Hyperspectral Image Segmentation by Spatialized Gaussian Mixtures and Model Selection

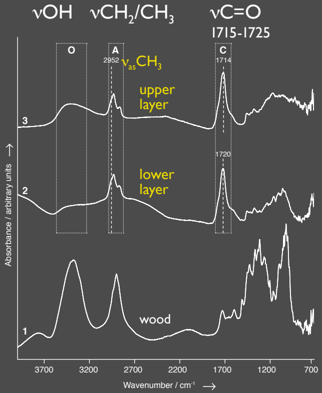
E. Le Pennec
(SELECT - Inria Saclay / Université Paris Sud)
and
S. Cohen (IPANEMA - CNRS / Soleil)

Marseille
25 November 2011

A. Stradivari (1644 - 1737)

Provigny (1716)

A. Giordan © Cité de la Musique

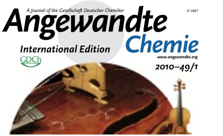


SOLEIL SYNCHROTRON

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 · C2RMF · CENTRE DE RECHERCHE ET DE RESTAURATION DES MUSÉES DE FRANCE · Laboratoire de Dynamique, Interactions et Réactivité



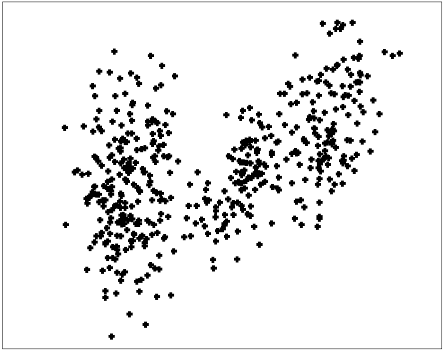
J.-P. Echard, L. Bertrand, A. von Bohlen, A.-S. Le Hô, C. Paris, L. Bellot-Gurlet, B. Soulier, A. Lattuati-Derieux, S. Thao, L. Robinet, B. Lavédrine, and S. Vaiedelich. *Angew. Chem. Int. Ed.*, 49(1), 197-201, 2010.

Hyperspectral Image Segmentation

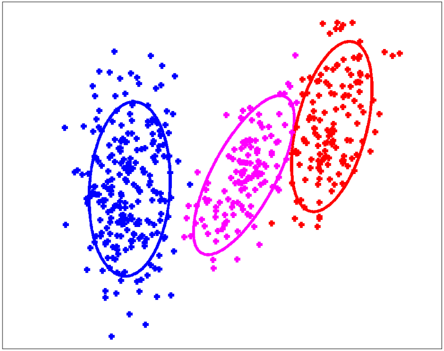
- Data :
  - image of size N between ~ 1000 and ~ 100000 pixels,
  - spectrums $\mathcal{S}$ of ~ 1024 points,
  - very good spatial resolution,
  - ability to measure a lot of spectrums per minute,
- Immediate goal :
  - automatic image segmentation,
  - without human intervention,
  - help to data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

A "Toy" Problem

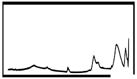
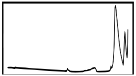
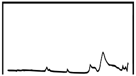
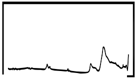
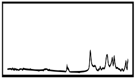
A "Toy" Problem

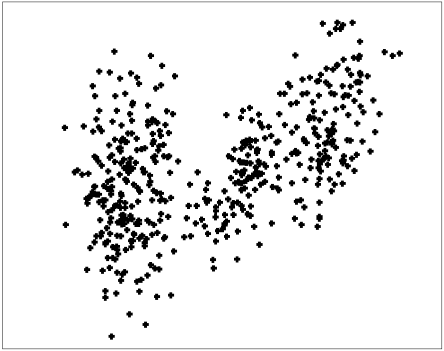


A "Toy" Problem

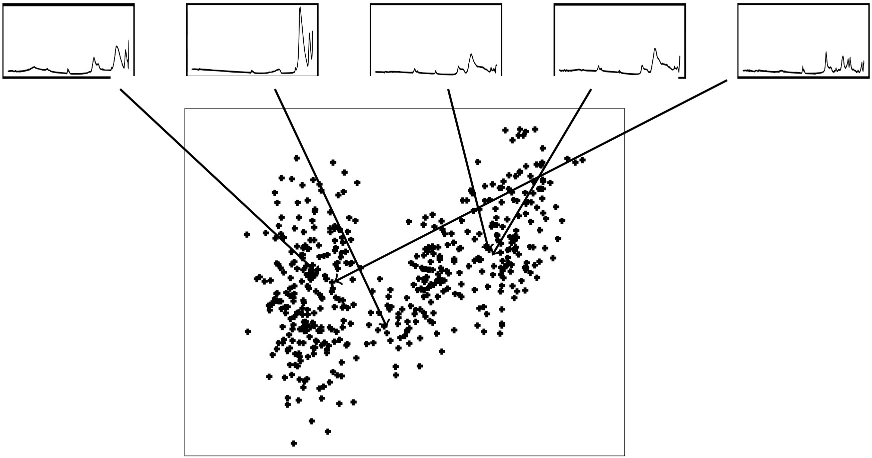


A "Toy" Problem

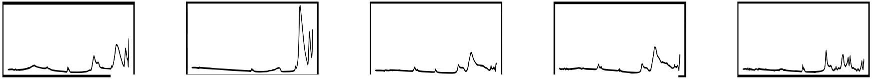






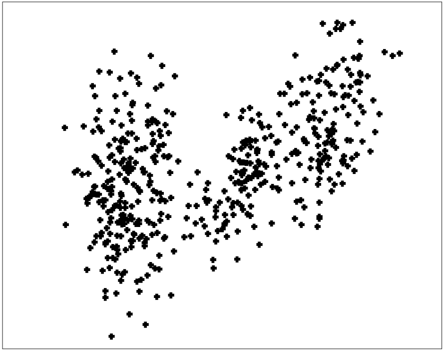


A “Toy” Problem

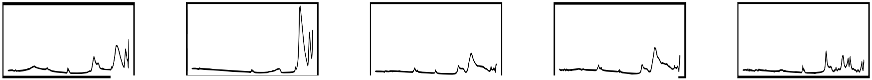


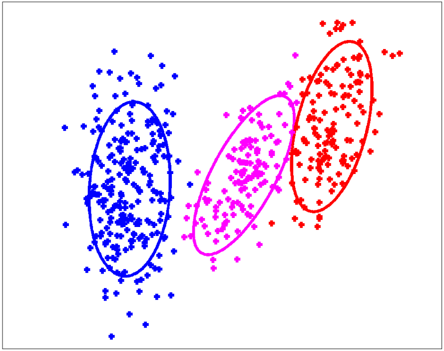
A "Toy" Problem



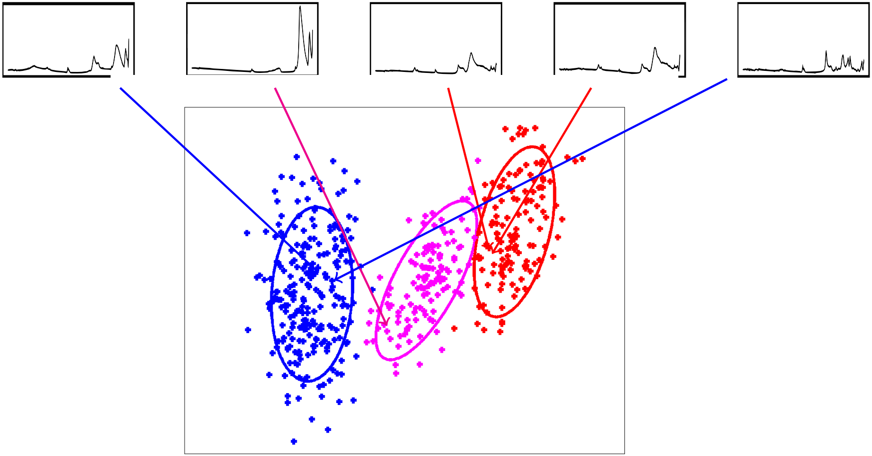


A “Toy” Problem





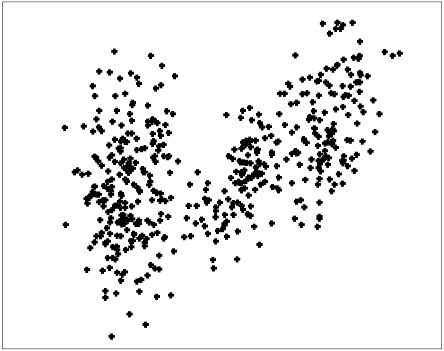
A "Toy" Problem



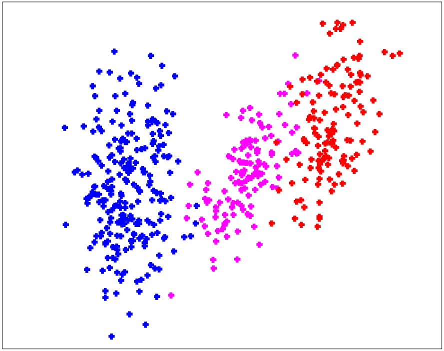
- Representation : mapping between spectrums and points in a large dimension space.
- Spectral method.

“Stochastic” Modelization

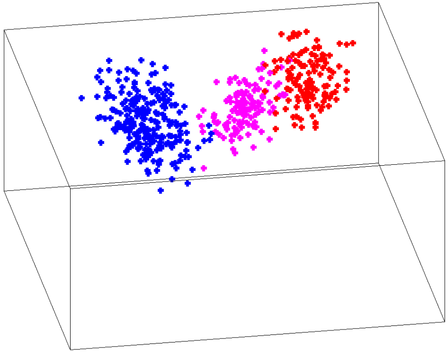
"Stochastic" Modelization



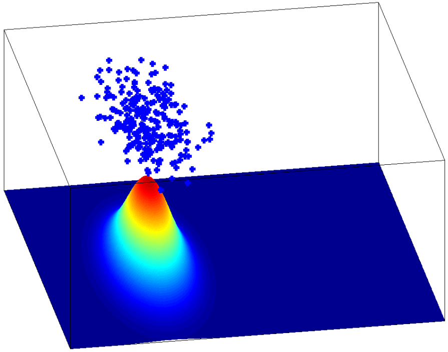
“Stochastic” Modelization



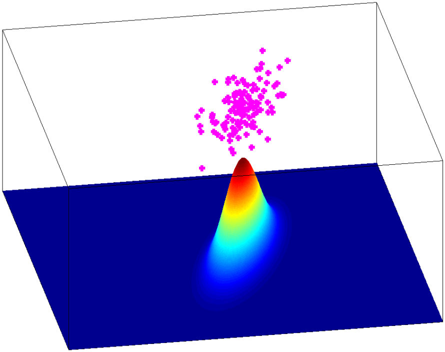
“Stochastic” Modelization



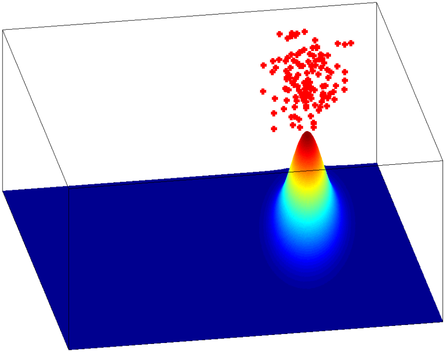
"Stochastic" Modelization



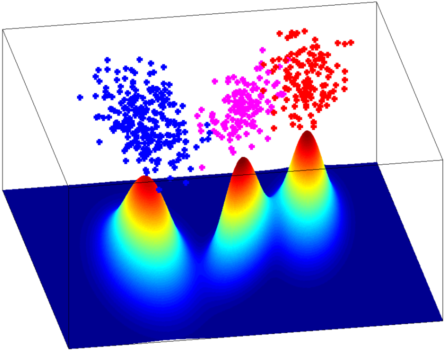
"Stochastic" Modelization



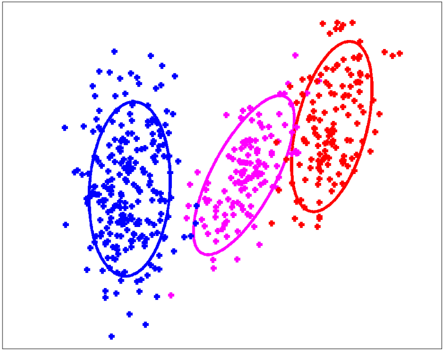
"Stochastic" Modelization



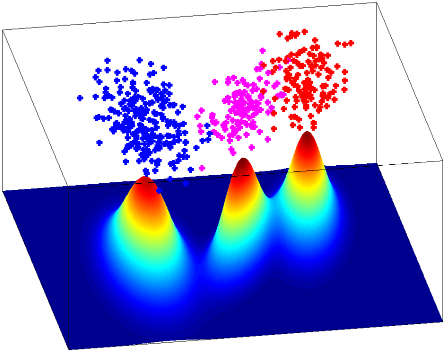
“Stochastic” Modelization



“Stochastic” Modelization



"Stochastic" Modelization

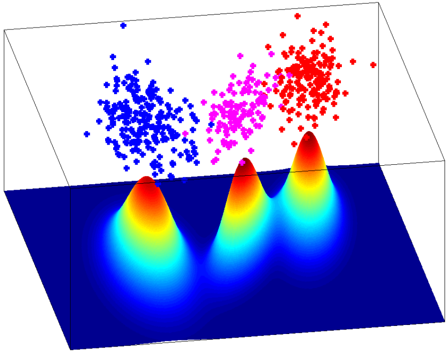


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

“Stochastic” Modelization

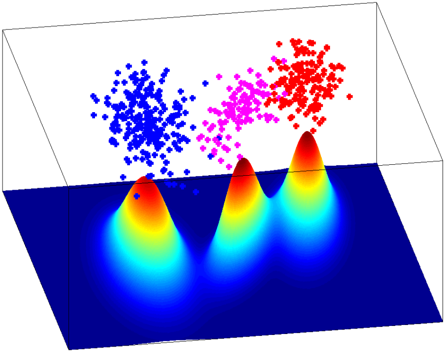


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

"Stochastic" Modelization

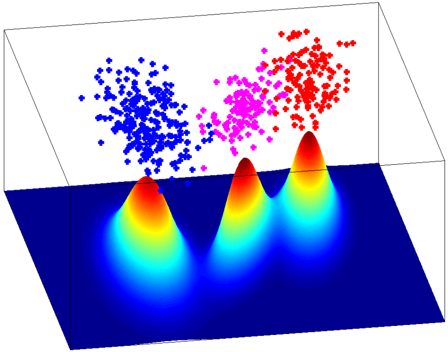


- Model : Gaussian Mixture with K classes.
- Mixture density :

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

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

“Stochastic” Modelization



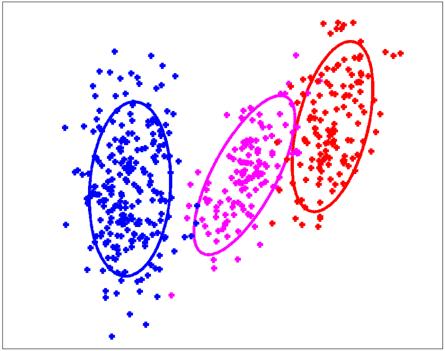
- Model : Gaussian Mixture with K classes.
- Mixture density :

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

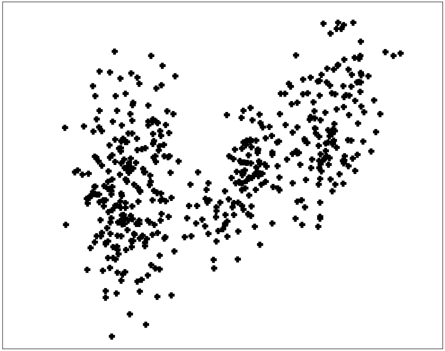
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$$

“Statistical” Estimation

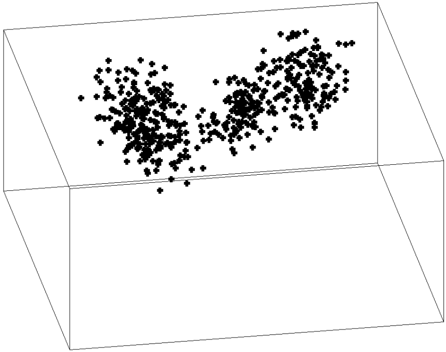
“Statistical” Estimation



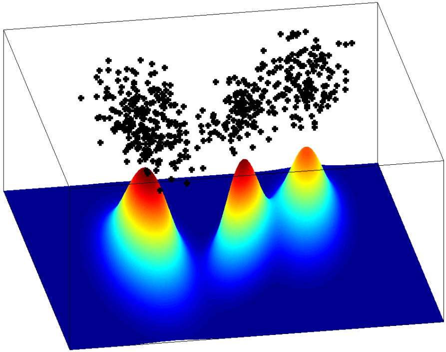
“Statistical” Estimation



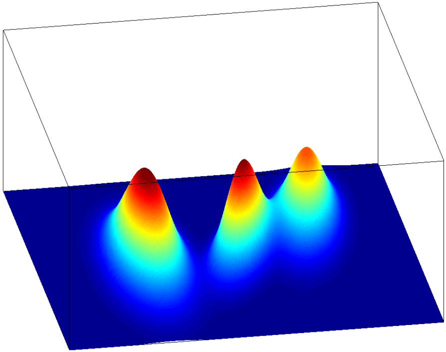
“Statistical” Estimation



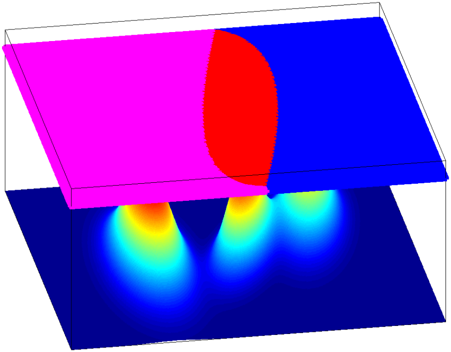
“Statistical” Estimation



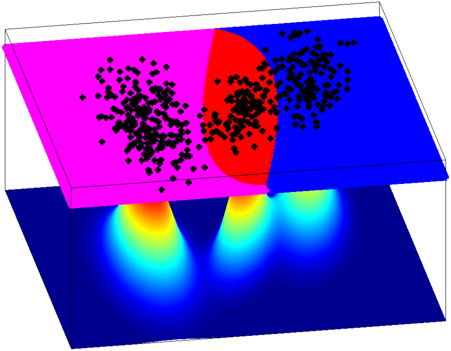
“Statistical” Estimation



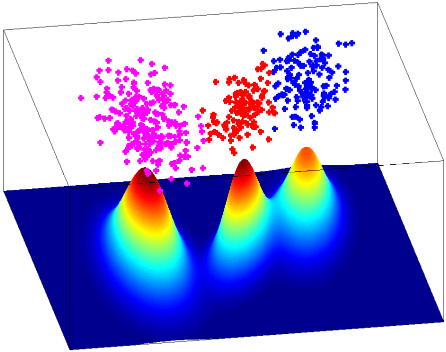
“Statistical” Estimation



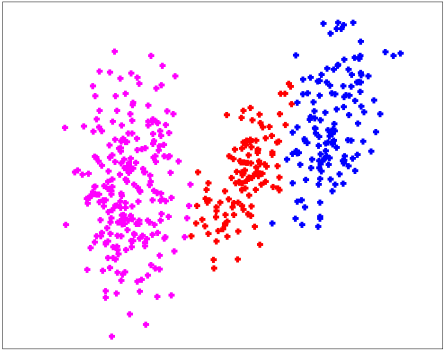
“Statistical” Estimation



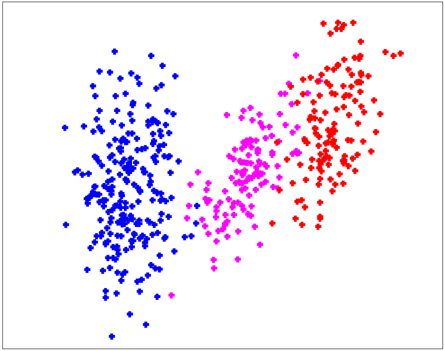
"Statistical" Estimation



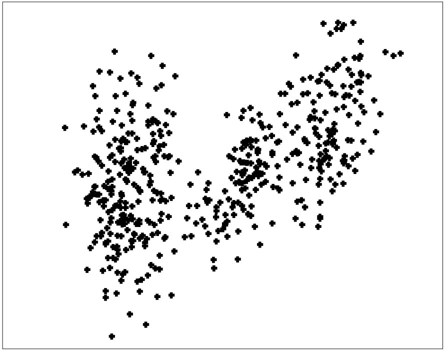
“Statistical” Estimation



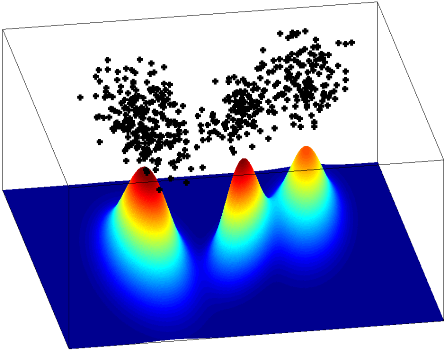
“Statistical” Estimation



“Statistical” Estimation



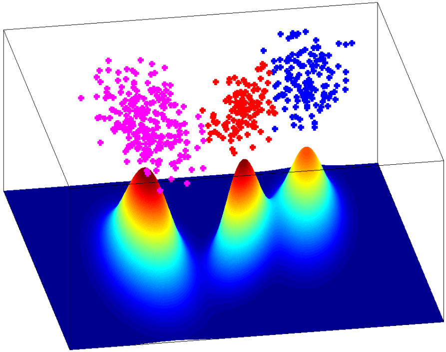
"Statistical" Estimation



- 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) :

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

Gaussian Mixture Modelization

- Stochastic modelization of the spectrums $\mathcal{S}$:
  - existence of K classes of spectrums,
  - proportion π_k for each class ($\sum_{k=1}^{K} \pi_k = 1$),
  - Gaussian law $\mathcal{N}_{\mu_k,\Sigma_k}$ on each class (strong assumption !)
- Density s_0 of $\mathcal{S}$ close to

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

- Goal : estimate all parameters K, π_k, μ_k, Σ_k from the data.
- Why ? : give possibility to assign a class to each observation by MAP

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

- Result in term of density estimation...

Gaussian Mixture Model

- Density s_0 of $\mathcal{S}$ close to $s_m(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$.
- Model $S_m = \{s_m\}$:
  - choice of a number of K,
  - choice of a structure for the means μ_k and the covariance matrices $\Sigma_k = L_k D_k A_k D'_k$
- Model $[\mu\, L\, D\, A]^K$: constraints (known, common or free values...) on the means μ_k, the volumes L_k, the diagonalization bases D_k and the eigenvalues A_k.
- Model S_m : parametric model of dimension $(K-1) + \dim([\mu\, L\, D\, A]^K)$ in a space of dimension p.
- Estimation by maximum likelihood of the parameters :
  - for each class, the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D'_k$
  - the mixing proportions π_k.
- Classical technique available : EM Algorithm.

Maximum Likelihood and MM

- "Maximum" likelihood for a given K :

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

- 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

$$P_k^{i,(n)} = P\left(k_i = k \middle| \mathcal{S}_i, \pi^{(n)}, \mu^{(n)}, \Sigma^{(n}\right) = \frac{\pi_k^{(n)} \mathcal{N}_{\mu_k^{(n)},\Sigma_k^{(n)}}(\mathcal{S}_i)}{\sum_{k'=1}^{K} \pi_{k'}^{(n)} \mathcal{N}_{\mu_{k'}^{(n)},\Sigma_{k'}^{(n)}}(\mathcal{S}_i)}$$

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

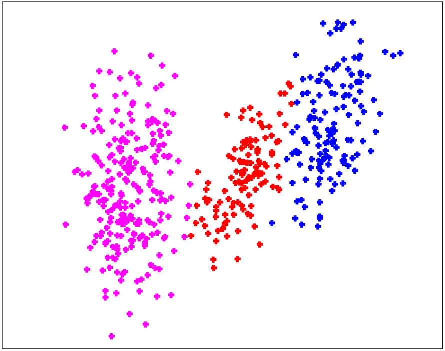
  - Kullback : $L^i \leq L^{i,(n)} + \mathrm{Cst}^{i,(n)}$ with equality at $(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$.
- Bonus :
  - Separability of $L^{i,(n)}$ in π and (μ, Σ) :

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

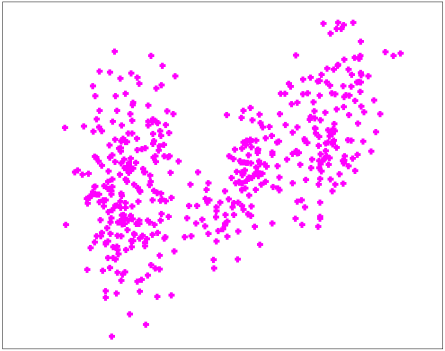
  - Close formulas for the Minimization of $L^{(n)}$ in π and (μ, Σ) !

How many classes ?

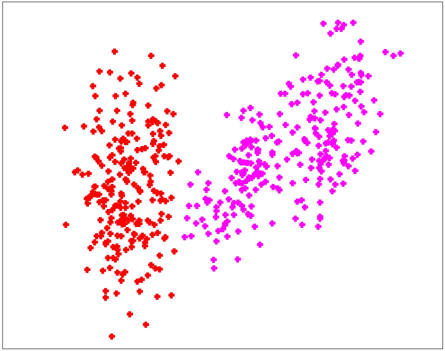
How many classes ?



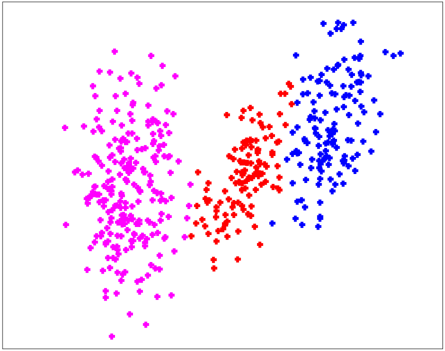
How many classes ?



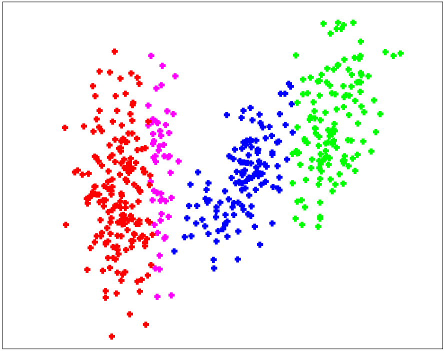
How many classes ?



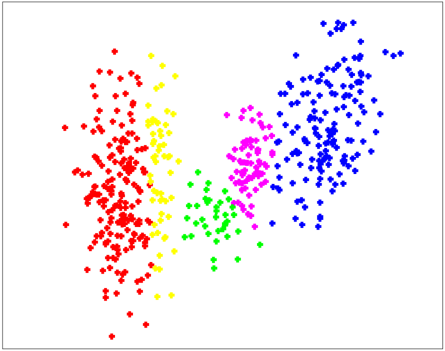
How many classes?



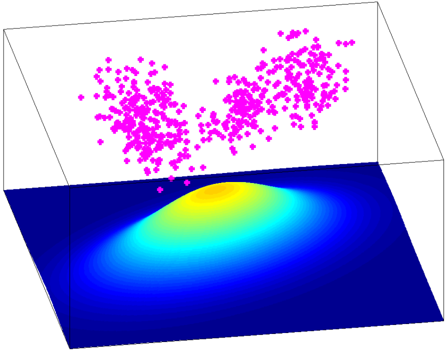
How many classes ?



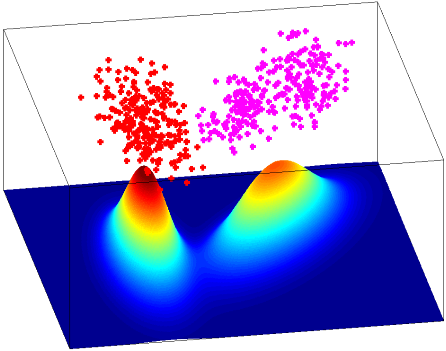
How many classes ?



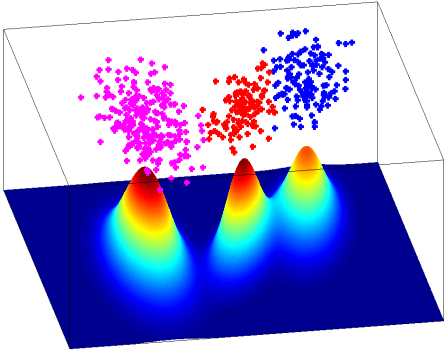
How many classes ?



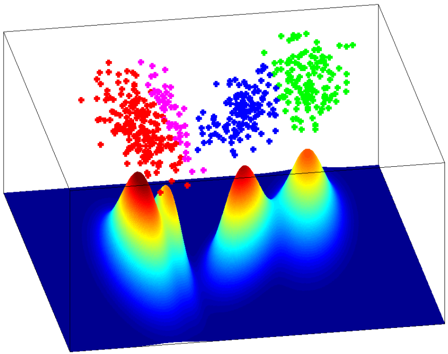
How many classes ?



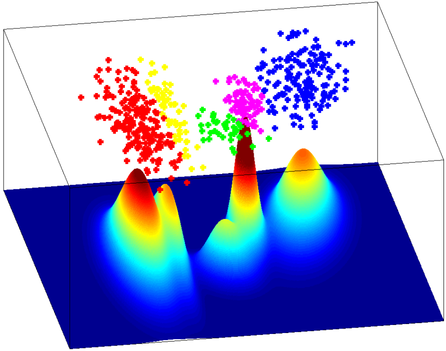
How many classes ?



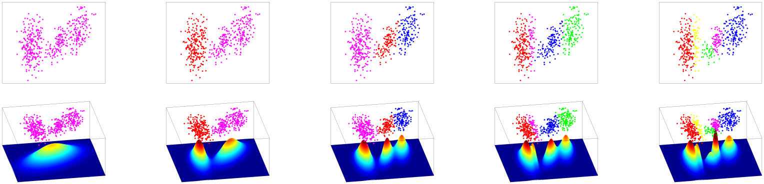
How many classes ?



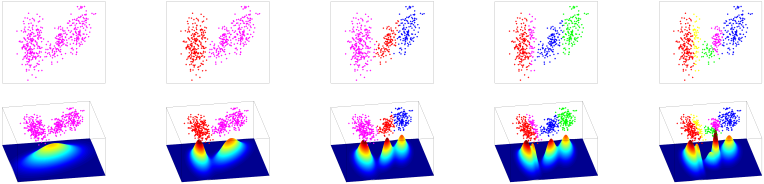
How many classes ?



How many classes ?

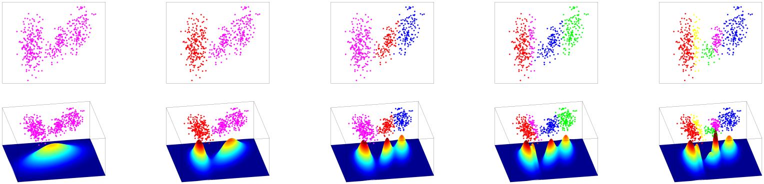


How many classes ?



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

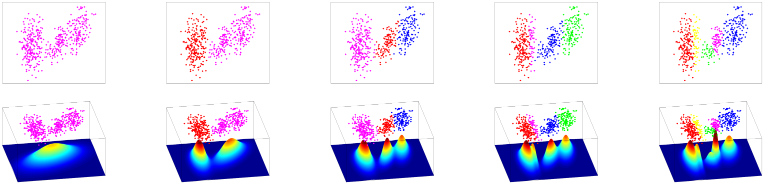
How many classes ?



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

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

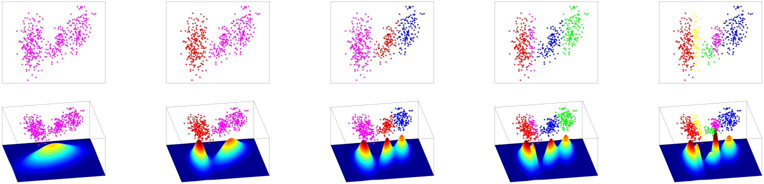
How many classes ?



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)

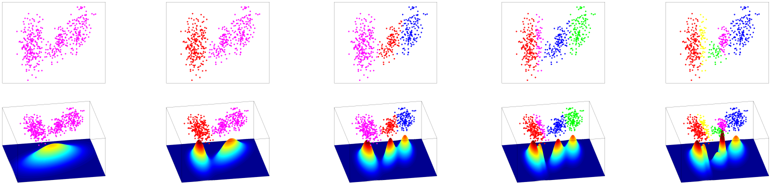
Ockham's Razor

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

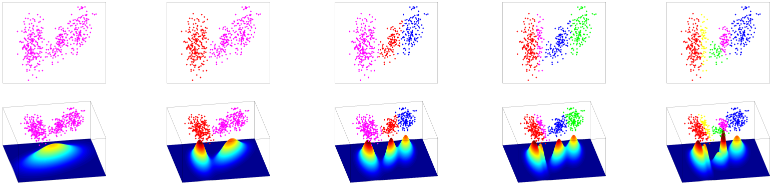
- Ockham's Razor (simplicity principle) : one should not add hypotheses, if the current ones are already sufficient !
- Balance between observation explanation power and simplicity.

Selection by Penalization

Selection by Penalization

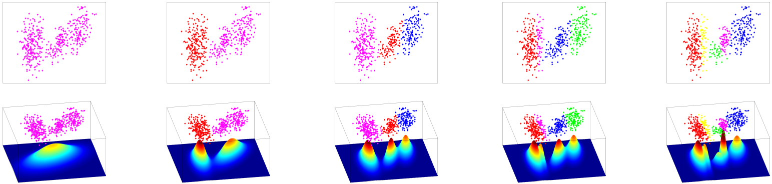


Selection by Penalization



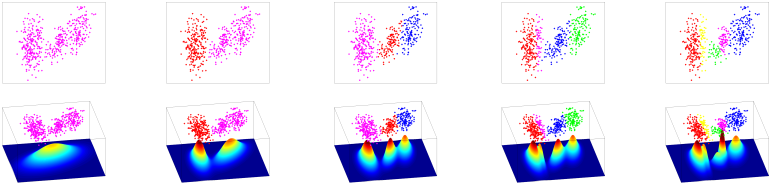
Likelihood	− −	+	+ + +	+ + +·	+ + + +
Simplicity	+ + + +	+ + +	+ +	+	−

Selection by Penalization



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

Selection by Penalization

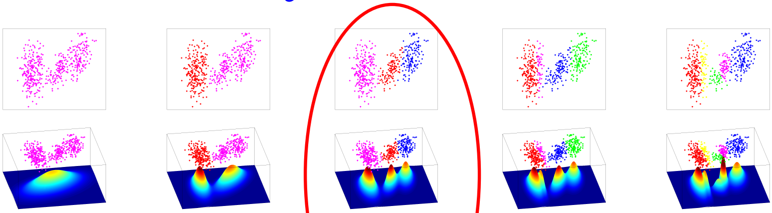


Likelihood	−−	+	+++	+++·	++++
+ Simplicity	++++	+++	++	+	−
= Tradeoff	++	++++	+++++	++++·	+++

- 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

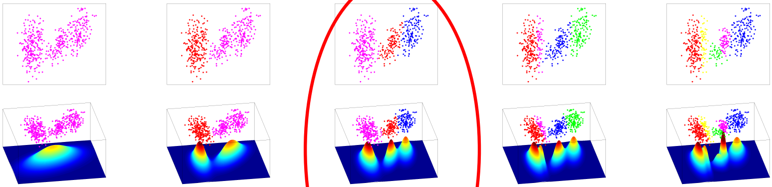


Likelihood	−−	+	+++	+++·	++++
+ Simplicity	++++	+++	++	+	−
= Tradeoff	++	++++	+++++	++++·	+++

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

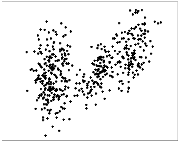
- Likelihood : $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicity : $-\lambda \mathrm{Dim}(S_K)$ (a lot of theory behind that).
- Penalized estimator :

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

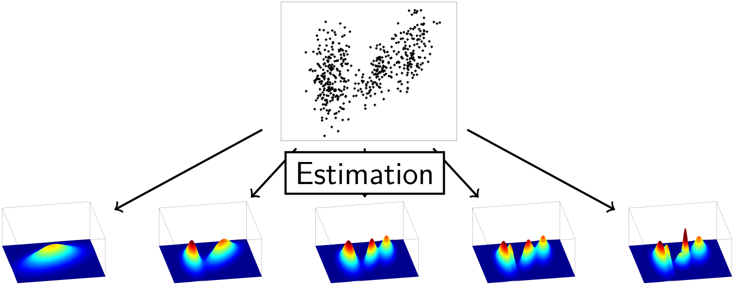
- Optimization in K by exhaustive exploration !

Methodology

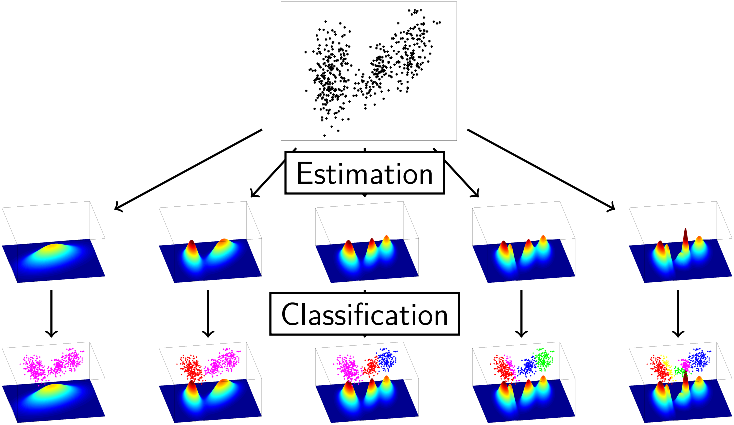
Methodology



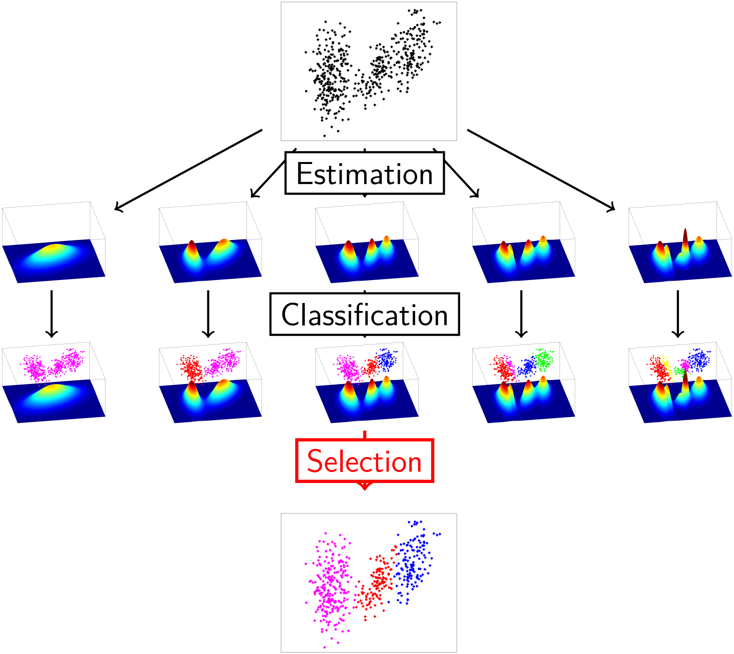
Methodology



Methodology



Methodology



Model Selection

- How to select the model S_m :
  - the number of classes K,
  - the model $[\mu\, L\, D\, A]^K$?
- Penalized selection principle :
  - choice of model collection $S_m = \{s_m\}$ with $m \in \mathcal{S}$,
  - estimation by maximum likelihood of a density s_m for each model S_m,
  - selection of a model $\widehat{m}$ by

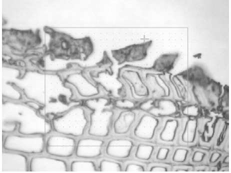
$$\widehat{m} = \mathrm{argmin} - \ln(\widehat{s}_m) + \mathrm{pen}(m).$$

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

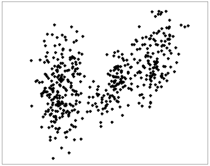
$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

  - numerical for unsupervised classification ($\neq$ segmentation),
  - classification consistency if $\ln \ln(n)$ in the penalties...

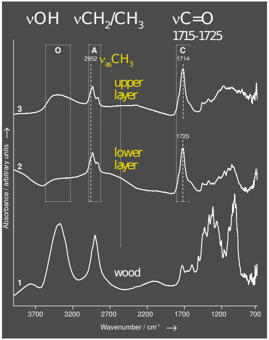
Back to our violins



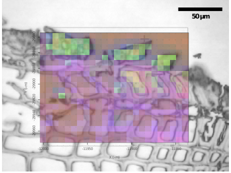
Representation



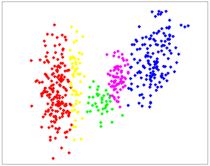
Segmentation



Classification



Spatial Info.



Segmentation and Gaussian Mixture

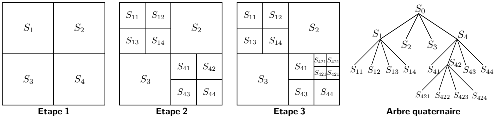
- Initial goal : unsupervised segmentation $\neq$ unsupervised classification.
- Take into account the spatial position x of the spectrums through the mixing proportions (Kolaczyk et al) : conditional density model

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

- Model mixing parametric and non-parametric setting...
- Estimation from the data :
  - for each class, the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D_k'$,
  - the mixing proportions $\pi_k(x)$.
- $\pi_k(x)$ function : regularization required.
- Model selection principle...

Gaussian Mixture and Hierarchical Partition

- How to select the model S_m ? :
  - the number of classes K,
  - the model $[\mu\, L\, D\, A]^K$,
  - the mixing proportions structure of $\pi_k(x)$.
- Simple structure : $\pi_k(x) = \sum_{\mathcal{R} \in \mathcal{P}} \pi_k[\mathcal{R}] \chi_{\{x \in \mathcal{R}\}} = \pi_k[\mathcal{R}(x)]$
  - piecewise constant on a hierarchical partition,
  - efficient optimization possible,
  - decent approximation property.



- $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
- Penalty $\mathrm{pen}(m) = \kappa \ln(n) \dim(S_m)$ sufficient for
  - a theoretical control in term of conditional density estimation,
  - numerical optimization (EM + dynamic programming).

Conditional Densities

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

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

  - With $\mathrm{pen}(m)$ suitably design, selection of a model $\widehat{m}$ by

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

- Conditional density estimation type result :

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

Numerical optimization

- Penalized Model Selection :

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

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

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

- Two tricks :
  - EM Algorithm
  - CART (dynamic programming)

EM Algorithm

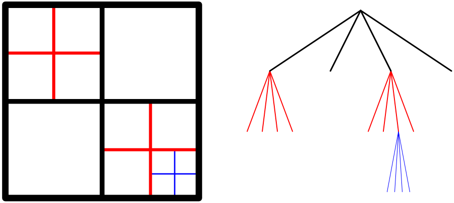
- E Step : with $P_k^{i,(n)} = P(k_i = k | x_i, \mathcal{S}_i, \mathcal{P}^{(n)}, \pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$

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

  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



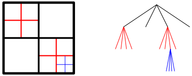
- Optimization in $(\mathcal{P},\pi)$ of

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

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

- 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 $\underset{\mathcal{P}}{\operatorname{argmin}} \sum_{\mathcal{R}\in\mathcal{P}} C[\mathcal{R}]$ 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, $\widehat{\mathcal{P}}[\mathcal{R}] = \cup_{\mathcal{R}'\in\text{Child}[\mathcal{R})}\widehat{\mathcal{P}}[\mathcal{R}']$

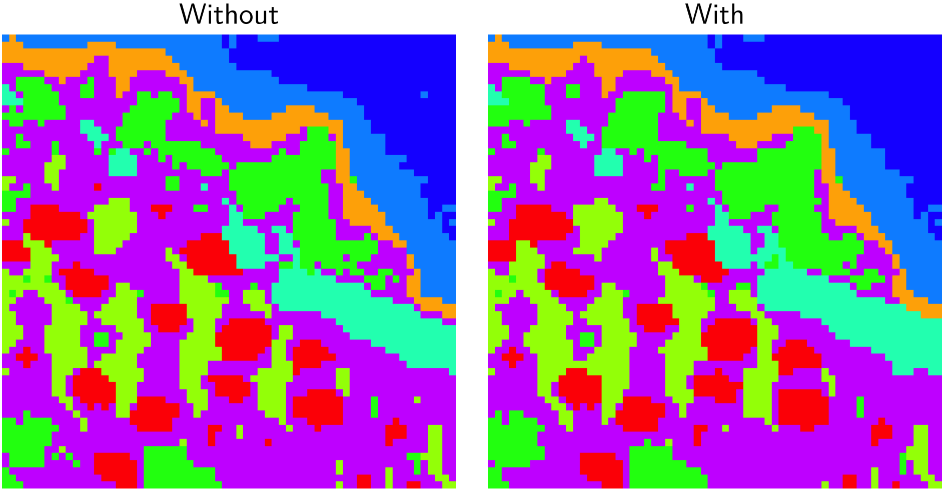
  with a decision based on

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

- 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.

Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

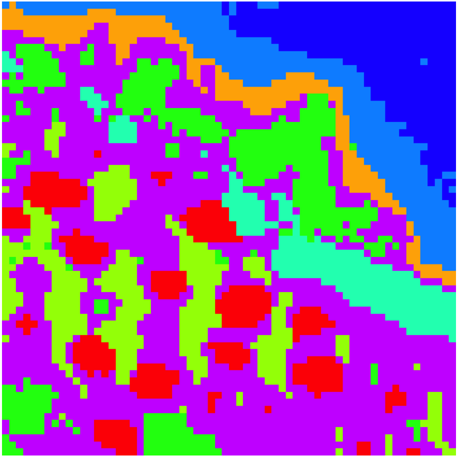


- $K = 8$, $[L_k\, D\, A]^K$ and optimal partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by (not so naive) ACP...

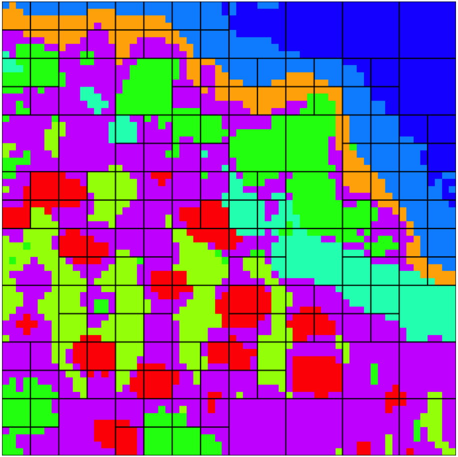
Unsupervised Segmentation

- Numerical result taking into account the spatial modeling :

Without

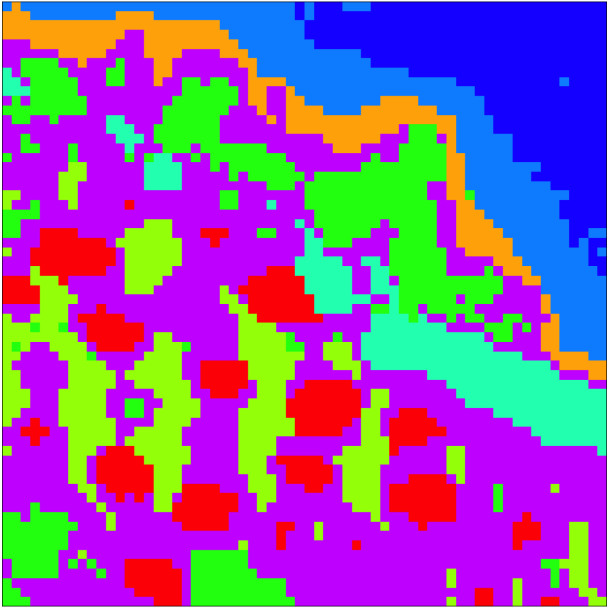


With

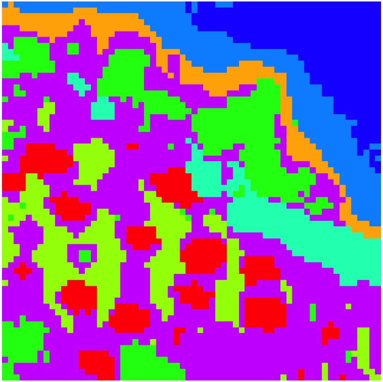


- $K = 8$, $[L_k\, D\, A]^K$ and optimal partition.
- Penalty calibration by slope heuristic.
- Dimension reduction by (not so naive) ACP...

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.
- 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).

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 model collection. 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 $m \in \mathcal{M}$ $\mathrm{pen}(m) \geq \kappa \left(n\sigma_m^2 + x_m\right)$ with $\kappa > \kappa_0$
where σ_m is 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 $$\mathbb{E}\left[JKL_{\rho}^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S_m\in\mathcal{S}} \left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + C_2\frac{\Sigma}{N} + \frac{\delta}{N}.$$

Theorem

- Oracle type 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

$$\mathrm{pen}(m) \geq \kappa\left(N\sigma_m^2 + x_m\right) \quad \text{with } \kappa > \kappa_0,$$

where $N\sigma_m^2$ measures the complexity of S_m (entropy) and x_m a coding cost within the collection (Kraft).

- « Distances » used $KL^{\otimes_n}$ and $JKL_\rho^{\otimes_n}$: « tensorized » Kullback divergence and Jensen-Kullback divergence.
- $N\sigma_m^2$ linked to the bracketing entropy of S_m measured with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.

Kullback, Hellinger and extensions

- Typical model selection oracle inequality :

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

- Density : Hellinger $d^2(s, s')$ (or affinity) (Kolaczyk, Barron, Bigot).
- Better result with $JKL(s, s') = 2KL(s, (s' + s)/2)$ (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 measure $sd\lambda$ and $td\lambda$ and all $\rho \in (0, 1)$

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

- $C_\rho \simeq 1/5$ if $\rho \simeq 1/2$.

Conditional densities

- Previous divergences should be adapted to the conditional density framework :
  - Divergence on the product density conditioned by the design (Kolaczyk, Bigot).
  - Tensorization principle and expectancy on a similar phantom 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 approaches but for Hellinger and JKL + Possibility to have result with expectancy on the design.
- Oracle inequality :

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

- Yield the classical density estimation theorem if $s(\cdot|X_i) = s(\cdot)$.

Penalization and complexity

- Penalty linked to the complexity of the model and of the collection.
- Complexity of the model S_m (entropy) :
  - $H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)$ bracketing entropy with respect to the tensorized Hellinger distance ($d^{\otimes_n} = \sqrt{d^{2\otimes_n}} = \sqrt{\mathbb{E}\left[\frac{1}{N}\sum d^2(s(\cdot|X_i), s'(\cdot|X_i))\right]}$).
  - Assumption (H) : for every model S_m, 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 such that for all $\sigma \in \mathbb{R}^+$ and all $s_m \in S_m$

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

  - Complexity measured by $N\sigma_m^2$ where σ_m is the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{N}\sigma$.
  - Often $N\sigma_m^2 \propto \dim(S_m)$
- Complexity of the collection (coding) :
  - measured by x_m satisfying a Kraft inequality $\sum_{m\in\mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- Classical constraint on the penalty

$$\mathrm{pen}(m) \geq \kappa\left(N\sigma_m^2 + x_m\right) \quad \text{with } \kappa > \kappa_0.$$

Spatialized Gaussian Mixture Case

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

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

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

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