Unsupervised hyperspectral image segmentation, Conditional density estimation and Penalized maximum likelihood model selection

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

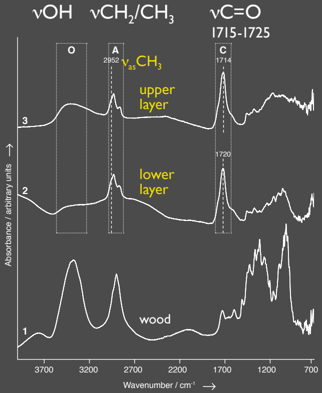
CLAPEM XII - Viña del Mar
27 Marzo 2012

informatiques mathématiques

A. Stradivari (1644 - 1737)

Provigny (1716)

A. Giordan © Cité de la Musique

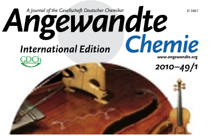


SOLEIL SYNCHROTRON

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.

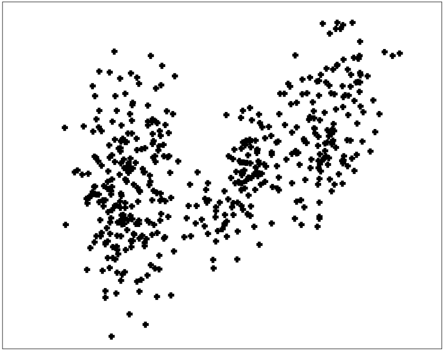
Outline

Hyperspectral Image Segmentation

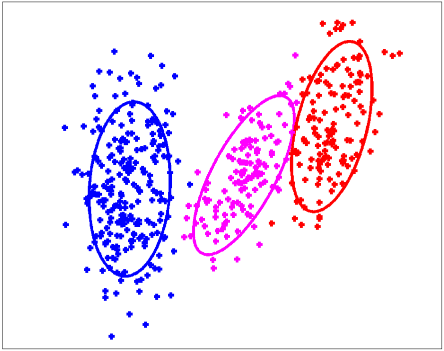
- Data:
  - image of size N between ~ 1000 and ~ 100000 pixels,
  - spectrums $\mathcal{S}$ of ~ 1024 points,
  - very good spatial resolution,
  - ability to measure a lot of spectrums per minute,
- Immediate goal:
  - automatic image segmentation,
  - without human intervention,
  - help to data analysis.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

A *Toy* Problem

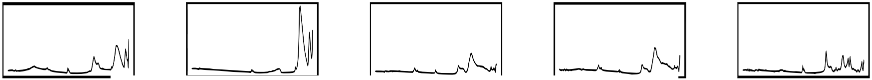
A *Toy* Problem

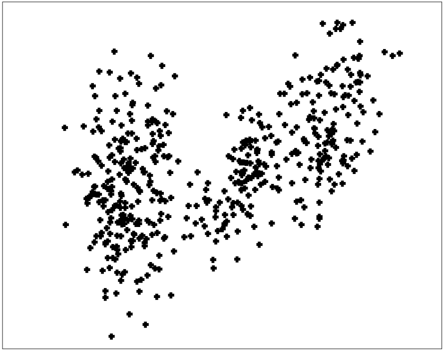


A *Toy* Problem

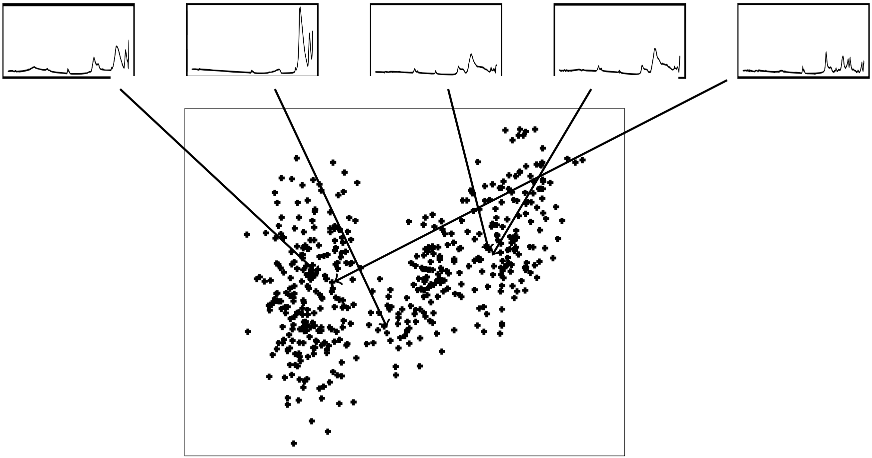


A *Toy* Problem

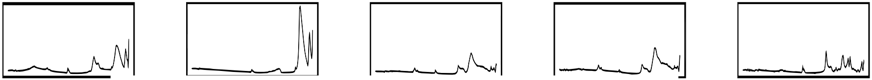


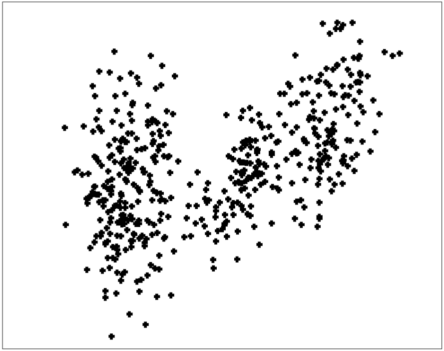


A *Toy* Problem

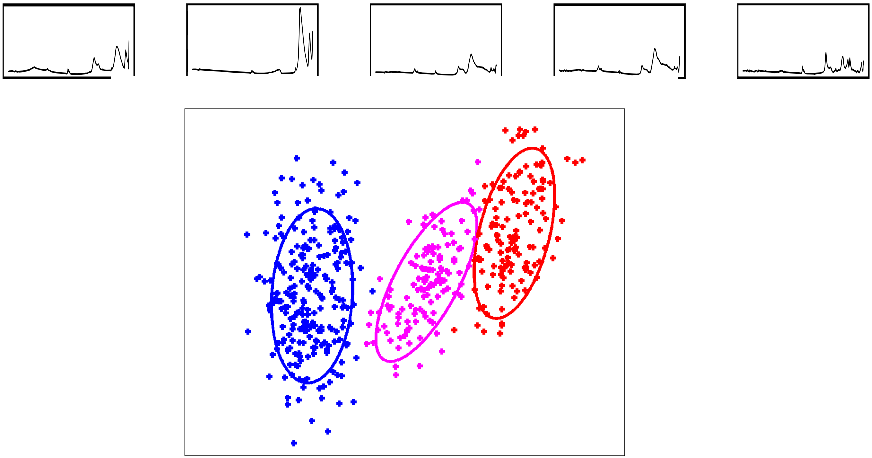


A *Toy* Problem

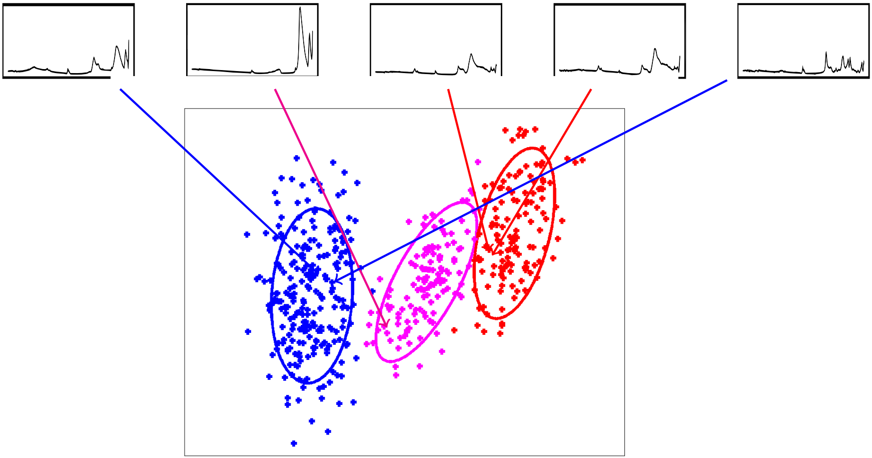




A *Toy* Problem



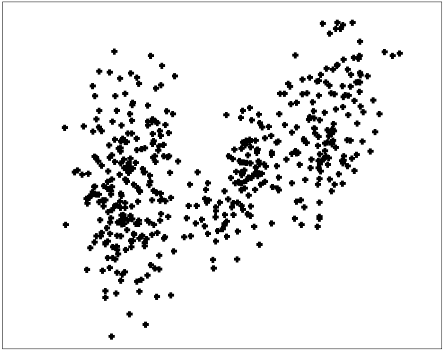
A *Toy* Problem



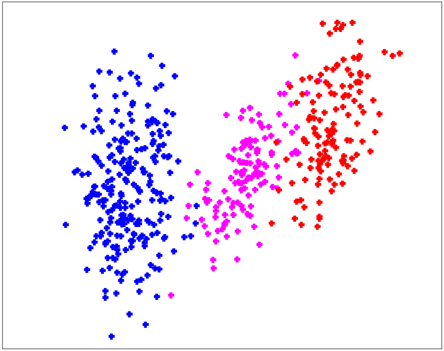
- Representation: mapping between spectrums and points in a large dimension space.
- Spectral method.

Stochastic Modelization

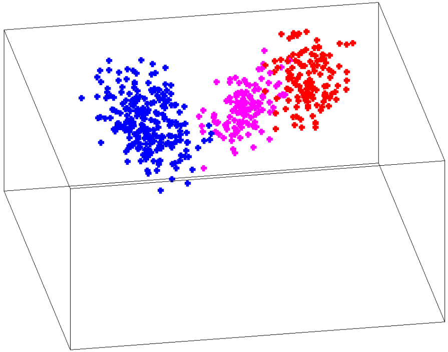
Stochastic Modelization



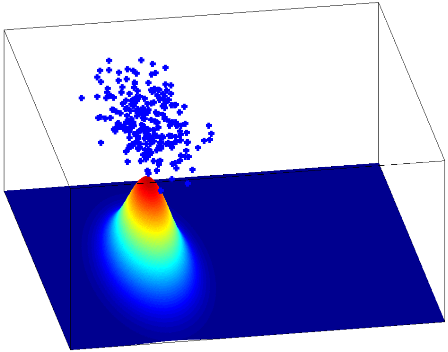
Stochastic Modelization



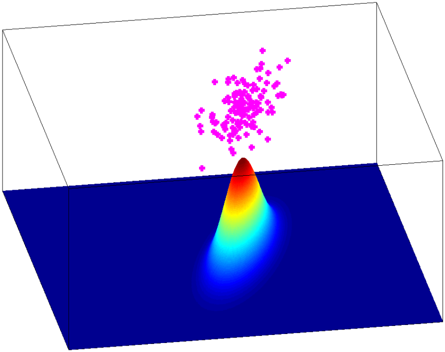
Stochastic Modelization



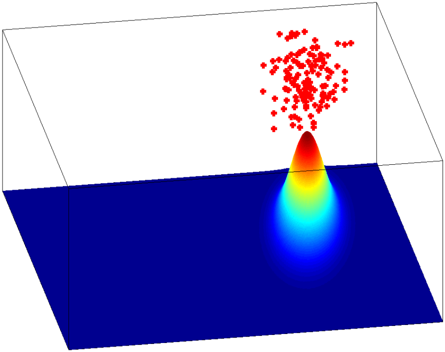
Stochastic Modelization



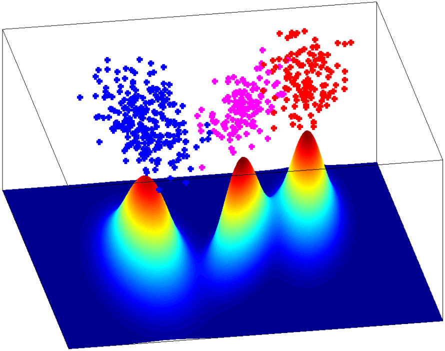
Stochastic Modelization



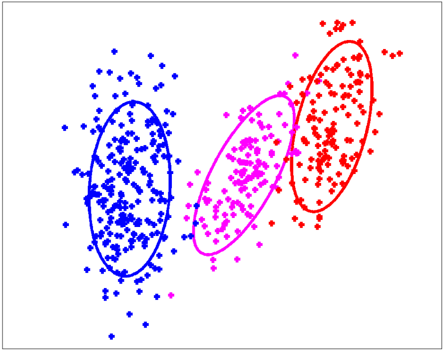
Stochastic Modelization



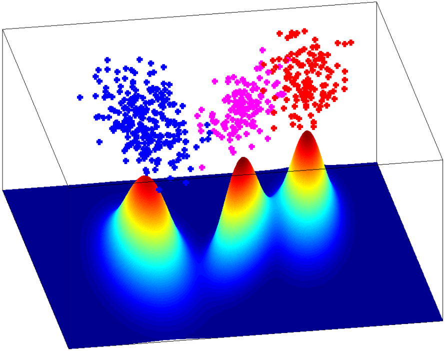
Stochastic Modelization



Stochastic Modelization



Stochastic Modelization

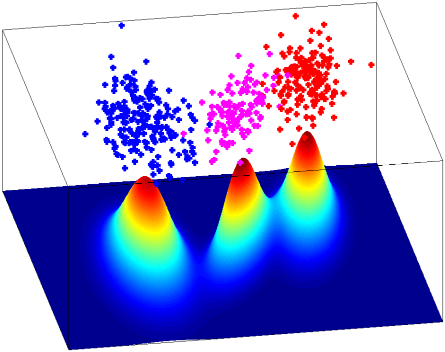


- Model : Gaussian Mixture with K classes.
- Mixture density:

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

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

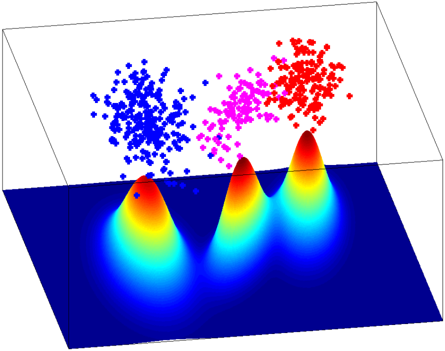
Stochastic Modelization



- 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 Modelization

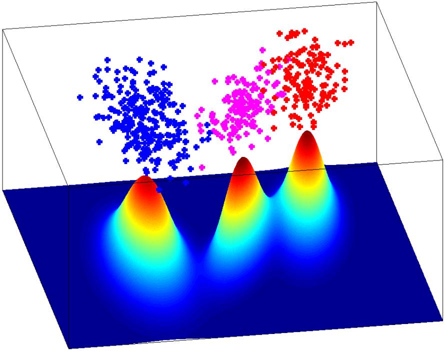


- Model : Gaussian Mixture with K classes.
- Mixture density:

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

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

Stochastic Modelization

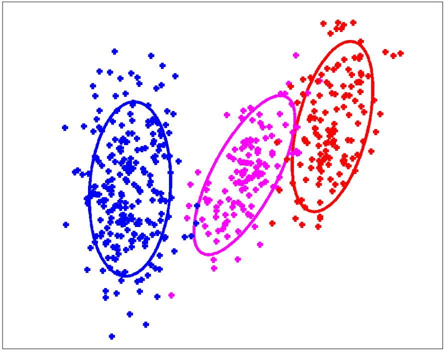


- Model : Gaussian Mixture with K classes.
- Mixture density:

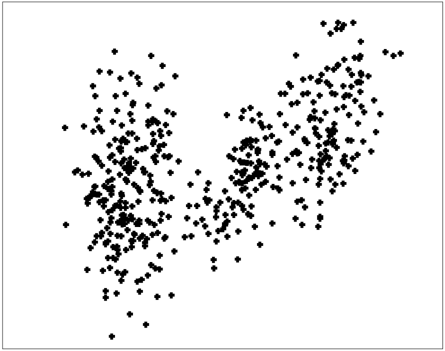
$$
\begin{aligned}
s_{K,\pi,\mu,\Sigma}(\mathcal{S}) &= \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(\mathcal{S}-\mu_k)^t \Sigma_k^{-1}(\mathcal{S}-\mu_k)} \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})
\end{aligned}
$$

Statistical Estimation

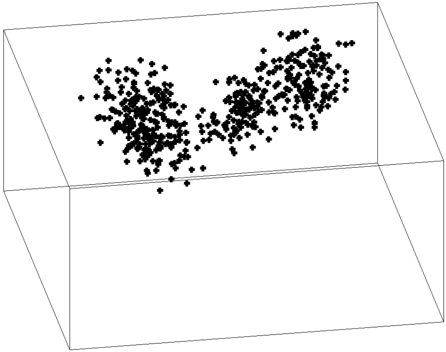
Statistical Estimation



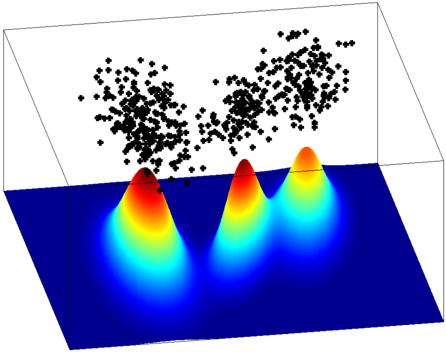
Statistical Estimation



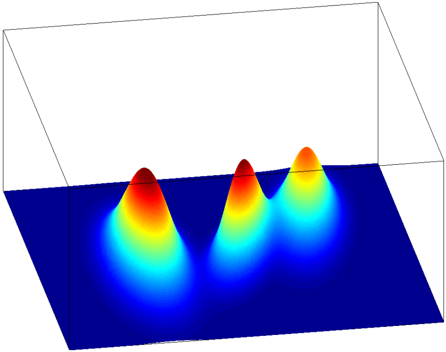
Statistical Estimation



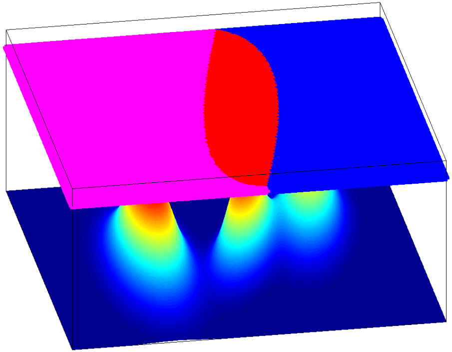
Statistical Estimation



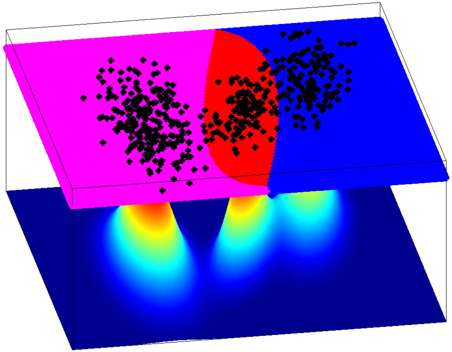
Statistical Estimation



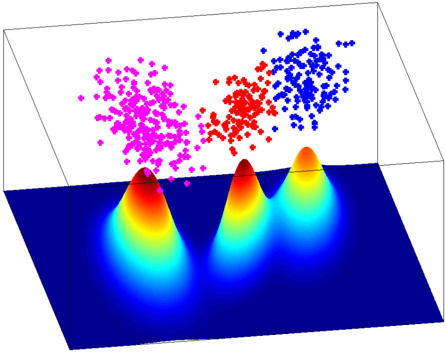
Statistical Estimation



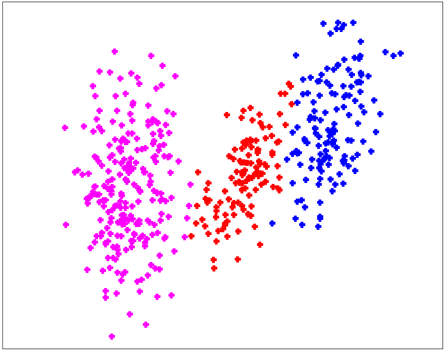
Statistical Estimation



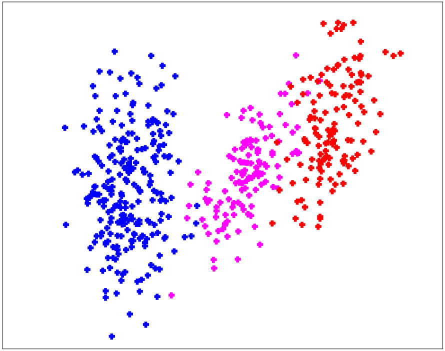
Statistical Estimation



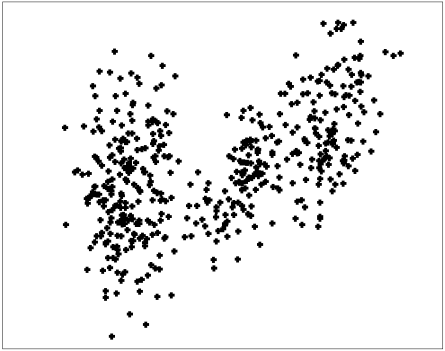
Statistical Estimation



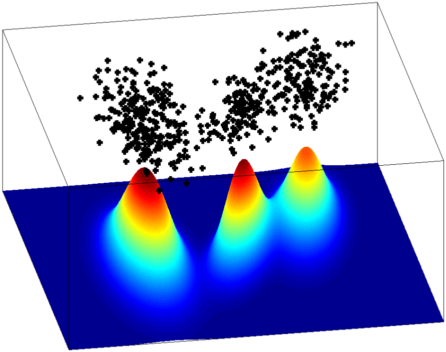
Statistical Estimation



Statistical Estimation



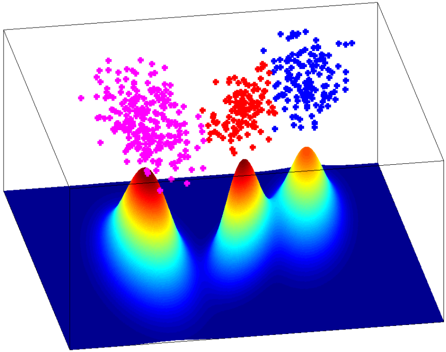
Statistical Estimation



- 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):

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

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

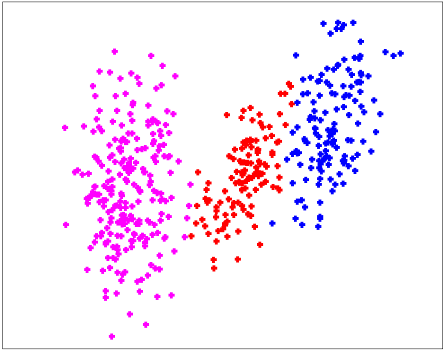
- True density s_0 of $\mathcal{S}$ close from

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

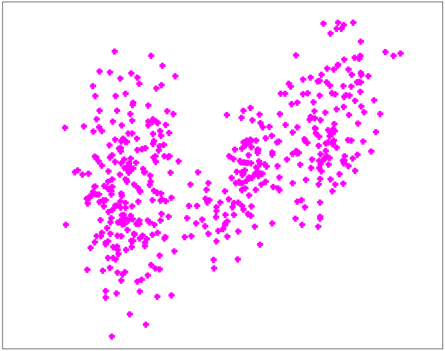
- Gaussian Mixture Model $S_m = \{s_m\}$ specified by
  - a number of classes K,
  - a structure for the means μ_k and the covariance matrices $\Sigma_k = L_k D_k A_k D_k'$
- Structure $[\mu\, L\, D\, A]^K$: structural constraints (know, common or free values...) on the means μ_k , the volumes L_k, the diagonalization basis D_k and the rescaled eigenvalues A_k 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?

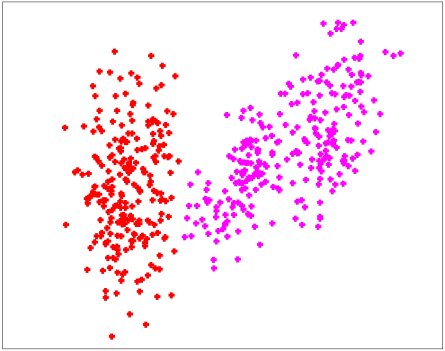
How many classes?



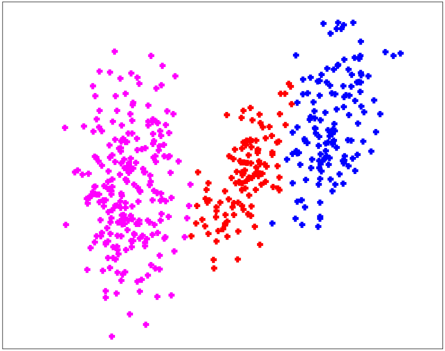
How many classes?



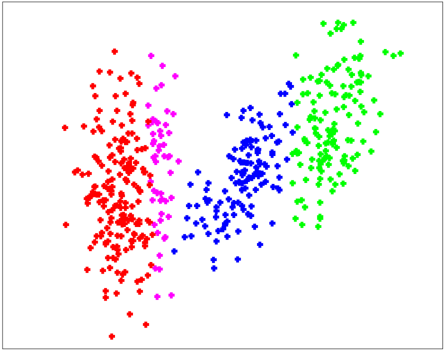
How many classes?



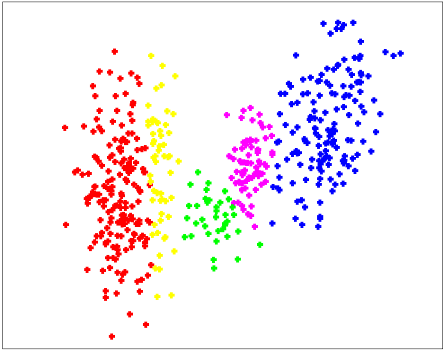
How many classes?



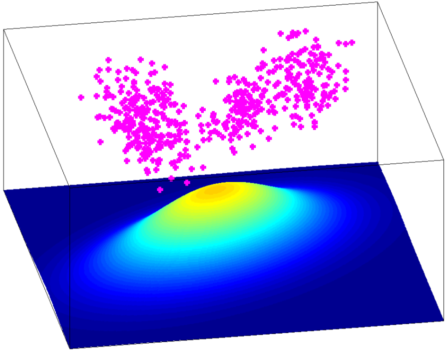
How many classes?



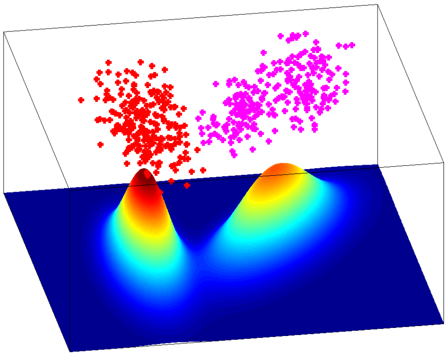
How many classes?



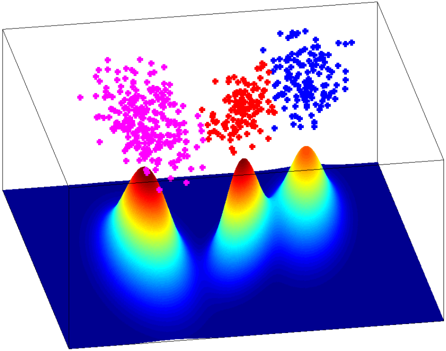
How many classes?



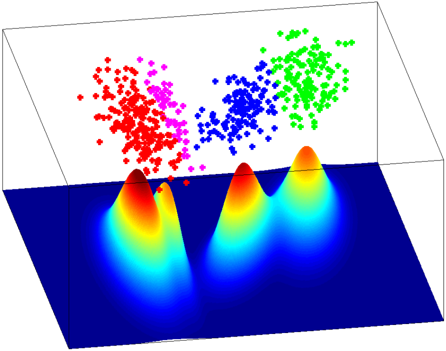
How many classes?



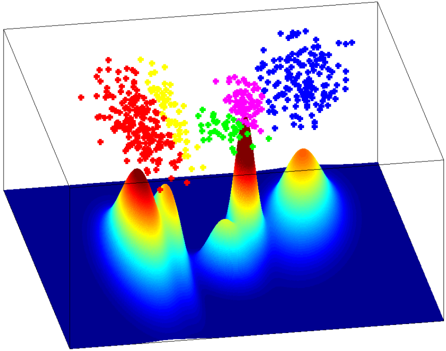
How many classes?



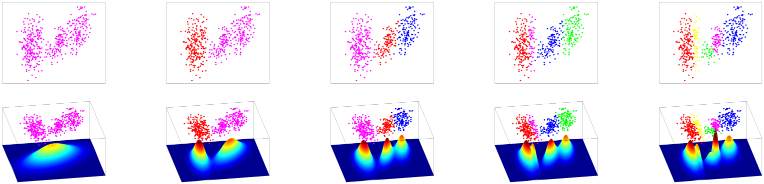
How many classes?



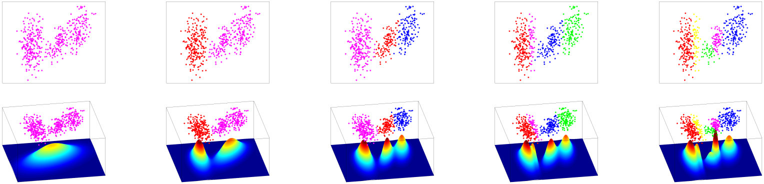
How many classes?



How many classes?

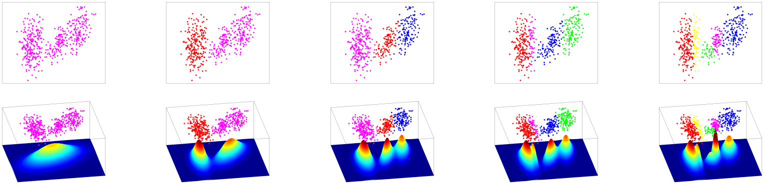


How many classes?



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

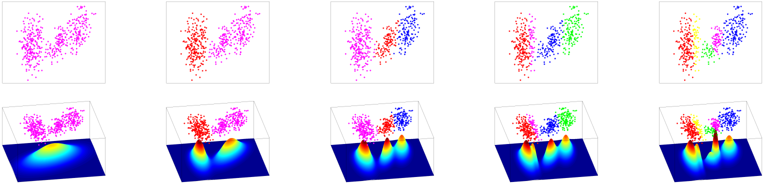
How many classes?



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

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

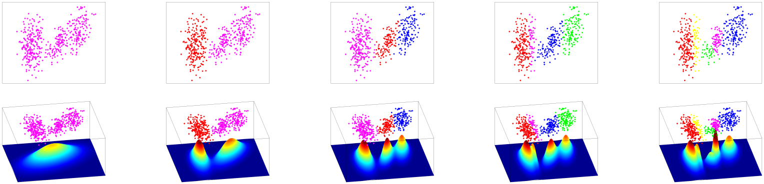
How many classes?



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

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

How 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)

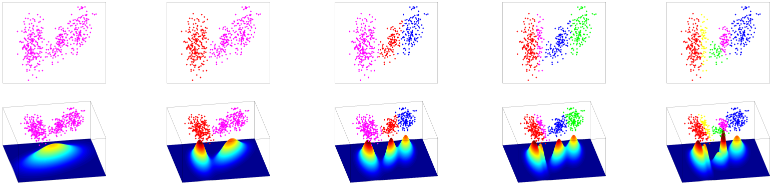
Ockham's Razor

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

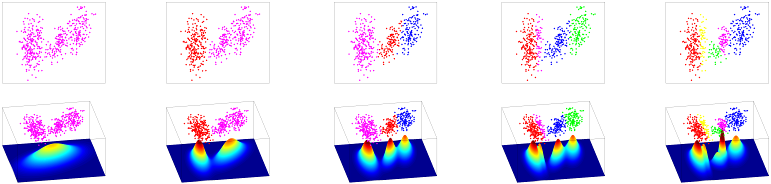
- Ockham's Razor (simplicity principle): one should not add hypotheses, if the current ones are already sufficient!
- Balance between observation explanation power and simplicity.

Selection by Penalization

Selection by Penalization

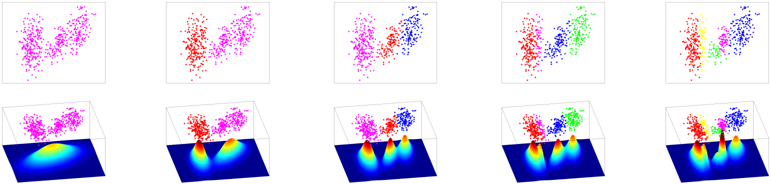


Selection by Penalization



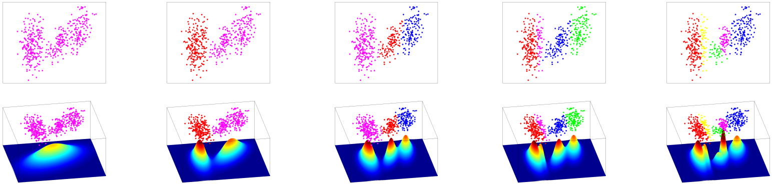
Likelihood	−−	+	+++	+++·	++++
Simplicity	++++	+++	++	+	−

Selection by Penalization



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

Selection by Penalization

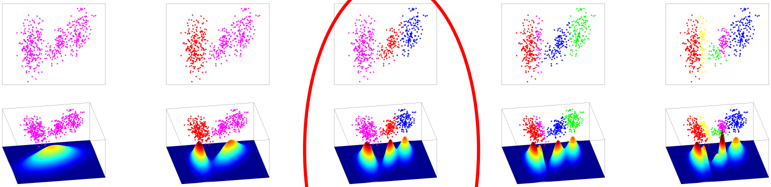


Likelihood	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicity	$++++$	$+++$	$++$	$+$	$-$
$=$ Tradeoff	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

- Likelihood: $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicity: $-\lambda \mathrm{Dim}(S_K)$ (a lot of theory behind that).
- 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

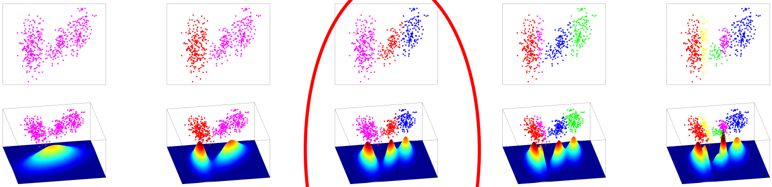


Likelihood	−−	+	+++	+++·	++++
+ Simplicity	++++	+++	++	+	−
= Tradeoff	++	++++	+++++	++++·	+++

- Likelihood: $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicity: $-\lambda \mathrm{Dim}(S_K)$ (a lot of theory behind that).
- 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	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

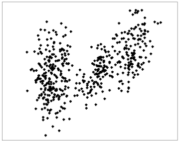
- Likelihood: $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicity: $-\lambda \mathrm{Dim}(S_K)$ (a lot of theory behind that).
- Penalized estimator:

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

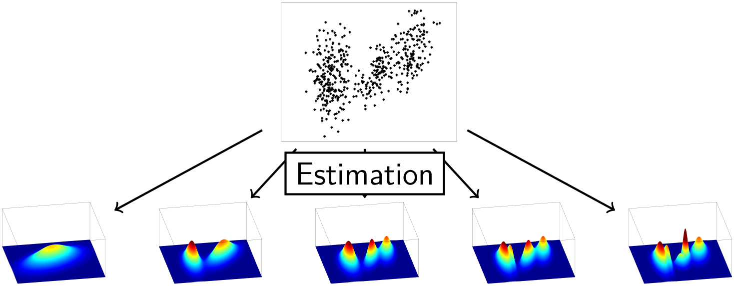
- Optimization in K by exhaustive exploration!

Methodology

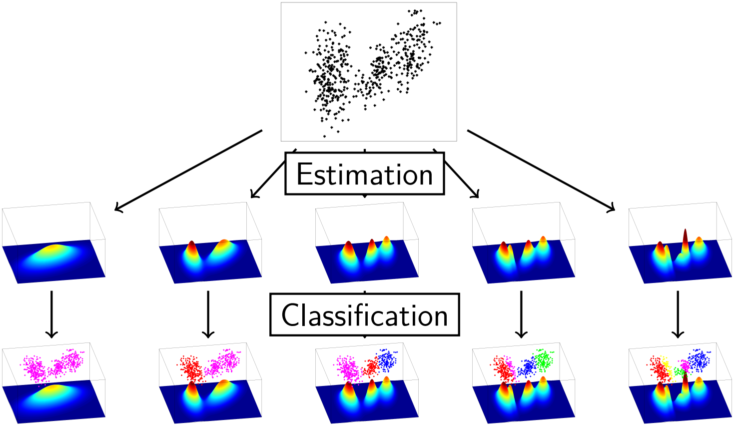
Methodology



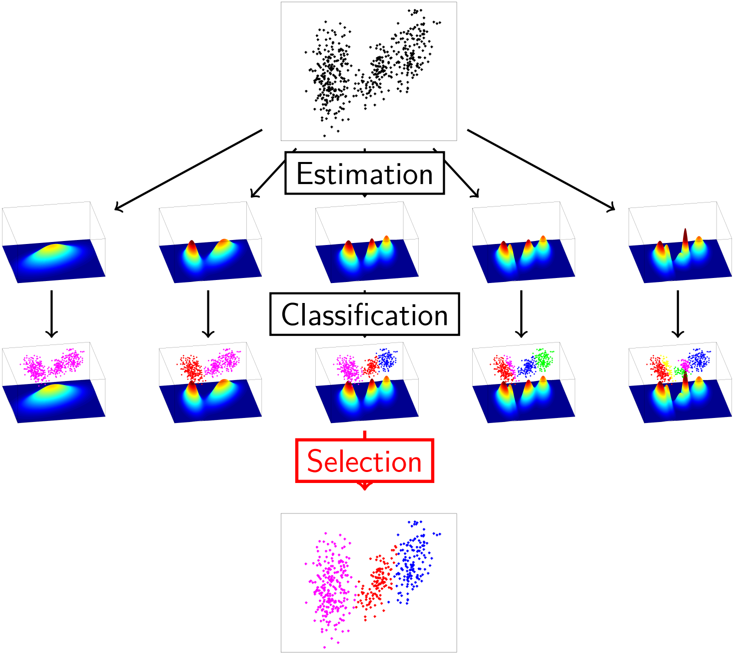
Methodology



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

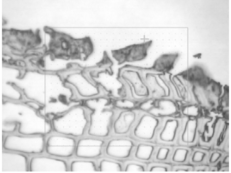
$$\widehat{m} = \mathrm{argmin} - \ln(\widehat{s}_m) + \mathrm{pen}(m).$$

  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,

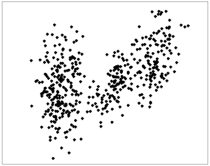
$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

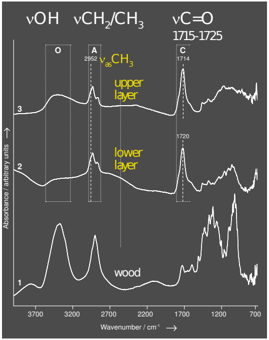
  - Clustering or unsupervised classification ($\neq$ segmentation): numerical results.
  - Consistency of the classification as soon as $\ln \ln(n)$ in the penalty...

Back to our violins



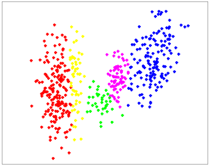
Representation

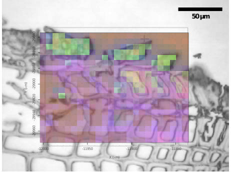




Segmentation

Classification





Spatial Info.

Segmentation and Spatialized GMM

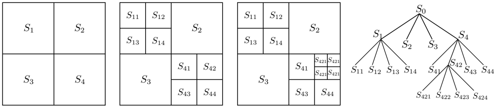
- Initial goal: segmentation $\neq$ clustering.
- Idea of Kolaczyk et al (cf Bigot): take into account the spatial position x of the spectrum in the mixing proportions .
- Conditional density model:

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

- 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 $\pi_k(x)$.
- Non parametric model ($\pi_k(x)$ function): regularization required!
- Model selection principle...

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 $s_m \in S_m$

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

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 are two constants κ_0 and C_2 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 $$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \right] \leq C_1 \inf_{S_m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + C_2 \frac{\Sigma}{n} + \frac{\delta}{n}.$$

Simplified Theorem...

- Oracle inequality:

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S_m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + C_2 \frac{\Sigma}{n} + \frac{\delta}{n}$$

  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 convexification 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 probability measures $s d\lambda$ and $t d\lambda$ and all $\rho \in (0, 1)$

$$C_\rho \, d^2_\lambda(s, t) \leq JKL_{\rho,\lambda}(s, t) \leq KL_\lambda(s, t)$$

  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 $\rho \simeq 1/2$.

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:

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

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

- Similar approach but difference for Jensen-Kullback-Leibler and Hellinger and possibility to have a result with expectation on the design.
- 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

- Oracle inequality:

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

- A good $\mathrm{pen}(m)$ should be of order $\mathbb{E}\left[\left|KL^{\otimes_n}(s_0,\widehat{s}_m) - \mathbb{E}\left[KL^{\otimes_n}(s_0,\widehat{s}_m)\right]\right|\right]$ (variance term).
- Control in expectation requires a larger $\mathrm{pen}(m)$:
  - with an intrinsic term corresponding to the complexity of the model (upper bound of the variance/deviation bound),
  - and with a term corresponding to the complexity of the collection (simultaneous control on all the collection/union bound)
- Complexity used here:
  - Model (entropy): $\mathfrak{D}_m$ 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 in the proof...)
  - Collection (coding): Kraft type inequality $\displaystyle\sum_{m\in\mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- Classical constraint on the penalty

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

Back to the spatialized GMM

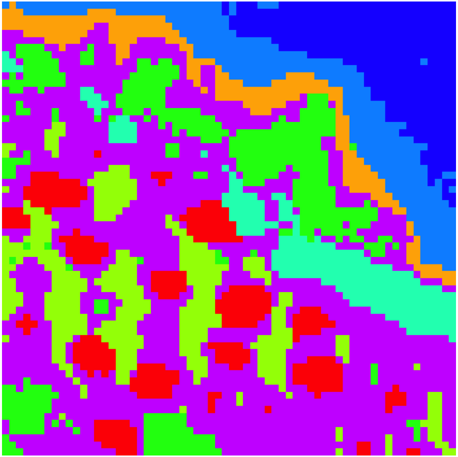
- Computation of an upper bound of $H_{[\cdot],d^{\otimes n}}(\epsilon, S_m)$ for the spatialized GMM (cf Maugis and Michel):
  - Bound on an upper bound of the entropy: $H_{[\cdot],d^{\sup}}(\epsilon, S_m)$ where $d^{\sup} = \sqrt{d^{2\,\sup}} = \sqrt{\sup_x d^2(s(\cdot|x), s'(\cdot|x))}$,
  - Result valid for every structure ($[\mu\, L\, D\, A]^K$) and every partition:
  $$H_{[\cdot],d^{\sup}}(\epsilon, S_m) \leq \dim(S_m)(C + \ln \frac{1}{\epsilon})$$
  with an (almost) explicit common C and $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
  - Consequence: $\mathfrak{D}_m \leq \kappa' \left(C' + \frac{1}{2} \left(\ln \left(\frac{n}{C' \dim(S_m)} \right) \right)_+ \right) \dim(S_m)$.
- Collection coding with $x_m \leq \kappa'' |\mathcal{P}| \leq \frac{\kappa''}{K-1} \dim(S_m)$.
- Condition on the penalty:

$$\mathrm{pen}(m) \geq \left(\kappa' \left(C' + \frac{1}{2} \left(\ln \left(\frac{n}{C' \dim(S_m)} \right) \right)_+ \right) + \frac{\kappa''}{K-1} \right) \dim(S_m).$$

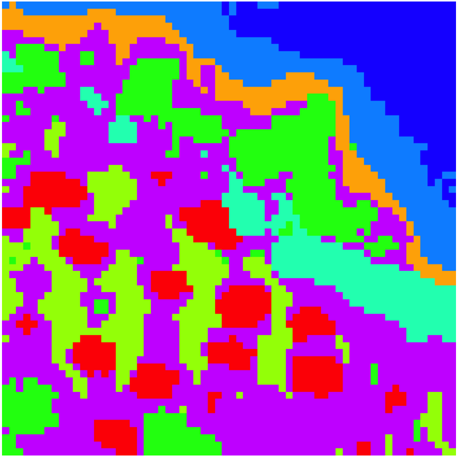
Unsupervised Segmentation

- Numerical result taking into account the spatial modeling:

Without



With

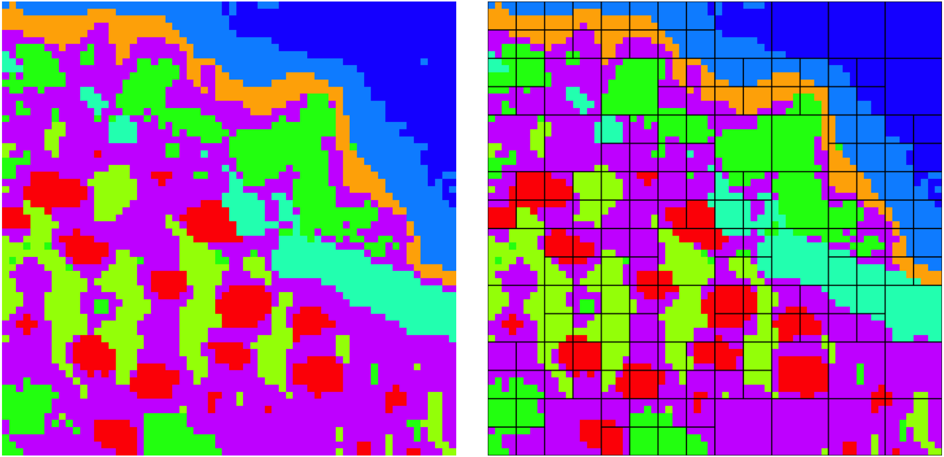


- $K = 8$, $[L_k\, D\, A]^K$ and optimal partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by (not so naive) PCA...

Unsupervised Segmentation

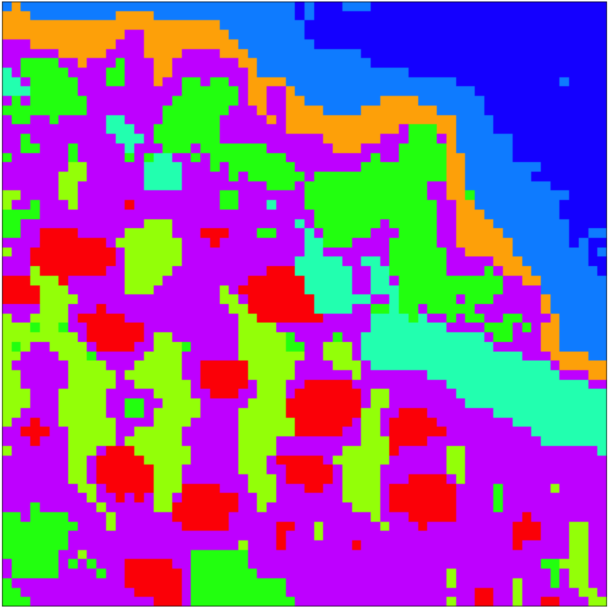
- Numerical result taking into account the spatial modeling:

Without | With

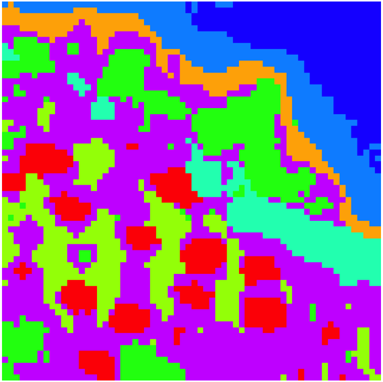


- $K = 8$, $[L_k\, D\, A]^K$ and optimal partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by (not so naive) PCA...

Segmentations



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.
- 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.
  - Other (partition based) conditional density estimators...
- Perspectives:
  - Formal link between conditional density estimation and unsupervised segmentation.
  - Penalty calibration by slope heuristic.
  - Dimension reduction adapted to unsupervised segmentation/classification.
  - Enh. Spatialized GMM with piecewise logistic weights (L. Montuelle).