Hyperspectral Image Segmentation by Spatialized Gaussian Mixtures and Model Selection

E. Le Pennec
(SELECT - Inria Saclay / Université Paris Sud)
and
S. Cohen (IPANEMA - CNRS / Soleil)

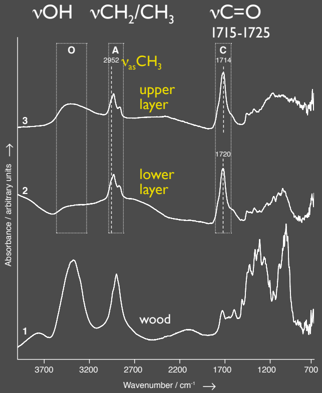
Parietal / Neurospin

Inria informatiques mathématiques

A. Stradivari (1644 - 1737)

Provigny (1716)

A. Giordan © Cité de la Musique

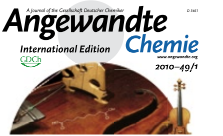


SOLEIL SYNCHROTRON

4 / 8 cm^{-1} resolution
64 / 128 scans
typ. 1 min/sp, 400sp

very simple process
no protein (amide I, amide II)
no gums, nor waxes
@SOLEIL: SMIS

cité de la musique · SOLEIL SYNCHROTRON · CRCC · CENTRE DE RECHERCHE ET DE RESTAURATION DES MUSÉES DE FRANCE · Laboratoire de Dynamique, Interactions et Réactivité



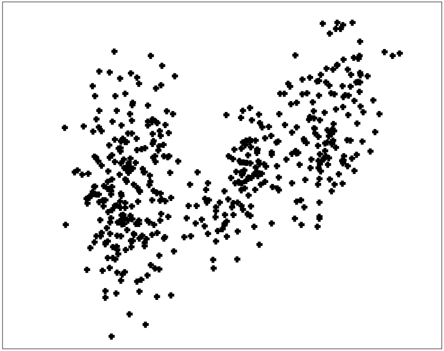
J.-P. Echard, L. Bertrand, A. von Bohlen, A.-S. Le Hô, C. Paris, L. Bellot-Gurlet, B. Soulier, A. Lattuati-Derieux, S. Thao, L. Robinet, B. Lavédrine, and S. Vaiedelich. *Angew. Chem. Int. Ed.*, 49(1), 197-201, 2010.

Hyperspectral Image Segmentation

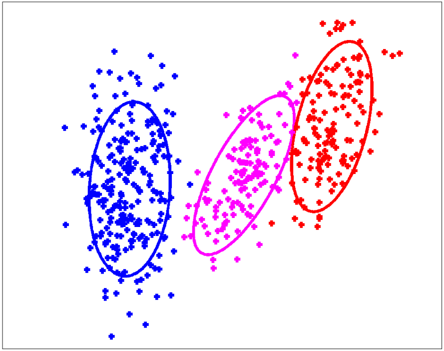
- Data :
  - image of size N between ~ 1000 and ~ 100000 pixels,
  - spectrums $\mathcal{S}$ of ~ 1024 points,
  - very good spatial resolution,
  - ability to measure a lot of spectrums per minute,
- Immediate goal :
  - automatic image segmentation,
  - without human intervention,
  - help to data analysis.
- Advanced goal :
  - automatic classification,
  - interpretation...

A “Toy” Problem

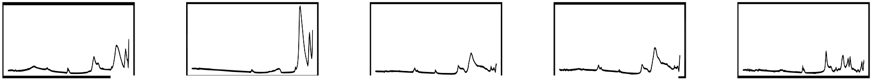
A "Toy" Problem

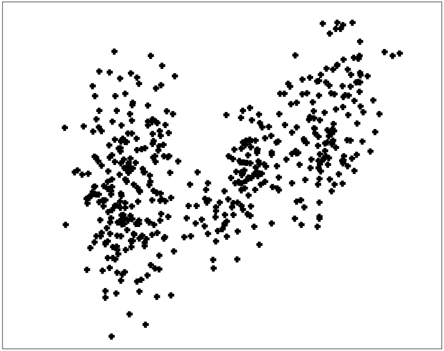


A “Toy” Problem

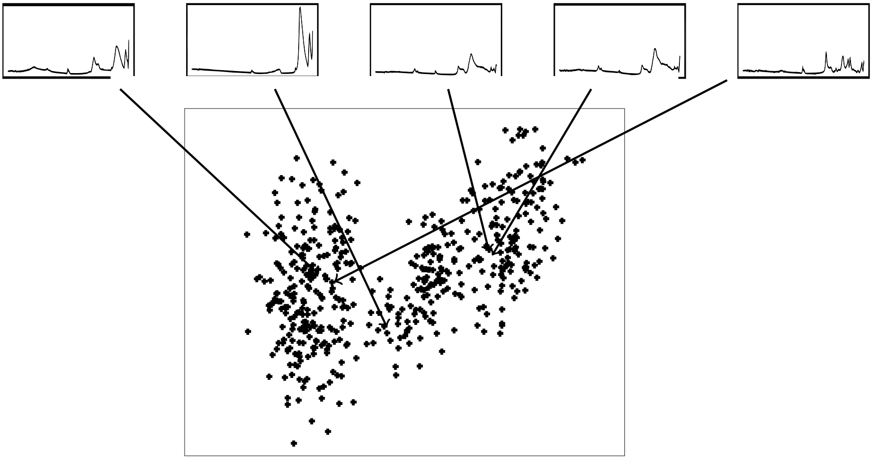


A “Toy” Problem

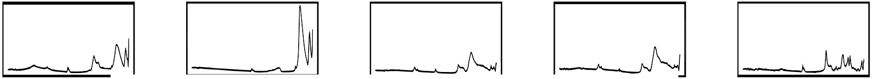


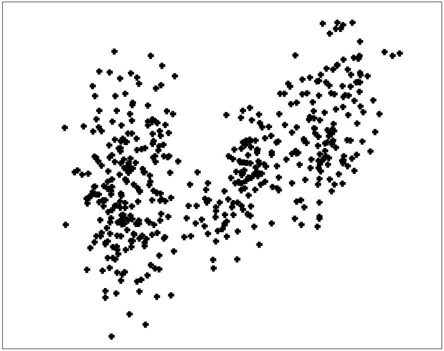


A “Toy” Problem

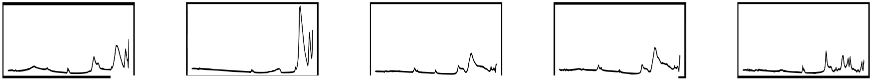


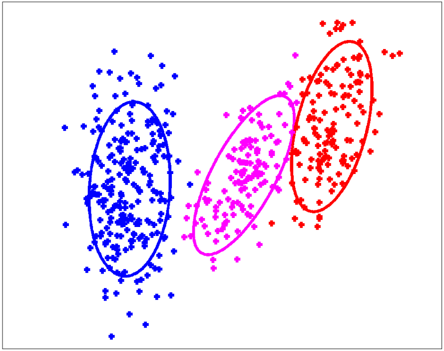
A “Toy” Problem



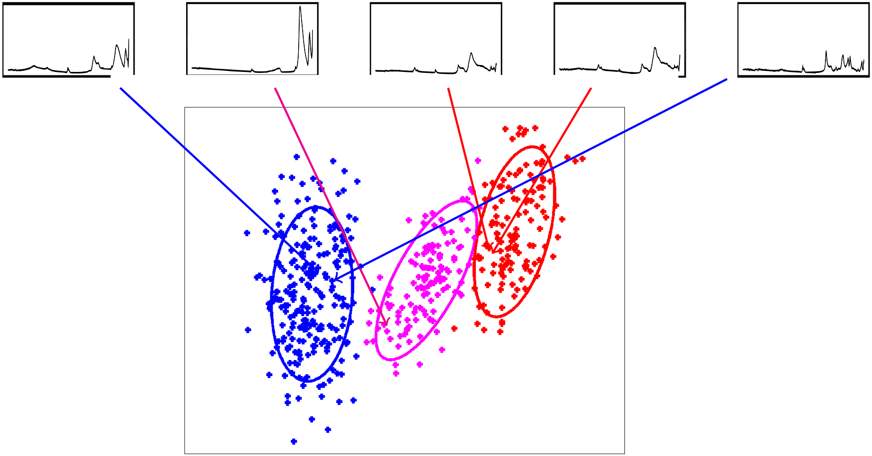


A "Toy" Problem





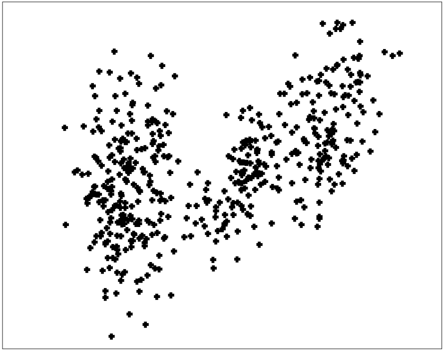
A “Toy” Problem



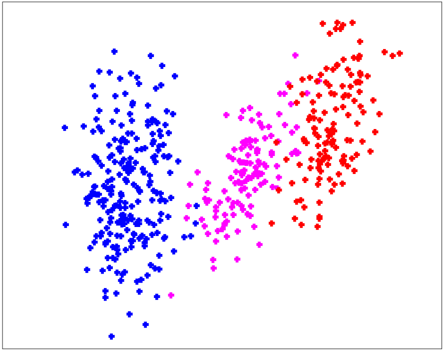
- Representation : mapping between spectrums and points in a large dimension space.
- Spectral method.

“Stochastic” Modeling

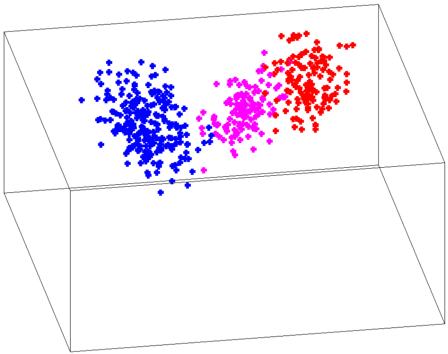
“Stochastic” Modeling



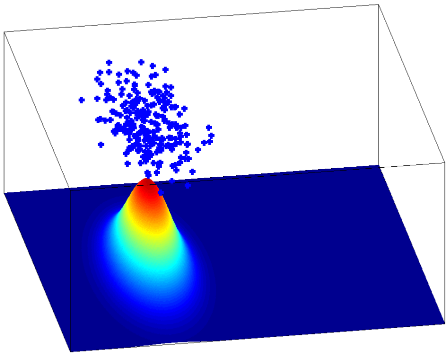
“Stochastic” Modeling



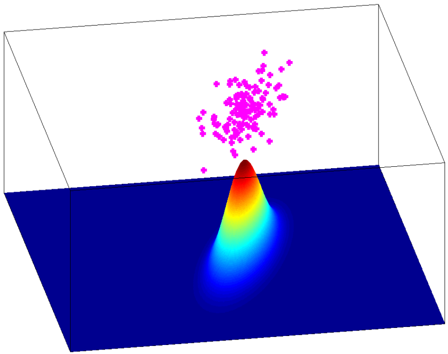
“Stochastic” Modeling



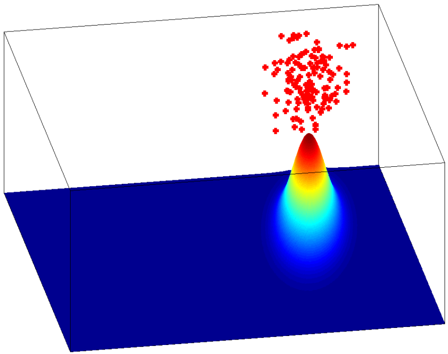
"Stochastic" Modeling



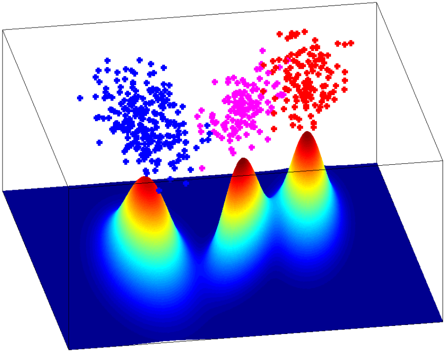
“Stochastic” Modeling



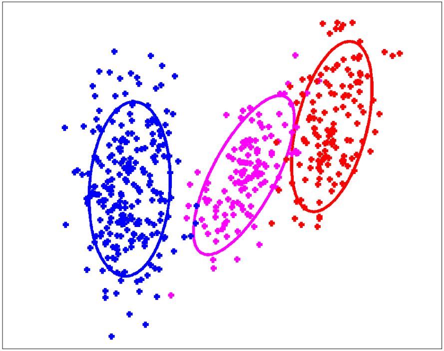
“Stochastic” Modeling



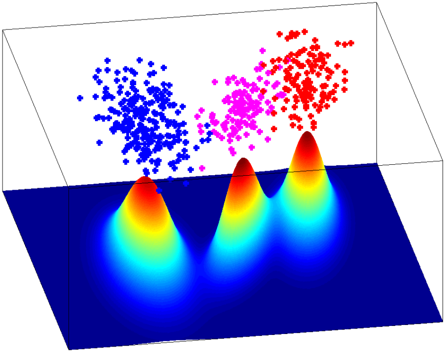
“Stochastic” Modeling



"Stochastic" Modeling



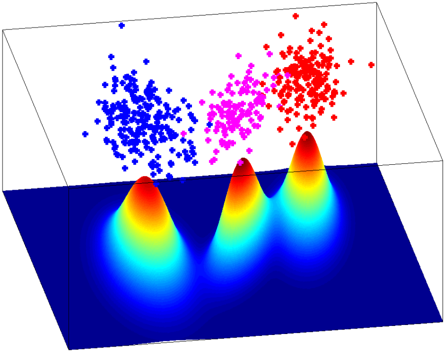
“Stochastic” Modeling



- Model : Gaussian Mixture with K classes.
- Mixture density :

$$
\begin{aligned}
s_{K,\pi,\mu,\Sigma}(\mathcal{S}) &= \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1} (\mathcal{S}-\mu_k)} \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})
\end{aligned}
$$

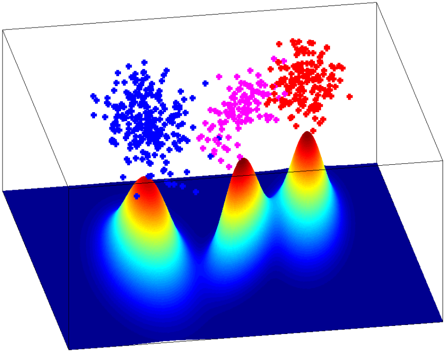
"Stochastic" Modeling



- Model : Gaussian Mixture with K classes.
- Mixture density :

$$
\begin{aligned}
s_{K,\pi,\mu,\Sigma}(\mathcal{S}) &= \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1} (\mathcal{S}-\mu_k)} \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})
\end{aligned}
$$

“Stochastic” Modeling

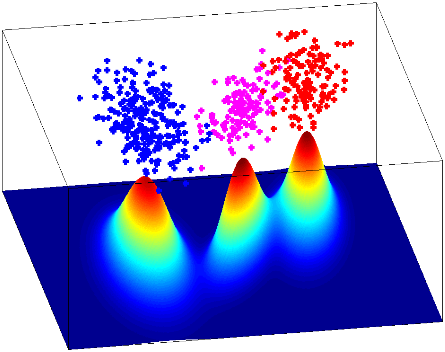


- Model : Gaussian Mixture with K classes.
- Mixture density :

$$s_{K,\pi,\mu,\Sigma}(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1} (\mathcal{S}-\mu_k)}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$$

"Stochastic" Modeling



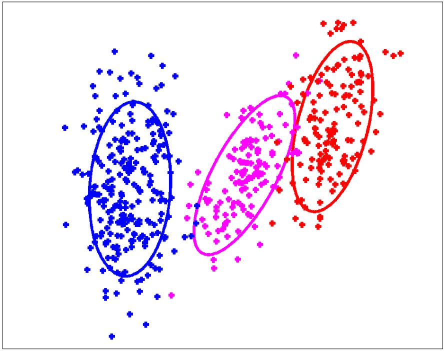
- Model : Gaussian Mixture with K classes.
- Mixture density :

$$s_{K,\pi,\mu,\Sigma}(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1} (\mathcal{S}-\mu_k)}$$

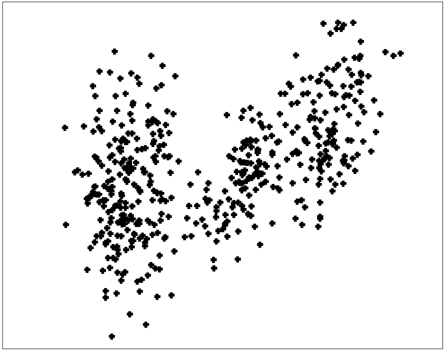
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$$

“Statistical” Estimation

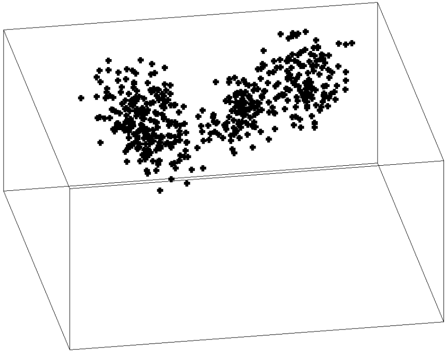
“Statistical” Estimation



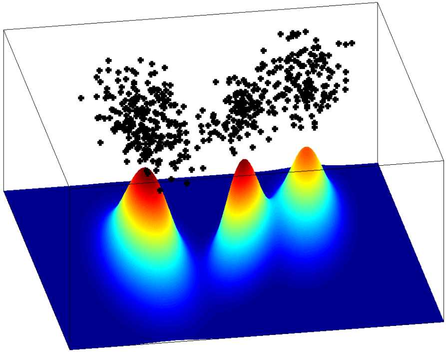
"Statistical" Estimation



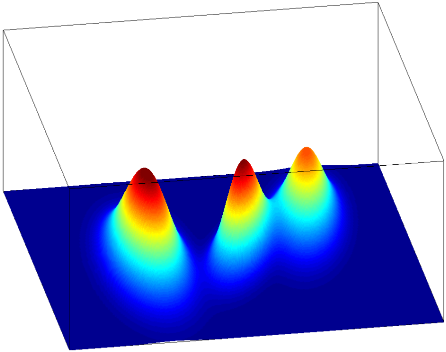
"Statistical" Estimation



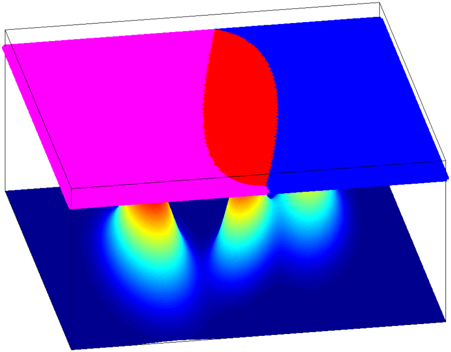
“Statistical” Estimation



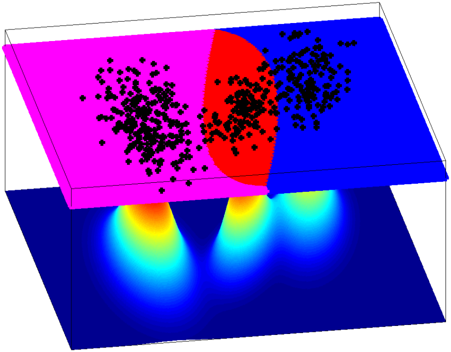
“Statistical” Estimation



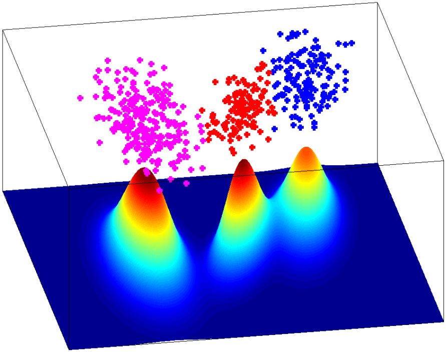
“Statistical” Estimation



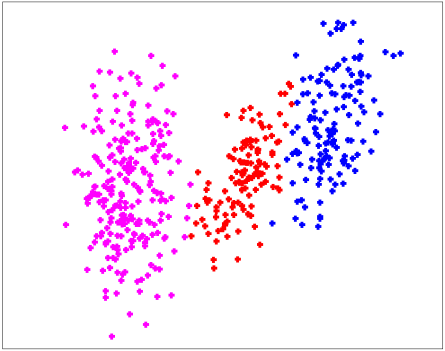
"Statistical" Estimation



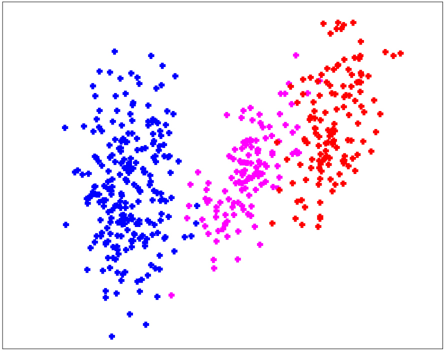
“Statistical” Estimation



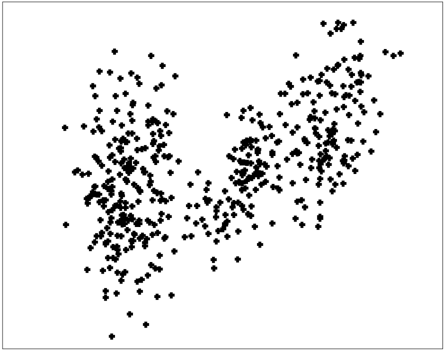
“Statistical” Estimation



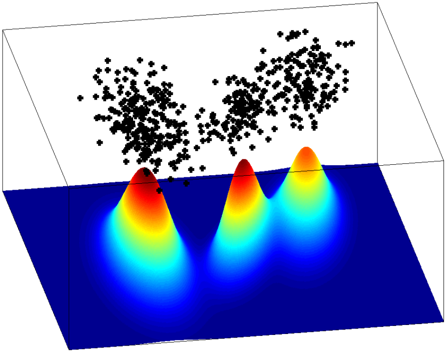
“Statistical” Estimation



"Statistical" Estimation



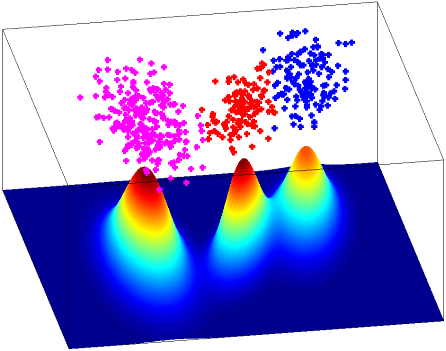
"Statistical" Estimation



- Estimation of π_k, $\widehat{\mu_k}$ and $\widehat{\Sigma_k}$ by maximum likelihood :

$$(\widehat{\pi_k}, \widehat{\mu_k}, \widehat{\Sigma_k}) = \operatorname{argmax} \sum_{i=1}^{N} \ln s_{K,(\pi_k,\mu_k,\Sigma_k)}(\mathcal{S}_i)$$

"Statistical" Estimation



- Estimation of π_k, $\widehat{\mu_k}$ and $\widehat{\Sigma_k}$ by maximum likelihood :

$$(\widehat{\pi_k}, \widehat{\mu_k}, \widehat{\Sigma_k}) = \operatorname{argmax} \sum_{i=1}^{N} \ln s_{K,(\pi_k,\mu_k,\Sigma_k)}(\mathcal{S}_i)$$

- Estimation of $\widehat{k}(\mathcal{S})$ by maximum a posteriori (MAP) :

$$\widehat{k}(\mathcal{S}) = \operatorname{argmax} \widehat{\pi_k} \mathcal{N}_{\widehat{\mu_k},\widehat{\Sigma_k}}(\mathcal{S})$$

Hyperspectral image segmentation with GMM

- *Classical* stochastic model of spectrum $\mathcal{S}$:
  - K spectrum classes,
  - with proportion π_k for each class ($\sum_{k=1}^{K} \pi_k = 1$),
  - Gaussian law $\mathcal{N}_{\mu_k,\Sigma_k}$ within each class (strong assumption !)
- Heuristic : true density s_0 of $\mathcal{S}$ close from

$$s(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}).$$

- Goal : estimate all parameters (K, π_k, μ_k and Σ_k) from the data.
- Why : yields a classification/segmentation by a maximum likelihood principle

$$\widehat{k}(\mathcal{S}) = \operatorname{argmax} \widehat{\pi_k} \mathcal{N}_{\widehat{\mu_k},\widehat{\Sigma_k}}(\mathcal{S})$$

- Typical result in term of density estimation and not classification...

Gaussian Mixture Model

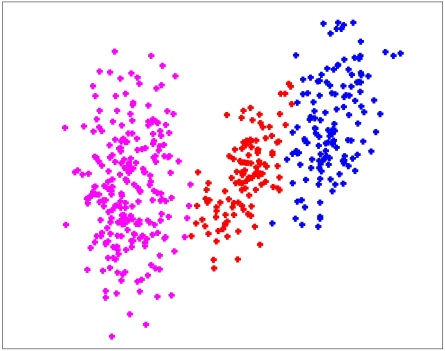
- True density s_0 of $\mathcal{S}$ close from

$$s(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}).$$

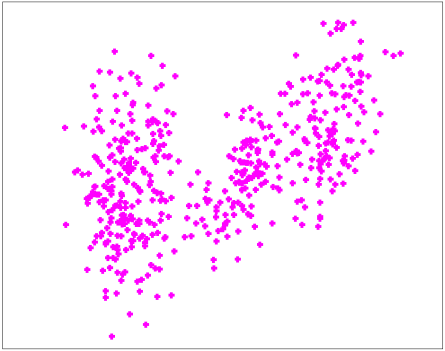
- Gaussian Mixture Model $S_m = \{s_m\}$ specified by
  - a number of classes K,
  - a structure for the means μ_k and the covariance matrices $\Sigma_k = L_k D_k A_k D'_k$ (Volume L_k, basis D_k and rescaled eigenvalues A_k)
- Structure $[\mu \, L \, D \, A]^K$ for the K-tuples of Gaussian parameters :
  - know, common or free values for each parameter
  - plus compactness and condition number assumptions.
- GMM S_m : parametric model of dimension $(K-1) + \dim([\mu \, L \, D \, A]^K)$.
- Maximum likelihood estimation by EM algorithm of :
  - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D'_k$ for each class
  - and the mixing proportions π_k

How many classes?

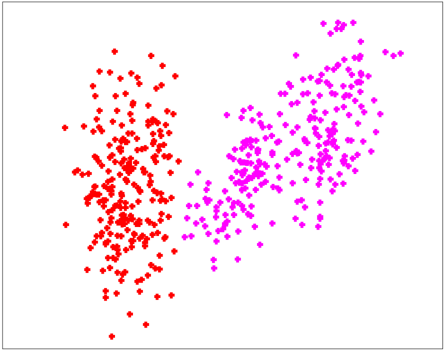
How many classes ?



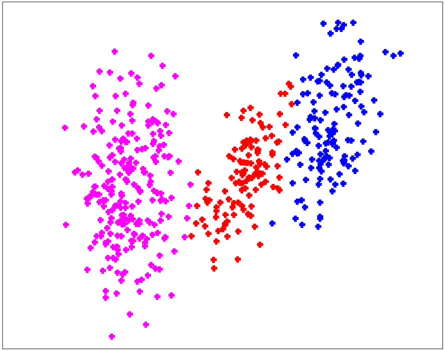
How many classes ?



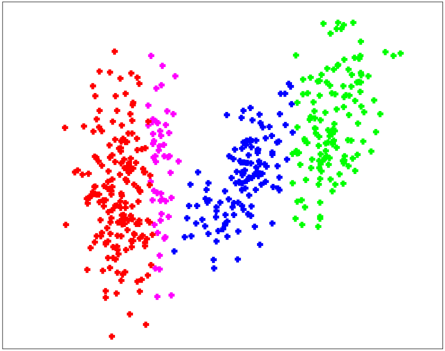
How many classes?



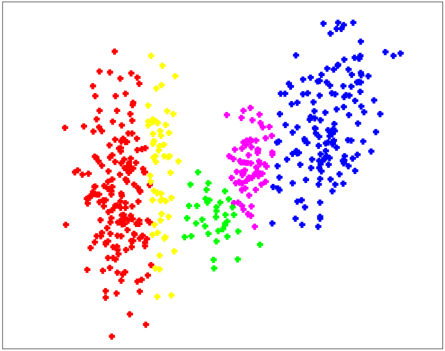
How many classes ?



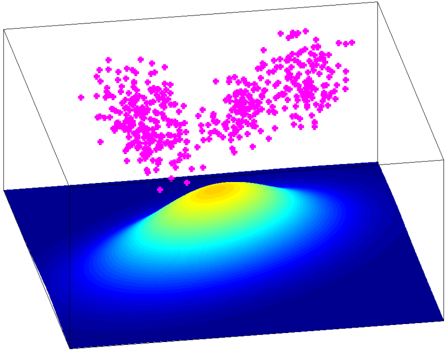
How many classes ?



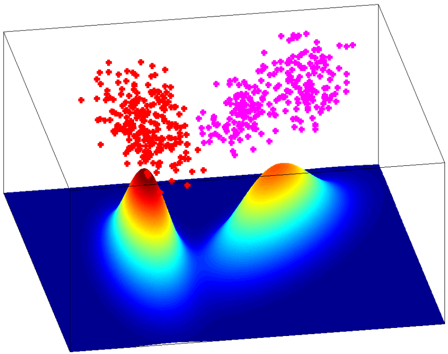
How many classes ?



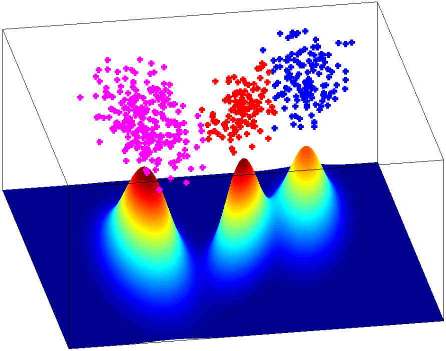
How many classes ?



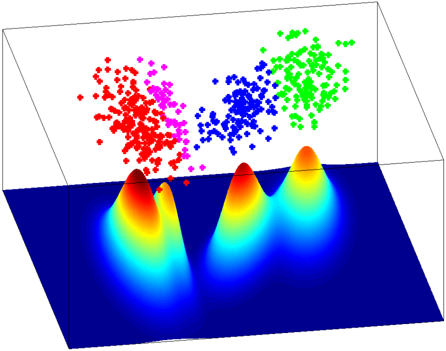
How many classes?



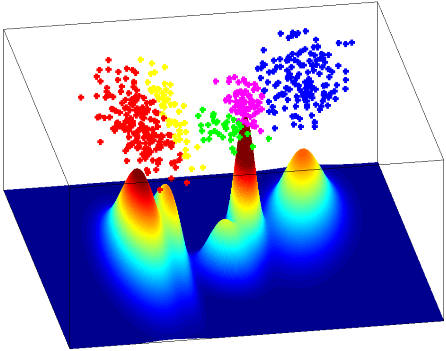
How many classes ?



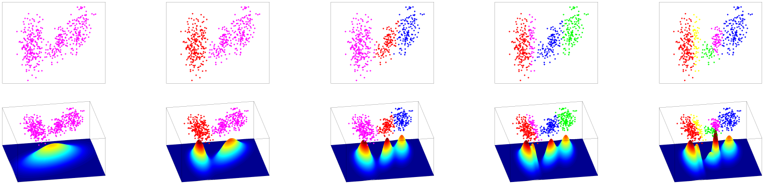
How many classes ?



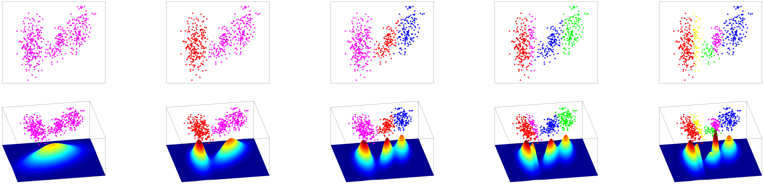
How many classes ?



How many classes ?

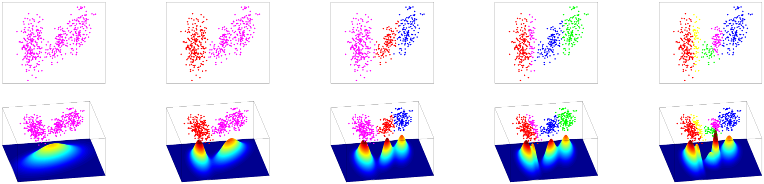


How many classes ?



Fidelity	−−	+	+++	+++·	++++

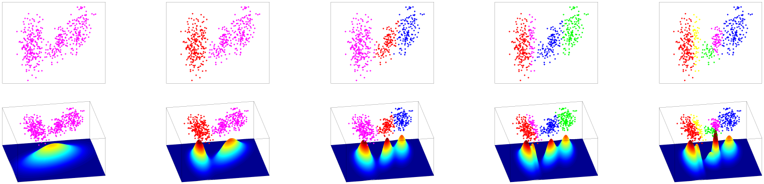
How many classes ?



Fidelity	$--$	$+$	$+++$	$+++\cdot$	$++++$

- Tough question for which the likelihood (the fidelity) is not sufficient !

How many classes ?



Fidelity	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicity	$++++$	$+++$	$++$	$+$	$-$

- Tough question for which the likelihood (the fidelity) is not sufficient !
- How to take into account the model complexity ?

Ockham's Razor

Ockham's Razor

entities must not be multiplied beyond necessity

William of Ockham ($\sim$ 1285 - 1347)

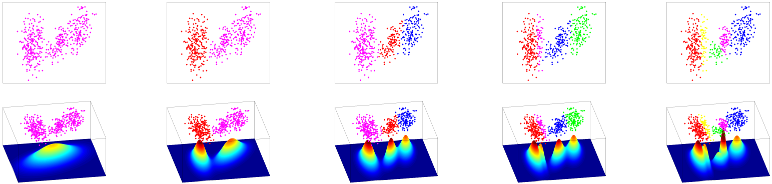
Ockham's Razor

entities must not be multiplied beyond necessity
William of Ockham ($\sim$ 1285 - 1347)

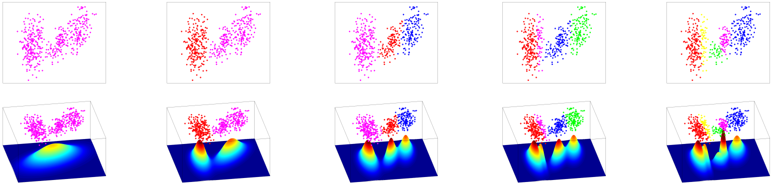
- Ockham's Razor (simplicity principle) : one should not add hypotheses, if the current ones are already sufficient !
- Balance between observation explanation power and simplicity.

Selection by Penalization

Selection by Penalization

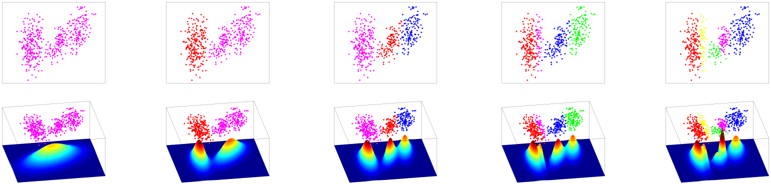


Selection by Penalization



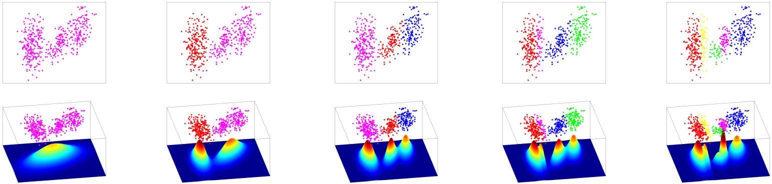
Likelihood	−−	+	+ + +	+ + +·	+ + ++
Simplicity	+ + ++	+ + +	++	+	−

Selection by Penalization



Likelihood	$--$	$+$	$+ + +$	$+ + +\cdot$	$+ + ++$
$+$ Simplicity	$+ + ++$	$+ + +$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$+ + ++$	$+ + + + +$	$+ + + + \cdot$	$+ + +$

Selection by Penalization

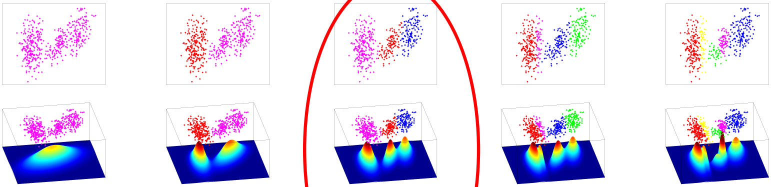


Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

- Likelihood : $\sum_{i=1}^{N} \ln \hat{s}_K(X_i)$.
- Simplicity : $-\lambda \mathrm{Dim}(S_K)$.
- Penalized estimator :

$$\operatorname{argmin} \underbrace{-\sum_{i=1}^{N} \ln \hat{s}_K(X_i)}_{\text{Likelihood}} + \underbrace{\lambda \mathrm{Dim}(S_K)}_{\text{Penalty}}$$

Selection by Penalization

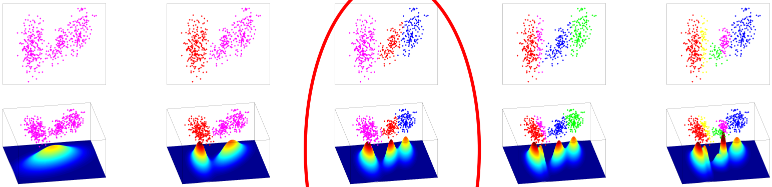


Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

- Likelihood : $\sum_{i=1}^{N} \ln \hat{s}_K(X_i)$.
- Simplicity : $-\lambda \mathrm{Dim}(S_K)$.
- Penalized estimator :

$$\operatorname{argmin} \underbrace{-\sum_{i=1}^{N} \ln \hat{s}_K(X_i)}_{\text{Likelihood}} + \underbrace{\lambda \mathrm{Dim}(S_K)}_{\text{Penalty}}$$

Selection by Penalization



Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

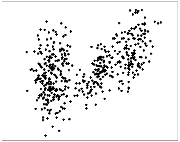
- Likelihood : $\sum_{i=1}^{N} \ln \hat{s}_K(X_i)$.
- Simplicity : $-\lambda \mathrm{Dim}(S_K)$.
- Penalized estimator :

$$\operatorname{argmin} \underbrace{-\sum_{i=1}^{N} \ln \hat{s}_K(X_i)}_{\text{Likelihood}} + \underbrace{\lambda \mathrm{Dim}(S_K)}_{\text{Penalty}}$$

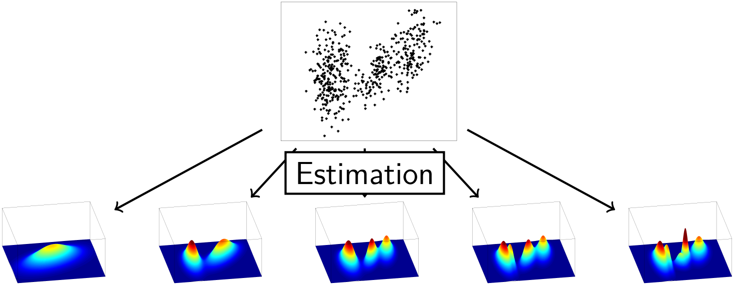
- Optimization in K by exhaustive exploration !

Methodology

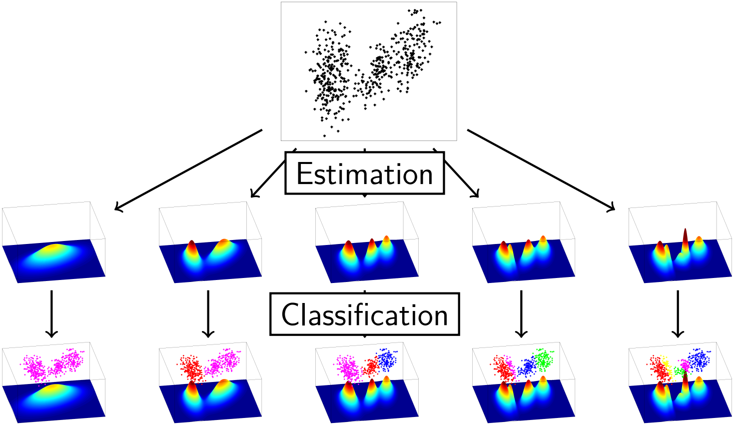
Methodology



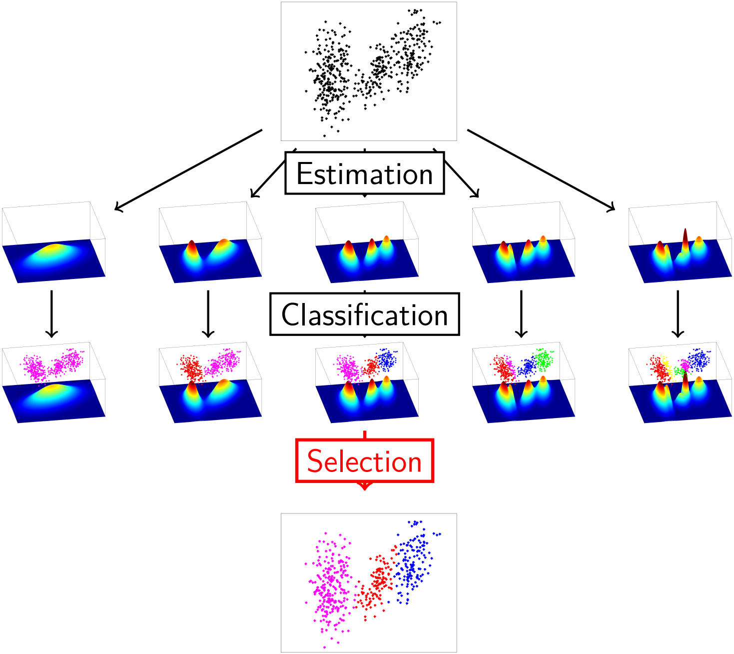
Methodology



Methodology



Methodology



Model selection

- How to choose the *good* model S_m :
  - the number of classes K,
  - the structure model $[\mu\, L\, D\, A]^K$?
- Penalized model selection principle :
  - Choice of a collection of models $S_m = \{s_m\}$ with $m \in \mathcal{S}$,
  - Maximum likelihood estimation of a density $\hat{s}_m$ for each model S_m,
  - Selection of a model $\widehat{m}$ by

$$\widehat{m} = \mathrm{argmin} - \ln(\widehat{s}_m) + \mathrm{pen}(m).$$

  with $\mathrm{pen}(m) = \kappa(\ln(n))\,\mathrm{dim}(S_m)$ (parametric dimension of S_m),
- Results (Birgé, Massart, Celeux, Maugis, Michel...) :
  - Density estimation : for κ large enough,

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

  - Clustering or unsupervised classification : numerical results.
  - Consistency of the classification as soon as $\ln \ln(n)$ in the penalty...

Probabilistic *distances*

- For sake of simplicity $dP = p d\mu$ and $dQ = q d\mu$
- Total variation :

$$TV(P,Q) = \sup \sum_i |P(A_i) - Q(A_i)| \quad \text{with } A_i \text{ disjoint measurable sets}$$

$$TV(P,Q) = 2 \sup_A |P(A) - Q(A)| = \int |p - q| d\mu \leq 2$$

- Hellinger :

$$d^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu \leq 2$$

- Kullback-Leibler divergence :

$$KL(P,Q) = \begin{cases} \int \ln(p/q) p d\mu & \text{if } q = 0 \implies p = 0 \\ +\infty & \text{otherwise} \end{cases}$$

Total variation

- Total variation :

$$TV(P,Q) = \sup \sum_i |P(A_i) - Q(A_i)| \quad \text{with } A_i \text{ disjoint measurable sets}$$

$$TV(P,Q) = 2 \sup_A |P(A) - Q(A)| = \int |p - q| d\mu \leq 2$$

- True distance !
- Associated affinity :

$$\mathcal{A}_{TV}(P,Q) = \int p \wedge q d\mu$$

- Prop : $TV(P,Q) = 2(1 - \mathcal{A}_{TV}(P,Q))$
- Minimax test : $\mathcal{A}_{TV}(P,Q) = \inf_{\Psi \in [0,1]} \int \Psi dP + \int (1 - \Psi) dQ$
- No simple formula when $P = \bigotimes P_i$ and $Q = \bigotimes Q_i$
- No natural M-estimator.

Hellinger distance

- Hellinger :

$$d^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2 d\mu \leq 2$$

- $d(P,Q)$ is a true distance.
- Associated affinity :

$$\mathcal{A}_d(P,Q) = \int \sqrt{p}\sqrt{q} d\mu$$

- Prop : $d^2(P,Q) = 2(1 - \mathcal{A}_d(P,Q)) d\mu$
- Prop : $d^2(P,Q) \leq TV(P,Q) \leq d(P,Q)\sqrt{4 - d^2(P,Q)}$
- Prop : $\mathcal{A}_d(\bigotimes_i P_i, \bigotimes_i Q_i) = \Pi_i \mathcal{A}_d(P_i, Q_i)$
- Battacharaya distance :
  $\mathcal{D}(P,Q) = -\ln \mathcal{A}_d(P,Q) \, (= \sum_i \mathcal{A}_d(P_i, Q_i))$
- No natural M-estimator

Kullback-Leibler divergence

- Kullback-Leibler divergence :

$$KL(P,Q) = \begin{cases} \int \ln(p/q) p d\mu & \text{if } q = 0 \implies p = 0 \\ +\infty & \text{otherwise} \end{cases}$$

- Not a true distance (lack of symmetry and triangular inequality !)
- Information theory interpretation (coding distance)
- Natural M-estimator (Maximum likelihood !)

$$-\frac{1}{n}\sum_{i=1}^{n} - \ln q(Y_i) \to KL(P,Q) - \int \ln(p) p d\mu$$

- Prop : $d^2(P,Q) \leq KL(P,Q) \leq (2 + \ln \|\frac{p}{q}\|_\infty) d^2(P,Q)$
- Prop : $KL\left(\bigotimes P_i, \bigotimes Q_i\right) = \sum_{i=1}^{n} KL(P_i, Q_i)$
- Issue for empirical control : unboundness...

Jensen-Kullback-Leibler

- Jensen-Kullback-Leibler : convexification
  $JKL_\rho(P,Q) = \frac{1}{\rho} KL(P, \rho Q + (1-\rho)P) \le -\frac{1}{\rho} \ln(1-\rho)$.
- Introduced explicitly with $\rho = 1/2$ by Massart but appears in works by van de Geer.
- Prop : for all $\rho \in (0,1)$

$$C_\rho \, d^2(P,Q) \le JKL_\rho(P,Q) \le KL(P,Q)$$

  with $C_\rho = \frac{1}{\rho} \min(\frac{1-\rho}{\rho}, 1) \left(\ln \left(1 + \frac{\rho}{1-\rho} \right) - \rho \right)$.
- $C_\rho \simeq 1/5$ if $\rho \simeq 1/2$.
- Prop (JKL moment bounds) : If Y of density s_0

$$\mathbb{E} \left(\left| \frac{1}{\rho} \ln \left(\frac{(1-\rho)s_0(Y) + \rho t(Y)}{(1-\rho)s_0(Y) + \rho u(Y)} \right) \right|^k \right) \le \frac{k!}{2} \left(\frac{9 d^2(t,u)}{8\rho(1-\rho)} \right) \left(\frac{2}{\rho} \right)^{k-2}$$

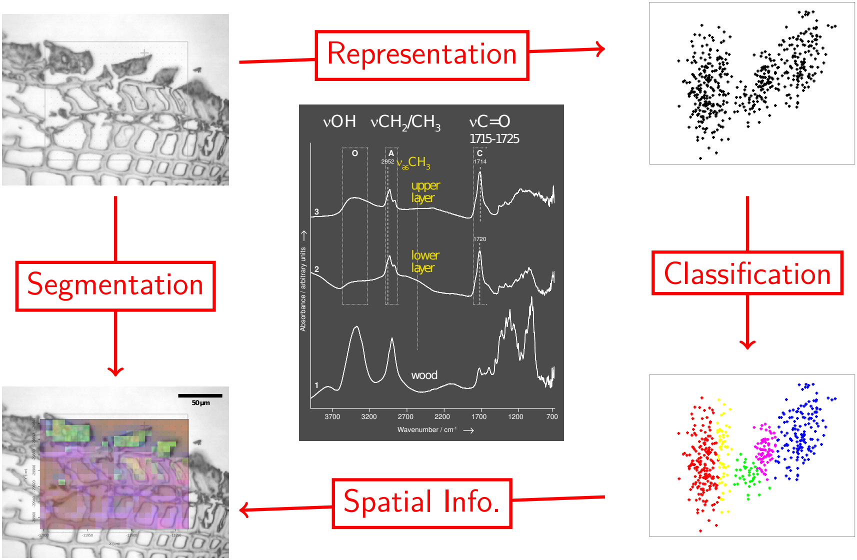
Oracle inequality and distances

- Model selection oracle inequality of type

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \left(\inf_{m \in \mathcal{S}} \inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

- Density : Hellinger $d^2(s, s')$ (or $\mathcal{D}$) (Kolaczyk, Barron, Bigot) on the left...
- Refinement with the *bounded* version of KL : $JKL(s, s') = 2KL(s, (s' + s)/2)$ (Massart, van de Geer)
- Kullback-Leibler divergence on the right...

Back to our violins



Segmentation and Spatialized GMM

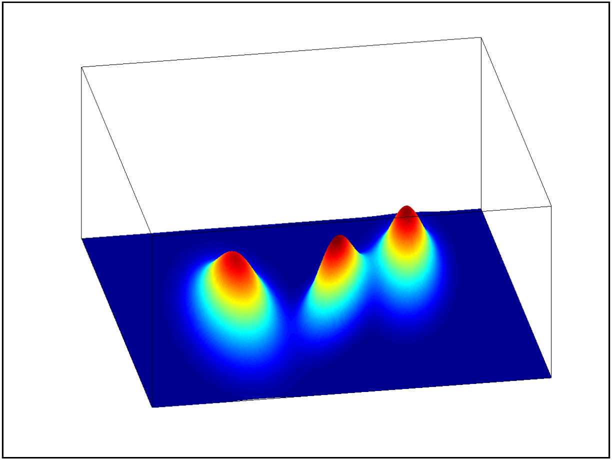
- Initial goal : segmentation $\neq$ clustering.
- Idea of Kolaczyk et al (cf Bigot) : take into account the spatial position x of the spectrum in the mixing proportions.
- Conditional density model :

$$s(\mathcal{S}|x) = \sum_{k=1}^{K} \pi_k(x)\,\mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}).$$

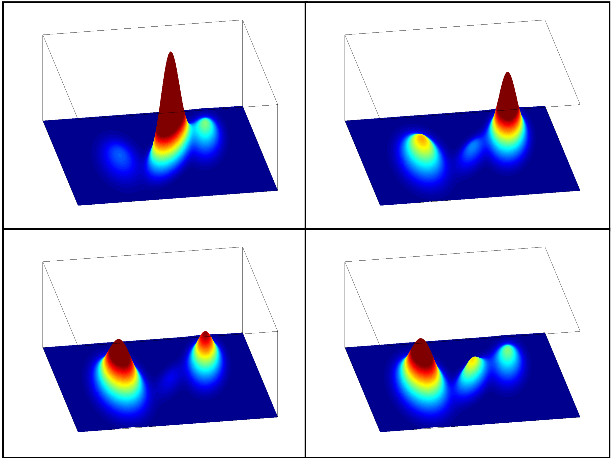
- Estimation from the data :
  - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D'_k$ for each class
  - and the mixing proportion functions $\pi_k(x)$.
- Segmentation by MAP principle :

$$\widehat{k}(\mathcal{S}|x) = \arg\max_k \widehat{\pi_k}(x)\mathcal{N}_{\widehat{\mu_k},\widehat{\Sigma_k}}(\mathcal{S})$$

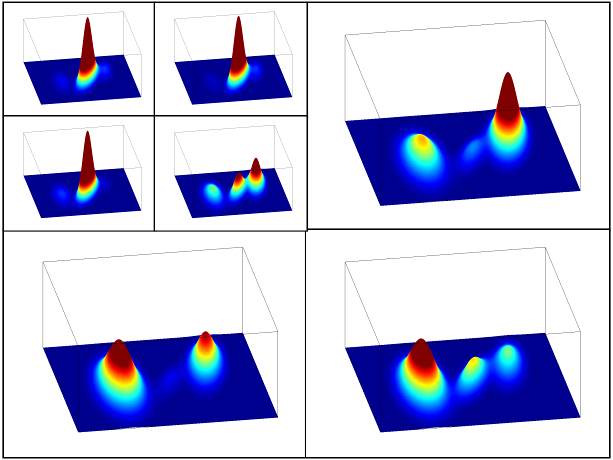
Segmentation and Spatialized GMM



Segmentation and Spatialized GMM

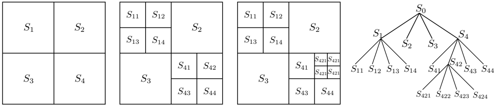


Segmentation and Spatialized GMM



Spat. GMM and hierarchical partition

- How to choose the *right* model S_m ? :
  - the number of classes K,
  - the structure model $[\mu\, L\, D\, A]^K$,
  - the structure of the mixing proportion functions $\pi_k(x)$.
- Simple structure for $\pi_k(x)$: $\pi_k(x) = \sum_{\mathcal{R}\in\mathcal{P}} \pi_k[\mathcal{R}]\chi_{\{x\in\mathcal{R}\}} = \pi_k[\mathcal{R}(x)]$
  - piecewise constant on a *hierarchical* partition,
  - efficient optimization algorithm,
  - good approximation properties.



- $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
- Penalty $\mathrm{pen}(m) = \kappa \ln(n) \dim(S_m)$ allows
  - a numerical optimization scheme (EM + dynamic programing)
  - a theoretical control : for κ large enough

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m\in\mathcal{S}} \left(\inf_{s_m\in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

Conditional density and selection

- General framework : observation of (X_i, Y_i) with X_i independent and Y_i cond. independent of law of density $s_0(y|X_i)$.
- Goal : estimation of $s_0(y|x)$.
- Penalized model selection principle :
  - choice of a collection of cond. dens. models $S_m = \{s_m(y|x)\}$ with $m \in \mathcal{S}$,
  - Maximum likelihood estimation of a cond. density $\hat{s}_m$ for each model S_m :

$$\hat{s}_m = \operatorname*{argmin}_{s_m \in S_m} - \sum_{i=1}^{n} \ln s_m(Y_i|X_i)$$

  - Selection of a model $\widehat{m}$ by

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{S}} - \sum_{i=1}^{n} \ln \widehat{s}_m(Y_i|X_i) + \mathrm{pen}(m).$$

    with $\mathrm{pen}(m)$ *well chosen*.
- Conditional density estimation result of type :

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

- Short biblio : Rosenblatt, Fan et al., de Gooijer and Zerom, Efromovitch, Brunel, Comte, Lacour... / Plugin, direct estimation, L^2, minimax, censure...

Theorem

Assumption (H) : For every model S_m in the collection $\mathcal{S}$, there is a non-decreasing function $\phi_m(\delta)$ such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0, +\infty)$ and for every $\sigma \in \mathbb{R}^+$ and every $s_m \in S_m$

$$\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}\left(\epsilon, S_m(s_m, \sigma)\right)}\, d\epsilon \leq \phi_m(\sigma).$$

Assumption (K) : There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative number such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty$$

Theorem

Assume we observe (X_i, Y_i) with unknown conditional s_0. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ a at most countable collection of conditional density sets. Assume Assumptions (H), (K) and (S) hold.
Let $\widehat{s}_m$ be a δ -log-likelihood minimizer in S_m :

$$\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^n -\ln(s_m(Y_i|X_i))\right) + \delta$$

Then for any $\rho \in (0, 1)$ and any $C_1 > 1$, there is a constant κ_0 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$ $\mathrm{pen}(m) \geq \kappa\left(\mathfrak{D}_m + x_m\right)$ with $\kappa > \kappa_0$
where $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$,
the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ defined by

$$\widehat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) + \mathrm{pen}(m)$$

satisfies $$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1\left(\inf_{S_m\in\mathcal{S}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n}\right) + \frac{\kappa_0\Sigma + \delta}{n}\right).$$

Simplified Theorem...

- Oracle inequality :

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0,\widehat{s}_{\widehat{m}})\right] \leq C_1\left(\inf_{S_m\in\mathcal{S}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0,s_m)+\frac{\mathrm{pen}\, m}{n}\right)+\frac{\kappa_0\Sigma+\delta}{n}\right)$$

  as soon as

$$\mathrm{pen}(m)\geq\kappa\left(\mathfrak{D}_m+x_m\right)\quad \text{with } \kappa>\kappa_0,$$

  where $\mathfrak{D}_m$ measure the complexity of the model S_m (entropy term) and x_m the coding cost within the collection.
- *Distances* used $KL^{\otimes_n}$ and $JKL_\rho^{\otimes_n}$: *tensorized* Kullback divergence and *Jensen-Kullback* divergence.
- $\mathfrak{D}_m$ linked to the *bracketing entropy* of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.
- Often $\mathfrak{D}_m \propto (\ln n)\dim(S_m)$...

Tensorized divergences

- Need to adapt to conditional density design :
  - Divergence on the product density conditioned on the design (Kolaczyk, Bigot).
  - *Tensorization* principle and expectation on the design : design :

$$KL \to KL^{\otimes_n}(s,s') = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} KL\left(s(\cdot|X_i), s'(\cdot|X_i)\right)\right],$$

$$JKL_\rho \to JKL_\rho^{\otimes_n} \quad \text{and} \quad d^2 \to d^{2\otimes_n}.$$

- Much more information using the second approach because losses used are *larger*.
- Ability to handle independent but non i.i.d. case and integrated loss.
- Oracle inequality of type

$$\mathbb{E}\left[JKL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m\in\mathcal{S}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0,s_m) + \frac{\mathrm{pen}(m)}{n}\right) + \frac{C'}{n}.$$

- Classical density estimation theorem if $s(\cdot|X_i) = s(\cdot)$.

Penalty and complexities

- Model selection : $\widehat{m} = \operatorname{argmin} KL^{\otimes_n}(s_0, \widehat{s}_m) + \frac{\operatorname{pen}(m)}{n}$.
- Ideally : $\operatorname{pen}(m)$ should be $n(\mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_m)\right] - KL^{\otimes_n}(s_0, \widehat{s}_m))$
- More reastically : $\operatorname{pen}(m)$ should be of order $\mathbb{E}\left[n(\mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_m)\right] - KL^{\otimes_n}(s_0, \widehat{s}_m))\right]$ (variance term).
- Control in expectation requires a larger $\operatorname{pen}(m)$ with two terms :
  - an intrinsic one related to the complexity of the model,
  - another one related to the complexity of the collection.
- Here :
  - Model complexity : entropic dimension $\mathfrak{D}_m$ defined from the *bracketing entropy* $H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)$ of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$. (Dudley integral and optimization of deviation bounds)
  - Collection (coding) : Kraft type inequality $\sum_{m \in \mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- Classical constraint on the penalty

$$\operatorname{pen}(m) \geq \kappa \left(\mathfrak{D}_m + x_m\right) \quad \text{with } \kappa > \kappa_0.$$

- Often $\mathfrak{D}_m \propto (\ln(n)) \dim(S_m)$ and thus classical penalization by dimension setting...

Sketch of proof

- Close from Theorem 7.11 of Massart's book.
- For all function $g(y)$, let $P_n^{\otimes_n}(g)$ be its empirical process

$$P_n^{\otimes_n}(g) = \frac{1}{n}\sum_{i=1}^{n} g(Y_i),$$

$P^{\otimes_n}(g)$ the expectation of this process

$$P^{\otimes_n}(g) = \mathbb{E}\left[P_n^{\otimes_n}(g)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i')\right]$$

with (Y_i') a phantom sample of same law than (Y_i) but independent and $\nu_n^{\otimes_n}(g) = P_n^{\otimes_n}(g) - P^{\otimes_n}(g)$ the recentred process.
- By definition,

$$KL^{\otimes_n}(s_0, t) = P^{\otimes_n}\left(-\ln\left(\frac{t}{s_0}\right)\right)$$

$$JKL_\rho^{\otimes_n}(s_0, t) = P^{\otimes_n}\left(-\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho t}{s_0}\right)\right)$$

Best(s) model(s)

- Define
  - $\widehat{s}_m = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n}(-\ln s_m) = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n}\left(-\ln \frac{s_m}{s_0}\right)$
  - $\overline{s}_m = \operatorname{argmin}_{s_m \in S_m} P^{\otimes n}\left(-\ln \frac{s_m}{s_0}\right) = \operatorname{argmin}_{s_m \in S_m} KL^{\otimes n}(s_0, s_m)$.
- Let

$$kl(\overline{s}_m) = -\ln\left(\frac{\overline{s}_m}{s_0}\right)$$

$$kl(\widehat{s}_m) = -\ln\left(\frac{\widehat{s}_m}{s_0}\right)$$

$$jkl(\widehat{s}_m) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho\widehat{s}_m}{s_0}\right)$$

- By convexity, $jkl(\widehat{s}_m) = -\frac{1}{\rho}\ln\frac{\rho\widehat{s}_m + (1-\rho)s_0}{s_0} \leq -\ln\frac{\widehat{s}_m}{s_0} = kl(\widehat{s}_m)$

Log-likelihood majorization

- Let $m \in \mathcal{S}$, for all m' such that

$$P_n^{\otimes_n}(kl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes_n}(kl(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n} :$$

$$\begin{aligned}
P_n^{\otimes_n}(jkl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} &\leq P_n^{\otimes_n}(kl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \\
&\leq P_n^{\otimes_n}(kl(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n} \\
&\leq P_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n}
\end{aligned}$$

- This implies

$$\begin{aligned}
&P^{\otimes_n}(jkl(\widehat{s}_{m'})) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \\
&\qquad \leq P^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n}
\end{aligned}$$

Oracle inequality up to deviation

- The previous inequality can be rewritten

$$\begin{aligned} JKL_\rho^{\otimes_n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\bar{s}_m)) \\ \leq KL^{\otimes_n}(s_0, \bar{s}_m) + \frac{\mathrm{pen}(m)}{n} \\ - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n} \end{aligned}$$

- Appear
  - the integrated loss of the estimate in the model m' : $JKL_\rho^{\otimes_n}(s_0, \hat{s}_{m'})$
  - a simple and centered process : $-\nu_n^{\otimes_n}(kl(\bar{s}_m))$,
  - the oracle $KL^{\otimes_n}(s_0, \bar{s}_m) + \frac{\mathrm{pen}(m)}{n}$
  - a random *remainder* $-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n}$
- It turns out that $\mathbb{E}\left[-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))\right]$ can be essentially bounded by $\epsilon JKL_\rho^{\otimes_n}(s_0, \hat{s}_{m'}) + \frac{\mathrm{pen}(m')}{n}$ as soon as $\mathrm{pen}(m') \geq \kappa(n\sigma_{m'}^2 + x_m)$...

Spatialized Gaussian Mixture Case

- Computation of an upper bound of the bracketing entropy possible (cf Maugis et Michel) implying :

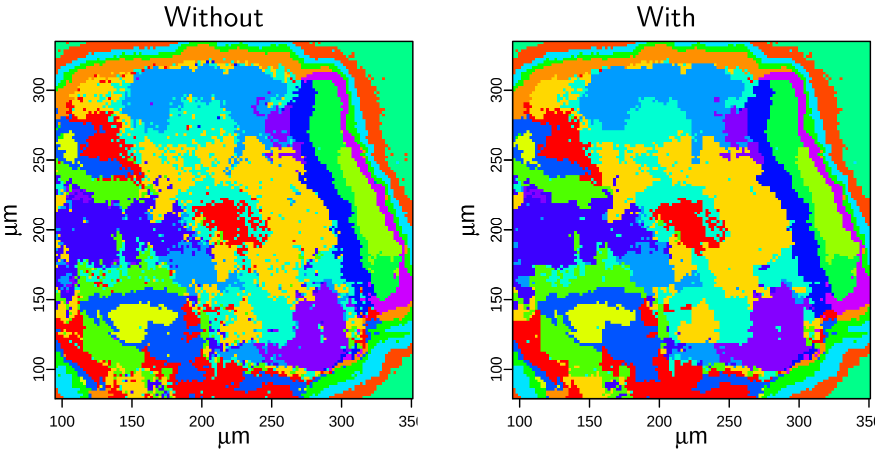
$$\mathfrak{D}_m \leq \kappa' \left(C' + \frac{1}{2} \left(\ln \left(\frac{N}{C' \dim(S_m)} \right) \right)_+ \right) \dim(S_m).$$

- Collection coding with $x_m \leq \kappa'' |\mathcal{P}| \leq \frac{\kappa''}{K-1} \dim(S_m)$.
- Constraint on the penalty :

$$\begin{aligned} \mathrm{pen}(m) &\geq \left(\kappa' \left(C' + \frac{1}{2} \left(\ln \left(\frac{N}{C' \dim(S_m)} \right) \right)_+ \right) + \frac{\kappa''}{K-1} \right) \dim(S_m) \\ &\geq \lambda_{0,N} |\mathcal{P}| (K-1) + \lambda_{1,N} \dim([\mu\, L\, D\, A]^K) \end{aligned}$$

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

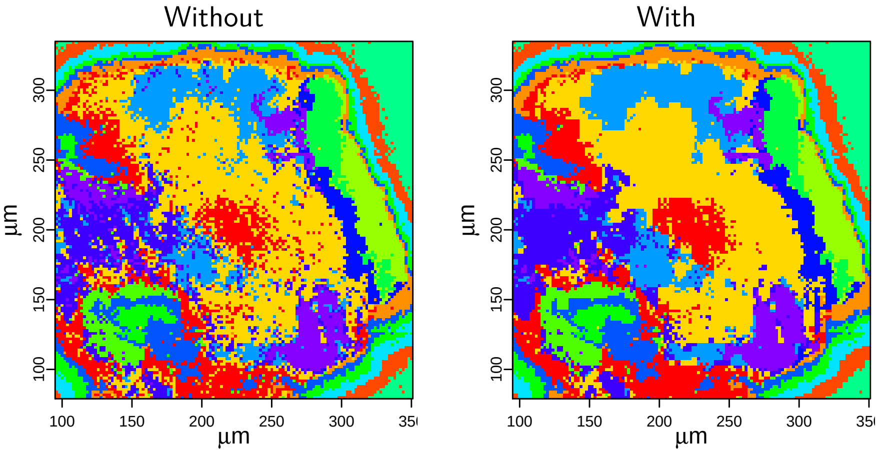


64 scan / 30 min.

- Automatic choice of K, $[L_k\, D\, A]^K$ and partition.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

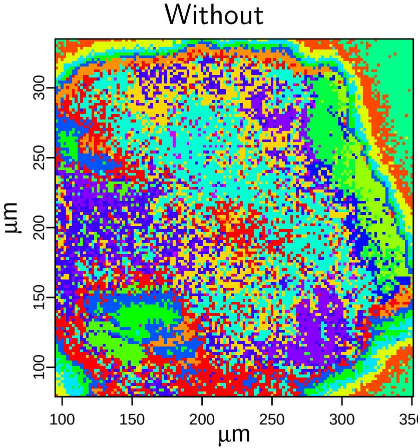


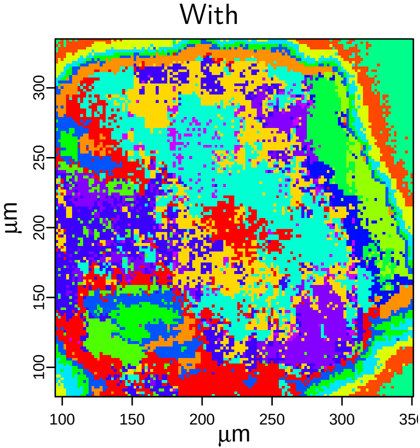
8 scan / 5 min.

- Automatic choice of K, $[L_k \, D \, A]^K$ and partition.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :



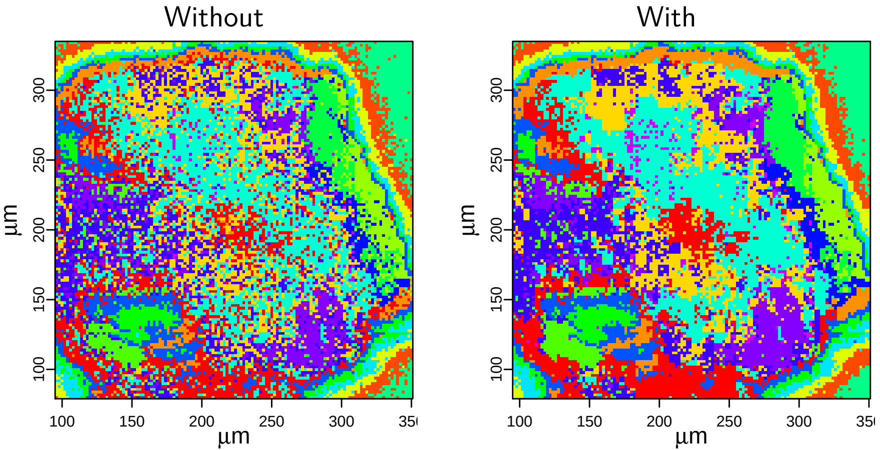


1 scan / 2 min.

- Automatic choice of K, $[L_k\, D\, A]^K$ and partition.

Unsupervised Segmentation

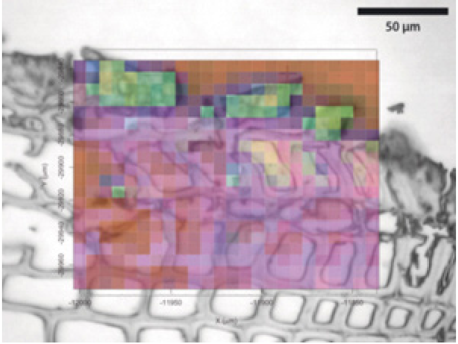
- Numerical result taking into account the spatial modeling :



1 scan / 2 min.

- Automatic choice of K, $[L_k\, D\, A]^K$ and partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by random projection.

Stradivari's Secret



- Two fine layers of varnish :
  - a first simple oil layer, similar to the painter's one, penetrating mildly the wood,
  - a second layer made from a mixture of oil, pine resin and red pigments.
- Classical technique up to the specific color choice (and a very good varnishing skill).
- Stradivari's secret was not his varnish !

Conclusion

- Framework :
  - Unsupervised segmentation problem.
  - Proposed tool : Spatialized Gaussian Mixture Model
  - Penalized maximum likelihood conditional density estimation.
- Results :
  - Theoretical guaranty for the conditional density estimation problem.
  - Direct application to the unsupervised segmentation problem.
  - Efficient minimization algorithm.
  - Unsupervised segmentation algorithm in between *spectral* methods and *spatial* ones.
- Perspectives :
  - Formal link between conditional density estimation and unsupervised segmentation.
  - Penalty calibration by slope heuristic.
  - Dimension reduction adapted to unsupervised segmentation/classification.
  - Enhanced Spatialized Gaussian Mixture Model with piecewise logistic weights (L. Montuelle).

Numerical optimization

- Penalized Model Selection :

$$\underset{K,[\mu\, L\, D\, A]^K,\mu,\Sigma,\mathcal{P},\pi}{\operatorname{argmin}} -\sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k[\mathcal{R}(x_i)]\,\mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i)\right) + \lambda_{0,N}|\mathcal{P}|(K-1) + \lambda_{1,N}\dim([\mu\, L\, D\, A]^K)$$

- Optimization on the number of classes K and the mean and covariance structure by exhaustive exploration.
- Model selection for a given number of classes K and a given structure $[\mu\, L\, D\, A]^K$:

$$\underset{\mu,\Sigma,\mathcal{P},\pi}{\operatorname{argmin}} -\sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k[\mathcal{R}(x_i)]\,\mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i)\right) + \lambda_{0,n}|\mathcal{P}|(K-1)$$

- Two tricks :
  - EM algorithm (a MM algorithm)
  - CART (dynamic programming)

Maximum likelidood estimation

- Model selection for a given number of classes K and a given structure $[\mu\, L\, D\, A]^K$:

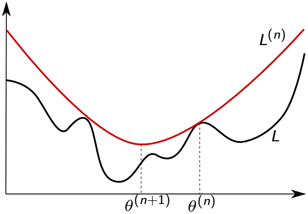
$$\underset{\mu,\Sigma,\mathcal{P},\pi}{\operatorname{argmin}} - \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k[\mathcal{R}(x_i)]\, \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i) \right) + \lambda_{0,n}|\mathcal{P}|(K-1)$$

- If one also fixes $\mathcal{P}$, back to maximum likelihood estimation :

$$\underset{\mu,\Sigma,\pi}{\operatorname{argmin}} - \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k[\mathcal{R}(x_i)]\, \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i) \right)$$

- Non convex minimization problem !
- Majorization/Minimization approach

MM approach



- Iterative approach to minimize $L(\theta)$ by minimizing a sequence of (convex) proxies of L.
- Majorization/Minimization :
  - Current estimate of the minimizer : $\theta^{(n)}$
  - Construction of a Majorization $L^{(n)}$ of L such that $L^{(n)}(\theta^{(n)}) = L(\theta^{(n)})$ with $L^{(n)}$ easy to minimize (convex for example).
  - Computation of a Minimizer

$$\theta^{(n+1)} = \operatorname{argmin} L^{(n)}(\theta)$$

- By construction, $L(\theta^{(n+1)}) \leq L(\theta^{(n)})$!
- Very generic methodology...
- Minimization can be replaced by a diminution...

Maximum Likelihood and EM

- Back to our maximum likelihood for a fixed partition :

$$L(\mu, \Sigma, \pi) = \sum_{i=1}^{N} - \ln \left(\sum_{k=1}^{K} \pi_k [\mathcal{R}(x_i)] \, \mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right)$$

- EM : specific case of MM for this type of mixture.
- (Conditional) Expectation : at step n, we let

$$P_k^{i,(n)} = P \left(k_i = k \middle| \mathcal{S}_i, \mu^{(n)}, \Sigma^{(n}, \pi^{(n)} \right)$$

$$\text{and} \quad L^{(n)}(\mu, \Sigma, \pi) = - \sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\pi_k [\mathcal{R}(x_i)] \mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right).$$

- Majorization prop. : $L \leq L^{(n)} + \mathrm{Cst}^{(n)}$ with equ. at $(\mu^{(n)}, \Sigma^{(n}, \pi^{(n)})$.
- Bonus :
  - Separability in (μ, Σ) and π :

$$L^{(n)}(\mu, \Sigma, \pi) = - \sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right) - \sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\pi_k [\mathcal{R}(x_i)] \right)$$

  - Close formulas for the Minimization of $L^{(n)}$ in (μ, Σ) and π !

Partition and EM Algorithm

- Maximum likelihood :

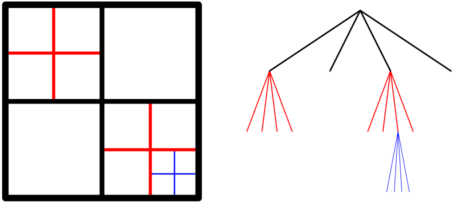
$$L(\mu, \Sigma, \mathcal{P}, \pi) = -\sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k[\mathcal{R}(x_i)]\,\mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i)\right) + \lambda_{0,n}|\mathcal{P}|(K-1)$$

- E Step : with $P_k^{i,(n)} = P(k_i = k | x_i, \mathcal{S}_i, \mu^{(n)}, \Sigma^{(n}, \mathcal{P}^{(n)}, \pi^{(n)})$

$$\begin{aligned} L(\mu, \Sigma, \mathcal{P}, \pi) & \\ & \leq -\sum_{i=1}^{N}\sum_{k=1}^{K} P_k^{i,(n)} \ln\left(\mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S}_i)\right) \\ & \qquad -\sum_{i=1}^{N}\sum_{k=1}^{K} P_k^{i,(n)} \ln\left(\pi_k[\mathcal{R}(x_i)]\right) + \lambda_{0,N}|\mathcal{P}|(K-1) + \mathrm{Cst}^{(n)} \end{aligned}$$

  with equality at $(\mu^{(n)}, \Sigma^{(n)}, \mathcal{P}^{(n)}, \pi^{(n)})$.
- M Step : Separate optimization in (μ, Σ) and $(\mathcal{P}, \pi)$ possible,
  - Optimization in (μ, Σ) : close formulas (classical...).
  - Optimization in $(\mathcal{P}, \pi)$ more interesting !

M Step and CART



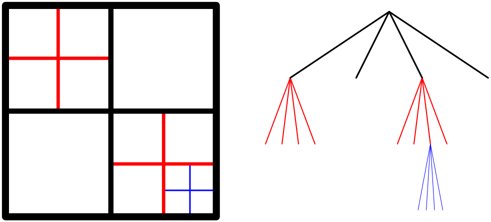
- Optimization in $(\mathcal{P},\pi)$ of

$$-\sum_{i=1}^{N}\sum_{k=1}^{K} P_k^{i,(n)} \ln\left(\pi_k[\mathcal{R}(x_i)]\right) + \lambda_{0,n}|\mathcal{P}|(K-1)$$

$$= -\sum_{\mathcal{R}\in\mathcal{P}}\left(\sum_{i|x_i\in\mathcal{R}}\sum_{k=1}^{K} P_k^{i,(n)} \ln\left(\pi_k[\mathcal{R}(x_i)]\right) + \lambda_{0,N}(K-1)\right)$$

- Two key properties :
  - For each $\mathcal{R}$, simple (classical) optimization of $\pi_k[\mathcal{R}]$.
  - Additivity in $\mathcal{R}$ of the cost structure.
- $\Rightarrow$ Fast CART optimization algorithm (Dynamic programming on tree structure).

CART Optimization



- Aim : compute efficiently $\underset{\mathcal{P}}{\operatorname{argmin}} \sum_{\mathcal{R}\in\mathcal{P}} C[\mathcal{R}]$ where $\mathcal{P}$ is a recursive dyadic partition (associated to quadtree) of limited depth.
- Key property : the optimal partition $\widehat{\mathcal{P}}[\mathcal{R}]$ of a dyadic square is
  - either this square, $\widehat{\mathcal{P}}[\mathcal{R}] = \{\mathcal{R}\}$
  - or the union of the opt. part. of its children, $\widehat{\mathcal{P}}[\mathcal{R}] = \cup_{\mathcal{R}'\in\mathsf{Child}[\mathcal{R})}\widehat{\mathcal{P}}[\mathcal{R}']$

  with a decision based on

$$C[\mathcal{R}] \leq \sum_{\mathcal{R}'\in\mathsf{Child}(\mathcal{R})} \sum_{\mathcal{R}''\in\widehat{\mathcal{P}}[\mathcal{R}']} C[\mathcal{R}'']$$

- Fast bottom-up algorithm.