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)

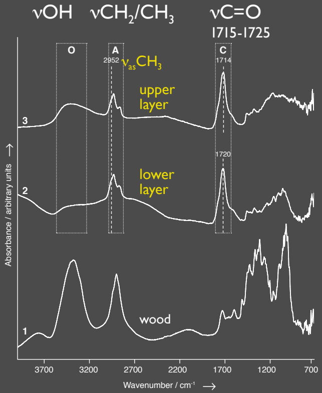
Santa Fe
?? March 2013

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

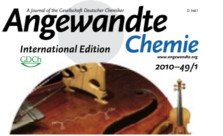
very simple process

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



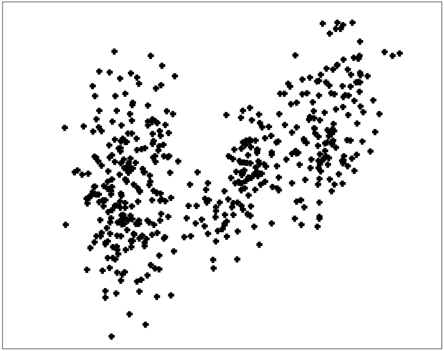
J.-P. Echard, L. Bertrand, A. von Bohlen, A.-S. Le Hô, C. Paris, L. Bellot-Gurlet, B. Soulier, A. Lattuati-Derieux, S. Thao, L. Robinet, B. Lavédrine, and S. Vaiedelich. *Angew. Chem. Int. Ed.*, 49(1), 197-201, 2010.

Hyperspectral Image Segmentation

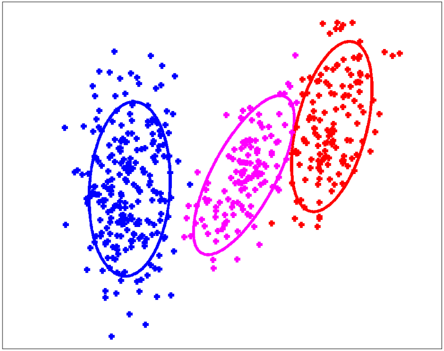
- Data :
  - image of size N between ~ 1000 and ~ 100000 pixels,
  - spectrums $\mathcal{S}$ of ~ 1024 points,
  - very good spatial resolution,
  - ability to measure a lot of spectrums per minute,
- Immediate goal :
  - automatic image segmentation,
  - without human intervention,
  - help to data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

A “Toy” Problem

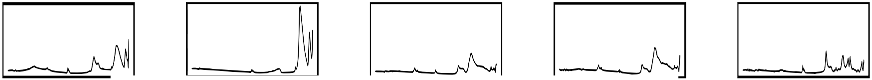
A "Toy" Problem

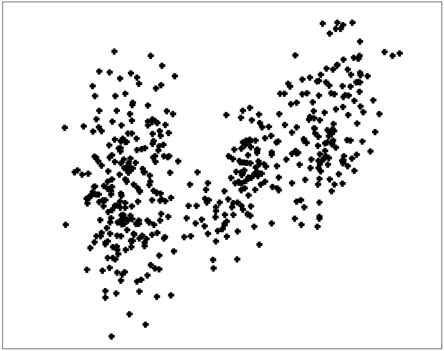


A "Toy" Problem

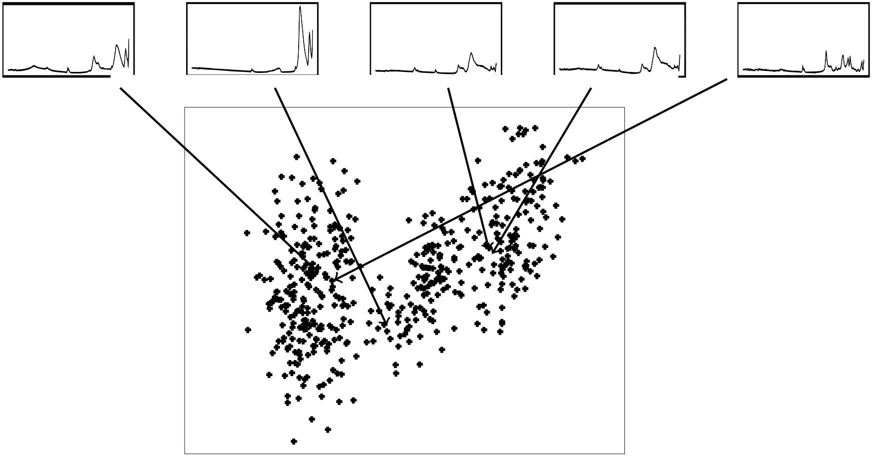


A "Toy" Problem

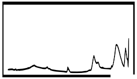
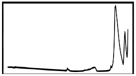
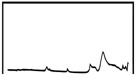
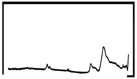
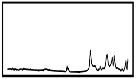


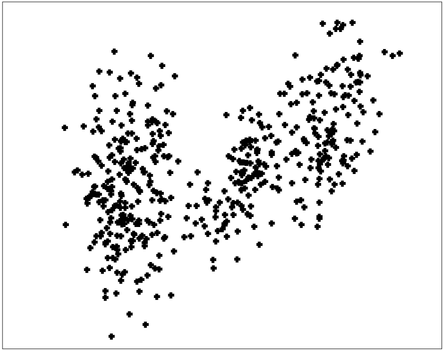


A “Toy” Problem

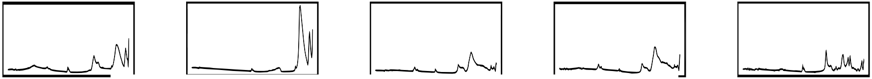


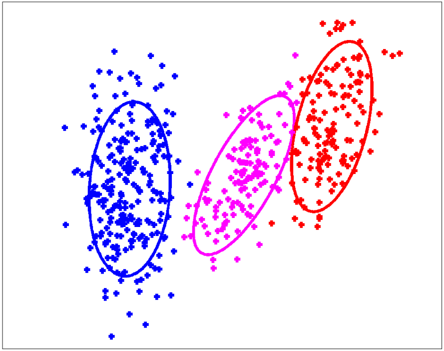
A "Toy" Problem

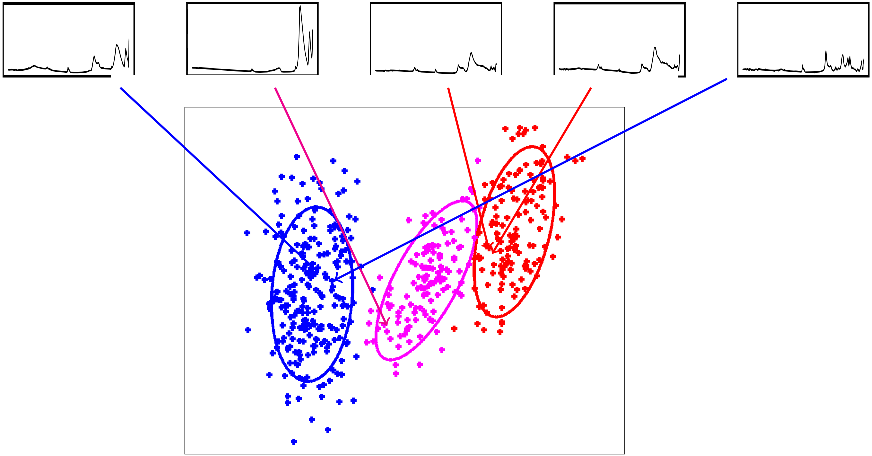


A “Toy” Problem





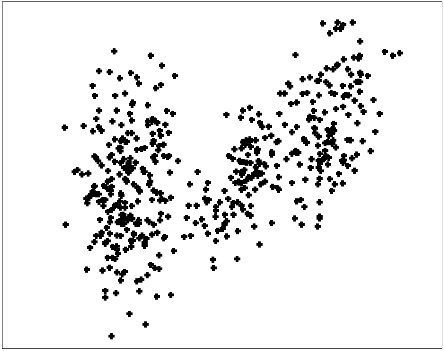
A “Toy” Problem



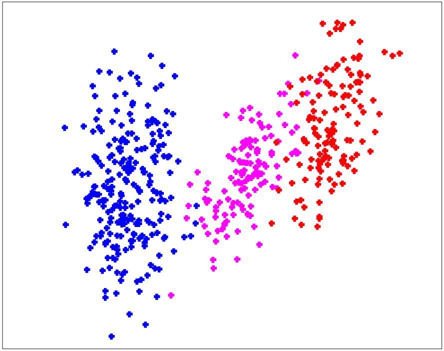
- Representation : mapping between spectrums and points in a large dimension space.
- Spectral method.

“Stochastic” Modeling

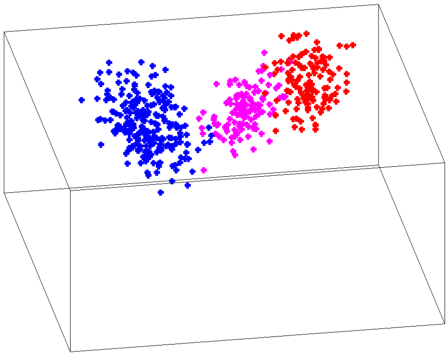
“Stochastic” Modeling



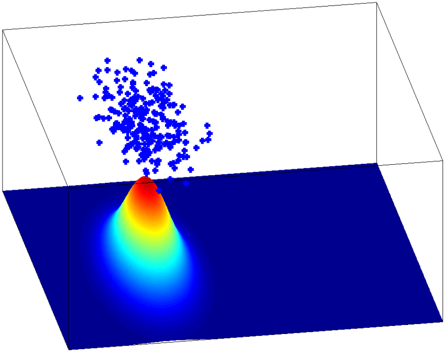
"Stochastic" Modeling



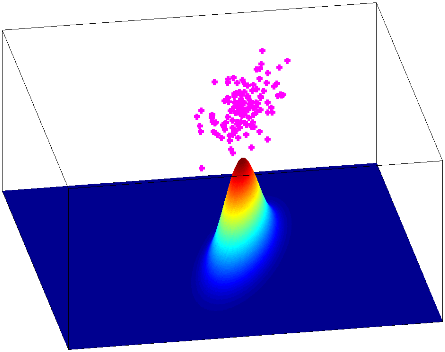
“Stochastic” Modeling



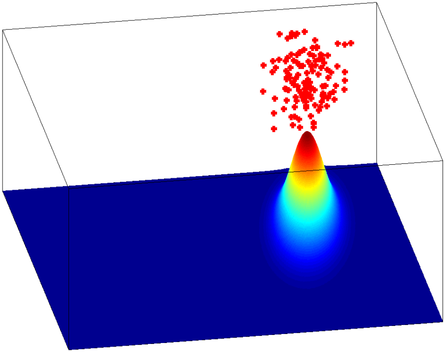
"Stochastic" Modeling



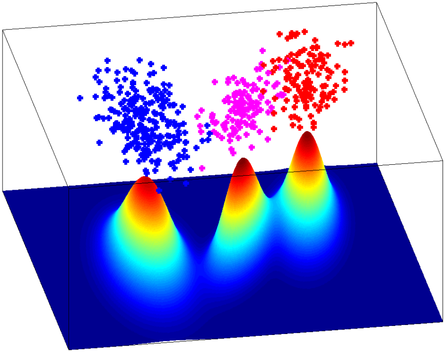
“Stochastic” Modeling



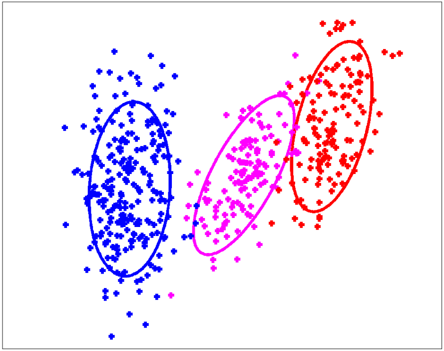
"Stochastic" Modeling



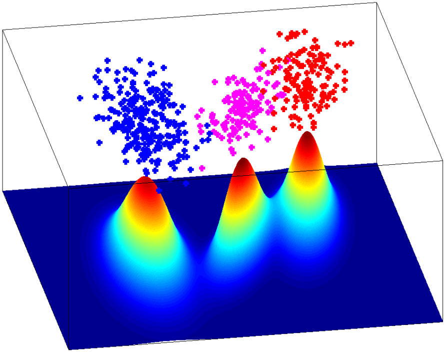
“Stochastic” Modeling



“Stochastic” Modeling



“Stochastic” Modeling

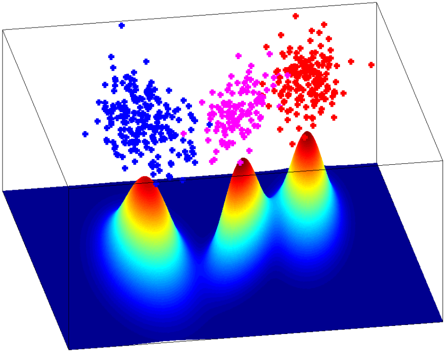


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

"Stochastic" Modeling

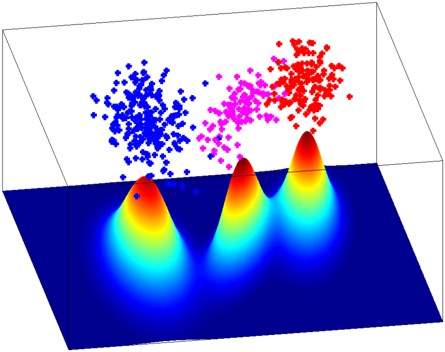


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

“Stochastic” Modeling

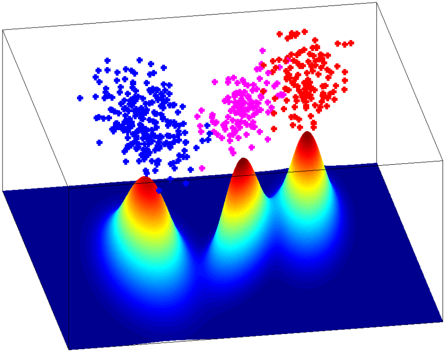


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

“Stochastic” Modeling



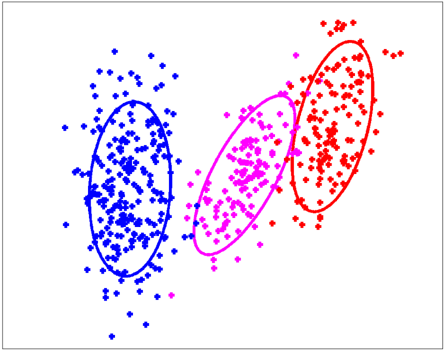
- Model : Gaussian Mixture with K classes.
- Mixture density :

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

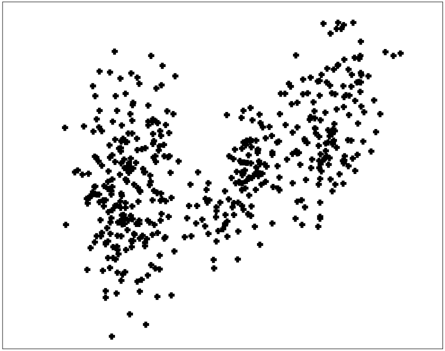
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$$

“Statistical” Estimation

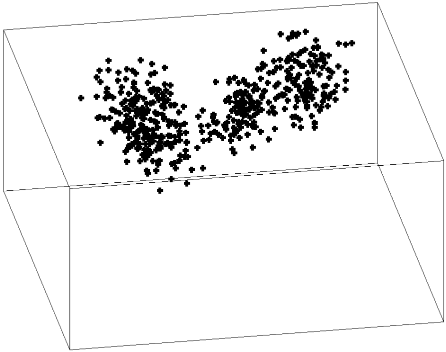
"Statistical" Estimation



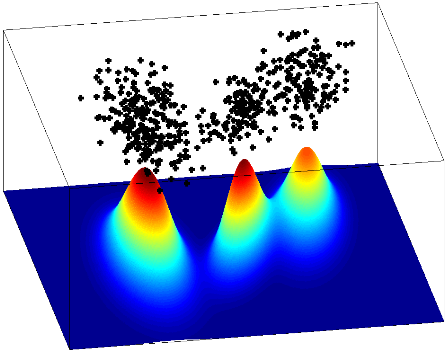
"Statistical" Estimation



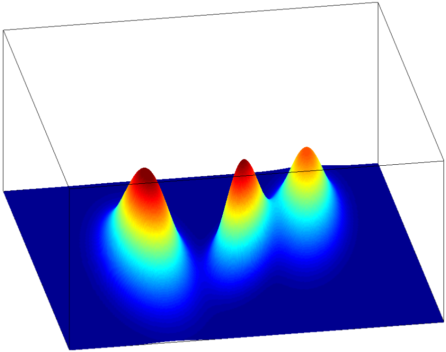
"Statistical" Estimation



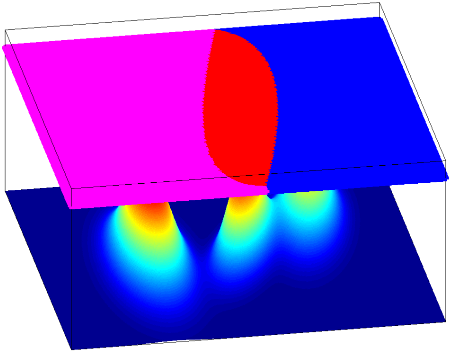
“Statistical” Estimation



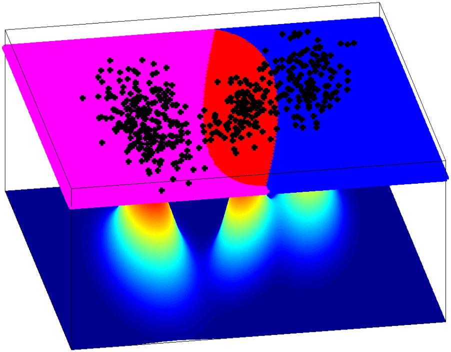
“Statistical” Estimation



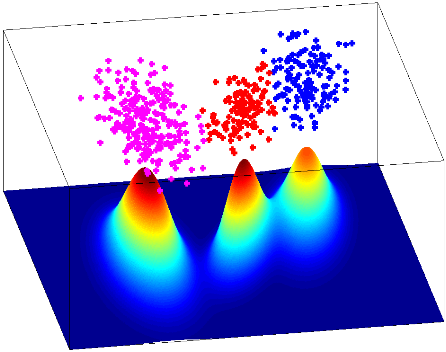
“Statistical” Estimation



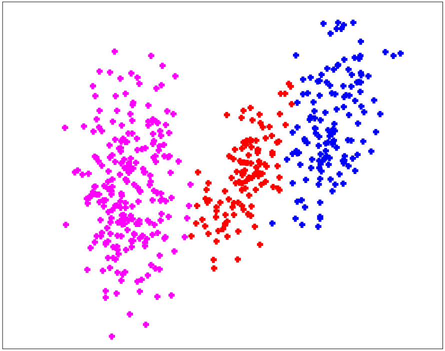
“Statistical” Estimation



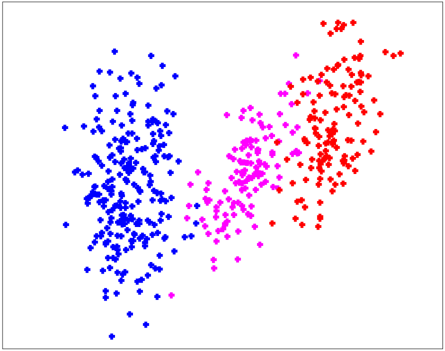
“Statistical” Estimation



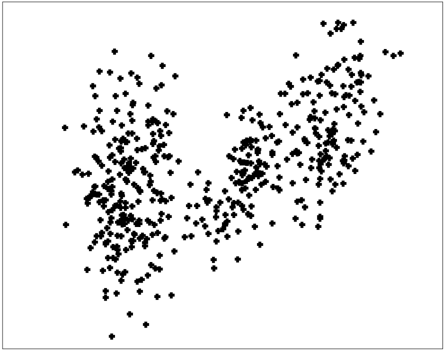
“Statistical” Estimation



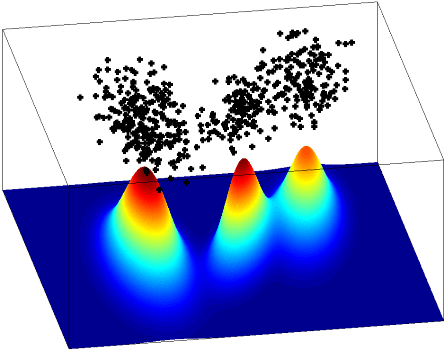
“Statistical” Estimation



“Statistical” Estimation



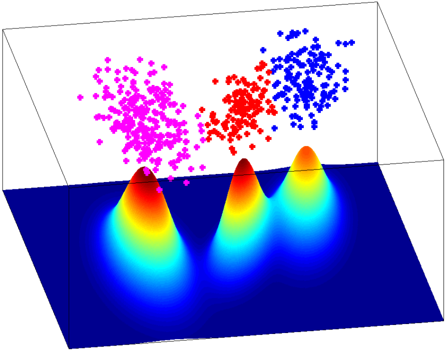
“Statistical” Estimation



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

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

"Statistical" Estimation



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

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

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

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

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

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

- 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

Maximum Likelihood and MM

- "Maximum" likelihood for a given K :

$$
\begin{aligned}
(\widehat{\pi_k}, \widehat{\mu_k}, \widehat{\Sigma_k}) &= \operatorname{argmin} \sum_{i=1}^{N} - \ln \left(\sum_{k=1}^{K} \pi_k \, \mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right) \\
&= \operatorname{argmin} L(\pi, \mu, \Sigma)
\end{aligned}
$$

- Function L rather complex !
- Iterative algorithm (MM) :
  - Current estimate : $(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$,
  - Construction of a Majorization $L^{(n)}$ of L such that

$$
L^{(n)}(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)}) = L(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)}).
$$

    and $L^{(n)}$ easy to minimize.
  - Computation of a Minimizer

$$
(\pi^{(n+1)}, \mu^{(n+1)}, \Sigma^{(n+1)}) = \operatorname{argmin} L^{(n)}(\pi, \mu, \Sigma)
$$

- Very generic methodology...
- Minimization can be replaced by a diminution...

Maximum Likelihood and EM

- Back to L :

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

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

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

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

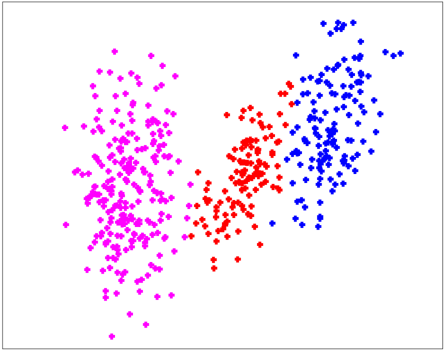
  - Majorization prop. : $L^i \leq L^{i,(n)} + \mathrm{Cst}^{i,(n)}$ with equality at $(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$.
- Bonus :
  - Separability of $L^{(n)} = \sum_{i=1}^{N} L^{i,(n)}$ in π and (μ, Σ) :

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

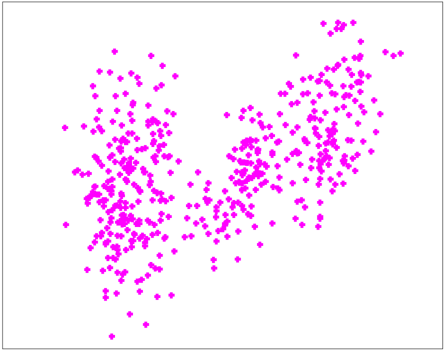
  - Close formulas for the Minimization of $L^{(n)}$ in π and (μ, Σ) !

How many classes ?

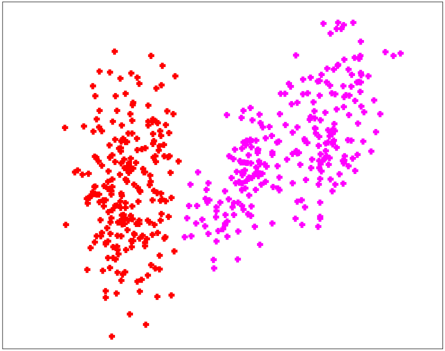
How many classes?



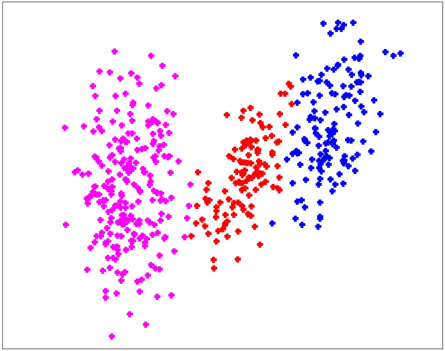
How many classes ?



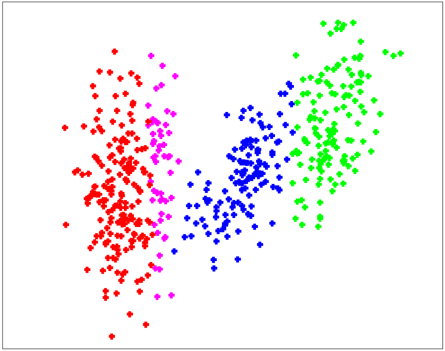
How many classes ?



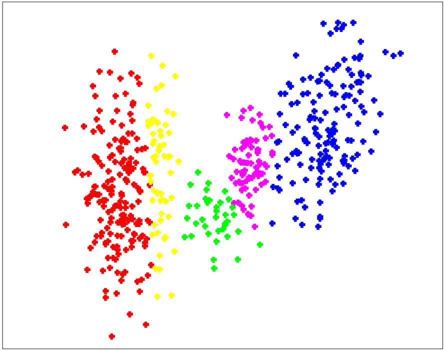
How many classes ?



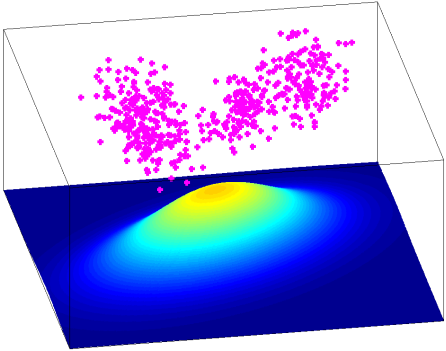
How many classes ?



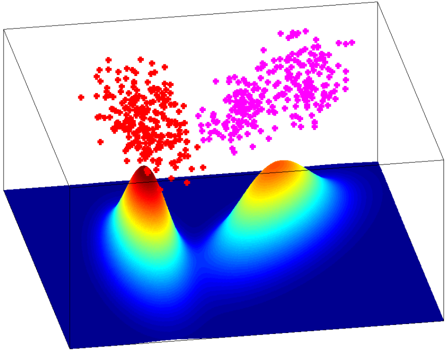
How many classes ?



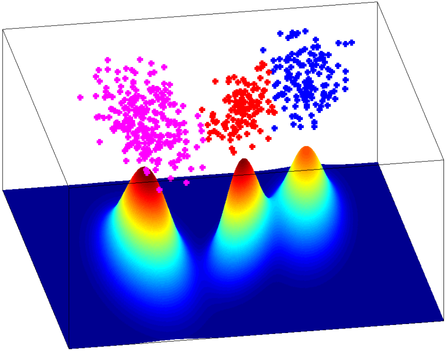
How many classes ?



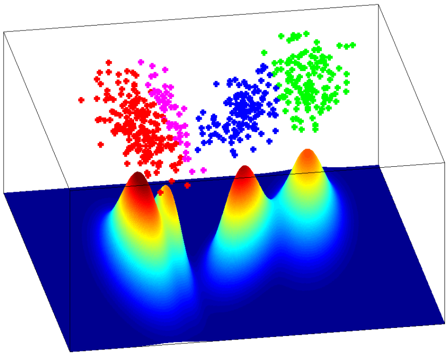
How many classes ?



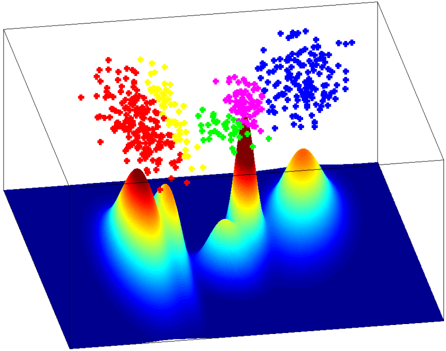
How many classes ?



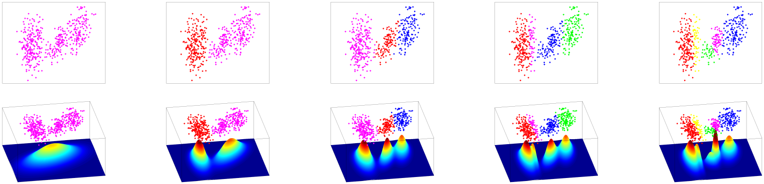
How many classes ?



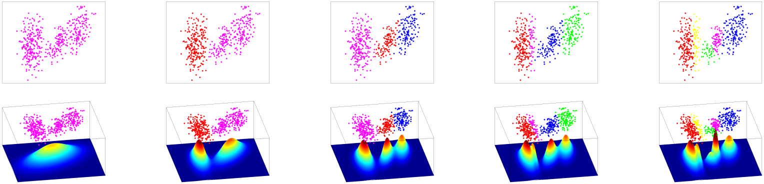
How many classes ?



How many classes ?

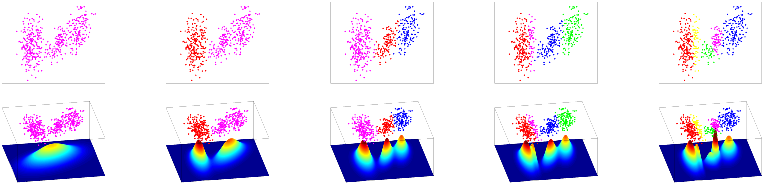


How many classes ?



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

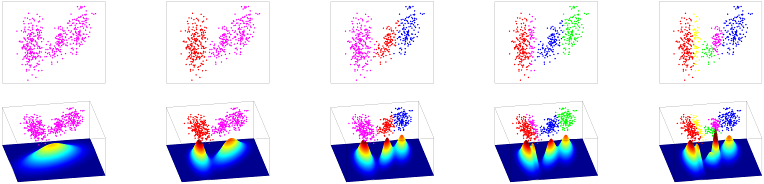
How many classes ?



Fidelity −− + + + + + + +· + + ++

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

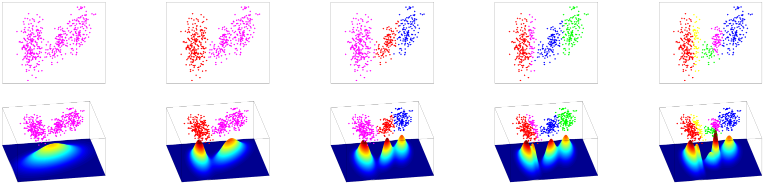
How many classes ?



Fidelity	−−	+	+ + +	+ + +·	+ + ++
Simplicity	+ + ++	+ + +	++	+	−

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

How many classes ?



Fidelity	− −	+	+ + +	+ + +·	+ + ++
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

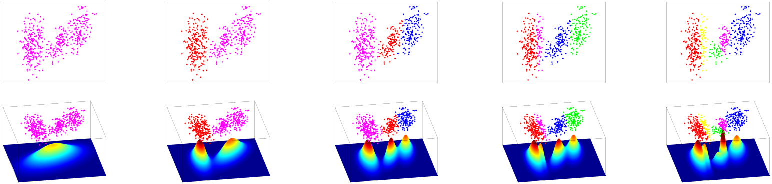
entities must not be multiplied beyond necessity

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

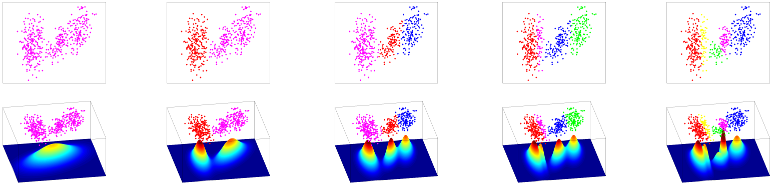
- Ockham's Razor (simplicity principle) : one should not add hypotheses, if the current ones are already sufficient !
- Balance between observation explanation power and simplicity.

Selection by Penalization

Selection by Penalization

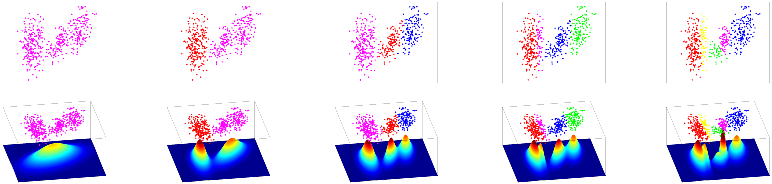


Selection by Penalization



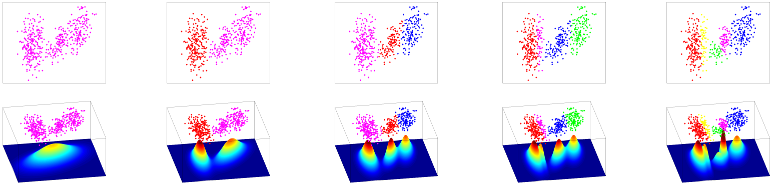
Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicity	$++++$	$+++$	$++$	$+$	$-$

Selection by Penalization



Likelihood	−−	+	+ + +	+ + +·	+ + ++
+ Simplicity	+ + ++	+ + +	++	+	−
= Tradeoff	++	+ + ++	+ + + + +	+ + + + ·	+ + +

Selection by Penalization

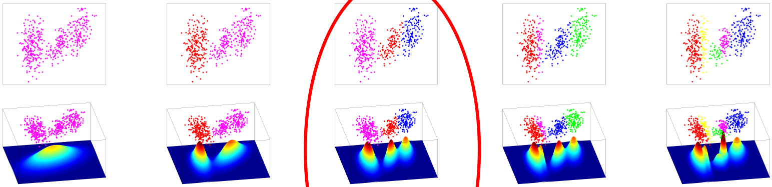


Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

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

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

Selection by Penalization

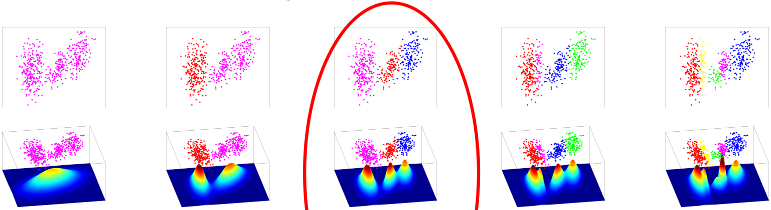


Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

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

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

Selection by Penalization



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

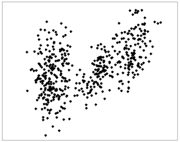
- Likelihood : $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicity : $-\lambda \mathrm{Dim}(S_K)$.
- Penalized estimator :

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

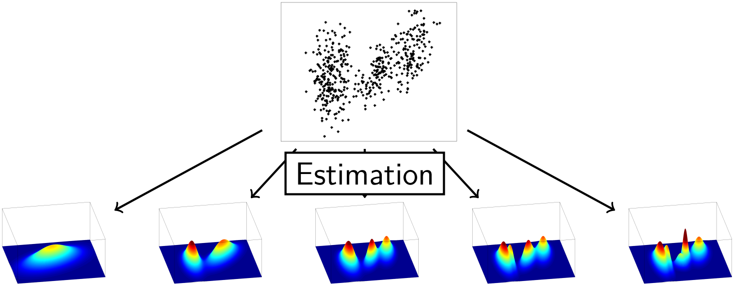
- Optimization in K by exhaustive exploration !

Methodology

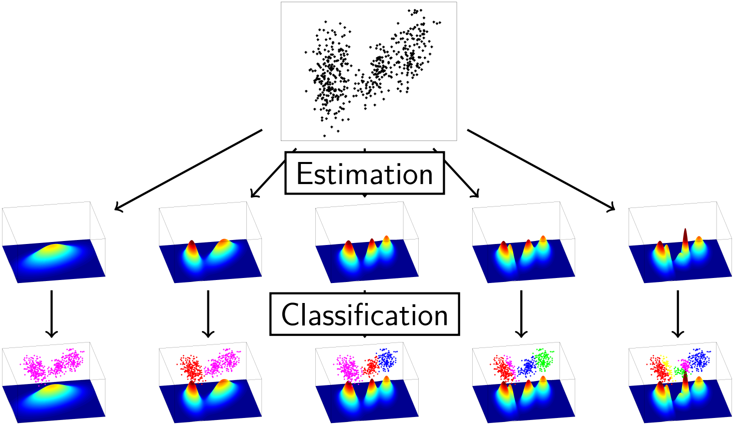
Methodology



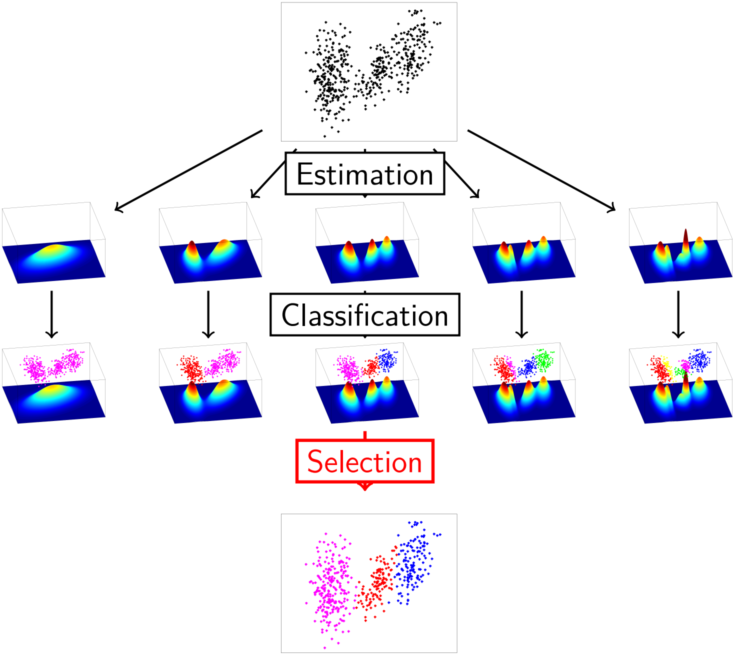
Methodology



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

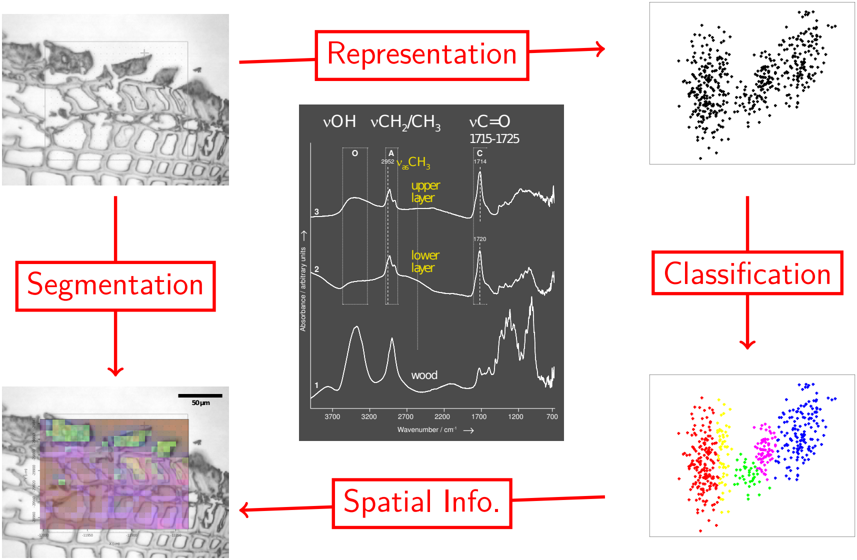
$$\widehat{m} = \mathrm{argmin} - \ln(\widehat{s}_m) + \mathrm{pen}(m).$$

  with $\mathrm{pen}(m) = \kappa(\ln(n))\,\dim(S_m)$ (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...

Back to our violins



Segmentation and Spatialized GMM

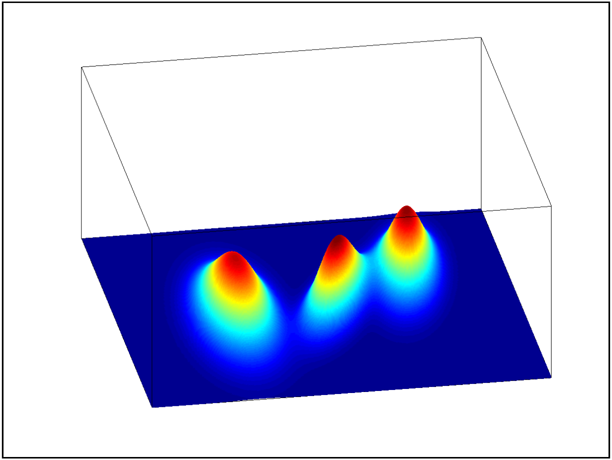
- Initial goal : segmentation $\neq$ clustering.
- Idea of Kolaczyk et al (cf Bigot) : take into account the spatial position x of the spectrum in the mixing proportions.
- Conditional density model :

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

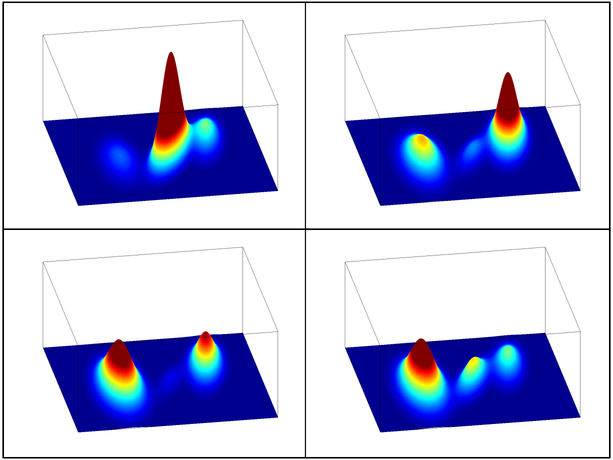
- Estimation from the data :
  - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D_k'$ for each class
  - and the mixing proportion functions $\pi_k(x)$.
- Segmentation by MAP principle :

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

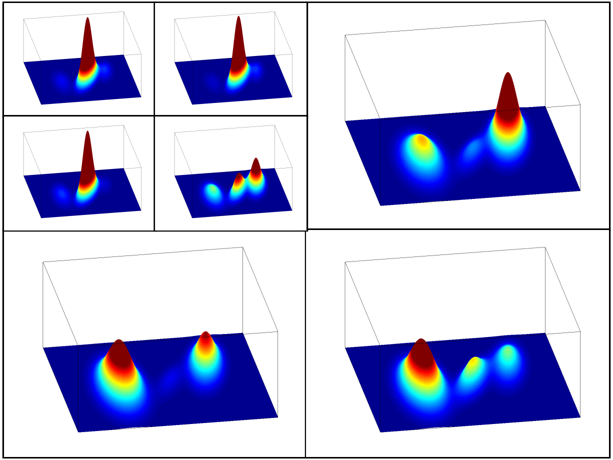
Segmentation and Spatialized GMM



Segmentation and Spatialized GMM

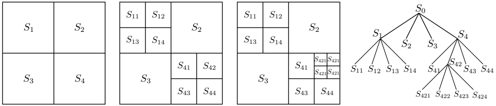


Segmentation and Spatialized GMM



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

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

Numerical optimization

- Penalized Model Selection :

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

- 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
  - CART (dynamic programming)

EM Algorithm

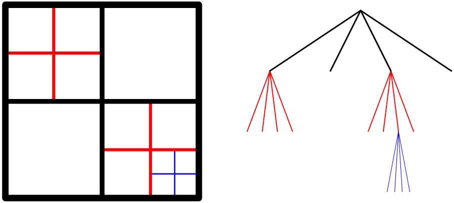
- E Step : with $P_k^{i,(n)} = P(k_i = k | x_i, \mathcal{S}_i, \mathcal{P}^{(n)}, \pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$

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

  with equality at $(\mathcal{P}^{(n)}, \pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$.

- M Step : Split optimization in $(\mathcal{P}, \pi)$ and (μ, Σ) possible,
  - Optimization in (μ, Σ) : close formulas (classical...).
  - Optimization in $(\mathcal{P}, \pi)$ more interesting !

M Step and CART



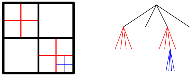
- Optimization in $(\mathcal{P},\pi)$ of

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

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

- 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 optimization algorithm of CART type (Dynamic programming on tree structure).

CART Optimization



- Aim : compute efficiently $\operatorname{argmin}_{\mathcal{P}} \sum_{\mathcal{R}\in\mathcal{P}} C[\mathcal{R}]$ where $\mathcal{P}$ belongs to the set of recursive dyadic partitions (associated to quadtree) of limited depth.
- Key observation : 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, $\widehat{\mathcal{P}}[\mathcal{R}] = \cup_{\mathcal{R}'\in\text{Child}[\mathcal{R})}\widehat{\mathcal{P}}[\mathcal{R}']$

  with a decision based on

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

- Algorithm : Precomputation of all $C[\mathcal{R}]$ then recursive determination of $\widehat{\mathcal{P}}[\mathcal{R}]$ and $\widehat{C}[\mathcal{R}] = \sum_{\mathcal{R}''\in\widehat{\mathcal{P}}} C[\mathcal{R}'']$ (either $C[\mathcal{R}]$ or the sum of the $\widehat{C}$ of its children) with stopping as soon as the square has no child.
- Non recursive version possible.

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 $s_m \in S_m$

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

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 $m \in \mathcal{M}$ $\mathrm{pen}(m) \geq \kappa(\mathfrak{D}_m + x_m)$ 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 $\mathfrak{D}_m \propto (\log n)\dim(S_m)$...

Kullback, Hellinger and extensions

- 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 affinity) (Kolaczyk, Barron, Bigot) on the left...
- Refinement with a *bounded* version of KL : $JKL(s, s') = 2KL(s, (s' + s)/2)$ (Massart, van de Geer)
- Jensen-Kullback-Leibler : generalization to $JKL_\rho(s, s') = \frac{1}{\rho} KL(s, \rho s' + (1-\rho)s)$.
- **Prop. :** For all $\rho \in (0, 1)$, there is a $C_\rho > 0$ such that

$$C_\rho\, d^2(s, t) \leq JKL_\rho(s, t) \leq KL(s, t).$$

- For $\rho \simeq 1/2$, $C_\rho \simeq 1/5$.

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 $\mathbb{E}\left[n(\mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_m)\right] - KL^{\otimes_n}(s_0, \widehat{s}_m))\right]$ (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 $d^{2\otimes_n}$.
  - 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...

Spatialized Gaussian Mixture Case

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

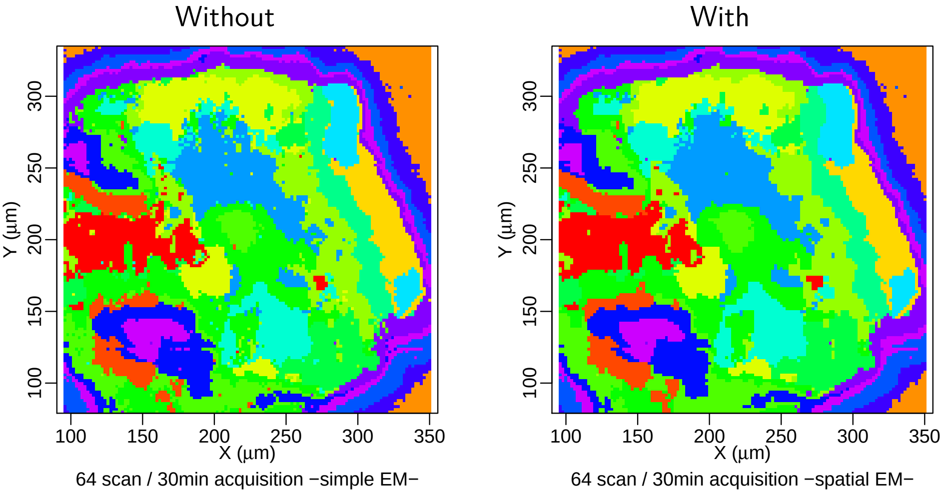
$$\mathfrak{D}_m \leq \kappa' \left(C' + \frac{1}{2} \left(\ln \left(\frac{N}{C' \dim(S_m)} \right) \right)_+ \right) \dim(S_m).$$

- 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 :

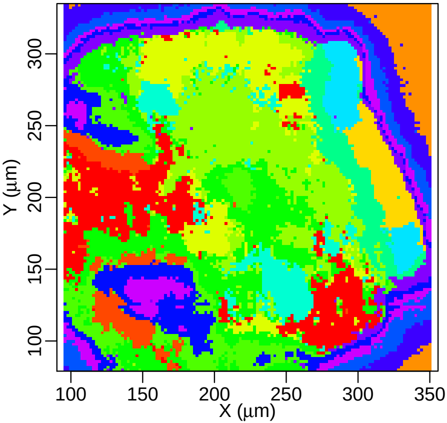


- Automatic choice of K, $[L_k\, D\, A]^K$ and partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by random projection.

Unsupervised Segmentation

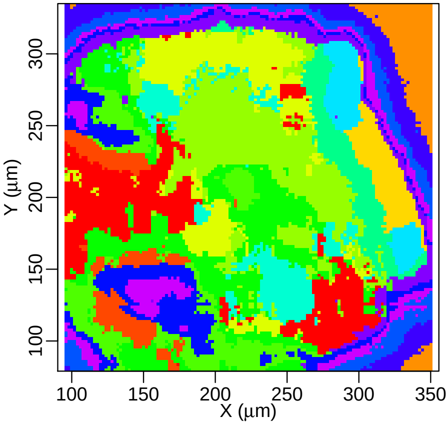
- Numerical result taking into account the spatial modeling :

Without



8 scan / 5min acquisition −simple EM−

With



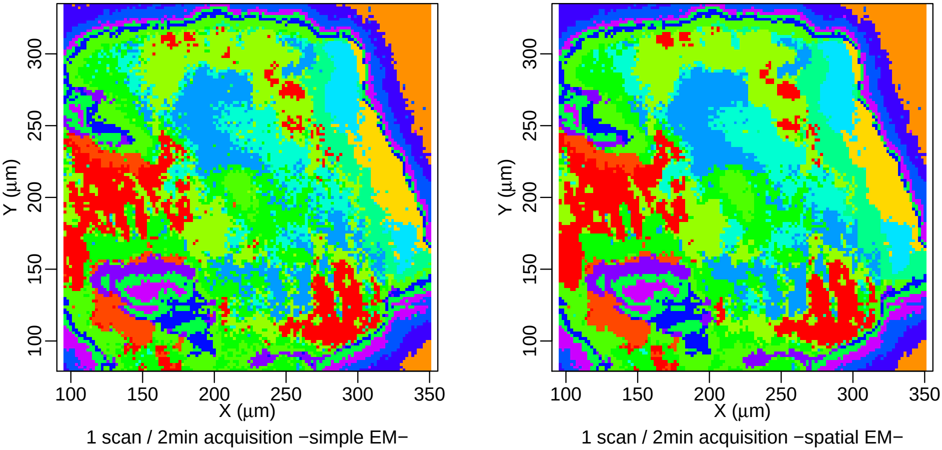
8 scan / 5min acquisition −spatial EM−

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

Unsupervised Segmentation

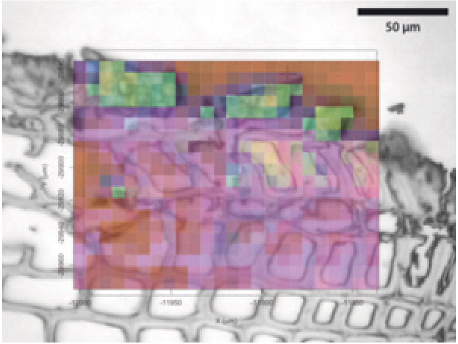
- Numerical result taking into account the spatial modeling :

Without / With



- 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.
  - 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).

Conclusion

- Framework :
  - Unsupervised segmentation problem.
  - 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).