

Hyperspectral Image Segmentation by Spatialized Gaussian Mixtures and Model Selection

E. Le Pennec

(CMAP - École Polytechnique)

and

S. Cohen (IPANEMA - CNRS / Soleil)



IMB
18 December 2014

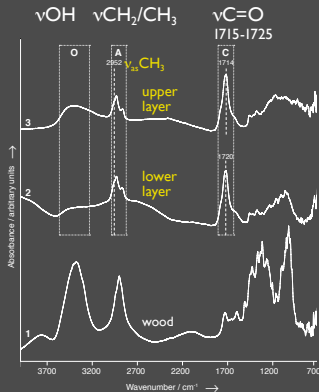


A. Stradivari (1644 - 1737)

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



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.

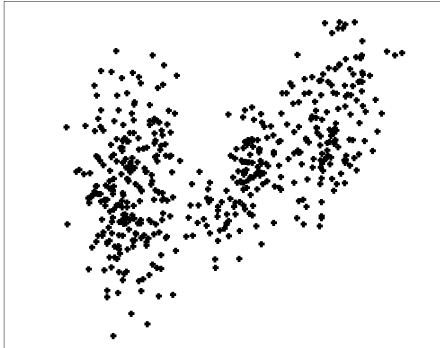


Hyperspectral Image Segmentation

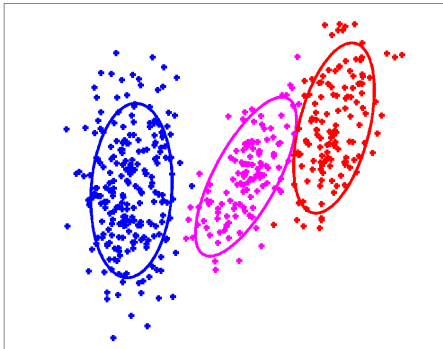
- Data :
 - image of size N between ~ 1000 and ~ 100000 pixels,
 - spectrums $\mathcal{S}$ of ~ 1024 points,
 - very good spatial resolution,
 - ability to measure a lot of spectrums per minute,
- Immediate goal :
 - automatic image segmentation,
 - without human intervention,
 - help to data analysis.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

A “Toy” Problem

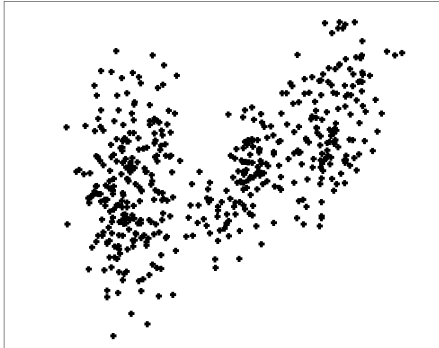
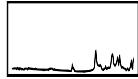
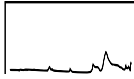
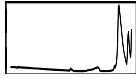
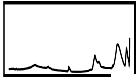
A “Toy” Problem



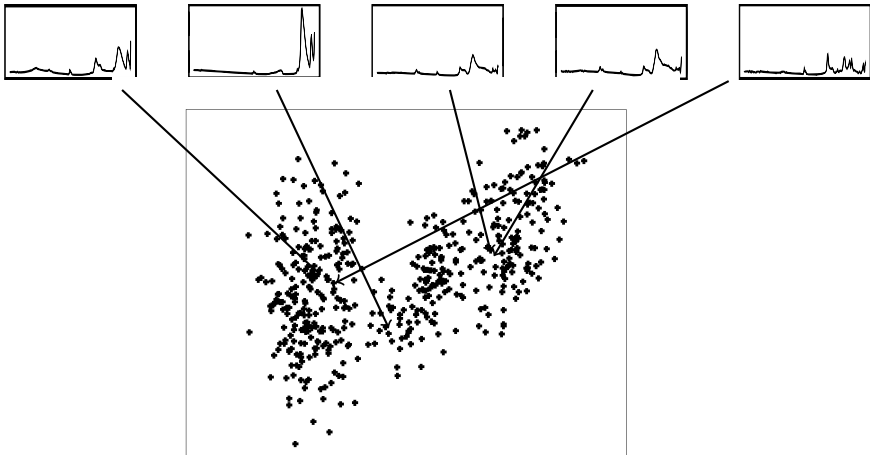
A “Toy” Problem



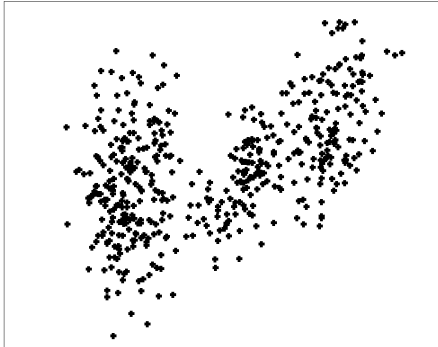
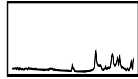
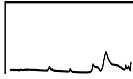
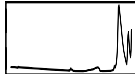
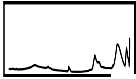
A “Toy” Problem



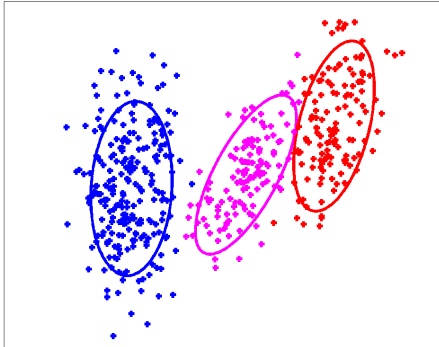
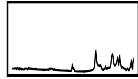
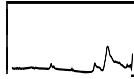
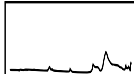
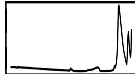
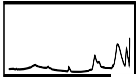
A “Toy” Problem



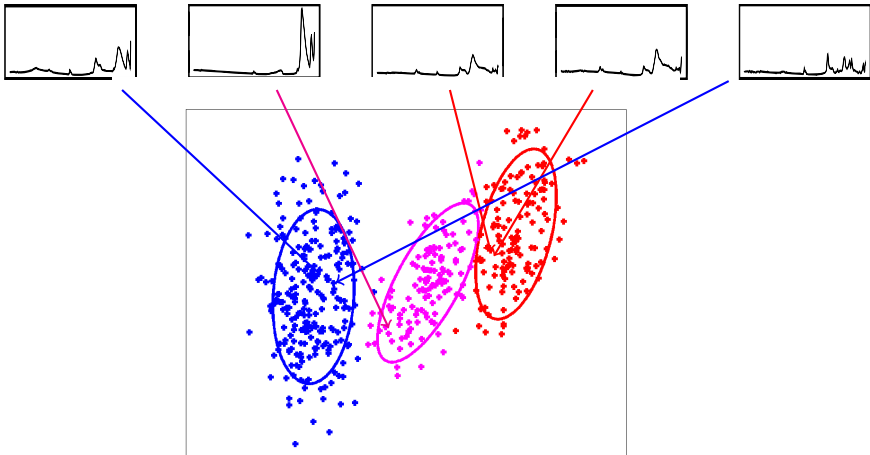
A “Toy” Problem



A “Toy” Problem



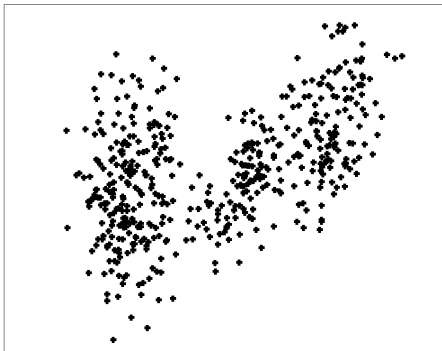
A “Toy” Problem



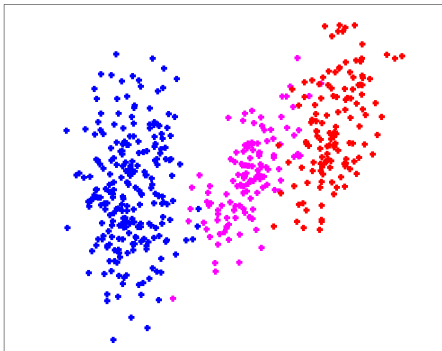
- Representation : mapping between spectrums and points in a large dimension space.
- Spectral method.

“Stochastic” Modeling

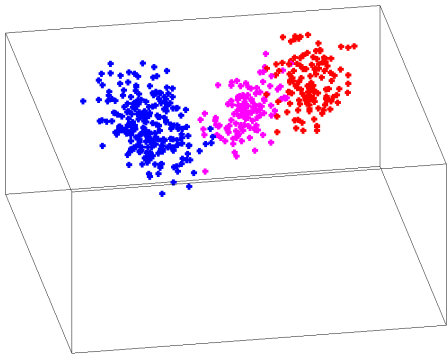
“Stochastic” Modeling



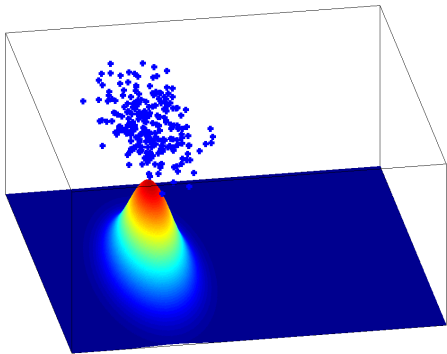
“Stochastic” Modeling



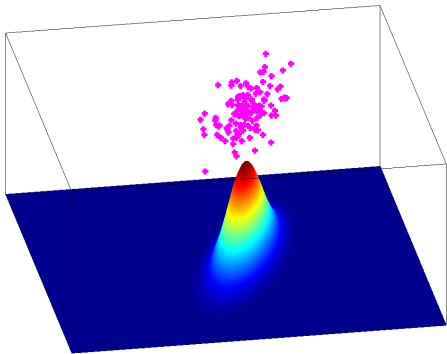
“Stochastic” Modeling



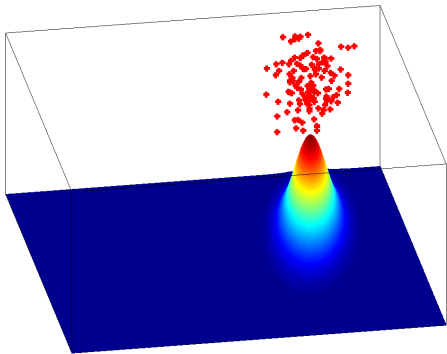
“Stochastic” Modeling



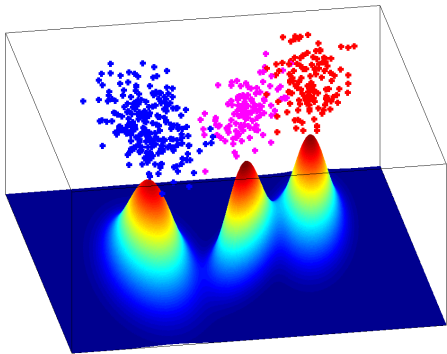
“Stochastic” Modeling



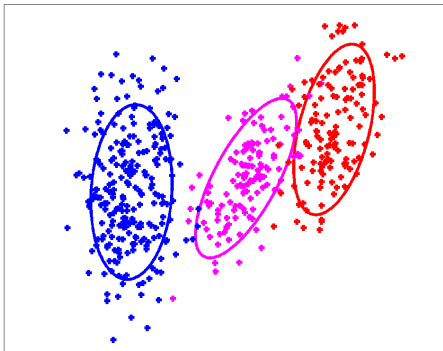
“Stochastic” Modeling



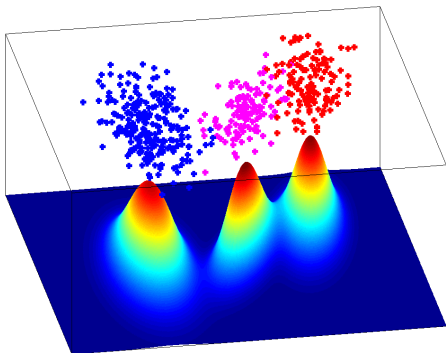
“Stochastic” Modeling



“Stochastic” Modeling



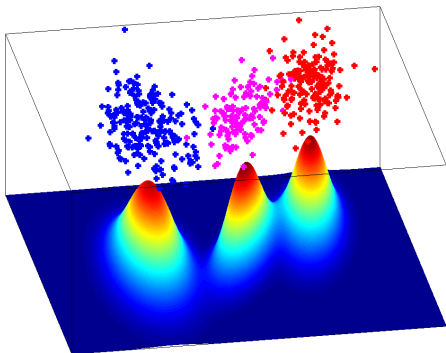
“Stochastic” Modeling



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

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

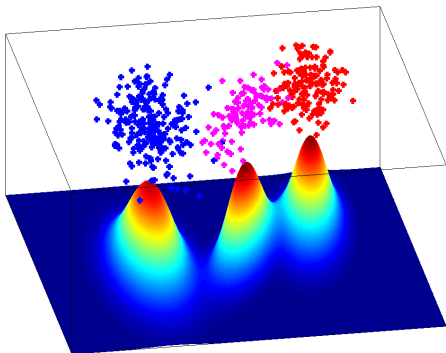
“Stochastic” Modeling



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

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

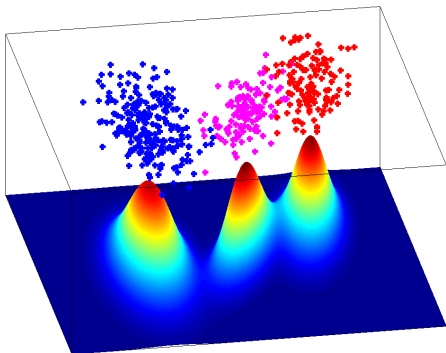
“Stochastic” Modeling



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

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

“Stochastic” Modeling

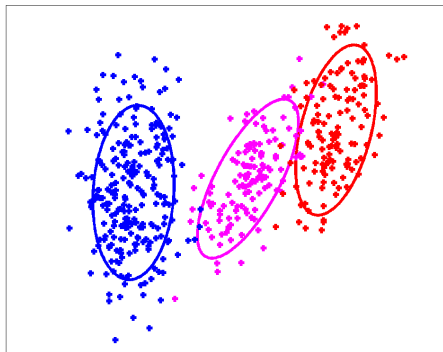


- Model : Gaussian Mixture with K classes.
- Mixture density :

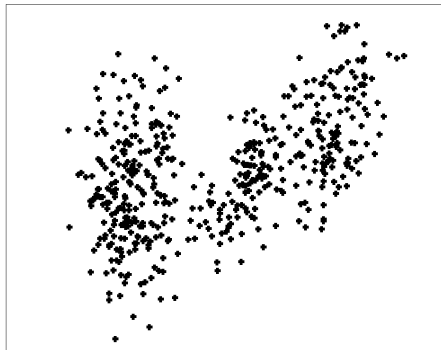
$$\begin{aligned} s_{K,\pi,\mu,\Sigma}(\mathcal{S}) &= \sum_{k=1}^K \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1} (\mathcal{S}-\mu_k)} \\ &= \sum_{k=1}^K \pi_k \mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}) \end{aligned}$$

“Statistical” Estimation

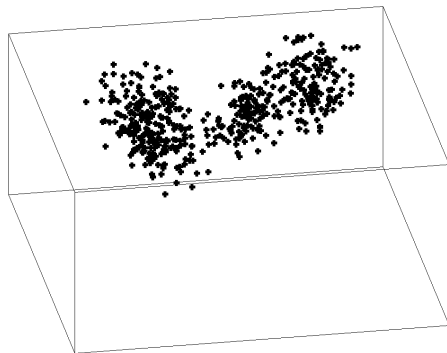
“Statistical” Estimation



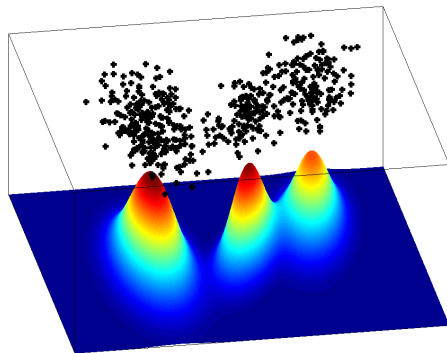
“Statistical” Estimation



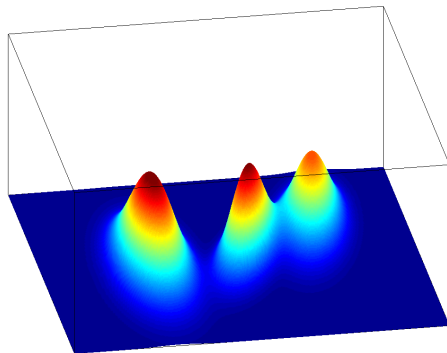
“Statistical” Estimation



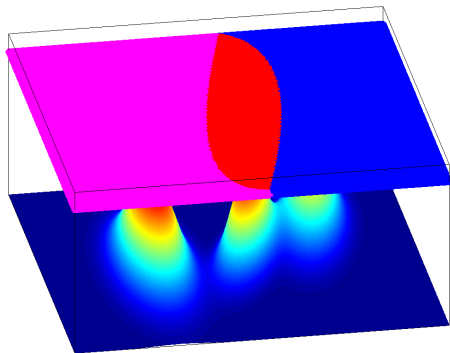
“Statistical” Estimation



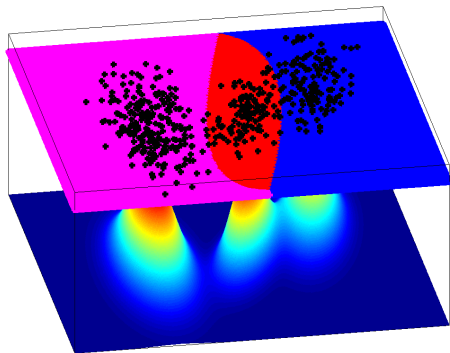
“Statistical” Estimation



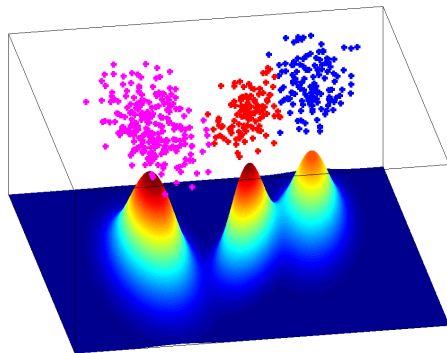
“Statistical” Estimation



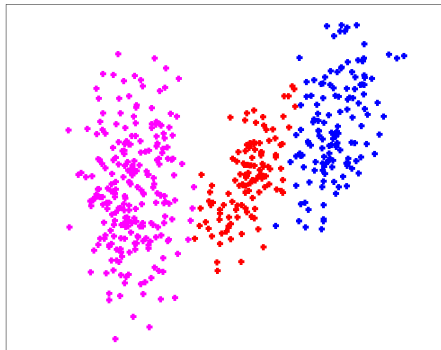
“Statistical” Estimation



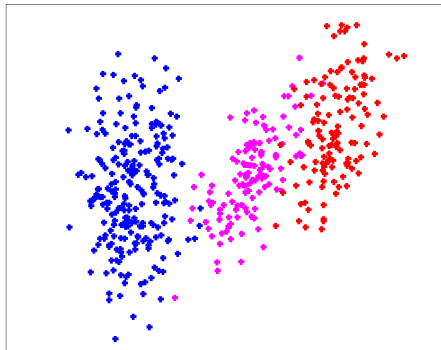
“Statistical” Estimation



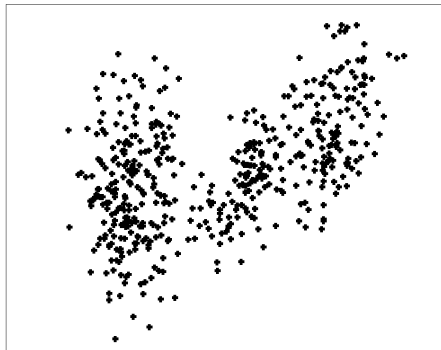
“Statistical” Estimation



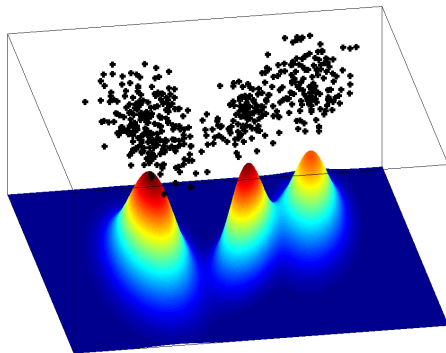
“Statistical” Estimation



“Statistical” Estimation



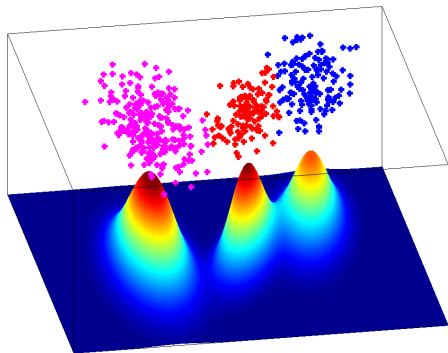
“Statistical” Estimation



- 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) :

$$\widehat{k}(\mathcal{S}) = \operatorname{argmax}_{\mu_k, \Sigma_k} \widehat{\pi}_k \mathcal{N}_{\widehat{\mu}_k, \widehat{\Sigma}_k}(\mathcal{S})$$

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

$$\hat{k}(\mathcal{S}) = \operatorname{argmax}_{\mu_k, \Sigma_k} \hat{\pi}_k \mathcal{N}_{\hat{\mu}_k, \hat{\Sigma}_k}(\mathcal{S})$$

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

Gaussian Mixture Model

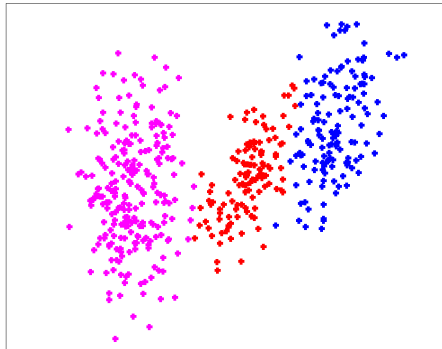
- True density s_0 of $\mathcal{S}$ close from

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

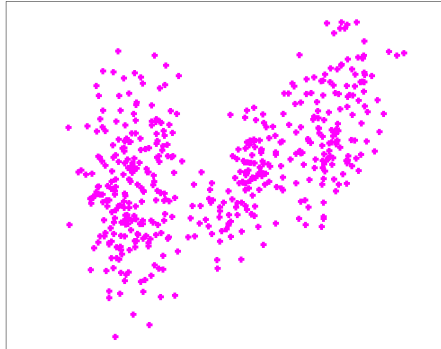
- Gaussian Mixture Model $S_m = \{s_m\}$ specified by
 - a number of classes K ,
 - a structure for the means μ_k and the covariance matrices $\Sigma_k = L_k D_k A_k D_k'$ (Volume L_k , basis D_k and rescaled eigenvalues A_k)
- Structure $[\mu \ L \ D \ A]^K$ for the K -tuples of Gaussian parameters :
 - 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 ?

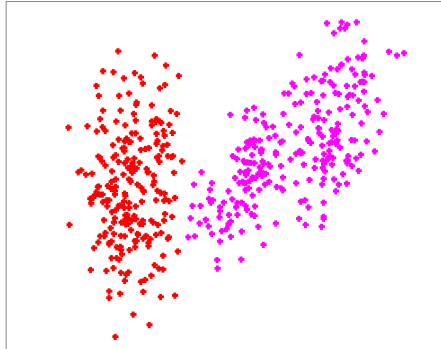
How many classes ?



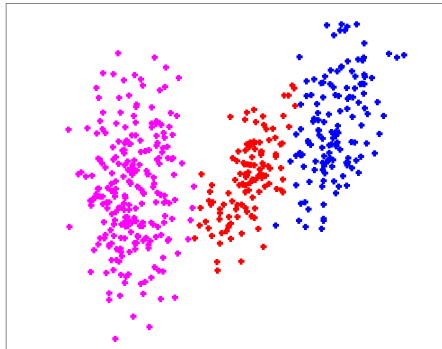
How many classes ?



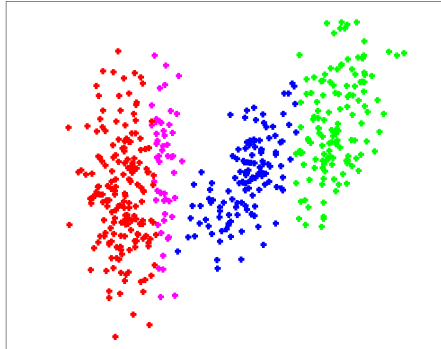
How many classes ?



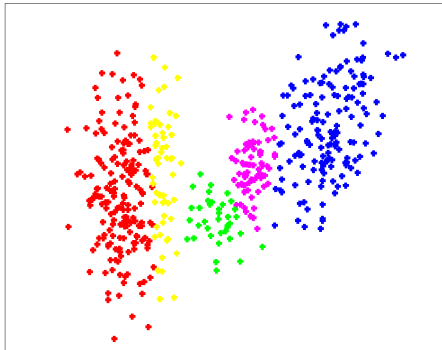
How many classes ?



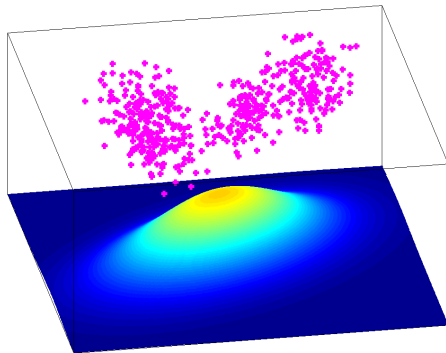
How many classes ?



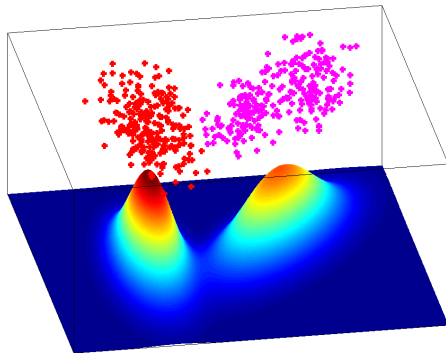
How many classes ?



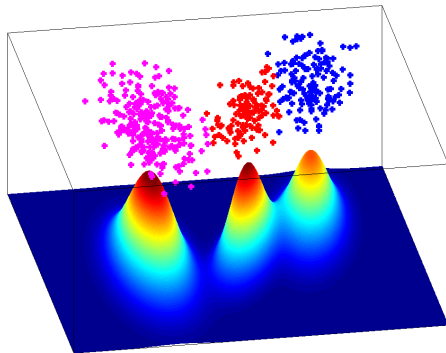
How many classes ?



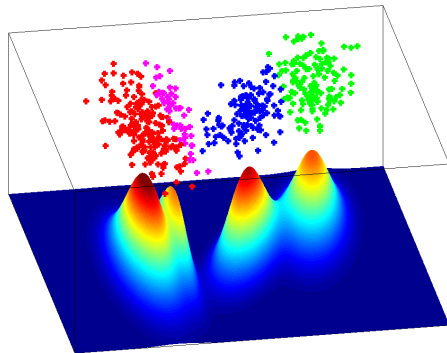
How many classes ?



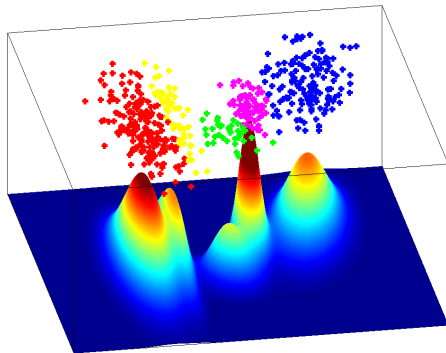
How many classes ?



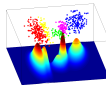
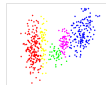
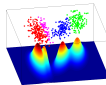
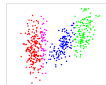
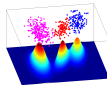
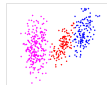
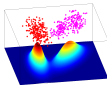
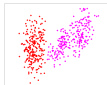
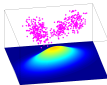
How many classes ?



How many classes ?

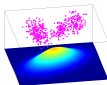


How many classes?

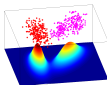
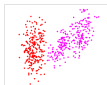


How many classes?

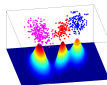
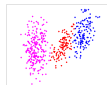
Fidelity



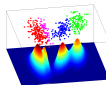
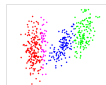
--



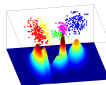
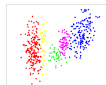
+



+++

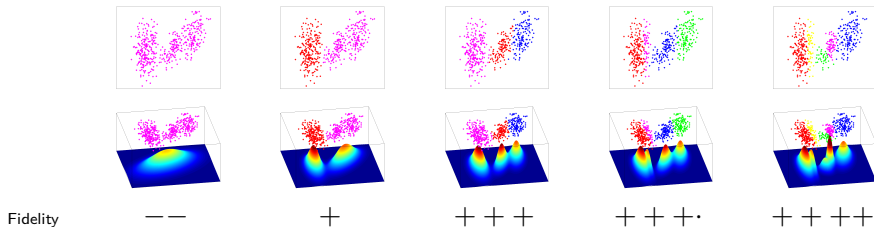


+++•



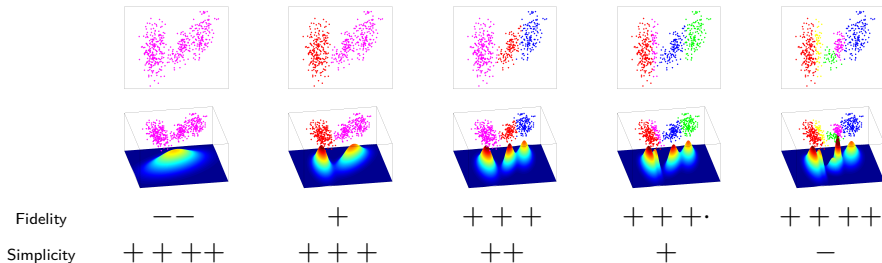
++++

How many classes?



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

How many classes?



- 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 (~ 1285 - 1347)

Ockham's Razor

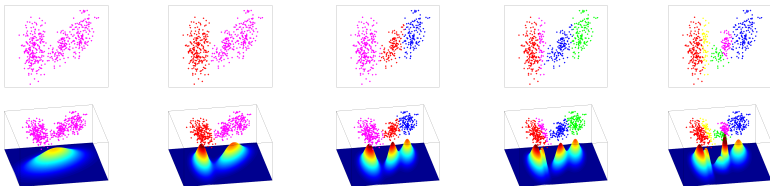


entities must not be multiplied beyond necessity
William of Ockham (~ 1285 - 1347)

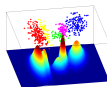
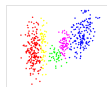
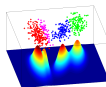
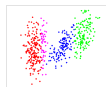
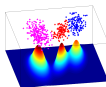
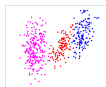
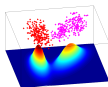
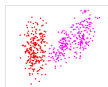
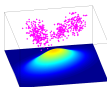
- Ockham's Razor (simplicity principle) : one should not add hypotheses, if the current ones are already sufficient !
- Balance between observation explanation power and simplicity.

Selection by Penalization

Selection by Penalization



Selection by Penalization



Likelihood

— —

+

++ +

+++ .

++++

Simplicity

++++

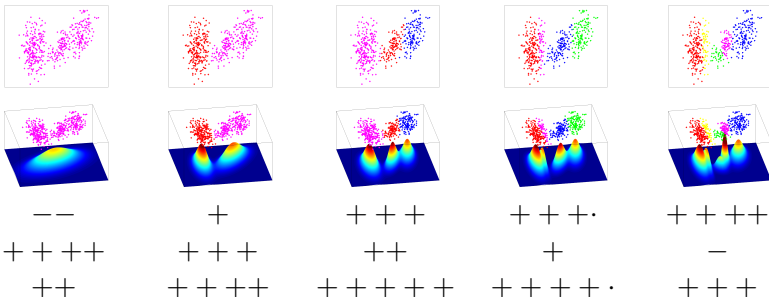
+++

++

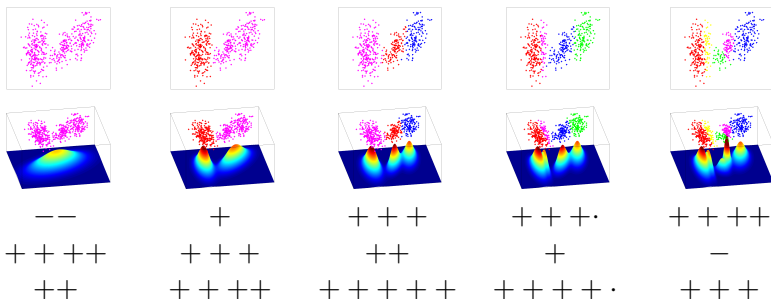
+

—

Selection by Penalization



Selection by Penalization



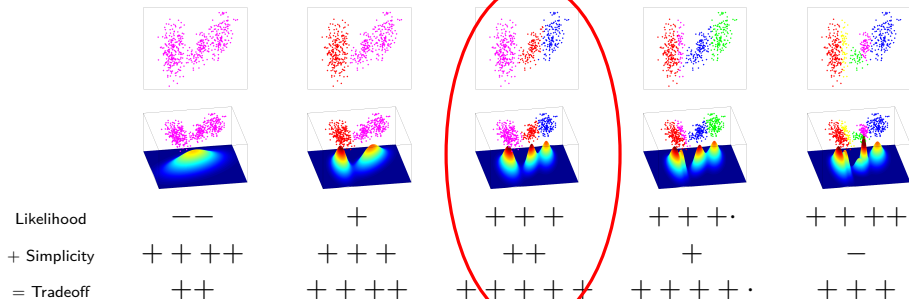
● Likelihood : $\sum_{i=1}^N \log \hat{s}_K(X_i)$.

● Simplicity : $-\lambda \text{Dim}(S_K)$.

● Penalized estimator :

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

Selection by Penalization



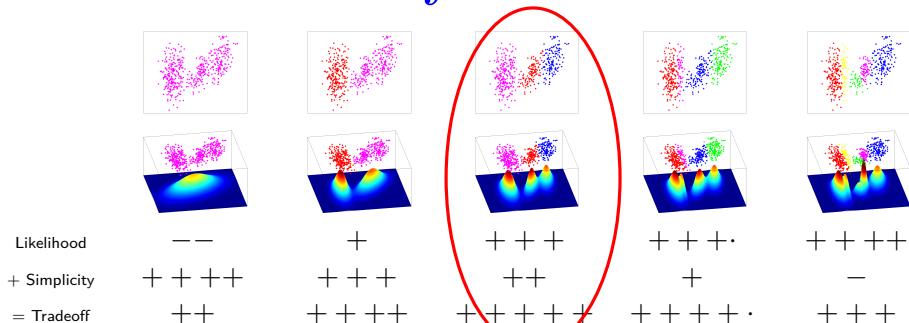
● Likelihood : $\sum_{i=1}^N \log \hat{s}_K(X_i)$.

● Simplicity : $-\lambda \text{Dim}(S_K)$.

● Penalized estimator :

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

Selection by Penalization



● Likelihood : $\sum_{i=1}^N \log \hat{s}_K(X_i)$.

● Simplicity : $-\lambda \text{Dim}(S_K)$.

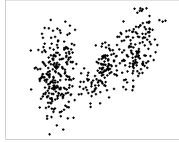
● Penalized estimator :

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

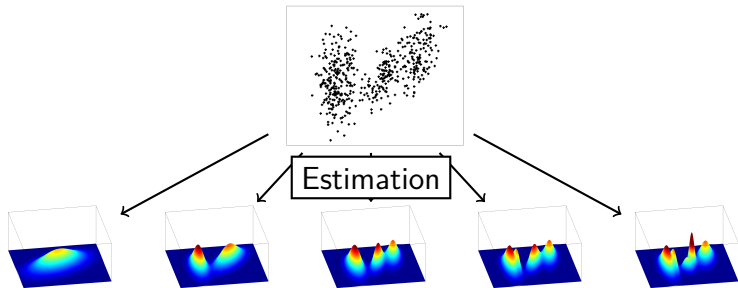
● Optimization in K by exhaustive exploration !

Methodology

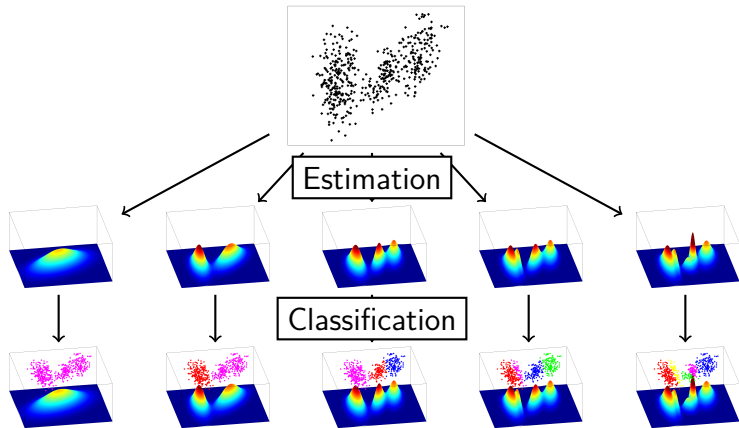
Methodology



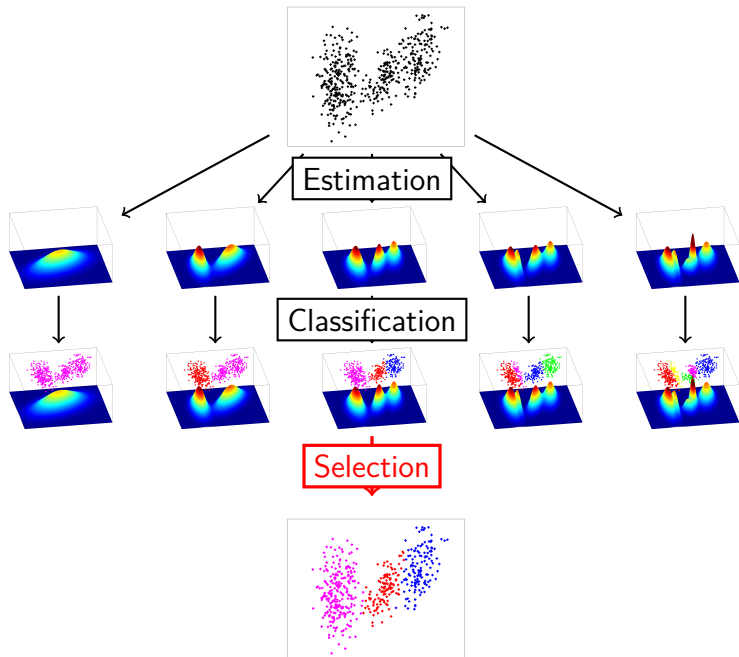
Methodology



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 $\hat{m}$ by

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

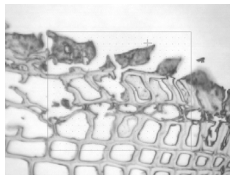
with $\operatorname{pen}(m) = \kappa(\ln(n)) \dim(S_m)$ (parametric dimension of S_m),

- Results (Birgé, Massart, Celeux, Maugis, Michel...) :
 - Density estimation : for κ large enough,

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

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

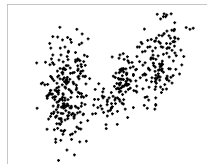
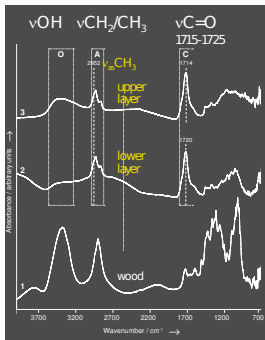
Back to our violins



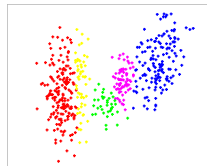
Segmentation



Representation



Classification



Spatial Info.

Segmentation and Spatialized GMM

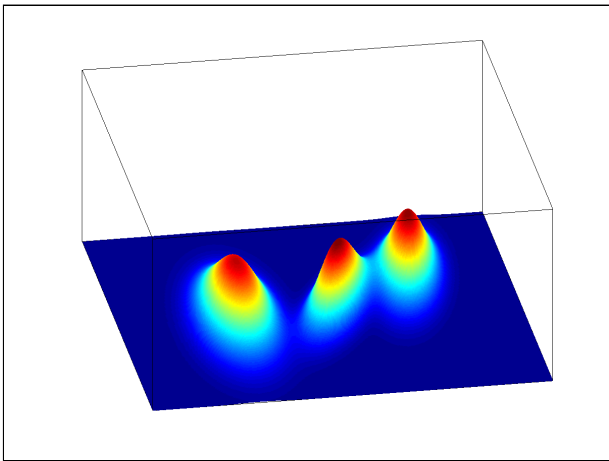
- Initial goal : segmentation $\neq$ clustering.
- Idea of Kolaczyk et al (cf Bigot) : take into account the spatial position x of the spectrum in the mixing proportions.
- Conditional density model :

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

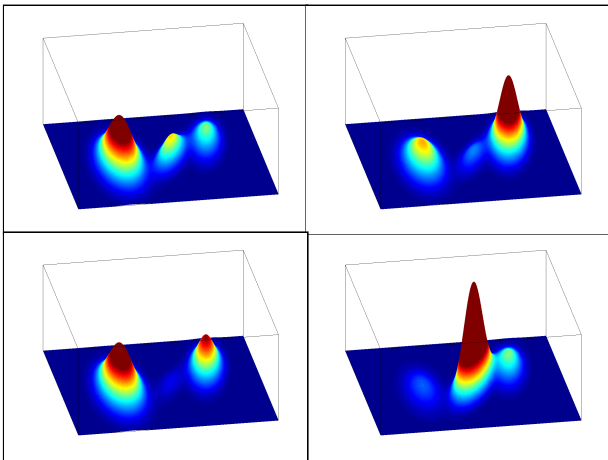
- Estimation from the data :
 - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D_k'$ for each class
 - and the mixing proportion functions $\pi_k(x)$.
- Segmentation by MAP principle :

$$\hat{k}(\mathcal{S}|x) = \arg \max_k \hat{\pi}_k(x) \mathcal{N}_{\hat{\mu}_k, \hat{\Sigma}_k}(\mathcal{S})$$

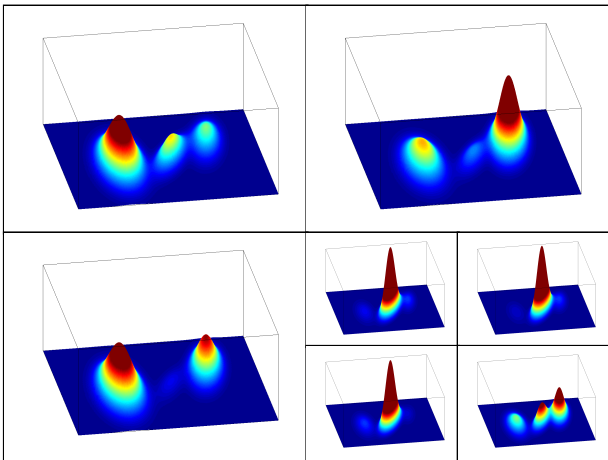
Segmentation and Spatialized GMM



Segmentation and Spatialized GMM

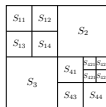
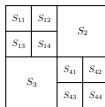


Segmentation and Spatialized GMM



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 $\text{pen}(m) = \kappa \ln(n) \dim(S_m)$ allows
 - a numerical optimization scheme (EM + dynamic programming)
 - a theoretical control : for κ large enough



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

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 $s_0(y|x)$.
- 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 $\hat{m}$ by
$$\hat{m} = \operatorname{argmin}_{m \in \mathcal{S}} - \sum_{i=1}^n \ln \hat{s}_m(Y_i|X_i) + \operatorname{pen}(m).$$

with $\operatorname{pen}(m)$ well chosen.

- Conditional density estimation result of type :

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

- 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 $s_m \in S_m$

$$\int_0^\sigma \sqrt{H_{[1, d^{\otimes n}]}(\epsilon, S_m(s_m, \sigma))} d\epsilon \leq \phi_m(\sigma).$$

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 $\hat{s}_m$ be a δ -log-likelihood minimizer in S_m :

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

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 $m \in \mathcal{M}$ $\text{pen}(m) \geq \kappa(\mathfrak{D}_m + x_m)$ 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 $\hat{s}_{\hat{m}}$ with $\hat{m}$ defined by

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^n -\ln(\hat{s}_m(Y_i|X_i)) + \text{pen}(m)$$

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

Simplified Theorem...

- Oracle inequality :

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

as soon as

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

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 $\mathfrak{D}_m \propto (\log n) \dim(S_m) \dots$

Kullback, Hellinger and extensions

- Model selection oracle inequality of type

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

- 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 :

$$KL \rightarrow KL^{\otimes n}(s, s') = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n KL(s(\cdot|X_i), s'(\cdot|X_i)) \right],$$
$$JKL_{\rho} \rightarrow JKL_{\rho}^{\otimes n} \quad \text{and} \quad d^2 \rightarrow d^{2 \otimes n}.$$

- 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

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

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

Sketch of proof

- Proof similar to the one of Theorem 7.11 of Massart's « Concentration Inequalities and Model Selection » book.
- For any function $g(x, y)$, we denote $P_n^{\otimes n}(g)$ its empirical process

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

$P^{\otimes n}(g)$ the corresponding expectation

$$P^{\otimes n}(g) = \mathbb{E} [P_n^{\otimes n}(g)] = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n g(X_i, Y_i) \right].$$

and $\nu_n^{\otimes n}(g) = P_n^{\otimes n}(g) - P^{\otimes n}(g)$ the recentered process.

- Let

$$\hat{s}_m = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n}(-\ln s_m) = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n} \left(-\ln \frac{s_m}{s_0} \right)$$

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

- Define

$$\hat{g}_m = -\ln \left(\frac{\hat{s}_m}{s_0} \right) \quad \bar{g}_m = -\ln \left(\frac{\bar{s}_m}{s_0} \right) \quad \hat{f}_m = -\frac{1}{\rho} \ln \frac{\rho \hat{s}_m + (1 - \rho) s_0}{s_0}$$

Upper bound of the likelihood

- By convexity, $\hat{f}_m = -\frac{1}{\rho} \ln \frac{\rho \hat{s}_m + (1-\rho)s_0}{s_0} \leq -\ln \frac{\hat{s}_m}{s_0} = \hat{g}_m$
- Let $m \in \mathcal{S}$, for any m' such that

$$P_n^{\otimes n}(\hat{g}_{m'}) + \frac{\text{pen}(m')}{n} \leq P_n^{\otimes n}(\hat{g}_m) + \frac{\text{pen}(m)}{n} :$$

$$\begin{aligned} P_n^{\otimes n}(\hat{f}_{m'}) + \frac{\text{pen}(m')}{n} &\leq P_n^{\otimes n}(\hat{g}_{m'}) + \frac{\text{pen}(m')}{n} \\ &\leq P_n^{\otimes n}(\hat{g}_m) + \frac{\text{pen}(m)}{n} \\ &\leq P_n^{\otimes n}(\bar{g}_m) + \frac{\text{pen}(m)}{n} \end{aligned}$$

- Let

$$\begin{aligned} P_n^{\otimes n}(\hat{f}_{m'}) - \nu_n^{\otimes n}(\bar{g}_m) \\ \leq P_n^{\otimes n}(\bar{g}_m) + \frac{\text{pen}(m)}{n} - \nu_n^{\otimes n}(\hat{f}_{m'}) - \frac{\text{pen}(m')}{n} \end{aligned}$$

Oracle inequality up to something

- Previous inequality can be rewritten as

$$\begin{aligned} JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(\bar{g}_m) \\ \leq KL^{\otimes n}(s_0, \bar{s}_m) + \frac{\text{pen}(m)}{n} \\ - \nu_n^{\otimes n}(\hat{f}_{m'}) - \frac{\text{pen}(m')}{n} \end{aligned}$$

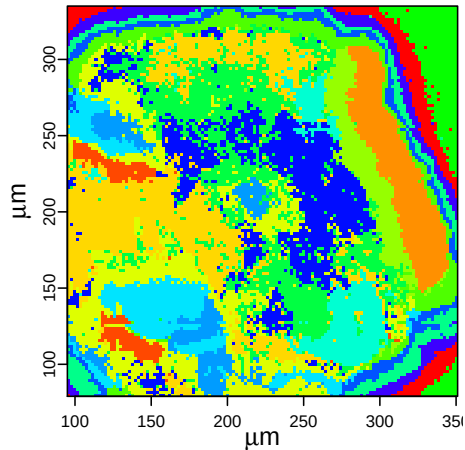
- Appear

- the integrated error for model m' : $JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'})$
 - a simple and centered process : $-\nu_n^{\otimes n}(\bar{g}_m)$,
 - the oracle $KL^{\otimes n}(s_0, \bar{s}_m) + \frac{\text{pen}(m)}{n}$
 - a random rest $-\nu_n^{\otimes n}(\hat{f}_{m'}) - \frac{\text{pen}(m')}{n}$
- It remains to control the deviation of $-\nu_n^{\otimes n}(\hat{f}_{m'})$ by $\epsilon JKL_{\rho}^{\otimes n}(s_0, \hat{s}_{m'}) + \frac{\text{pen}(m')}{n}$ as soon as $\text{pen}(m') \geq \kappa(n\sigma_{m'}^2 + x_m) \dots$

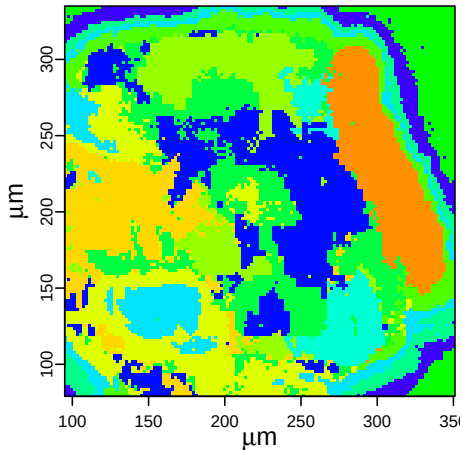
Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

Without



With



30 min.

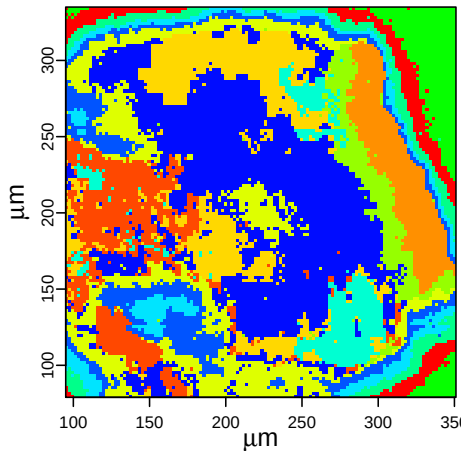
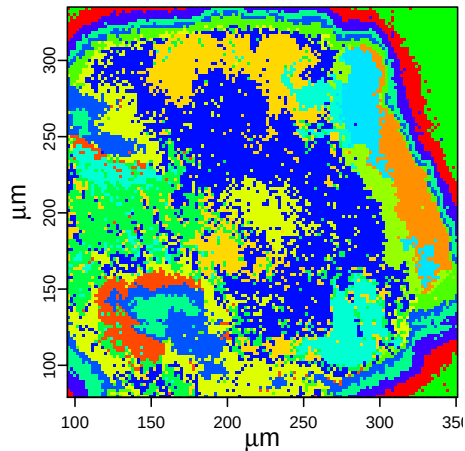
- Automatic choice of K , $[L_k D A]^K$ and partition.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

Without

With



5 min.

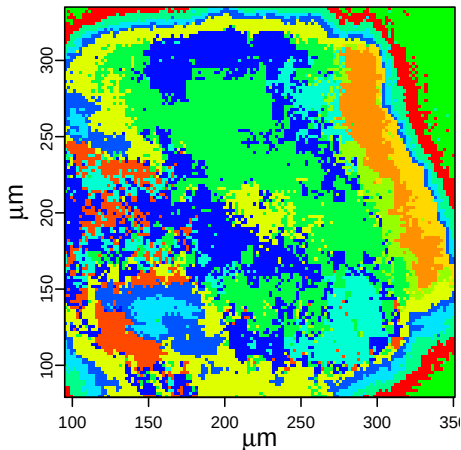
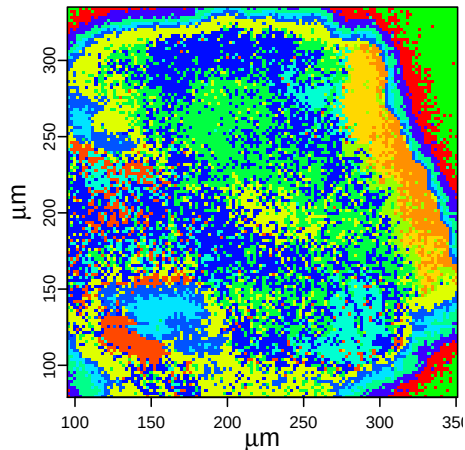
- Automatic choice of K , $[L_k D A]^K$ and partition.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

Without

With



1 min.

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

Numerical optimization

- Penalized Model Selection :

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

- 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 $[\mu \ L \ D \ A]^K$:

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

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

Maximum likelihood estimation

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

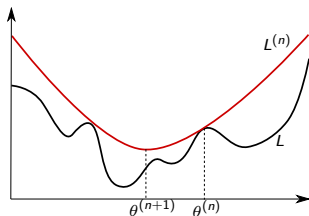
$$\operatorname{argmin}_{\mu, \Sigma, \mathcal{P}, \pi} - \sum_{i=1}^N \ln \left(\sum_{k=1}^K \pi_k [\mathcal{R}(x_i)] \mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right) + \lambda_{0,n} |\mathcal{P}| (K - 1)$$

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

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

- 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

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

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

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

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

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

Partition and EM Algorithm

- Maximum likelihood :

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

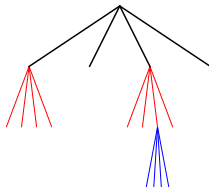
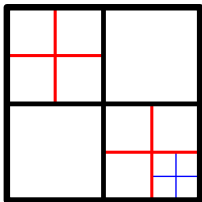
- E Step : with $P_k^{i,(n)} = P(k_i = k | x_i, \mathcal{S}_i, \mu^{(n)}, \Sigma^{(n)}, \mathcal{P}^{(n)}, \pi^{(n)})$

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

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

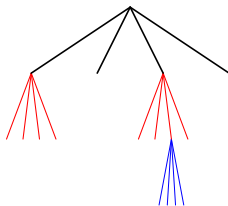
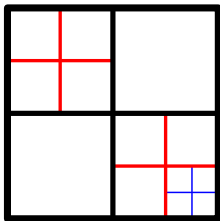


- Optimization in $(\mathcal{P}, \pi)$ of

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

- 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 $\operatorname{argmin}_{\mathcal{P}} \sum_{\mathcal{R} \in \mathcal{P}} C[\mathcal{R}]$ where $\mathcal{P}$ is a recursive dyadic partition (associated to quadtree) of limited depth.
- Key property : the optimal partition $\hat{\mathcal{P}}[\mathcal{R}]$ of a dyadic square is
 - either this square, $\hat{\mathcal{P}}[\mathcal{R}] = \{\mathcal{R}\}$
 - or the union of the opt. part. of its children, $\hat{\mathcal{P}}[\mathcal{R}] = \cup_{\mathcal{R}' \in \text{Child}[\mathcal{R}]} \hat{\mathcal{P}}[\mathcal{R}']$ with a decision based on

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

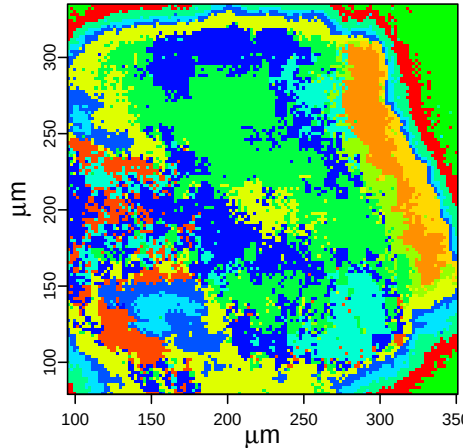
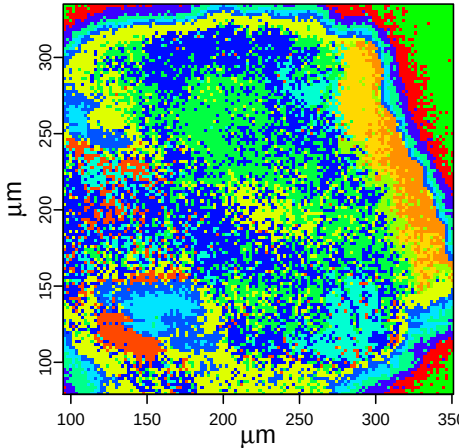
- Fast bottom-up algorithm.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

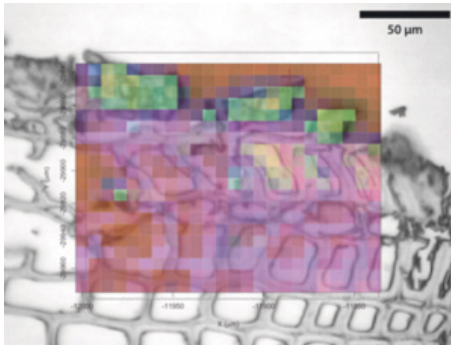
Without

With



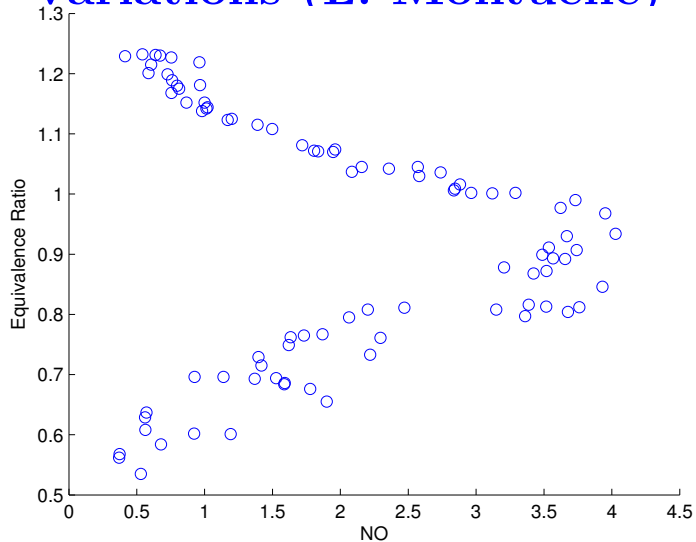
- Automatic choice of K , $[L_k D A]^K$ and partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by random projection.

Stradivari's Secret

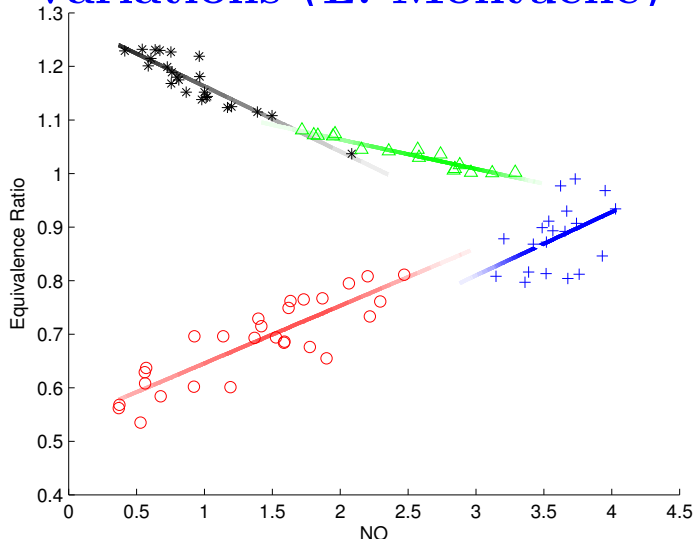


- Two fine layers of varnish :
 - a first simple oil layer, similar to the painter's one, penetrating mildly the wood,
 - a second layer made from a mixture of oil, pine resin and red pigments.
- Classical technique up to the specific color choice (and a very good varnishing skill).
- Stradivari's secret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

Variations (L. Montuelle)

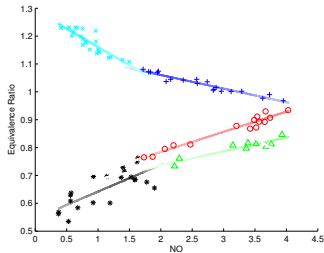
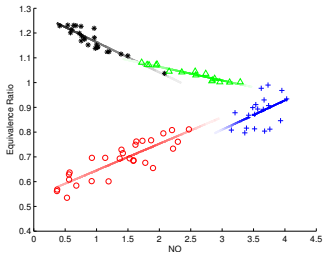
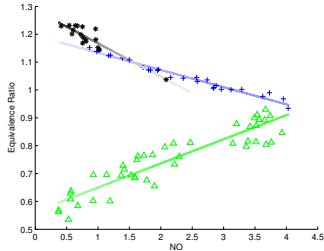
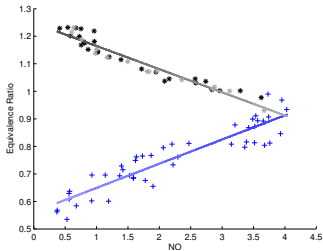


Variations (L. Montuelle)



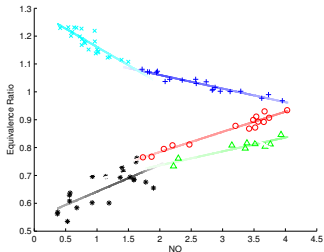
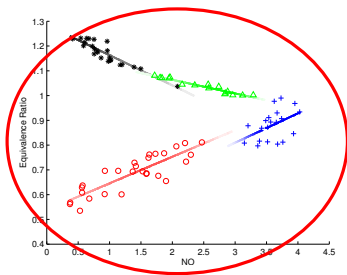
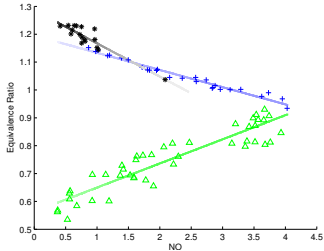
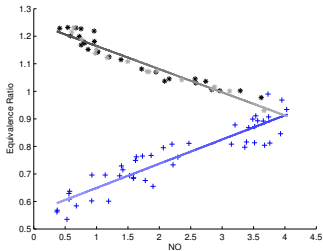
● Regression Mixture

Variations (L. Montuelle)



- Regression Mixture
- Mixtures of Gaussian regressions with logistic weights

Variations (L. Montuelle)



- Regression Mixture
- Mixtures of Gaussian regressions with logistic weights
- Model selection

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.
- Enhanced Spatialized Gaussian Mixture Model with piecewise logistic weights (L. Montuelle).
- Perspectives :
 - Formal link between conditional density estimation and unsupervised segmentation.
 - Penalty calibration by slope heuristic.
 - Dimension reduction adapted to unsupervised segmentation/classification.