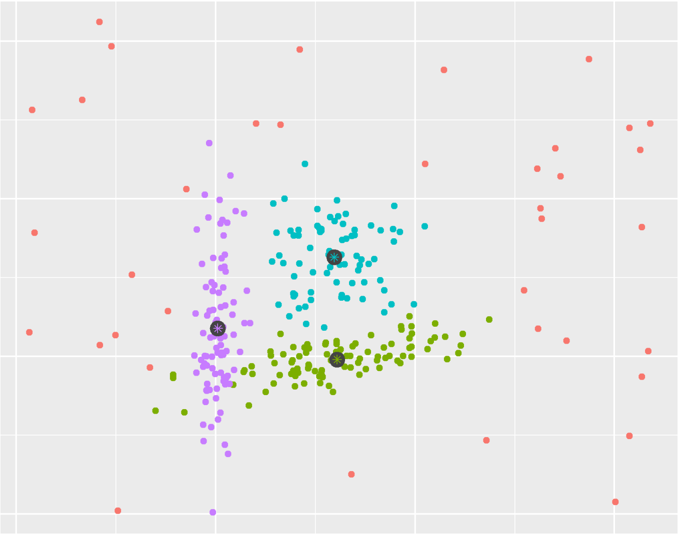
$$VC(\mathscr{C}^*_{d,d}) = \frac{d^2+3d}{2}$$



k-PDTM



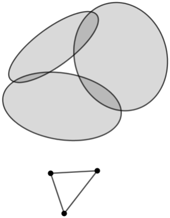
k-PLM

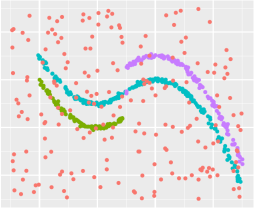


We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.

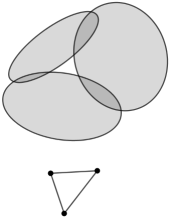


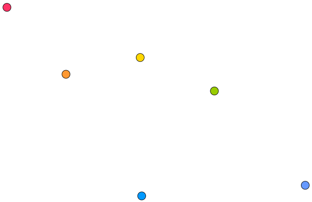


We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.

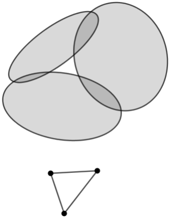


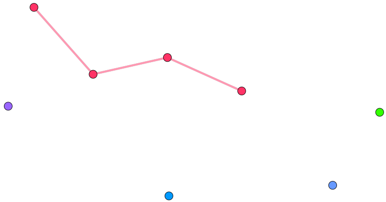


We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.



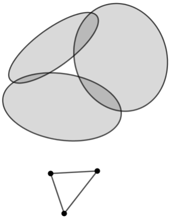


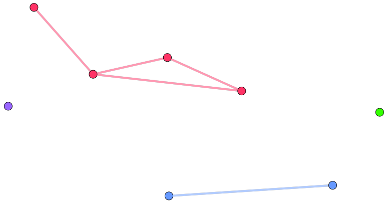
Clustering with unions of ellipsoids

We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.



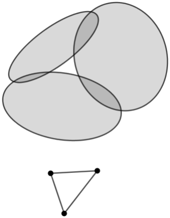


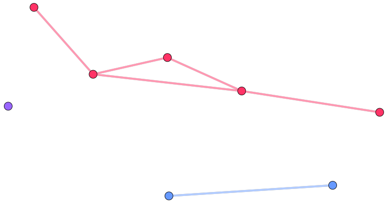
Clustering with unions of ellipsoids

We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.

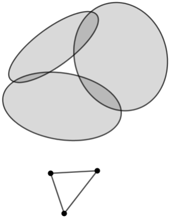


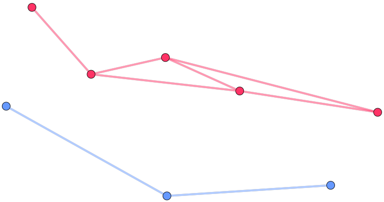


We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.



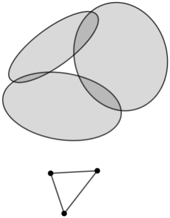


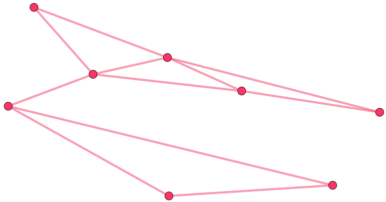
Clustering with unions of ellipsoids

We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.

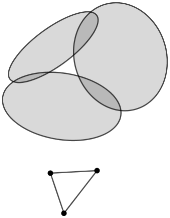


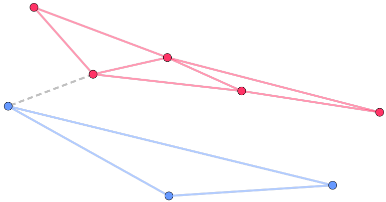


We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.



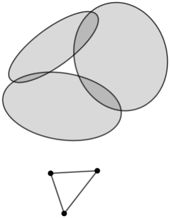


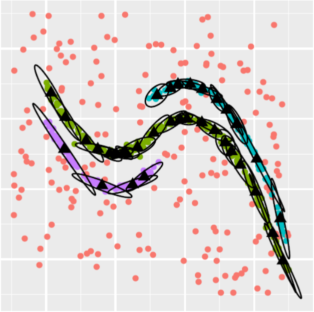
Clustering with unions of ellipsoids

We consider unions of ellipsoids indexed by $\alpha \in \mathbb{R}$:

$$\bigcup_{i=1..k} \{x \mid \|x - c_i\|^2_{\Sigma_i^{-1}} + \omega_i \leq \alpha\},$$

with $\omega_i \in \mathbb{R}$, $c_i \in \mathbb{R}^d$ and covariance matrices Σ_i.





Partitionnement (B. 2020)

Thank you !