Segmentation non supervisée, mélange de gaussiennes et algorithme E.M.

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

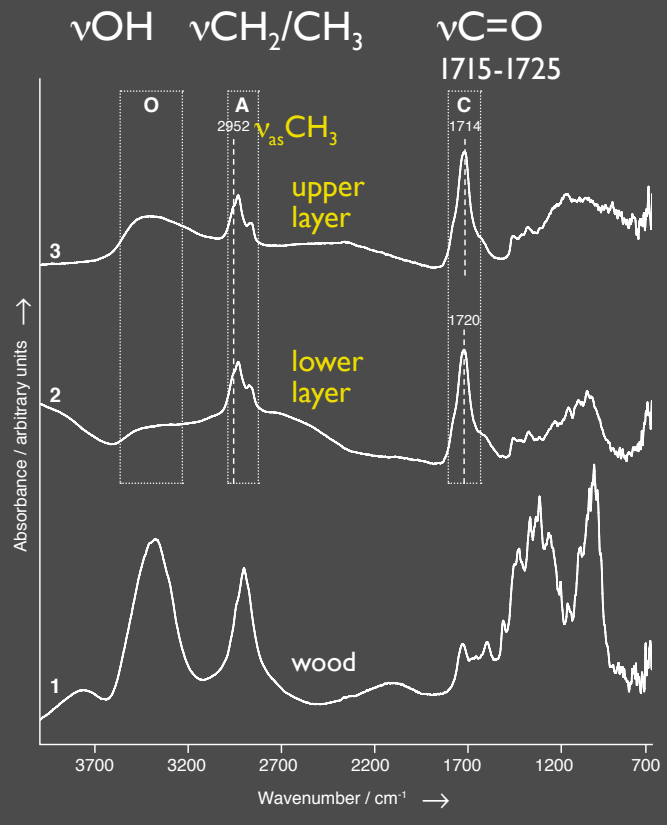
Gdr MOA et MSPC
07 juin 2011

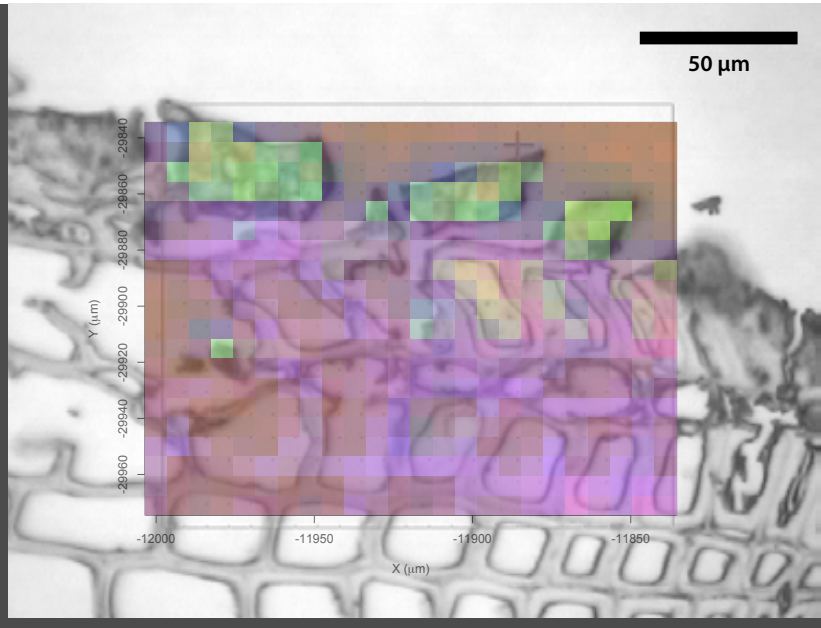
INRIA

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

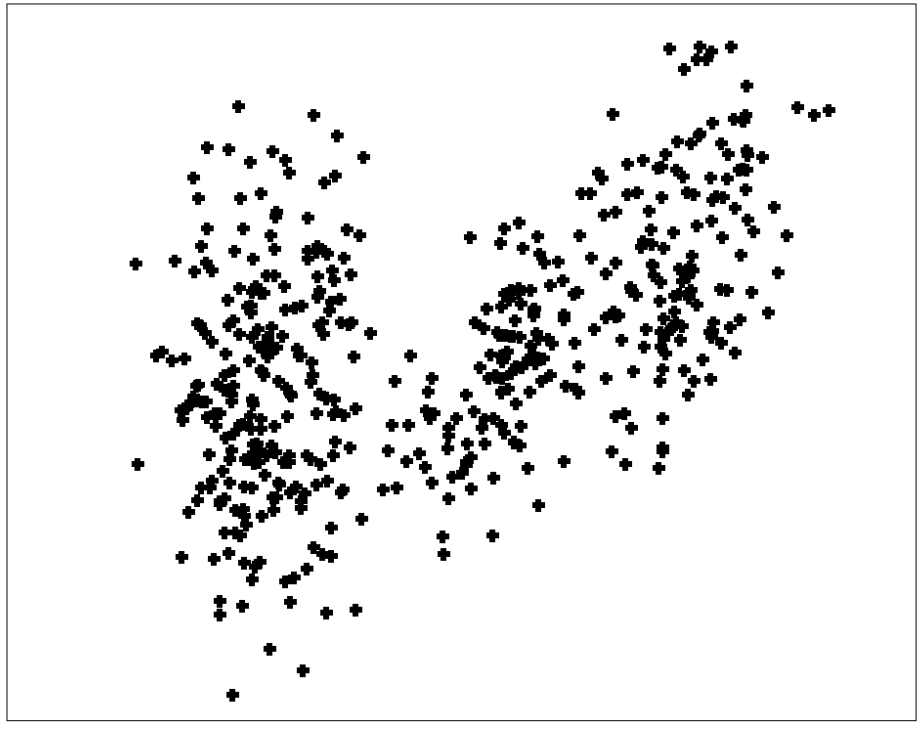
J.-P. Echard, L. Bertrand, A. von Bohlen, A.-S. Le Hô, C. Paris, L. Bellot-Gurlet, B. Soulier, A. Lattuati-Derieux, S. Thao, L. Robinet, B. Lavédrine, and S. Vaiedelich. *Angew. Chem. Int. Ed.*, 49(1), 197-201, 2010.

Segmentation d'images hyperspectrales

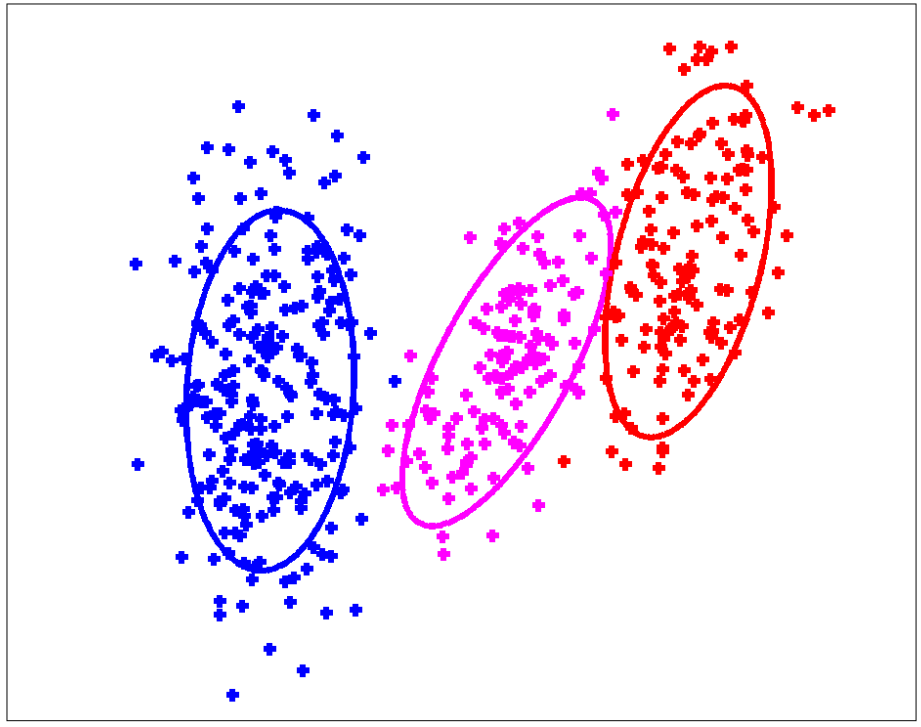
- Données :
  - image de taille N comprise entre ~ 1000 et ~ 100000 pixels,
  - spectres $\mathcal{S}$ de ~ 1024 points,
  - résolution $\sim 4/8$ cm^{-1} (10 fois meilleure dans le visible),
  - possibilité de mesurer de très nombreux spectres par minute...
- Objectifs immédiats :
  - segmentation automatique de ces images,
  - sans intervention humaine,
  - aide à l'analyse des résultats.
- Objectifs lointains :
  - classification automatique,
  - interprétation...

Un problème “jouet”

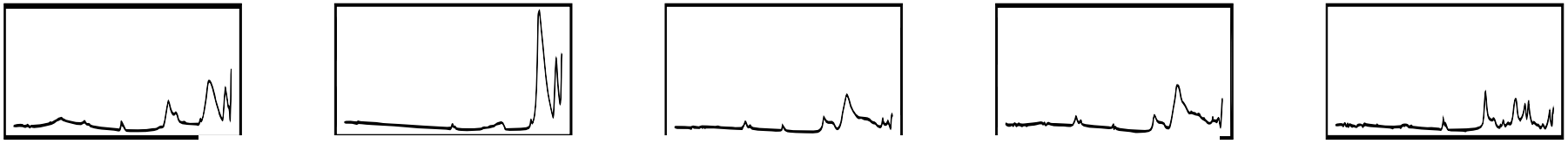
Un problème “jouet”

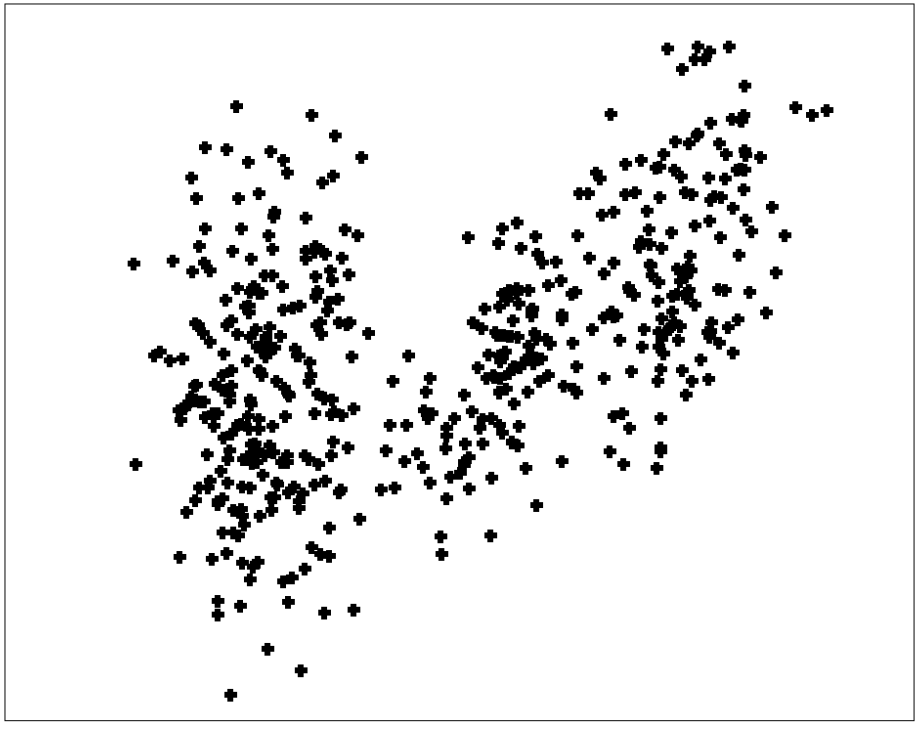


Un problème “jouet”

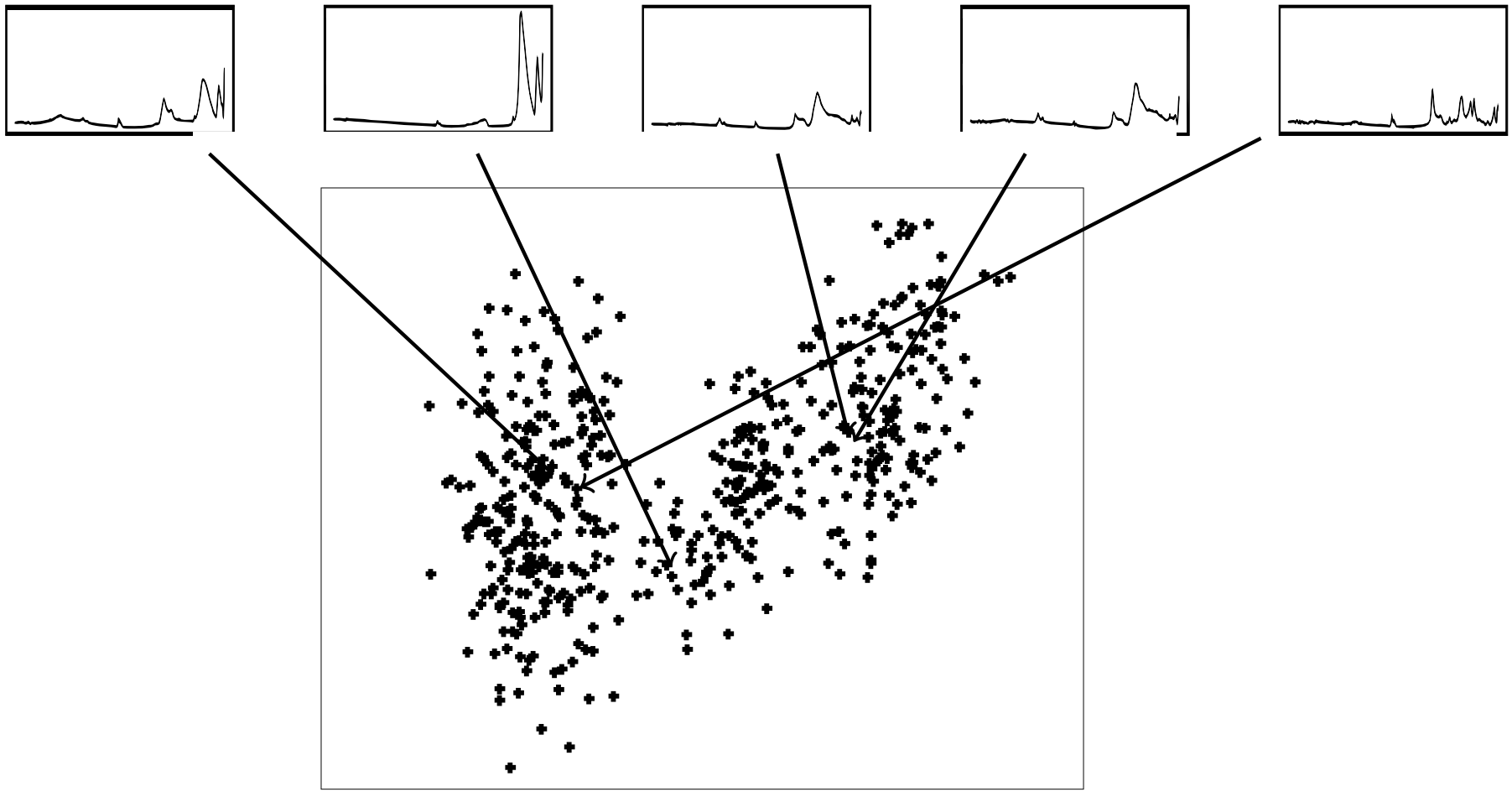


Un problème “jouet”

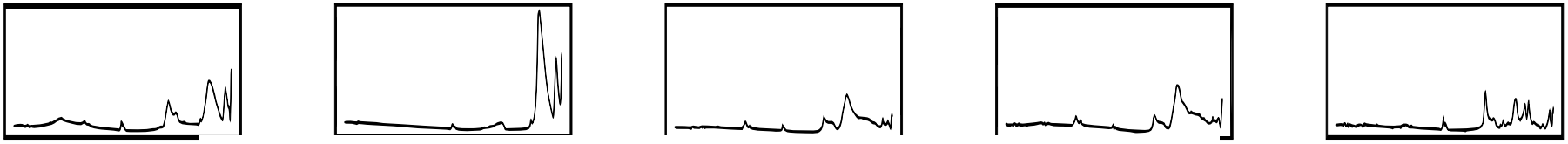


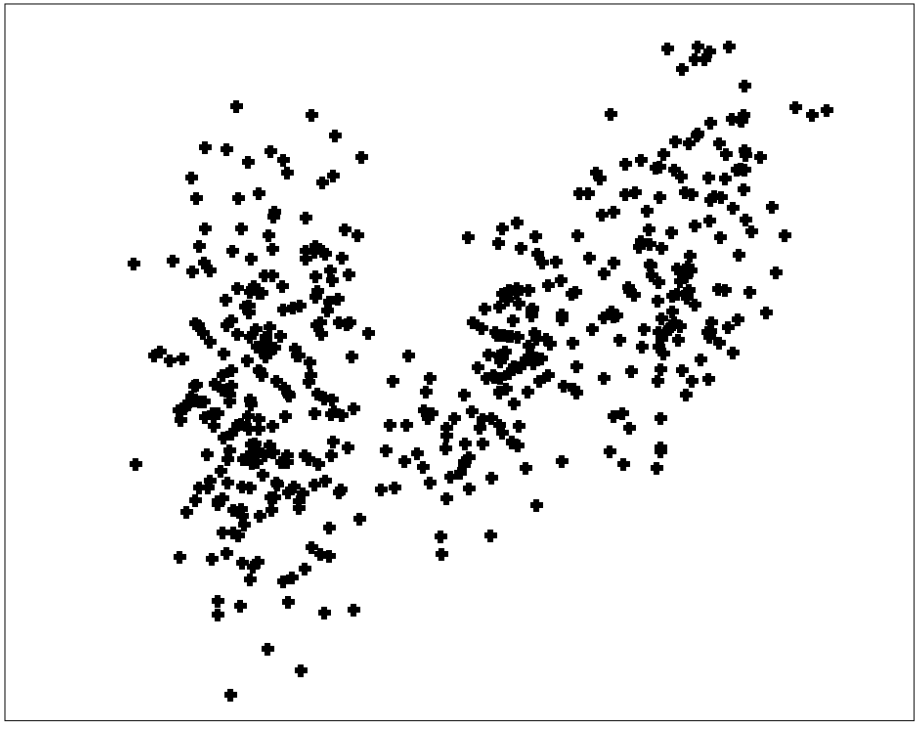


Un problème “jouet”

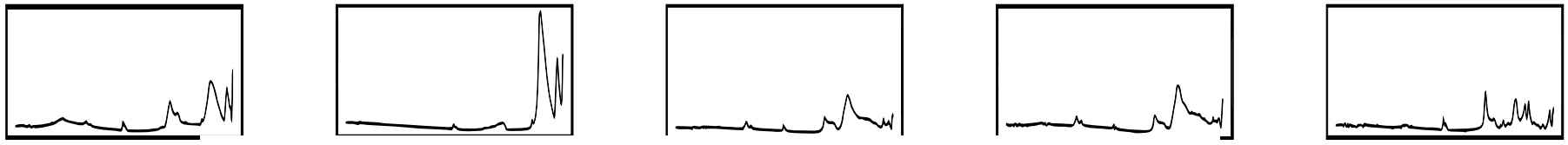


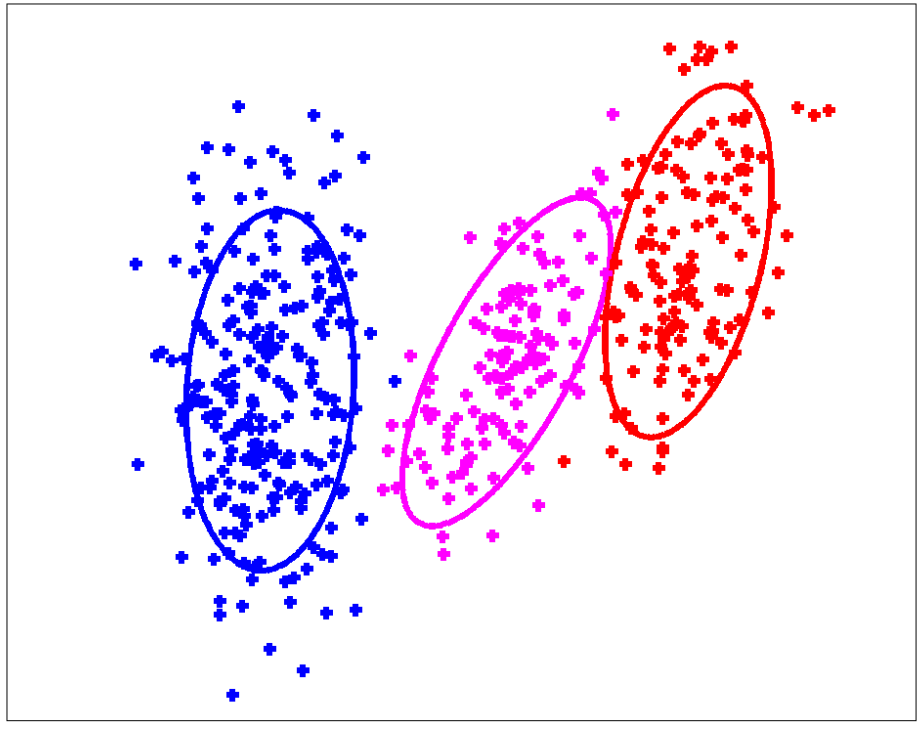
Un problème “jouet”



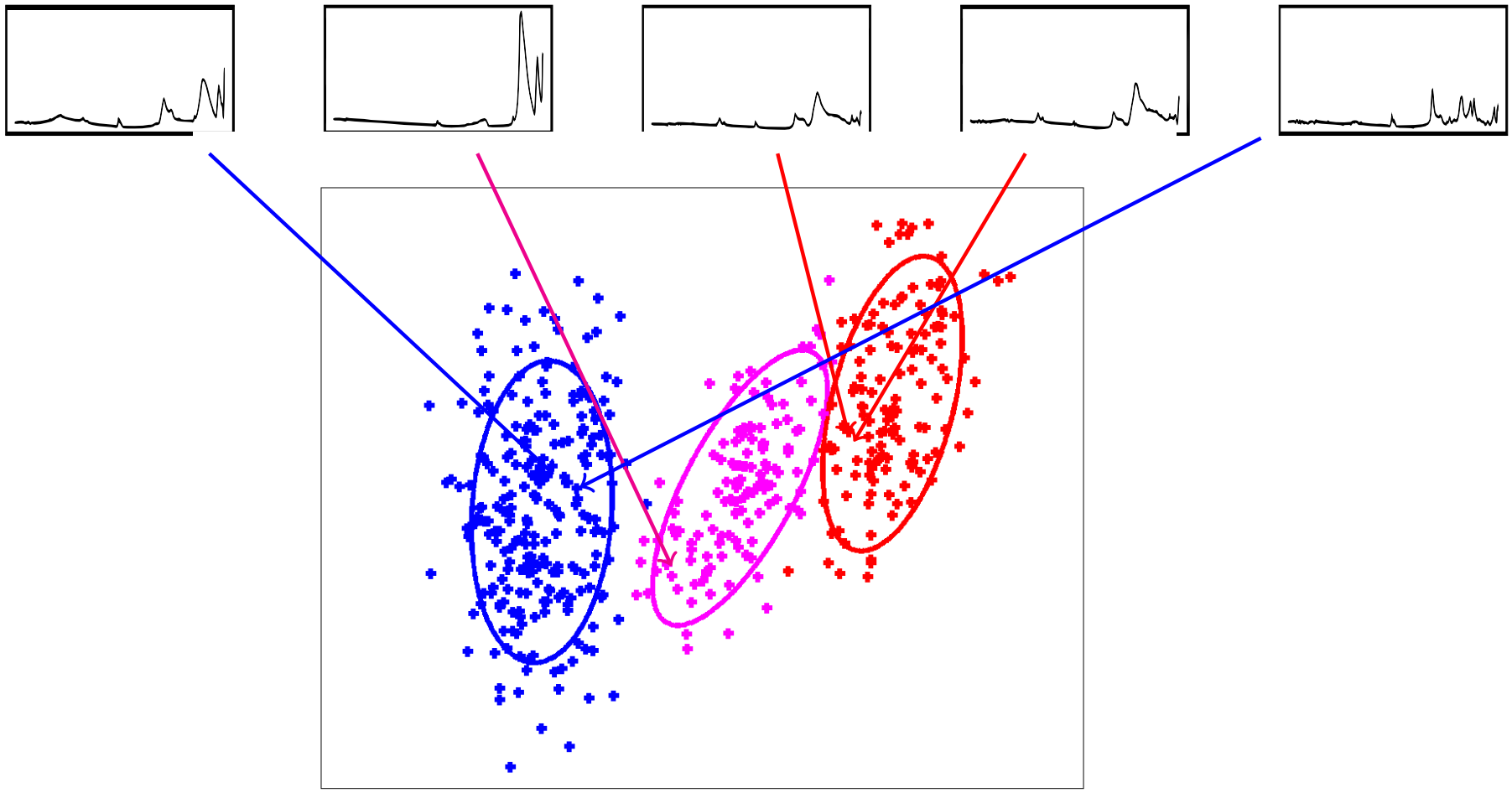


Un problème “jouet”





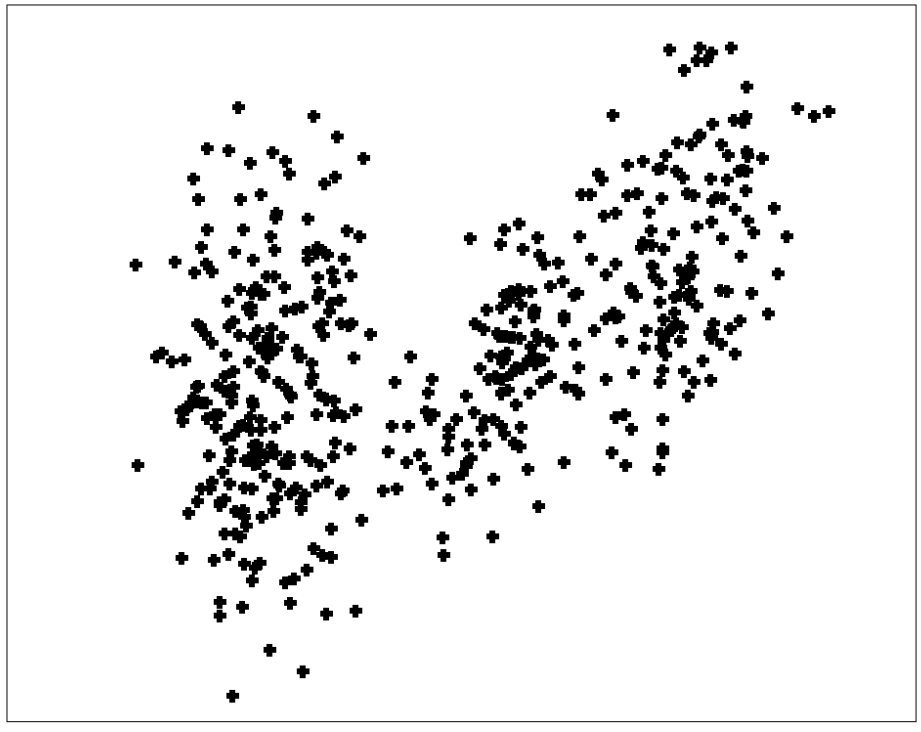
Un problème “jouet”



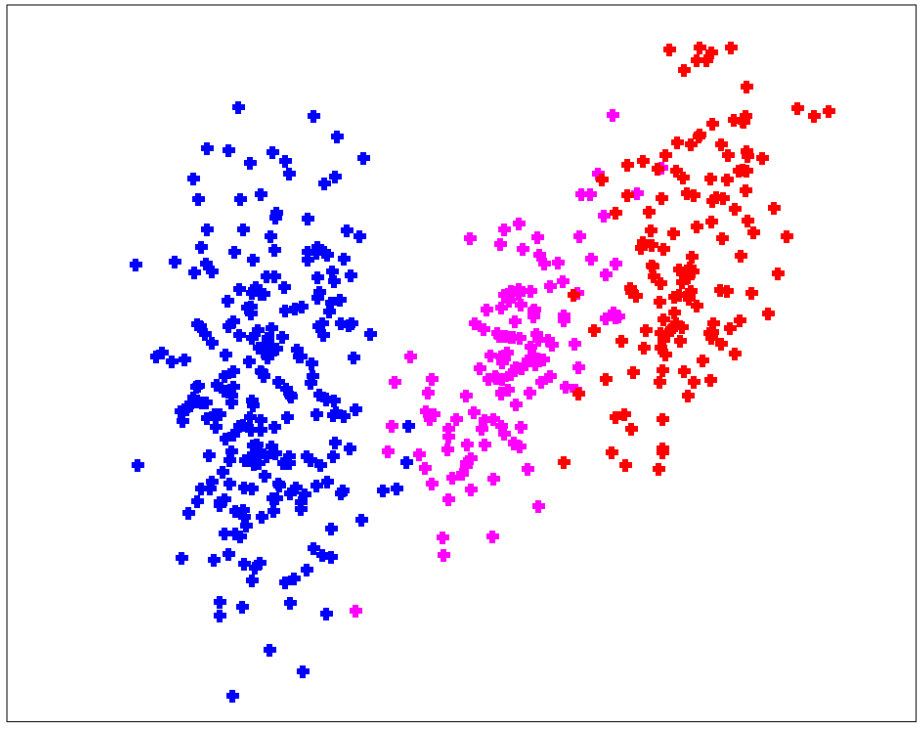
- Représentation : correspondance entre les spectres et des points dans un espace de grande dimension.

Modélisation “stochastique”

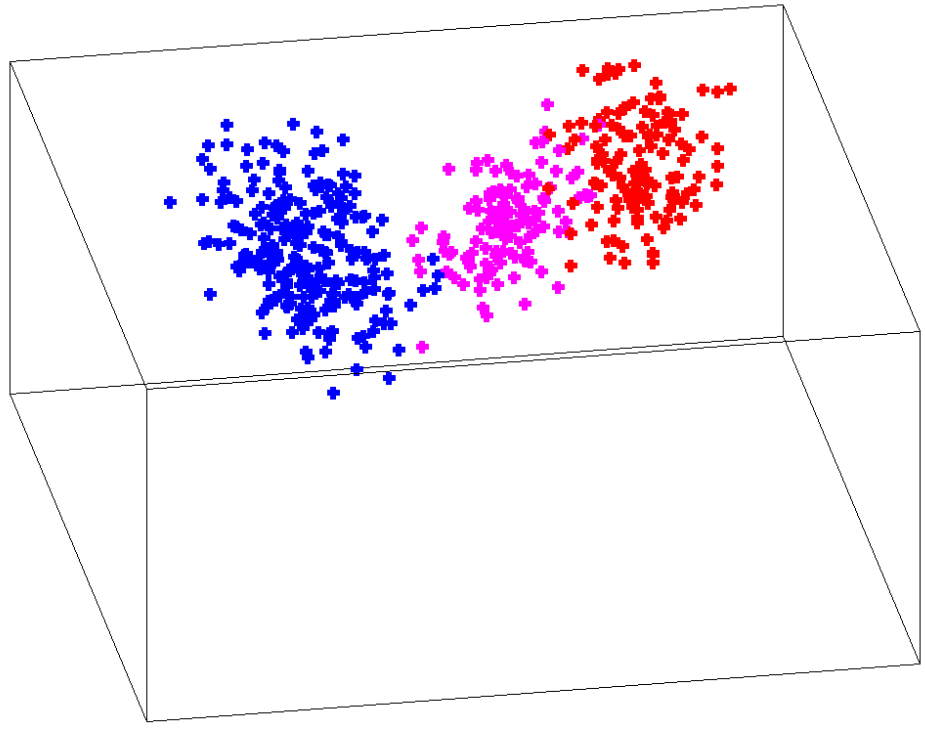
Modélisation “stochastique”



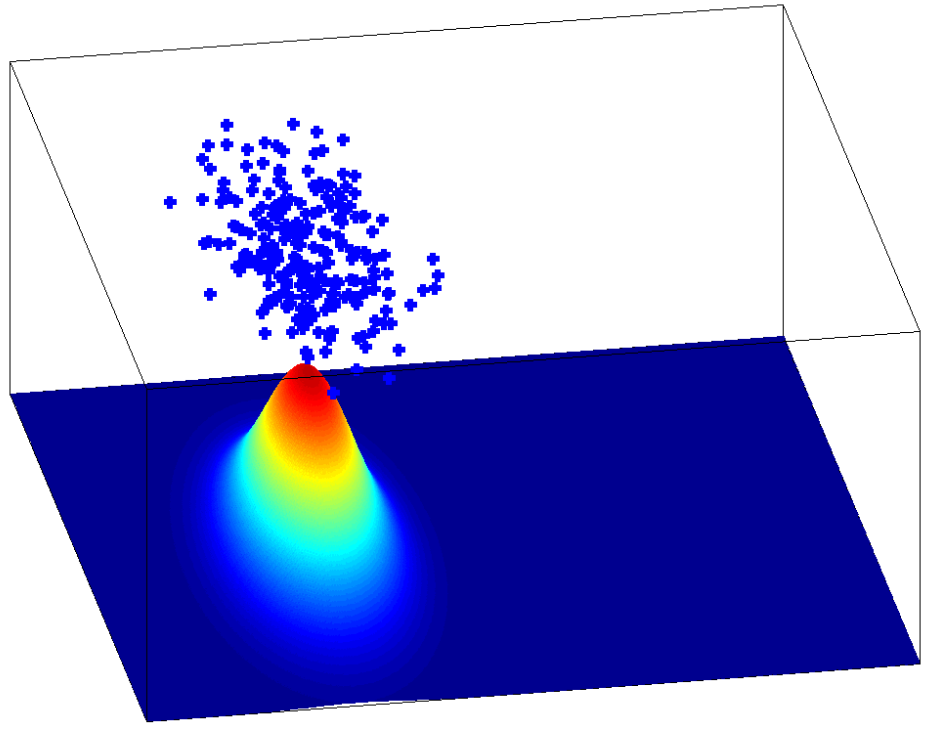
Modélisation “stochastique”



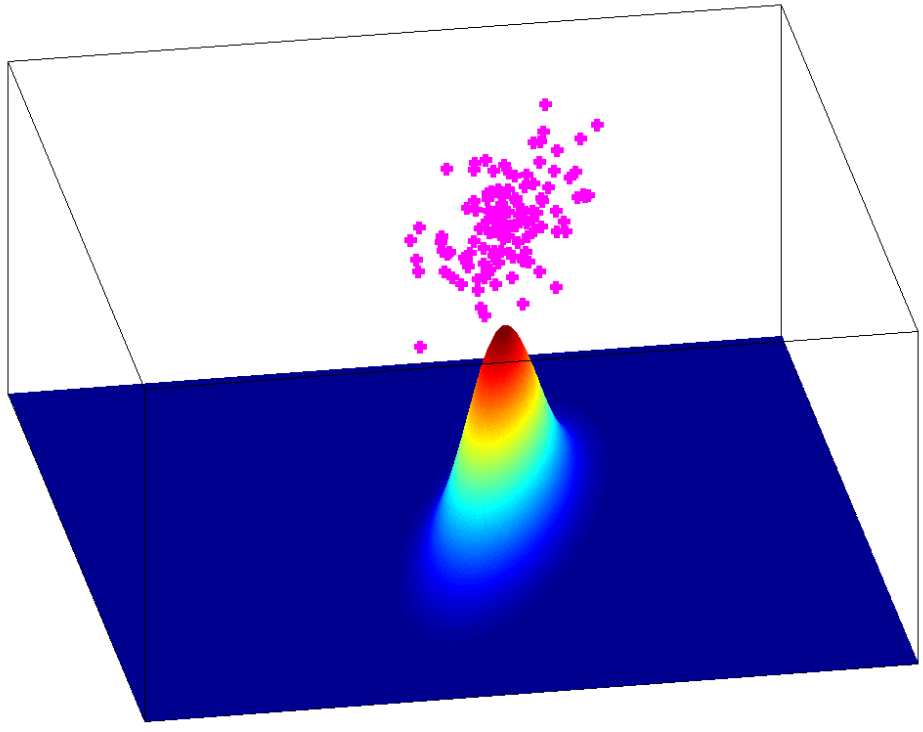
Modélisation “stochastique”



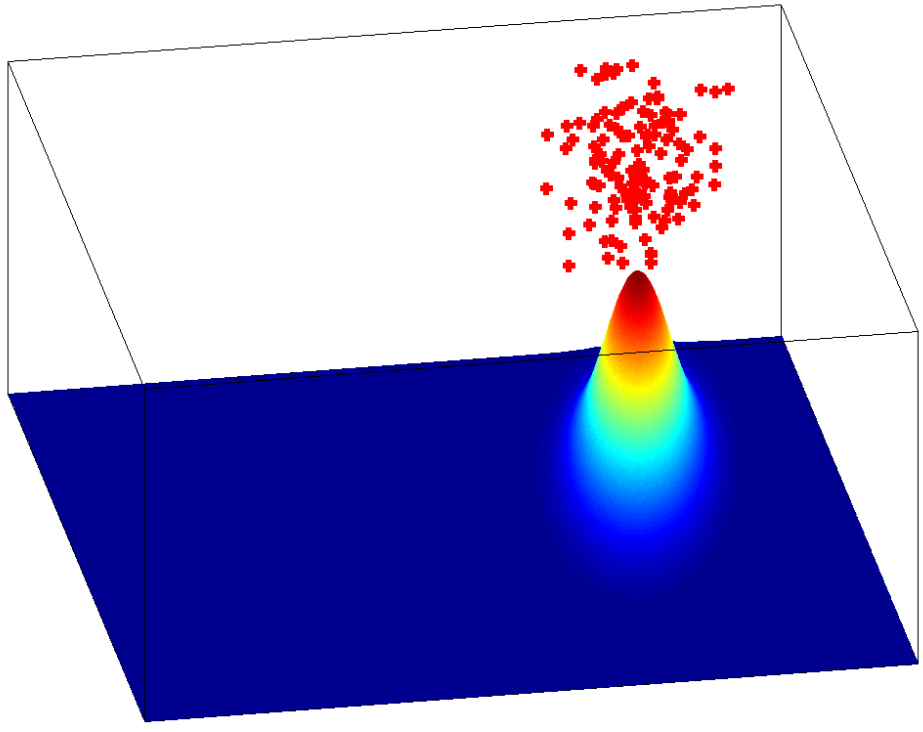
Modélisation “stochastique”



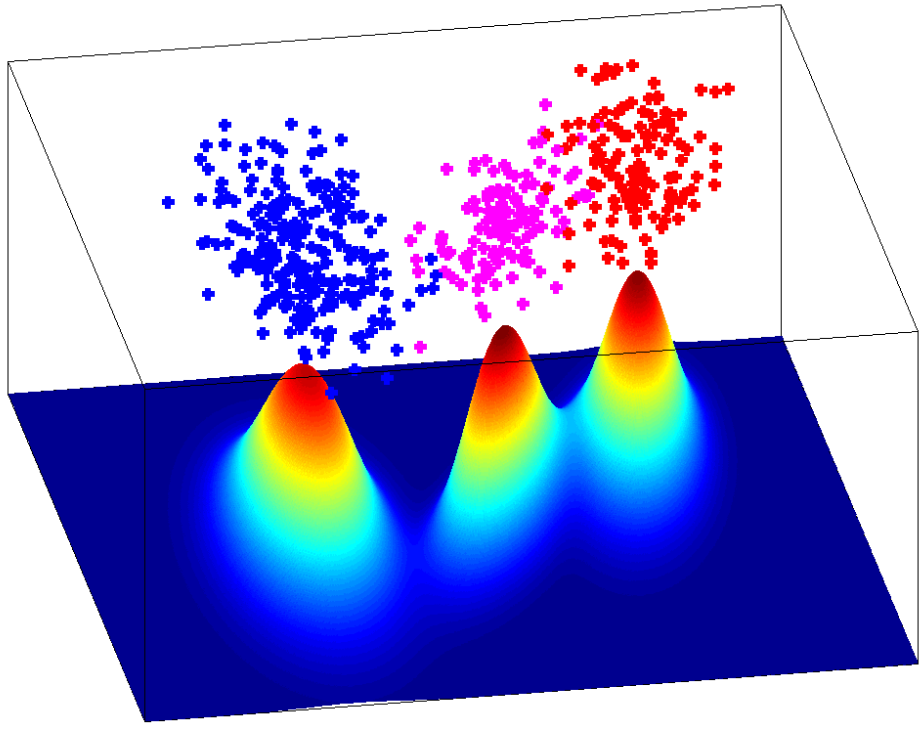
Modélisation “stochastique”



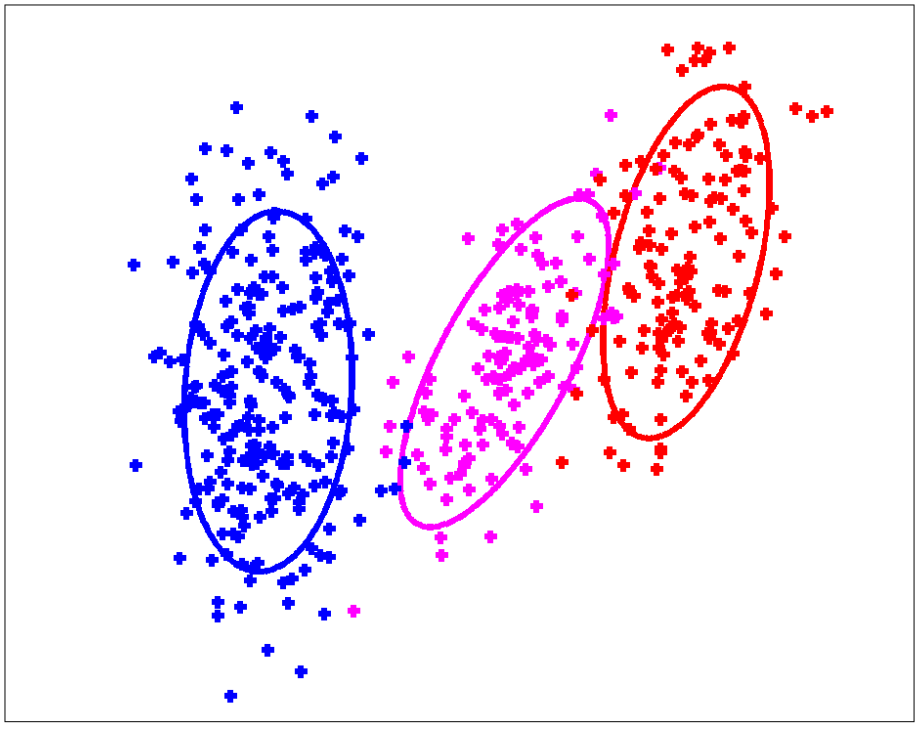
Modélisation “stochastique”



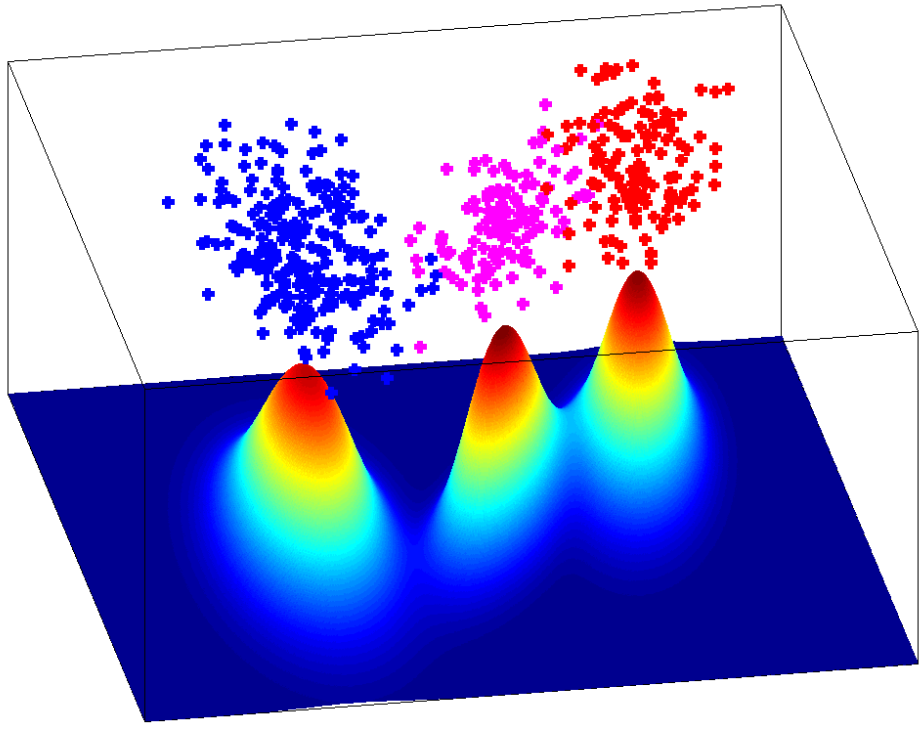
Modélisation “stochastique”



Modélisation “stochastique”



Modélisation “stochastique”

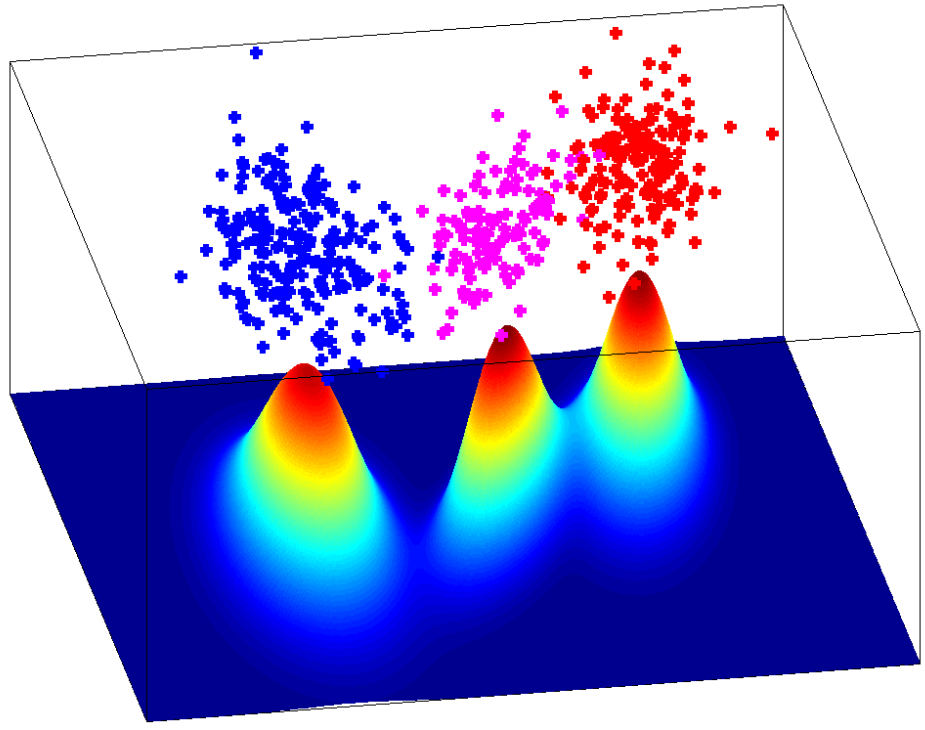


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”

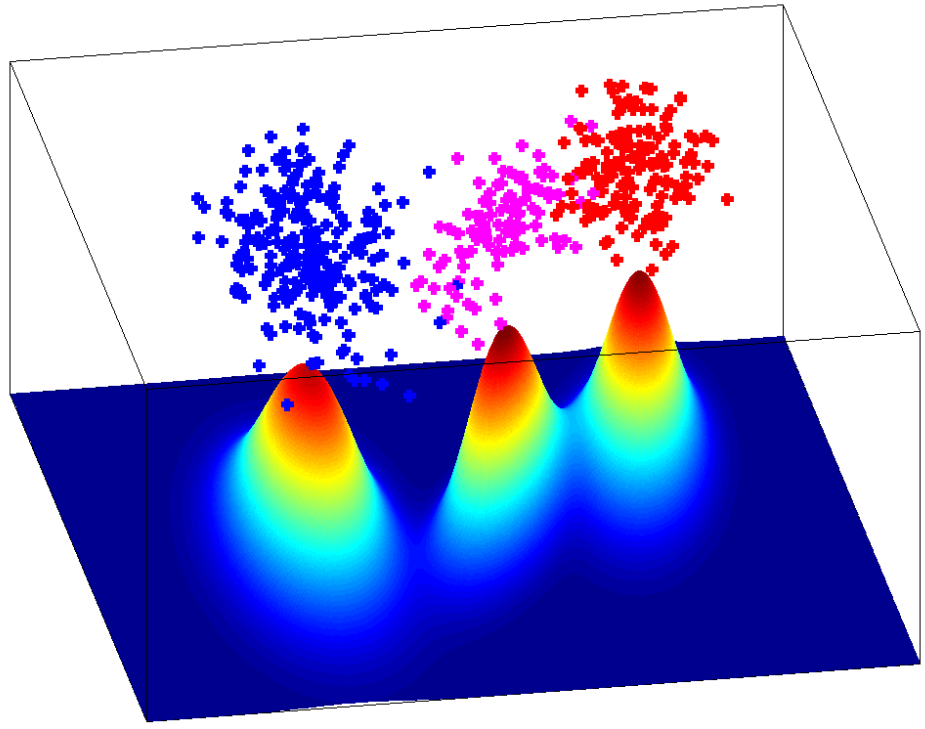


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”

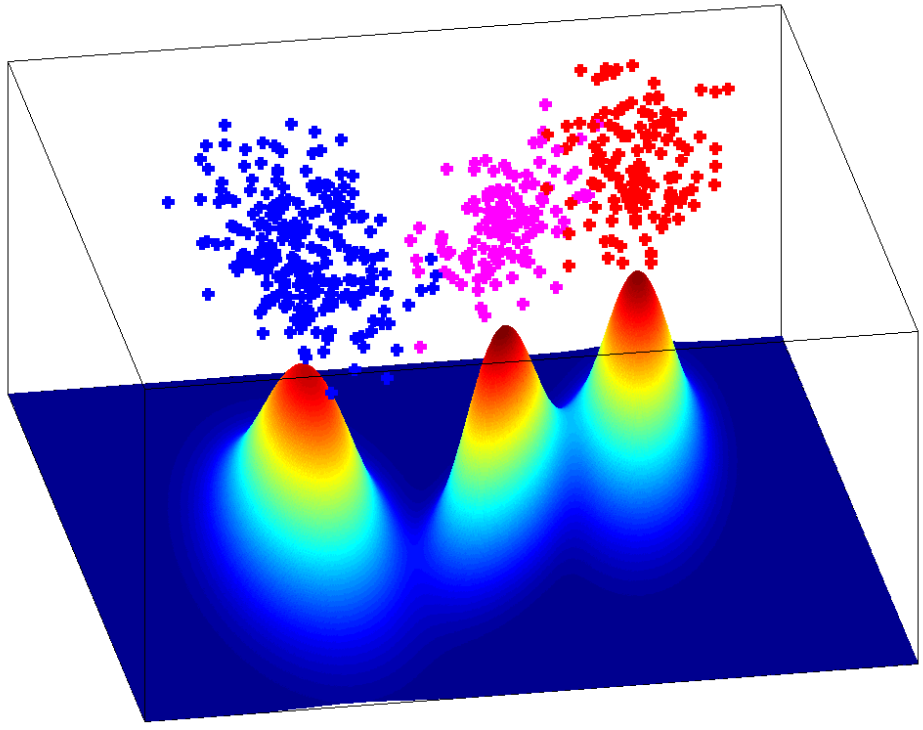


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”



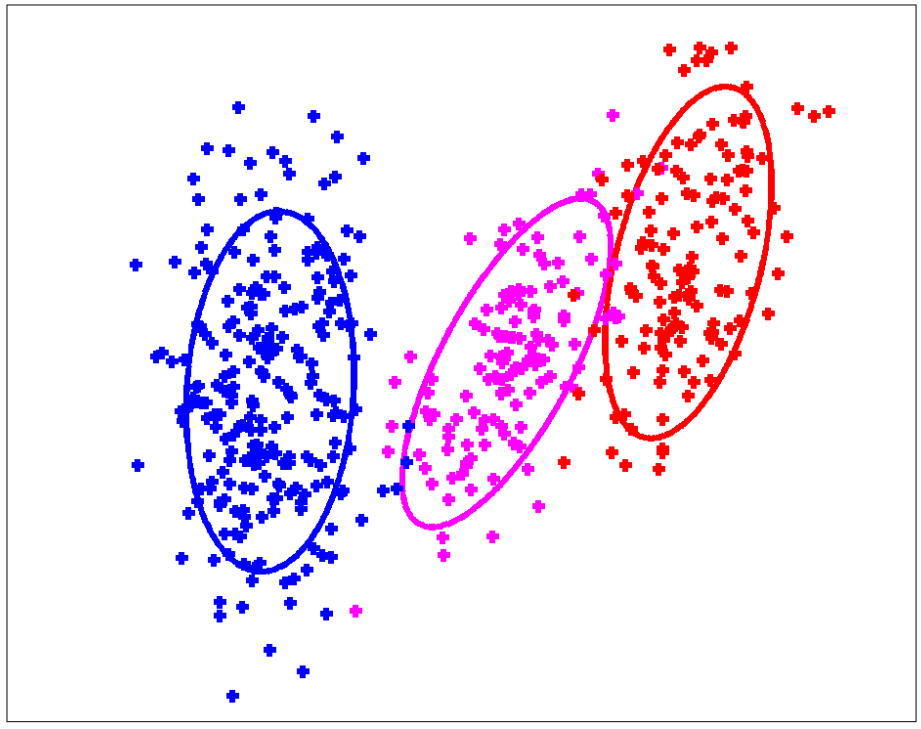
- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

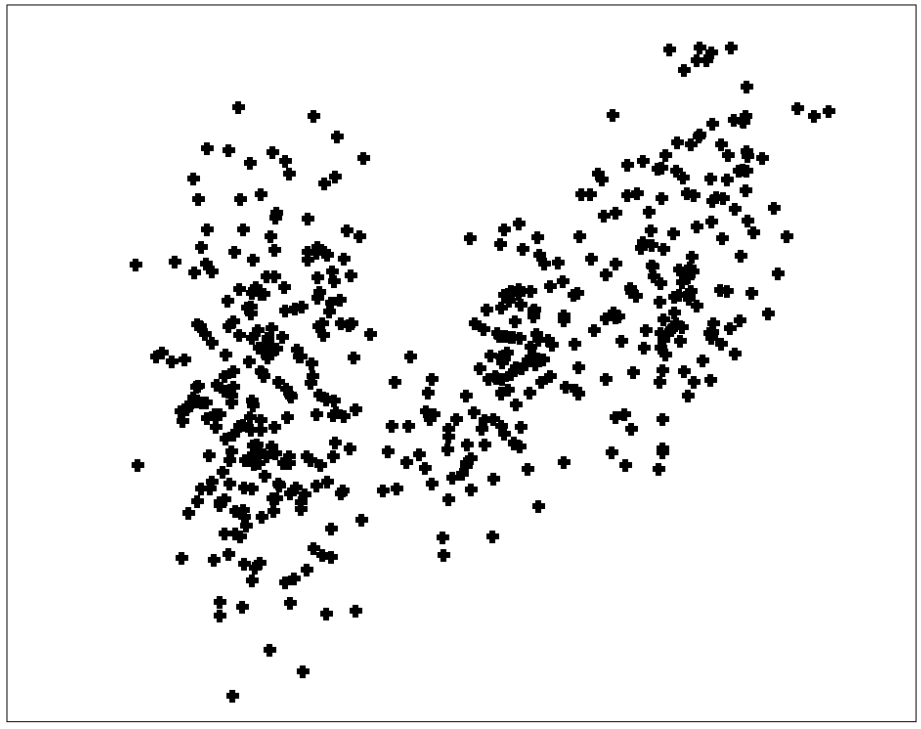
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$$

Estimation “statistique”

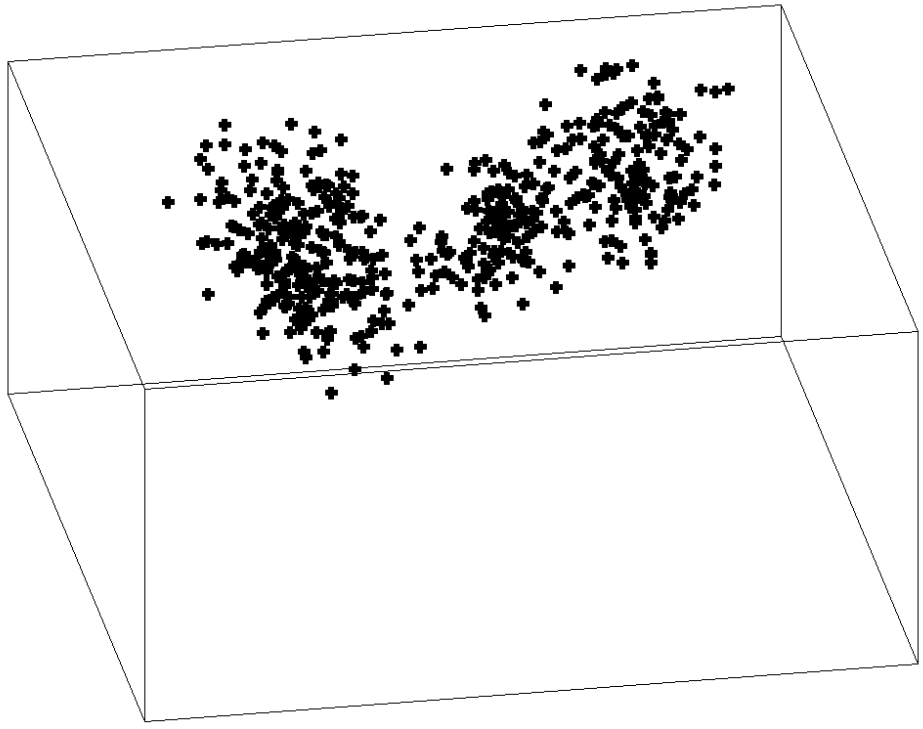
Estimation “statistique”



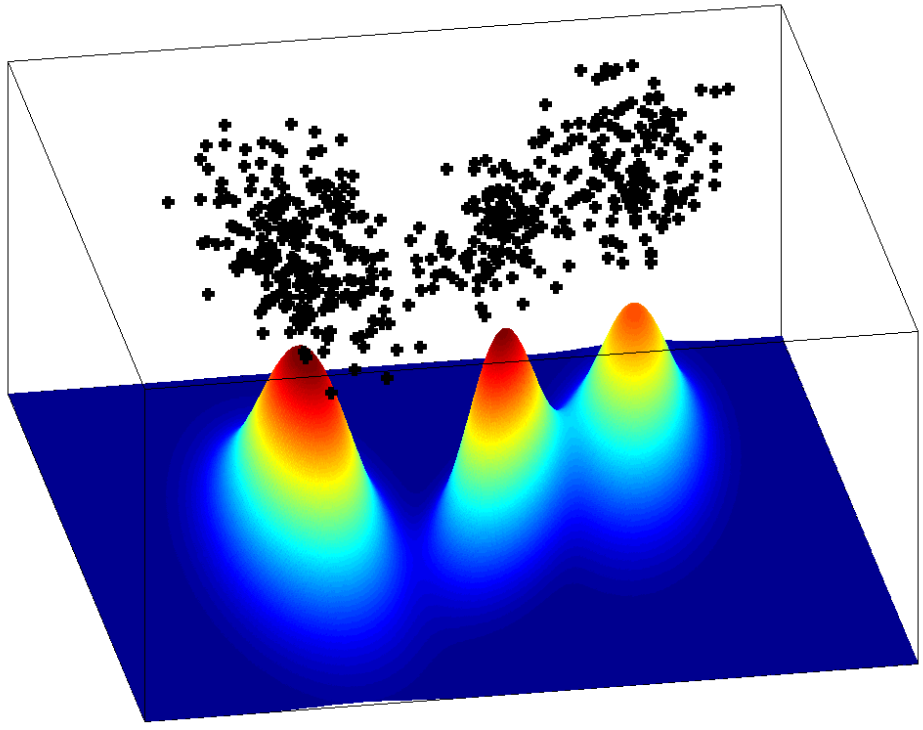
Estimation “statistique”



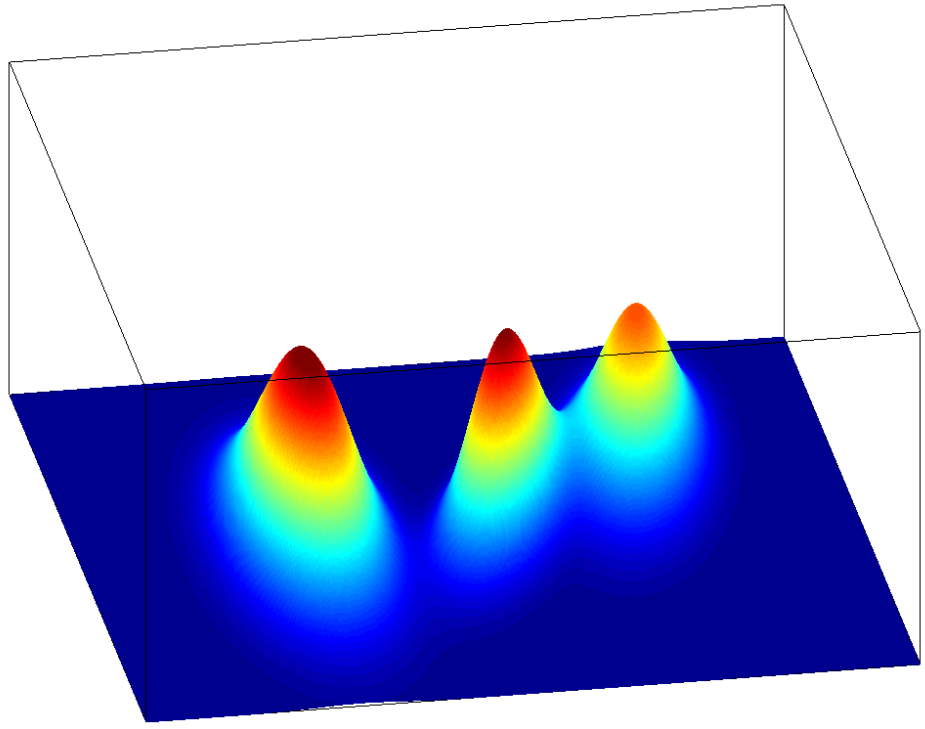
Estimation “statistique”



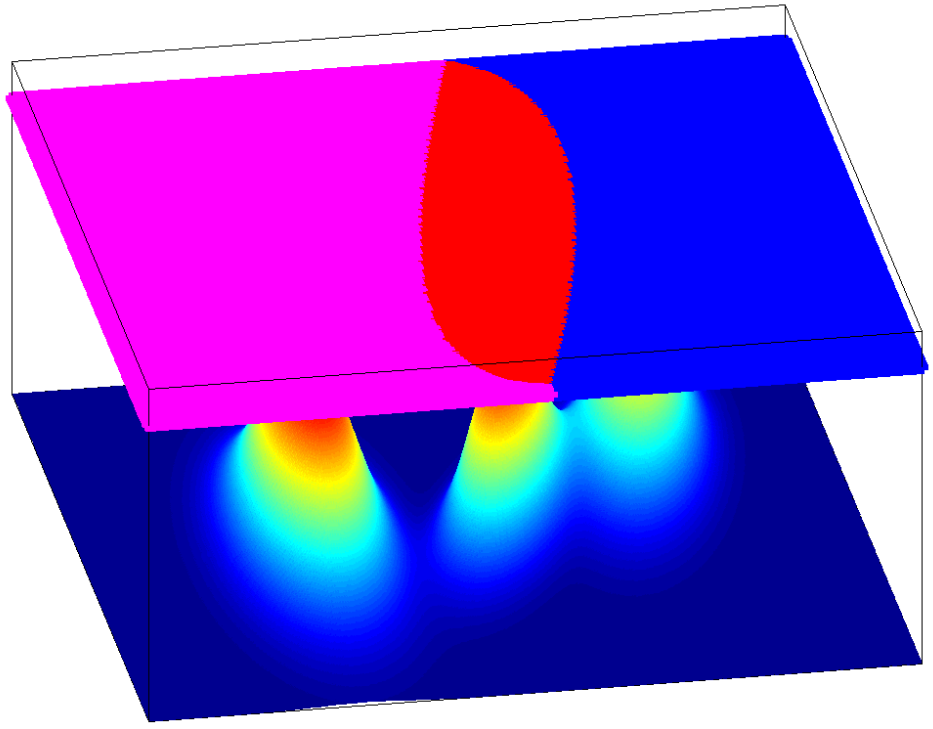
Estimation “statistique”



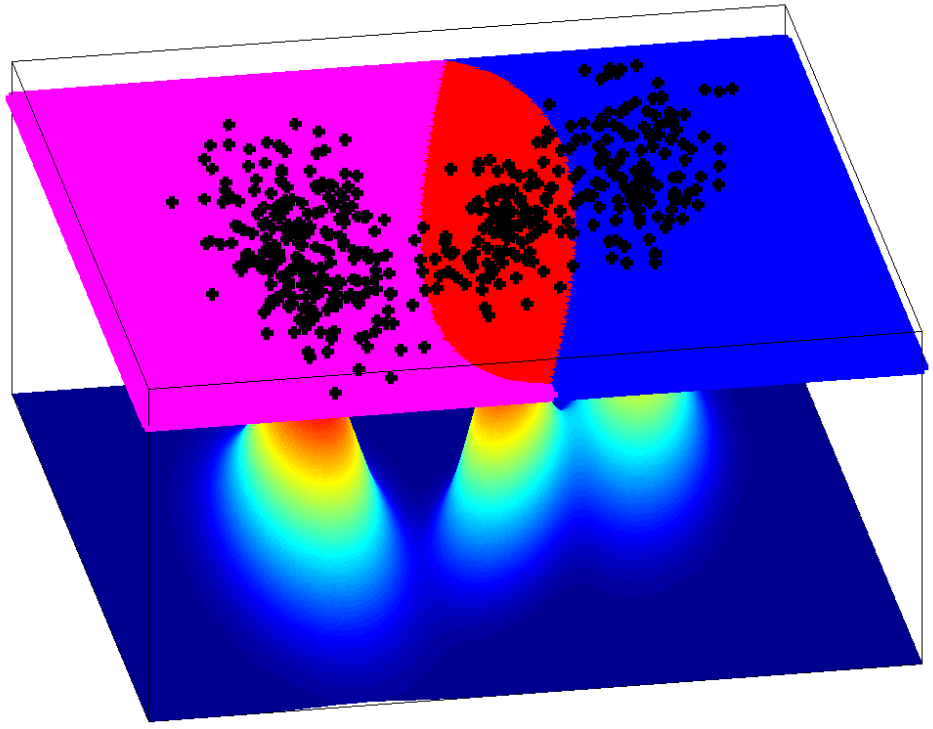
Estimation “statistique”



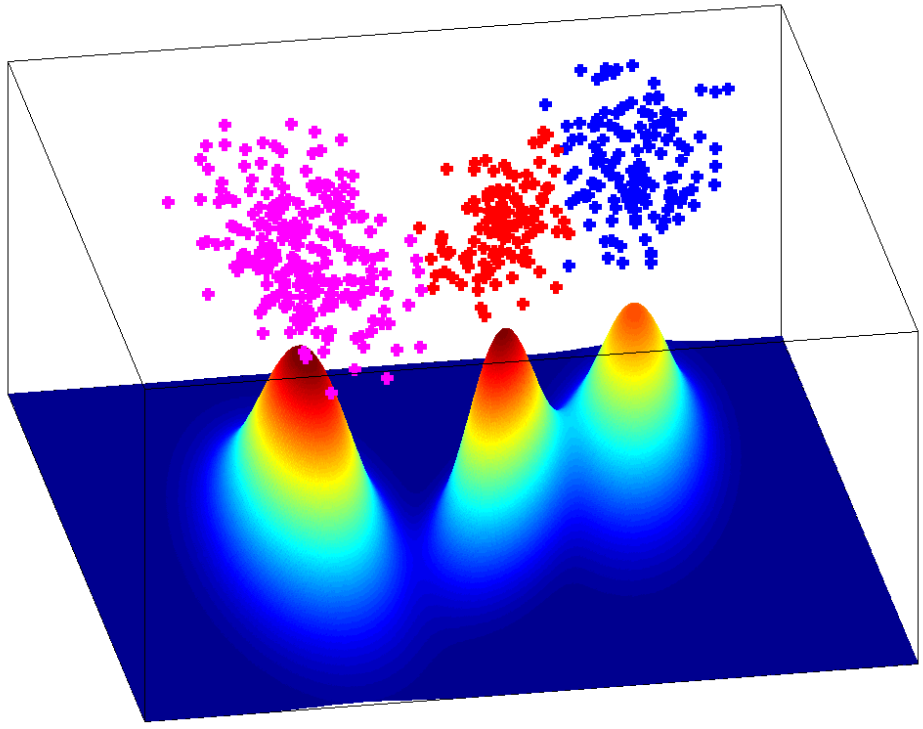
Estimation “statistique”



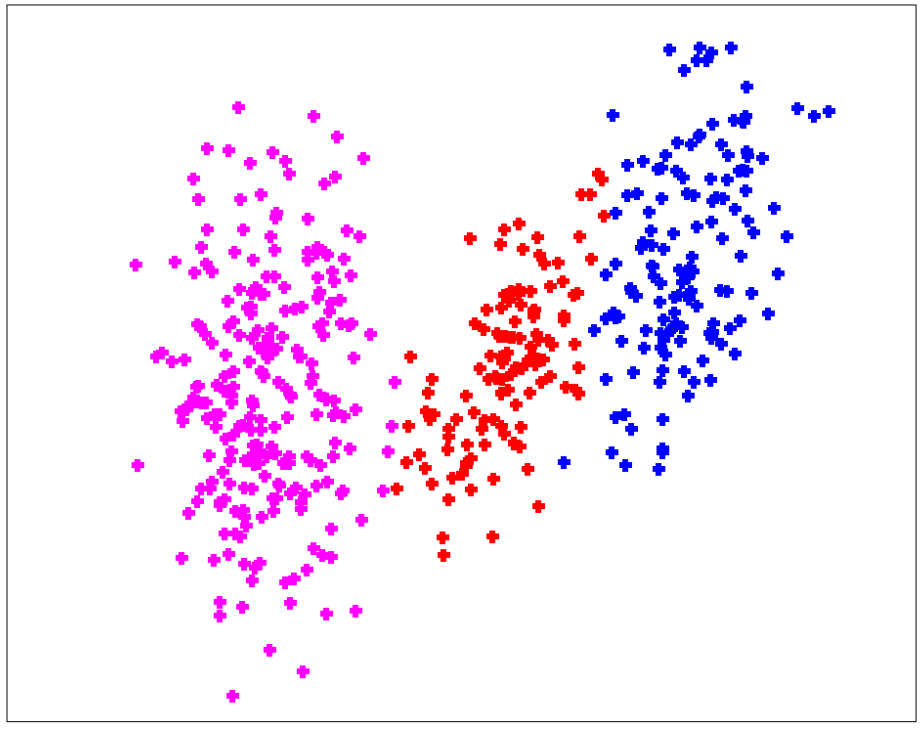
Estimation “statistique”



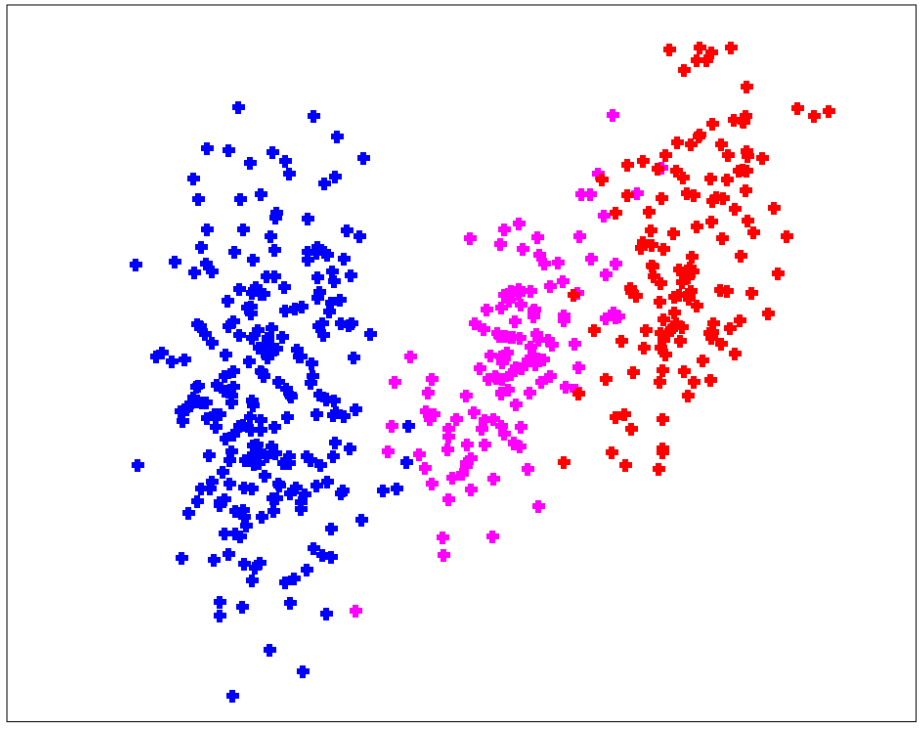
Estimation “statistique”



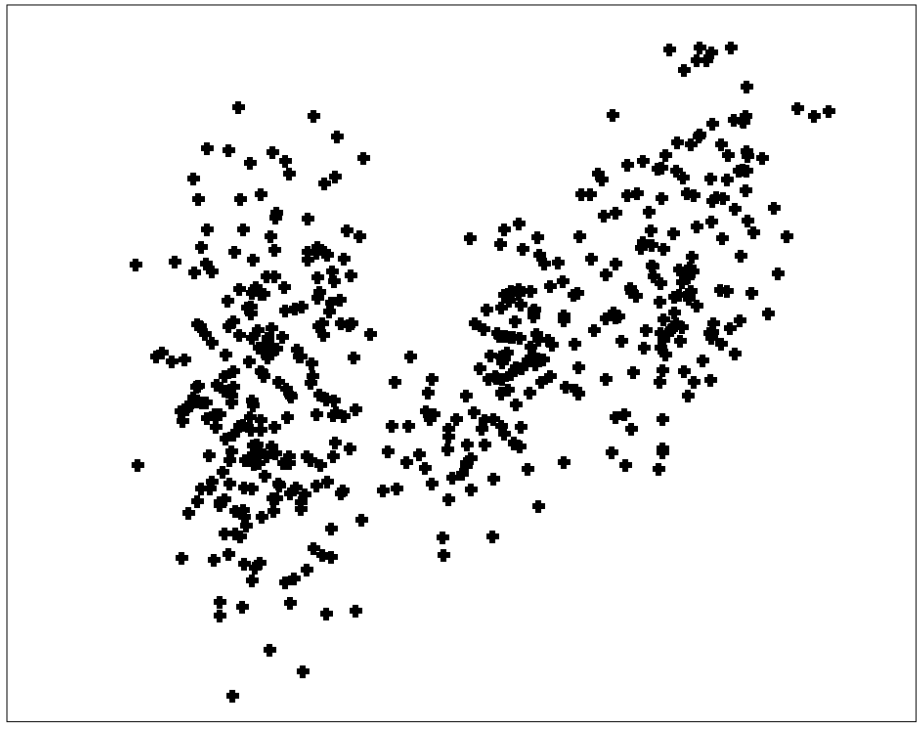
Estimation “statistique”



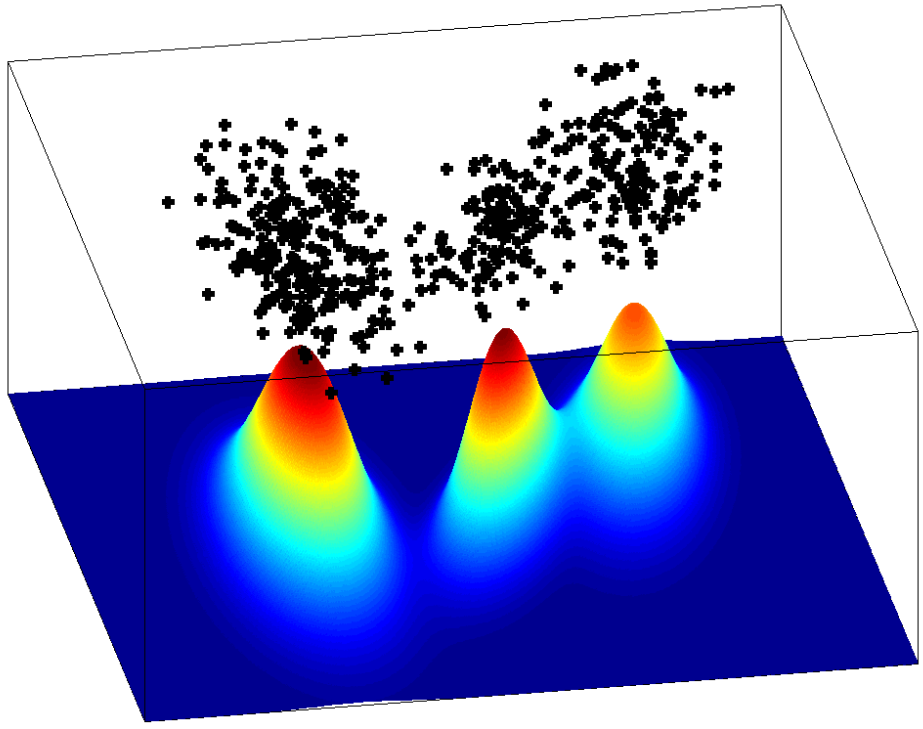
Estimation “statistique”



Estimation “statistique”



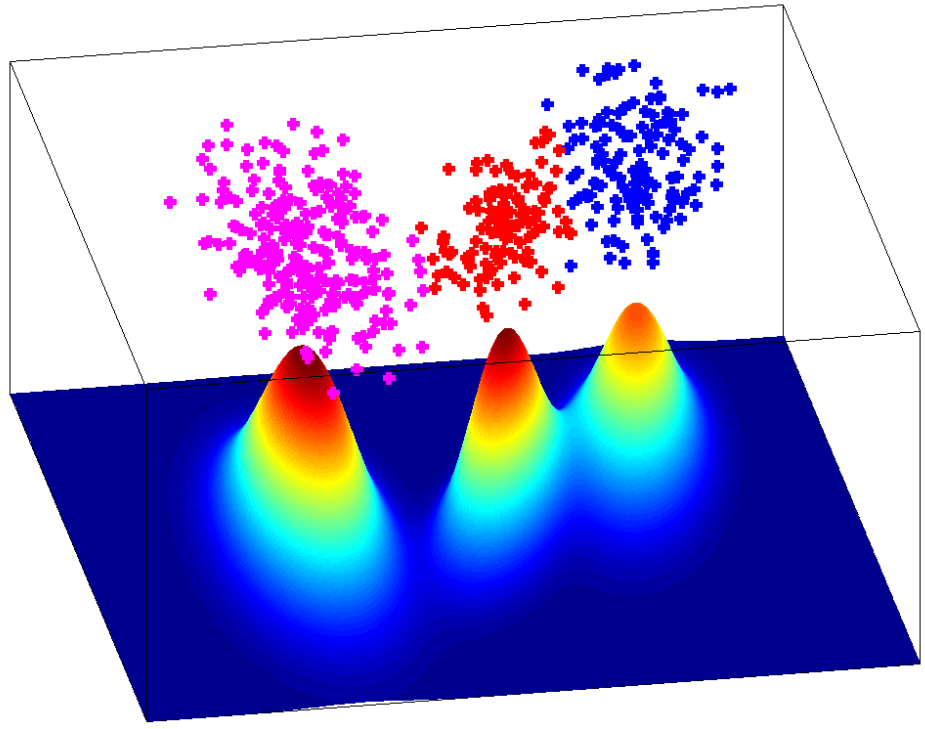
Estimation “statistique”



- Estimation des π_k, $\widehat{\mu_k}$ et $\widehat{\Sigma_k}$ par maximum de vraisemblance :

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

Estimation “statistique”



- Estimation des π_k, $\widehat{\mu_k}$ et $\widehat{\Sigma_k}$ par maximum de vraisemblance :

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

- Estimation de $\widehat{k}(\mathcal{S})$ par maximum à posteriori :

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

Modélisation par un mélange de gaussiennes

- Modélisation stochastique des spectres $\mathcal{S}$:
  - existence de K classes de spectres,
  - proportion π_k pour chacune des classes ($\sum_{k=1}^{K} \pi_k = 1$),
  - loi gaussienne $\mathcal{N}_{\mu_k,\Sigma_k}$ sur chacune des classes (restriction forte !)
- Densité s_0 de $\mathcal{S}$ proche de

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

- Objectif : estimer les paramètres K, π_k, μ_k, Σ_k à partir des données.
- Pourquoi ? : possibilité d'assigner ensuite une classe à une observation par maximum de vraisemblance

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

Modèle de mélange de gaussiennes

- Densité s_0 de $\mathcal{S}$ proche de $s_m(\mathcal{S}) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(\mathcal{S})$.
- Modèle $S_m = \{s_m\}$:
  - choix d'un nombre de classe K,
  - choix d'une structure pour les moyennes μ_k et les covariances $\Sigma_k = L_k D_k A_k D_k'$
- Modèles $[\mu\, L\, D\, A]^K$: contraintes (valeurs connues, communes ou libres...) sur les moyennes μ_k, les volumes L_k, les bases de diagonalisation D_k et les valeur propres A_k.
- Modèle S_m : modèle paramétrique de dimension $(K-1) + \dim([\mu\, L\, D\, A]^K)$ dans un espace de dimension p.
- Estimation par maximum de vraisemblance des paramètres :
  - pour chaque classe, la moyenne μ_k et la covariance $\Sigma_k = L_k D_k A_k D_k'$
  - les proportions π_k du mélange.
- Technique classique avec algorithme (EM) efficace disponible.

Max. de vraisemblance et MM

- "Maximum" de vraisemblance à K fixé :

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

- Fonctionnelle L compliquée !
- Algorithme itératif (MM) :
  - Estimée courante : $(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$,
  - Construction d'un Majorant $L^{(n)}$ de L tel que

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

    et $L^{(n)}$ facile à minimiser.
  - Calcul d'un Minimiseur

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

- Méthode très générique...
- La minimisation peut être remplacée par une simple diminution...

Max. de vraisemblance et EM

- Retour vers 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 : cas particulier de MM pour ce type de mélange,
  - Espérance (conditionnelle) : à l'étape n, on pose

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

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

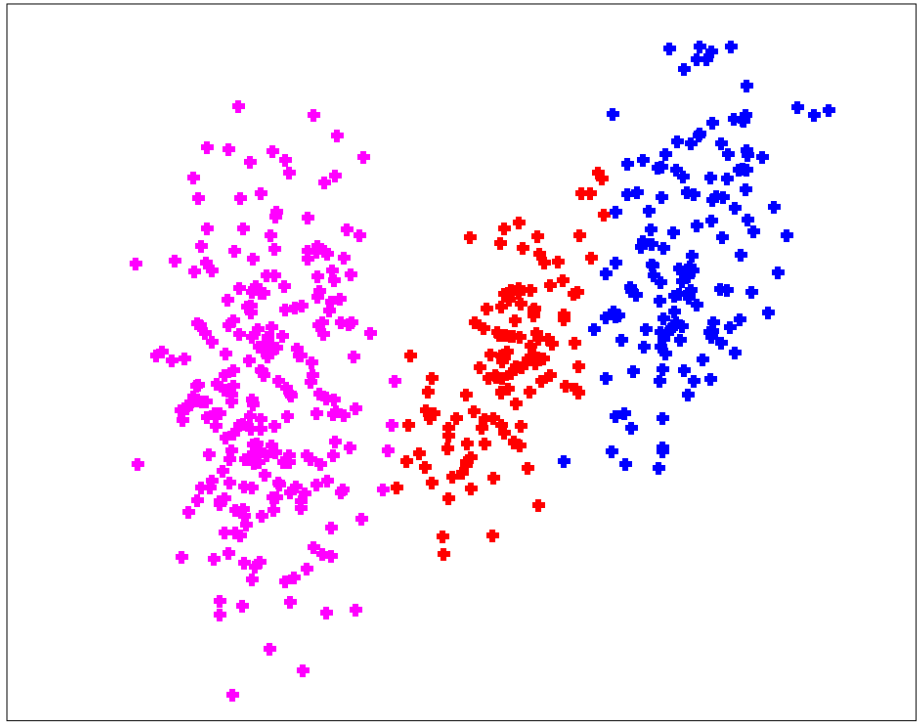
  - Kullback : $L^i \leq L^{i,(n)} + \text{Cst}^{i,(n)}$ avec égalité en $(\pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$.
- Bonus :
  - Séparabilité de $L^{i,(n)}$ en π et (μ, Σ) :

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

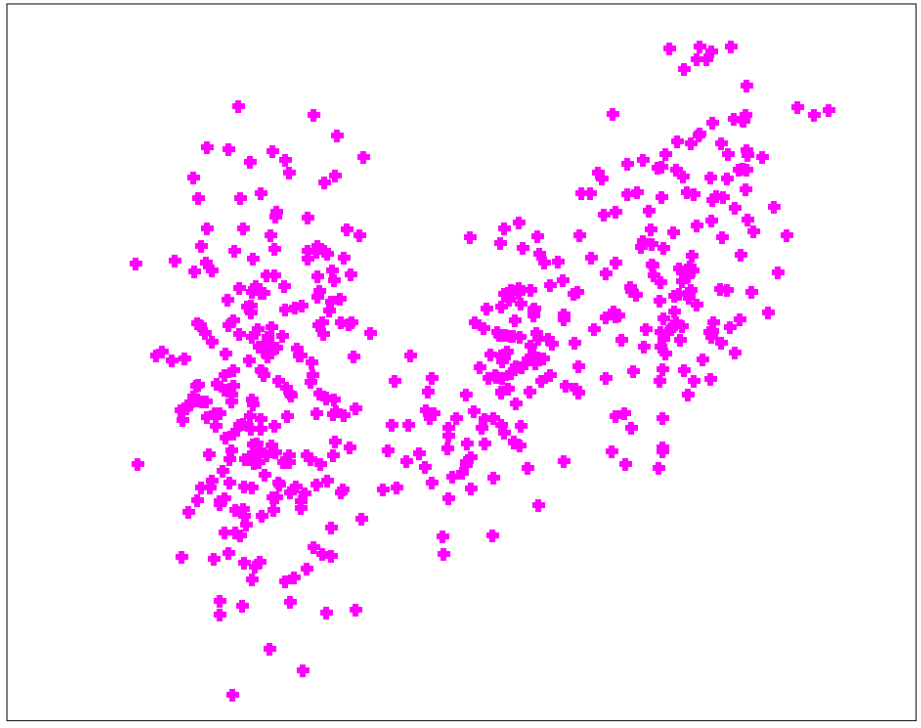
  - Formules closes pour la Minimisation de $L^{(n)}$ en π et (μ, Σ) !

Combien de classes ?

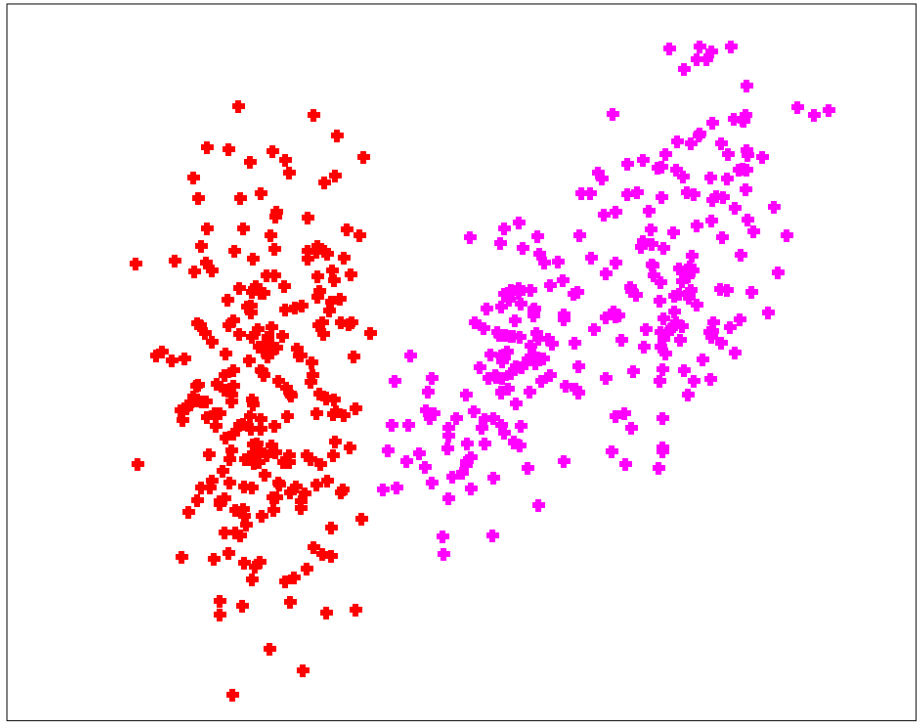
Combien de classes ?



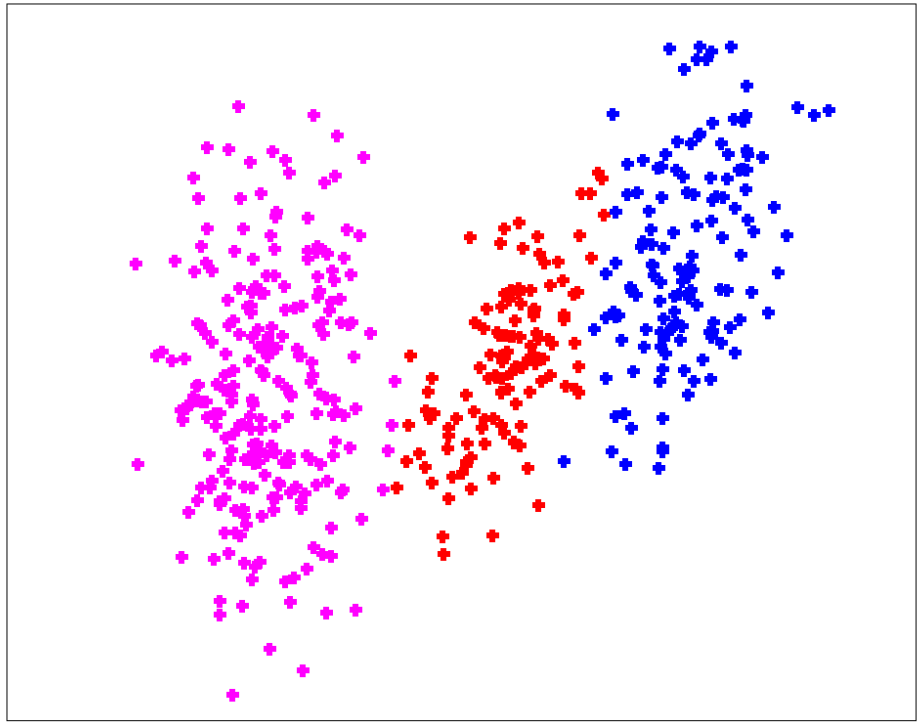
Combien de classes ?



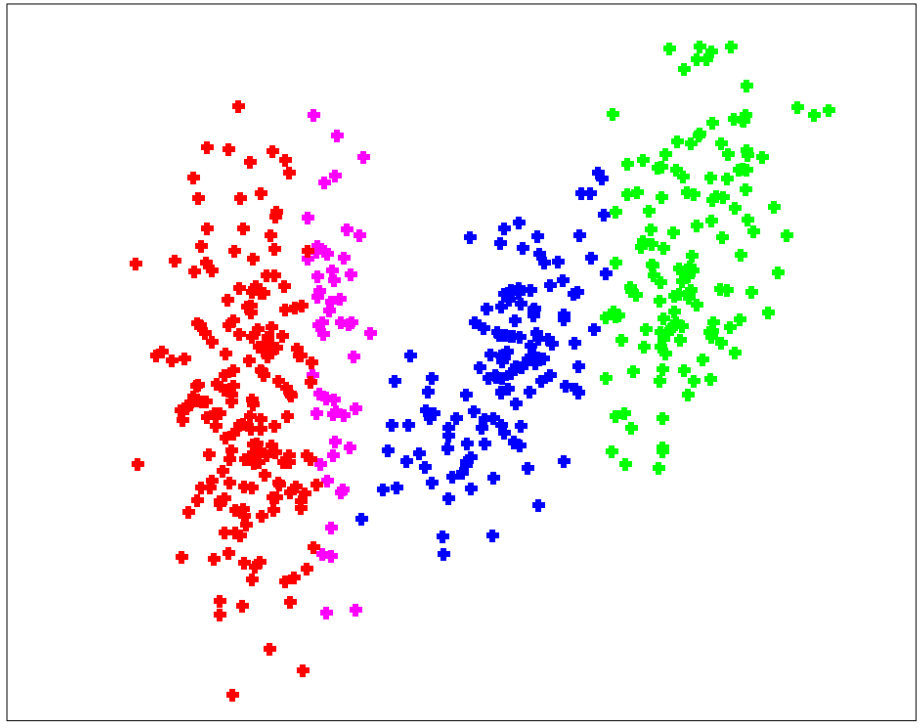
Combien de classes ?



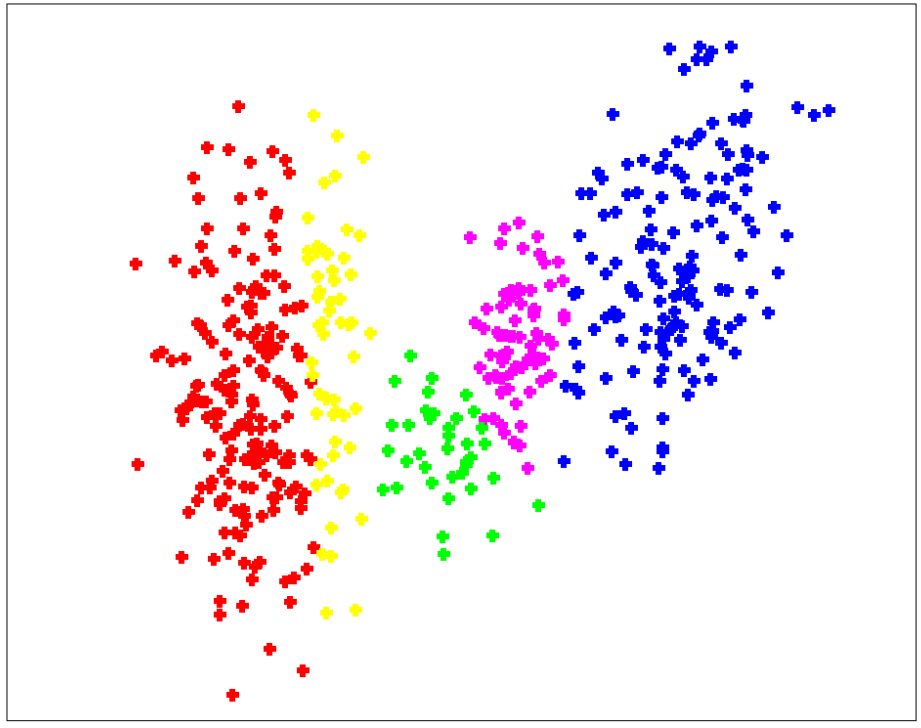
Combien de classes ?



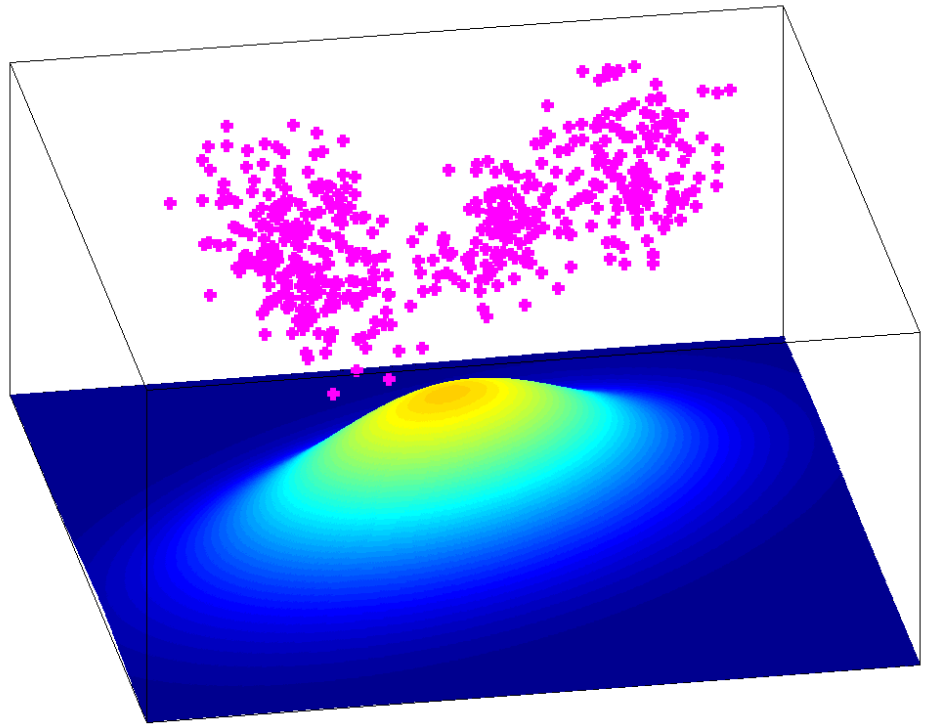
Combien de classes ?



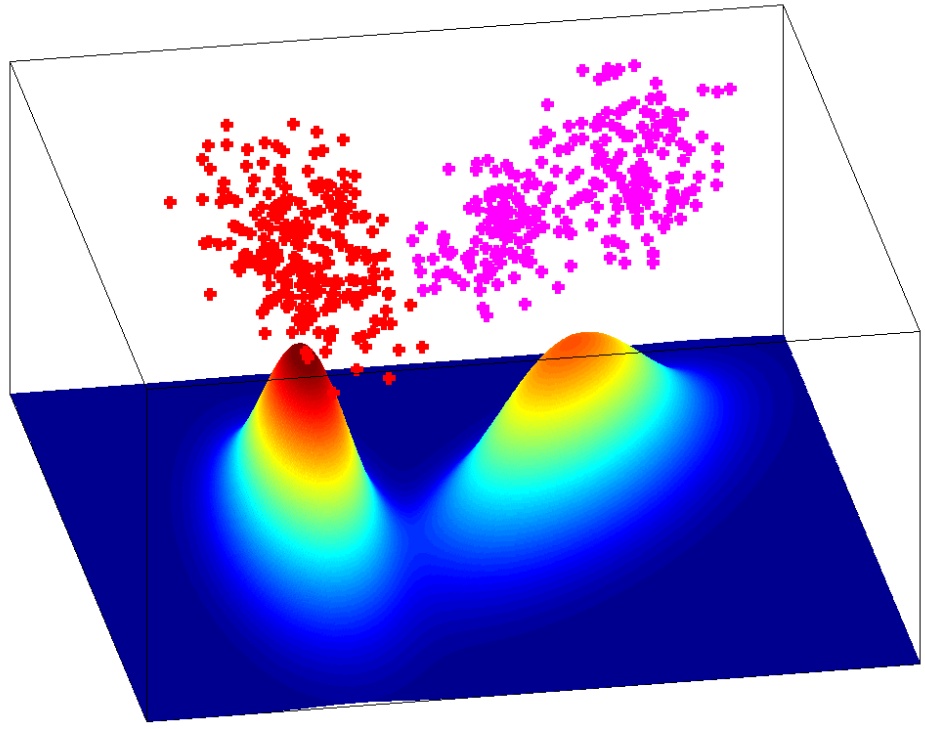
Combien de classes ?



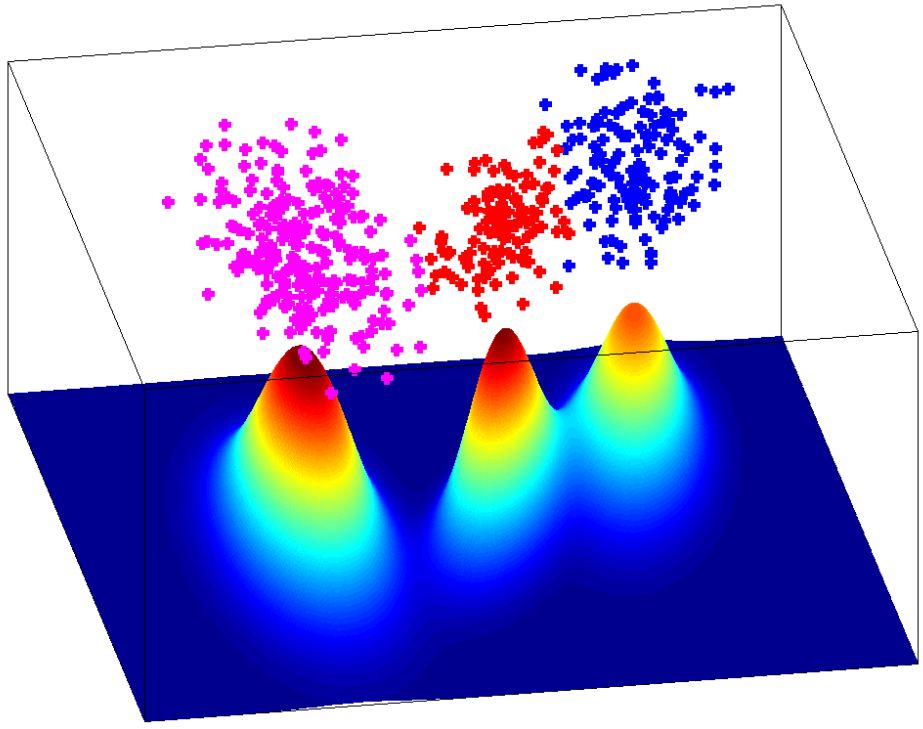
Combien de classes ?



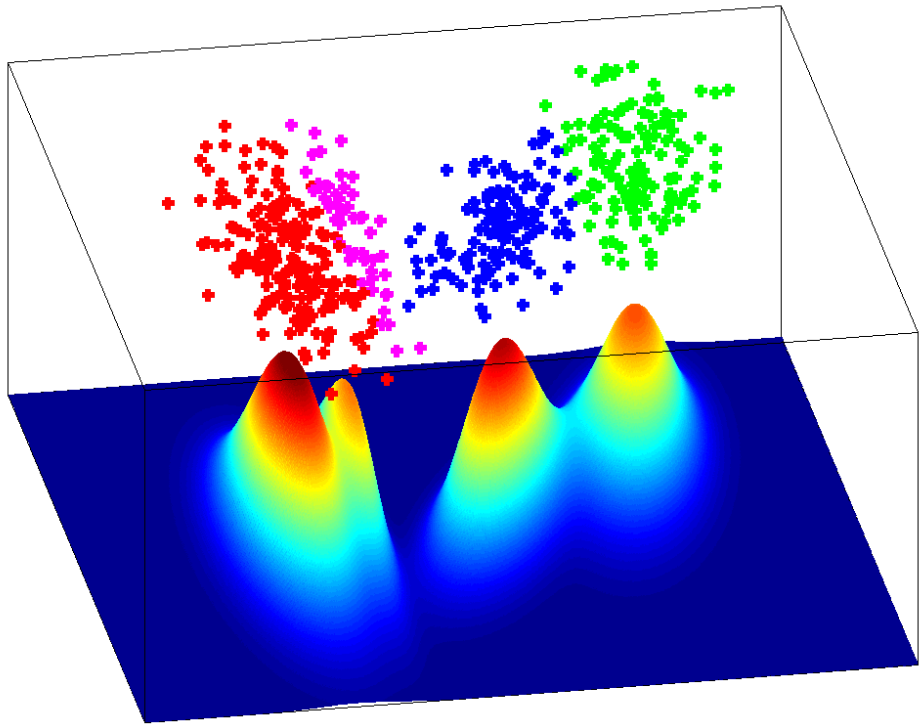
Combien de classes ?



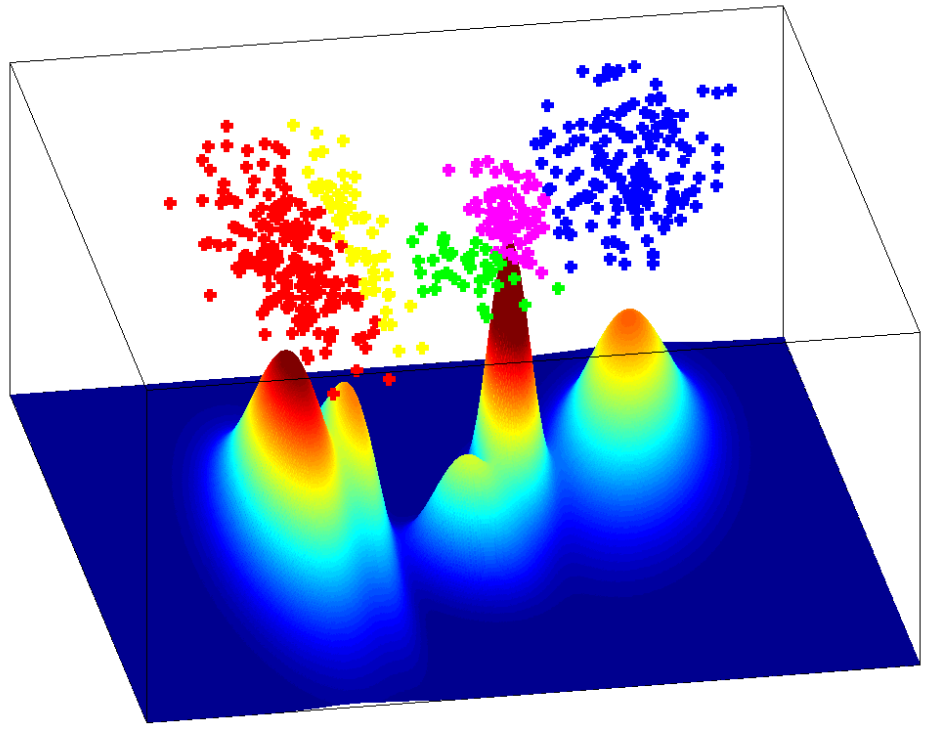
Combien de classes ?



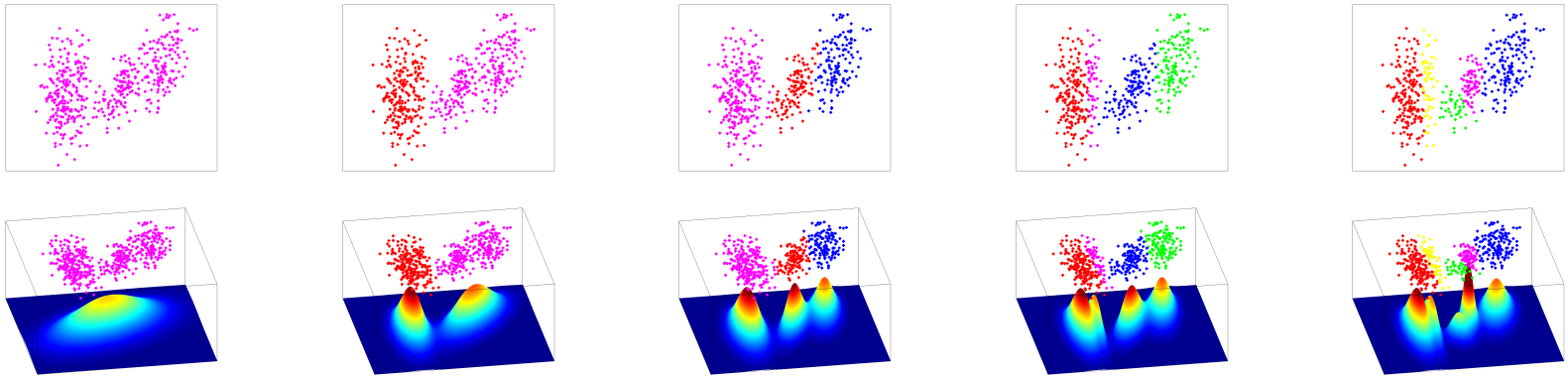
Combien de classes ?



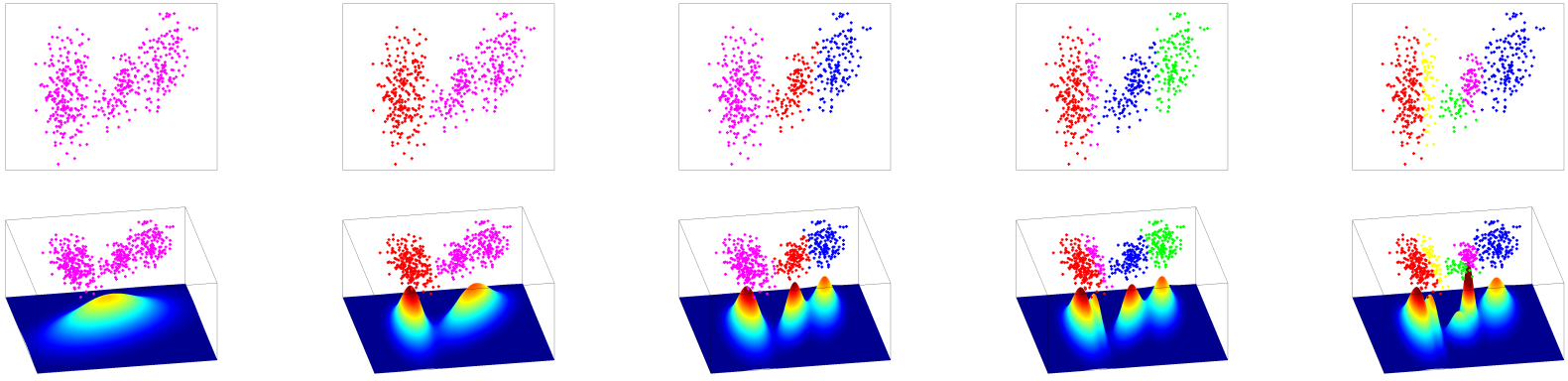
Combien de classes ?



Combien de classes ?

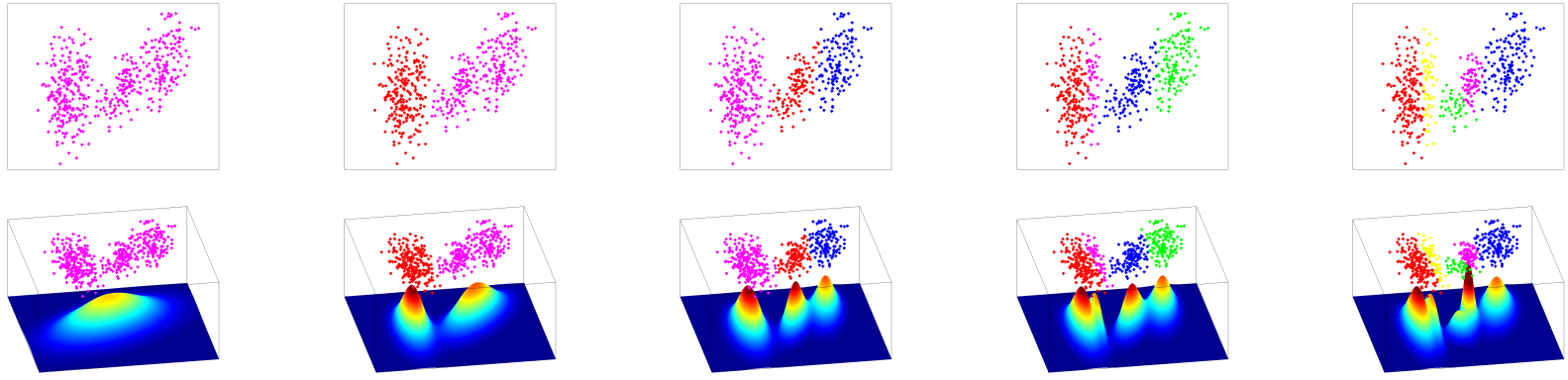


Combien de classes ?



Fidélité $\quad -- \quad + \quad +++ \quad +++\cdot \quad ++++$

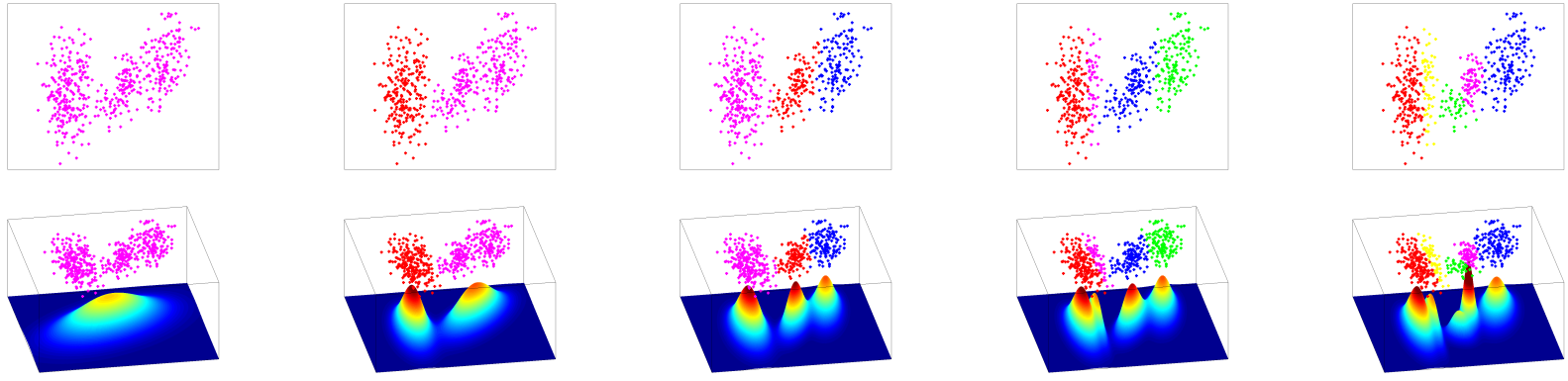
Combien de classes ?



Fidélité $\quad -- \quad + \quad +++ \quad +++\cdot \quad ++++$

- Question difficile où la vraisemblance (la fidélité) ne suffit pas !

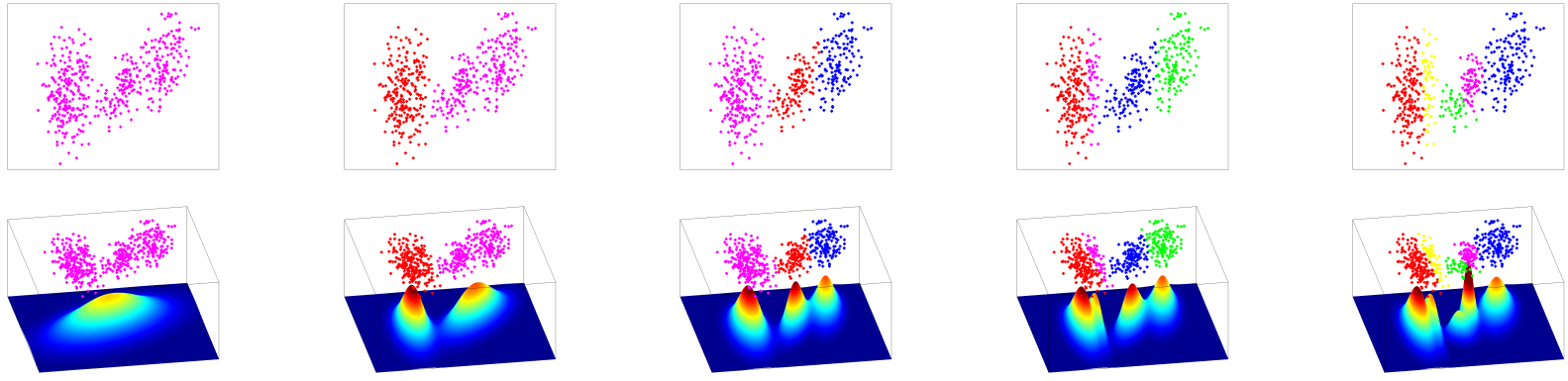
Combien de classes ?



Fidélité	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicité	$+++++$	$+++$	$++$	$+$	$-$

- Question difficile où la vraisemblance (la fidélité) ne suffit pas !

Combien de classes ?



Fidélité	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicité	$+++++$	$+++$	$++$	$+$	$-$

- Question difficile où la vraisemblance (la fidélité) ne suffit pas !
- Prise en compte de la complexité du modèle ?

Le rasoir d'Ockham

Le rasoir d'Ockham

Les multiples ne doivent pas être utilisés sans nécessité.
Guillaume d'Ockham ($\sim$ 1285 - 1347)

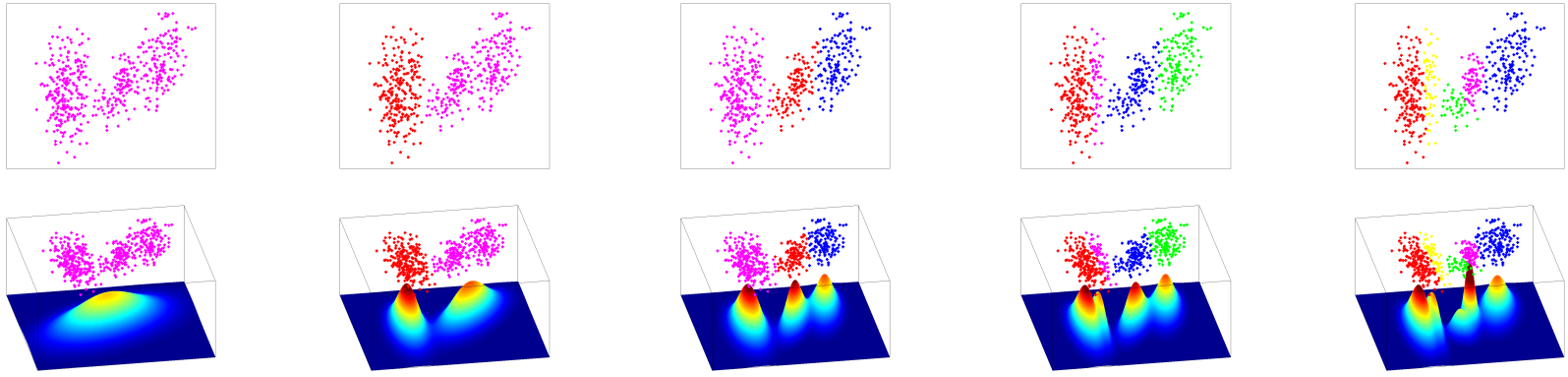
Le rasoir d'Ockham

Les multiples ne doivent pas être utilisés sans nécessité.
Guillaume d'Ockham ($\sim$ 1285 - 1347)

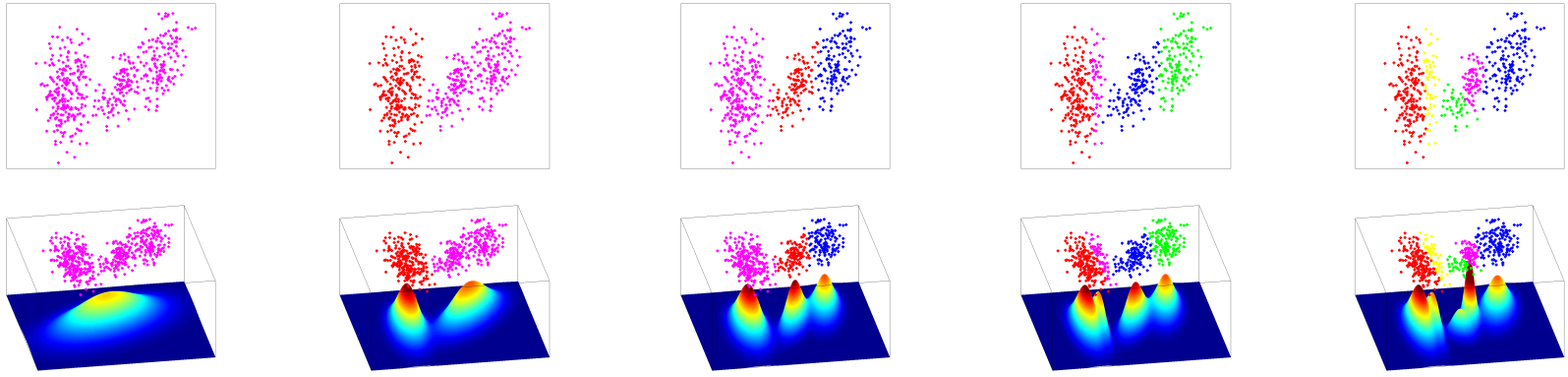
- Rasoir d'Ockham (principe de simplicité) : il ne faut pas ajouter des hypothèses, si celles utilisées suffisent déjà !
- Compromis entre pouvoir d'explication et simplicité.

Sélection par pénalisation

Sélection par pénalisation

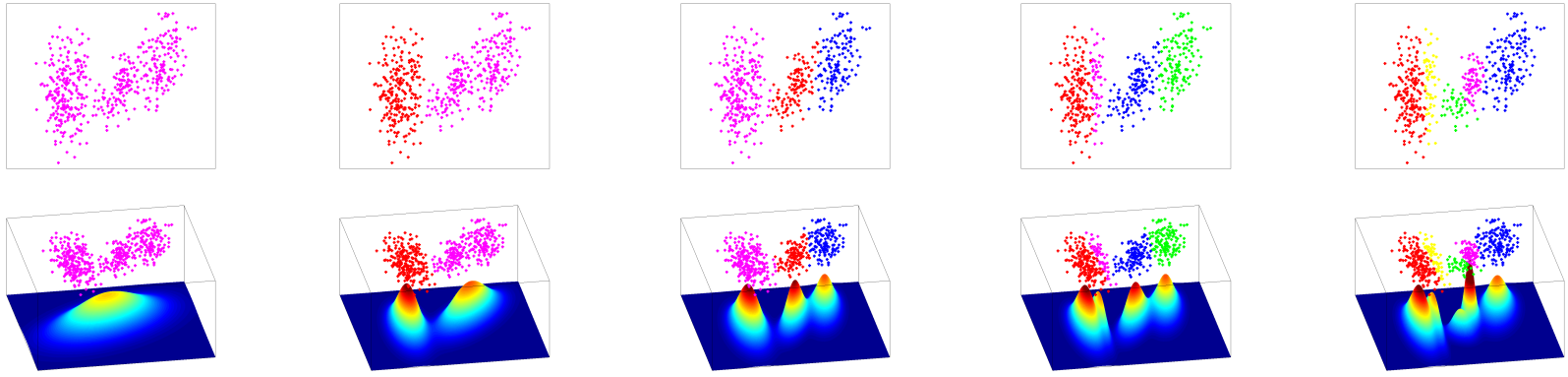


Sélection par pénalisation



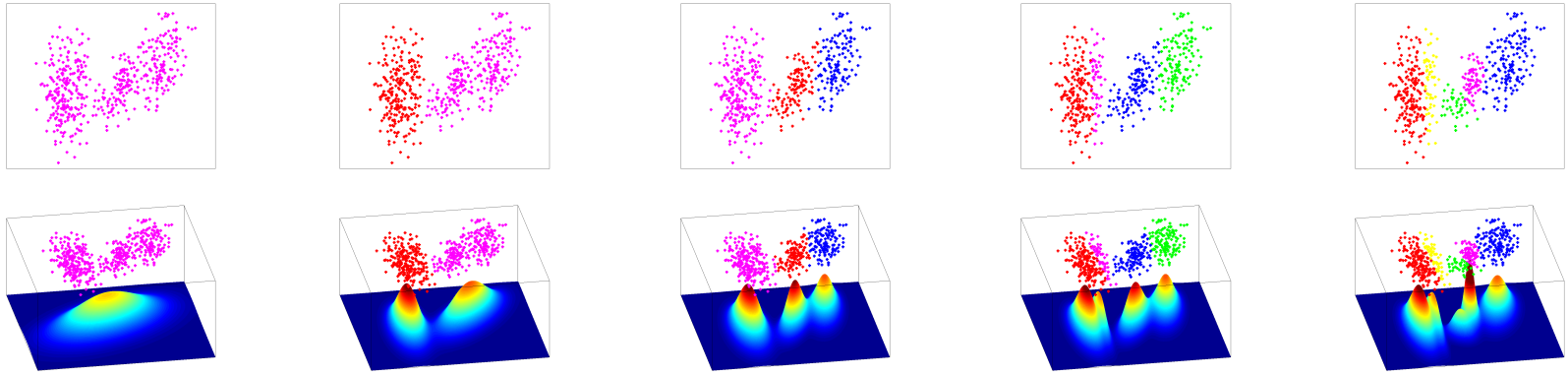
Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicité	$+++++$	$+++$	$++$	$+$	$-$

Sélection par pénalisation



Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicité	$+++++$	$+++$	$++$	$+$	$-$
$=$ Compromis	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

Sélection par pénalisation

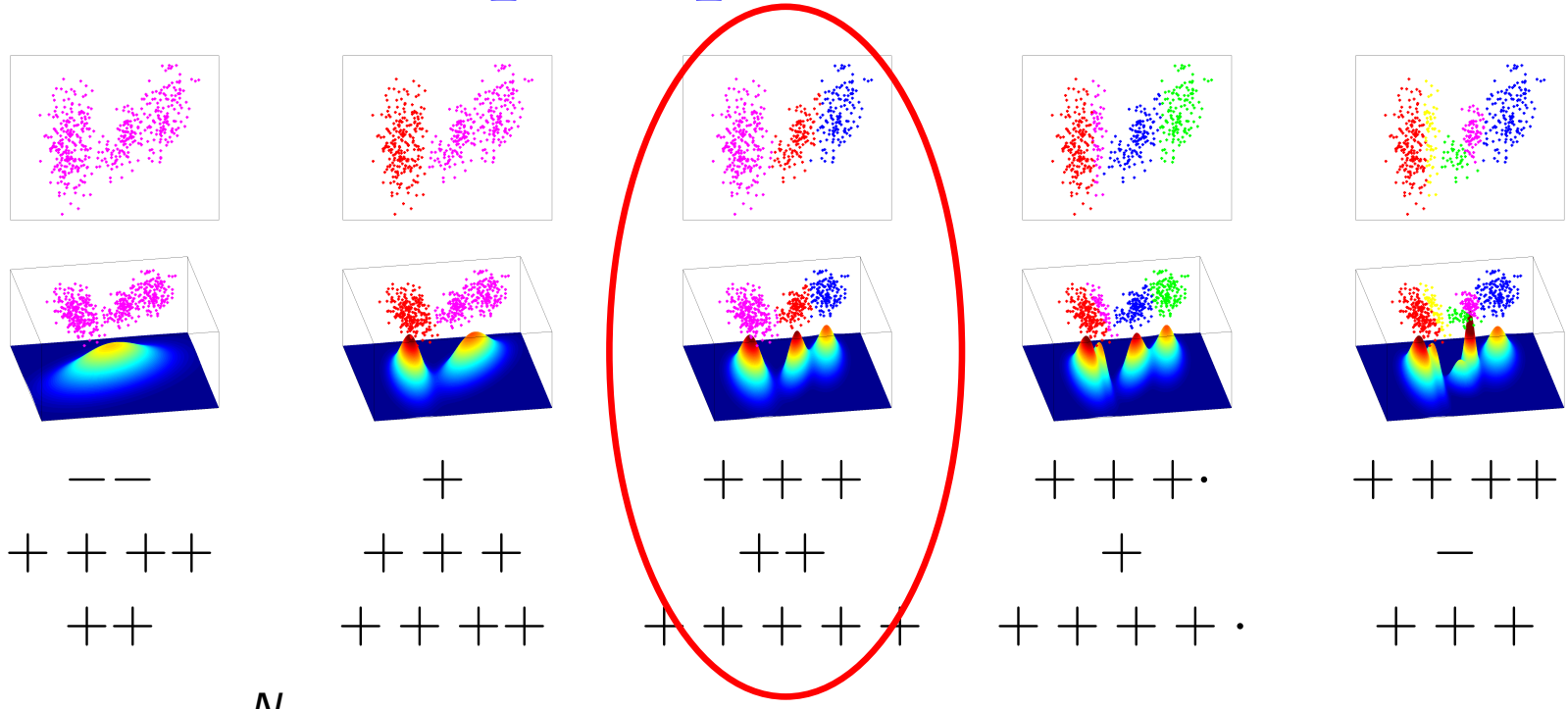


Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicité	$+++++$	$+++$	$++$	$+$	$-$
$=$ Compromis	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

- Vraisemblance : $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicité : $-\lambda \mathrm{Dim}(S_K)$ (beaucoup de théorie derrière).
- Estimateur pénalisé :

$$\mathrm{argmax} \underbrace{\sum_{i=1}^{N} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \mathrm{Dim}(S_K)}_{\text{Pénalité}}$$

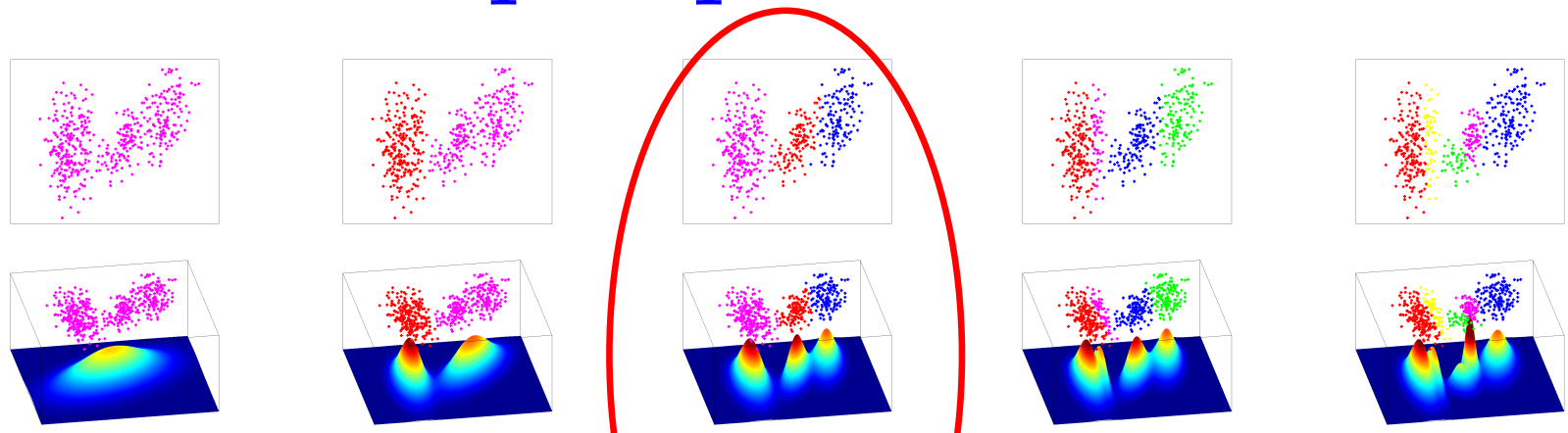
Sélection par pénalisation



- Vraisemblance : $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicité : $-\lambda \mathrm{Dim}(S_K)$ (beaucoup de théorie derrière).
- Estimateur pénalisé :

$$\mathrm{argmax} \underbrace{\sum_{i=1}^{N} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \mathrm{Dim}(S_K)}_{\text{Pénalité}}$$

Sélection par pénalisation



Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicité	$+++++$	$+++$	$++$	$+$	$-$
$=$ Compromis	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

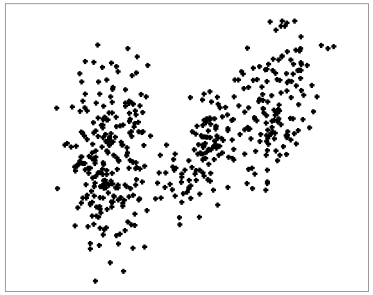
- Vraisemblance : $\sum_{i=1}^{N} \log \hat{s}_K(X_i)$.
- Simplicité : $-\lambda \mathrm{Dim}(S_K)$ (beaucoup de théorie derrière).
- Estimateur pénalisé :

$$\mathrm{argmax} \underbrace{\sum_{i=1}^{N} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \mathrm{Dim}(S_K)}_{\text{Pénalité}}$$

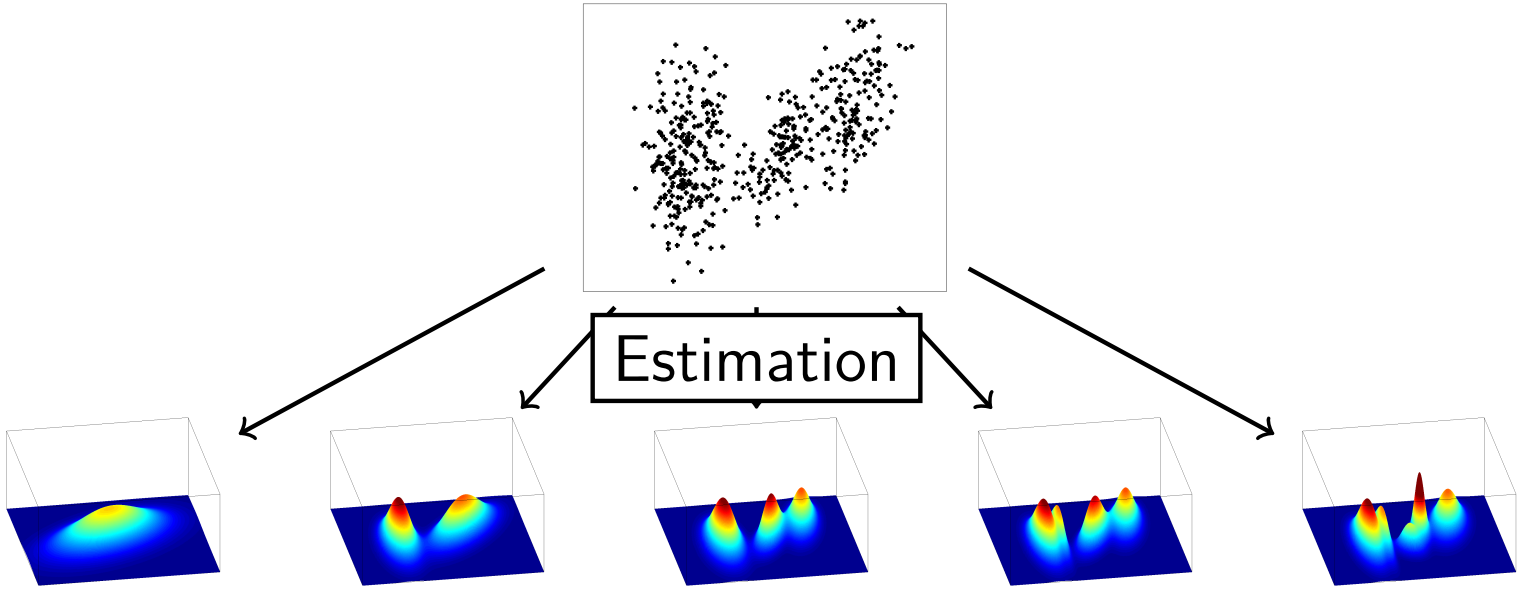
- Optimisation en K par exploration exhaustive !

Méthodologie

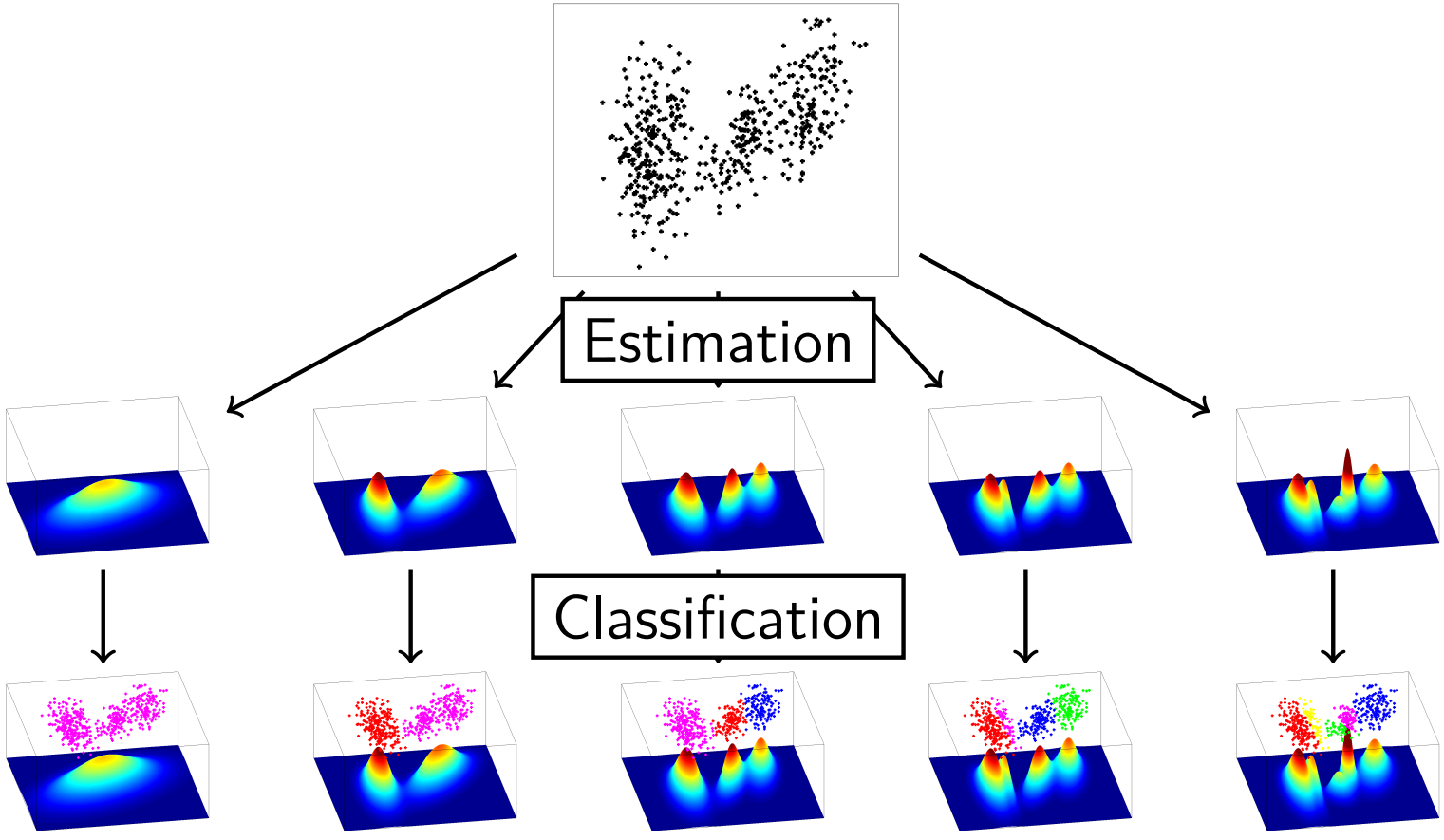
Méthodologie



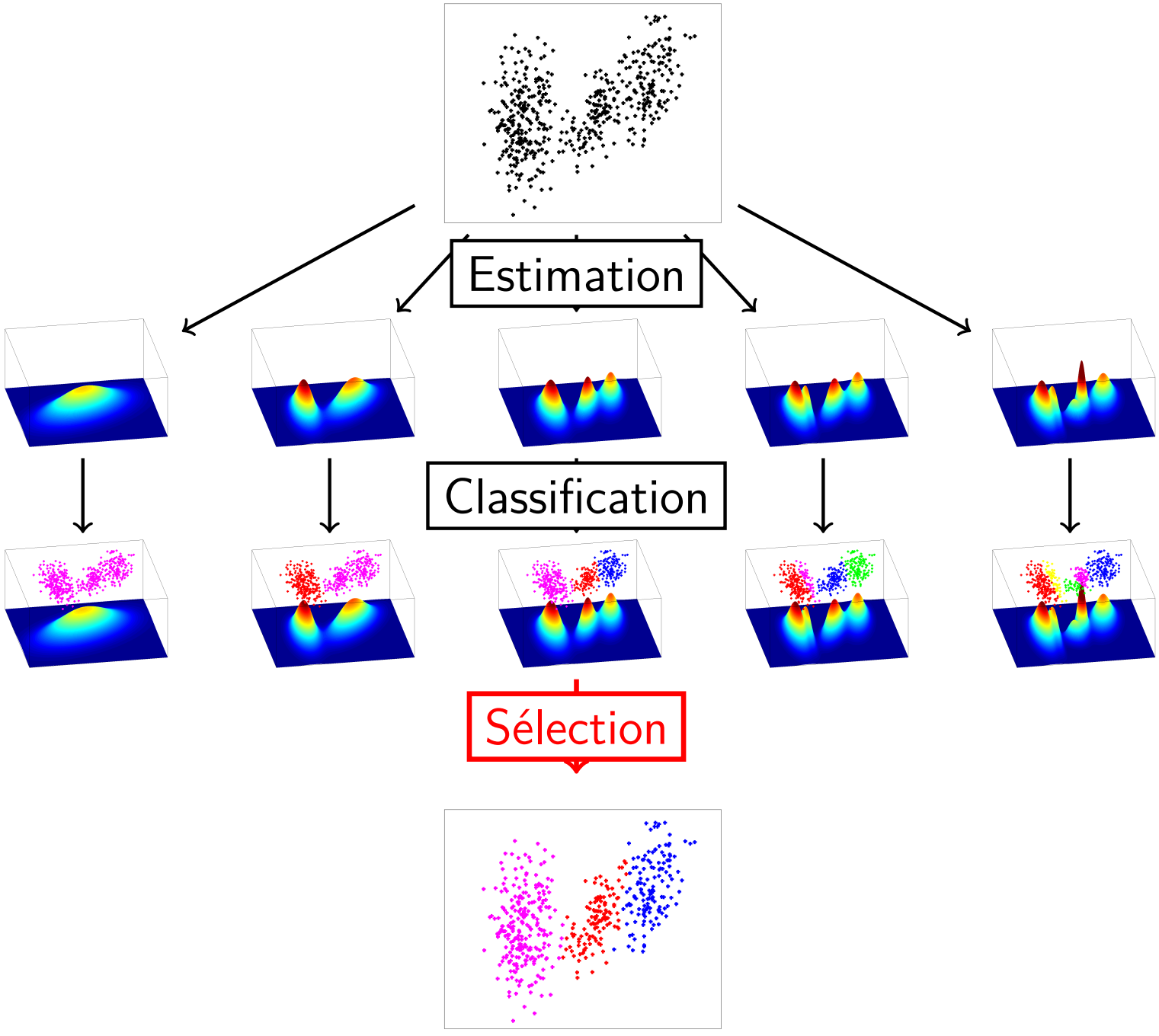
Méthodologie



Méthodologie



Méthodologie



Sélection de modèles

- Comment choisir le "modèle" S_m :
  - le nombre de classe K,
  - le modèle $[\mu\, L\, D\, A]^K$?
- Principe de sélection de modèles par pénalisation :
  - choix d'une collection de modèles $S_m = \{s_m\}$ avec $m \in \mathcal{S}$,
  - estimation par maximum de vraisemblance d'une densité $\hat{s}_m$ pour chaque modèle S_m,
  - sélection d'un modèle $\widehat{m}$ par

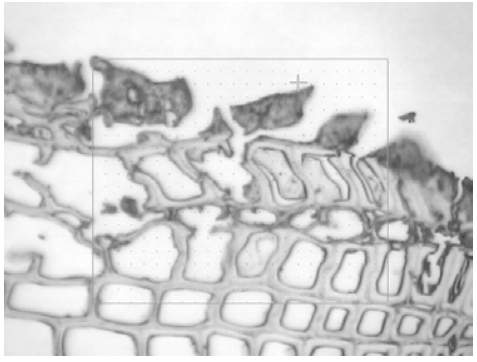
$$\widehat{m} = \operatorname{argmin} - \ln(\widehat{s}_m) + \operatorname{pen}(m).$$

  avec $\operatorname{pen}(m) = \kappa(\ln(n))\dim(S_m)$ (dimension intrinsèque de S_m),
- Résultats (Birgé, Massart, Celeux, Maugis, Michel...) :
  - théorique d'estimation du mélange : pour κ assez grand,

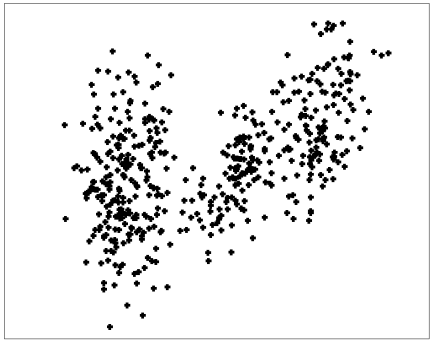
$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\operatorname{pen}(m)}{n} \right) + \frac{C'}{n}.$$

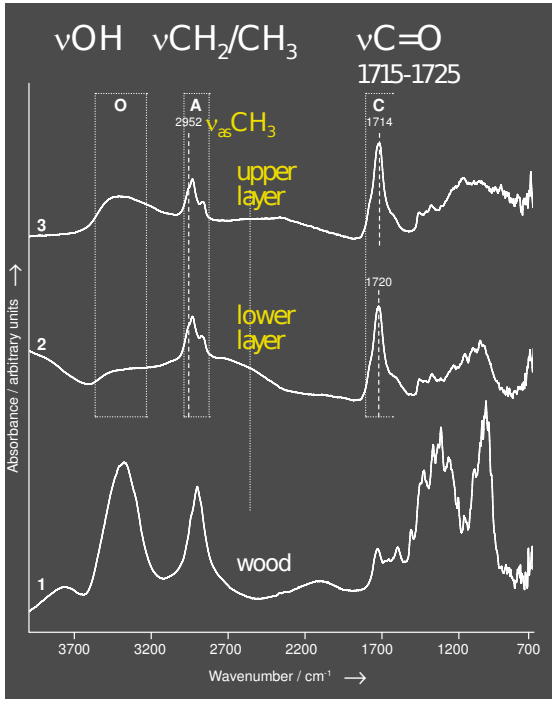
  - pratique de classification non supervisée ($\neq$ segmentation),
  - consistance de la classification si $\ln\ln(n)$ dans la pénalité...

Retour à nos violons



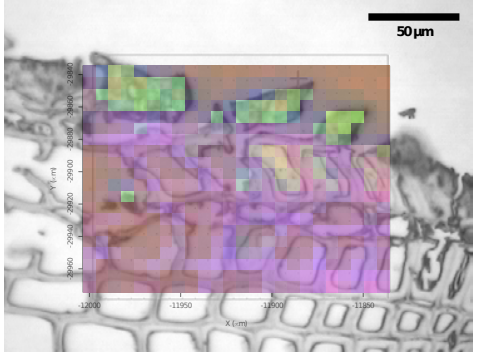
Représentation

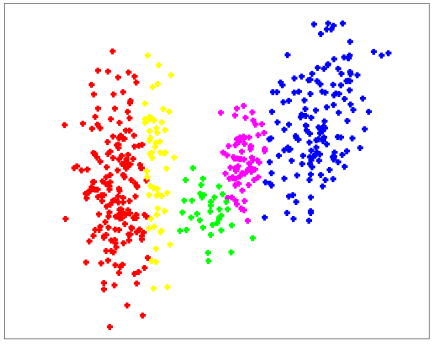




Segmentation

Classification





Info. Spatiale

Segmentation et mélange de gaussiennes

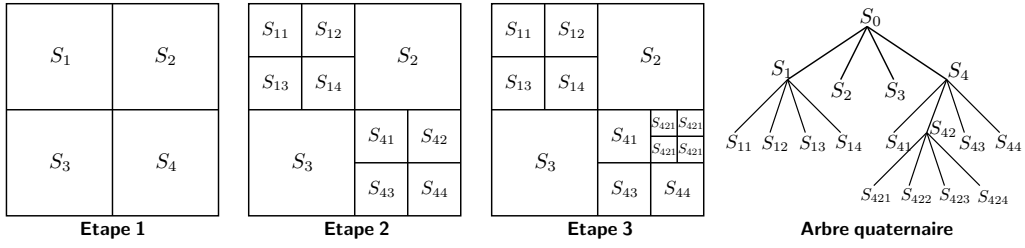
- Objectif initial : segmentation $\neq$ classification non supervisée.
- Prise en compte de la position spatiale x du spectre à travers les proportions du mélange (Kolaczyk et al) : modèle de densités conditionnelles

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

- Modèle mélangeant paramétrique et "non-paramétrique"...
- Estimation à partir des données :
  - pour chaque classe, la moyenne μ_k et la covariance $\Sigma_k = L_k D_k A_k D_k'$,
  - de la fonction de mélange $\pi_k(x)$.
- $\pi_k(x)$ fonction : régularisation nécessaire.
- Principe de sélection de modèles...

Mélange de gaussiennes et partition hiérarchique

- Comment choisir le "modèle" S_m ? :
  - le nombre de classe K,
  - le modèle $[\mu\, L\, D\, A]^K$,
  - la structure des paramètres de mélange $\pi_k(x)$.
- Structure simple : $\pi_k(x) = \sum_{\mathcal{R}\in\mathcal{P}} \pi_k[\mathcal{R}]\chi_{\{x\in\mathcal{R}\}} = \pi_k[\mathcal{R}(x)]$
  - constant par morceau sur une partition "hiérarchique",
  - optimisation efficace possible,
  - performance d'approximation raisonnable.



- $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
- Pénalité $\mathrm{pen}(m) = \kappa \ln(n) \dim(S_m)$ suffisante pour
  - le contrôle théorique en terme d'estimation de densité,
  - l'optimisation numérique (EM + programmation dynamique).

Densités conditionnelles

- Cadre plus général : observation de (X_i, Y_i) avec X_i indépendants et Y_i indépendants de loi de densité $s_0(y|X_i)$.
- Objectif : estimation de $s_0(y|x)$.
- Principe de sélection de modèles par pénalisation :
  - choix d'une collection de modèles $S_m = \{s_m(y|x)\}$ avec $m \in \mathcal{S}$,
  - estim. par max. de vraisemblance d'une dens. $\hat{s}_m$ pour chaque modèle S_m :

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

  - avec $\operatorname{pen}(m)$ à bien choisir, sélection d'un modèle $\widehat{m}$ par

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

- Résultat d'estimation de densité du type

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

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

Théorème

- Inégalité oracle

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

dès que

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

où $n\sigma_m^2$ mesure la complexité du modèle S_m (entropie) et x_m le coût de codage dans la collection (Kraft).

- « Distances » utilisées $KL^{\otimes n}$ et $JKL_{\rho}^{\otimes n}$: divergence de Kullback et divergence de Jensen-Kullback « tensorisées ».
- $n\sigma_m^2$ lié à l'entropie à crochet de S_m mesurée par rapport à la distance de Hellinger tensorisée $d^{2\otimes n}$.

Le cas des modèles de mélanges spatiaux

- Calcul d'un majorant des entropies à crochet possible (cf Maugis et Michel) impliquant :

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

- Codage de la collection avec $x_m \leq \kappa'' |\mathcal{P}| \leq \frac{\kappa''}{K-1} \dim(S_m)$.
- Condition sur la pénalité :

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

Optimisation numérique

- Sélection de modèle par pénalisation :

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

- Optimisation du nombre de classe K et de la structure des moyennes et des covariances $[\mu\,L\,D\,A]^K$ par exploration exhaustive.
- Sélection de modèle à nombre de classes K et structure $[\mu\,L\,D\,A]^K$ fixés :

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

- Deux astuces :
  - Algorithme EM
  - CART (programmation dynamique)

Algorithme EM

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

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

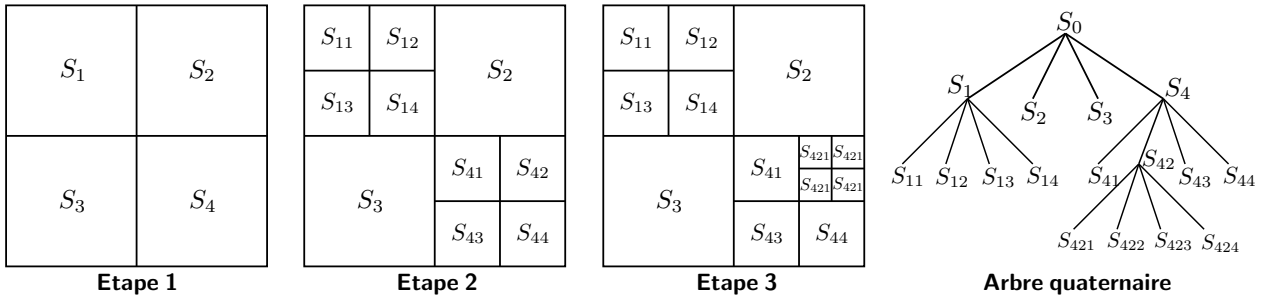
$$\leq -\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\pi_k [\mathcal{R}(x_i)] \right) + \lambda_{0,n} |\mathcal{P}| (K-1)$$

$$+ \left(-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\mathcal{N}_{\mu_k, \Sigma_k}(\mathcal{S}_i) \right) \right) + \text{Cst}^{(n)}$$

  avec égalité en $(\mathcal{P}^{(n)}, \pi^{(n)}, \mu^{(n)}, \Sigma^{(n)})$.

- Étape M : optimisation séparée en $(\mathcal{P}, \pi)$ et (μ, Σ) possible,
  - Optimisation en (μ, Σ) : formule close (et classique..).
  - Optimisation en $(\mathcal{P}, \pi)$ plus intéressante !

Étape M et CART



- Optimisation en $(\mathcal{P}, \pi)$ de

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

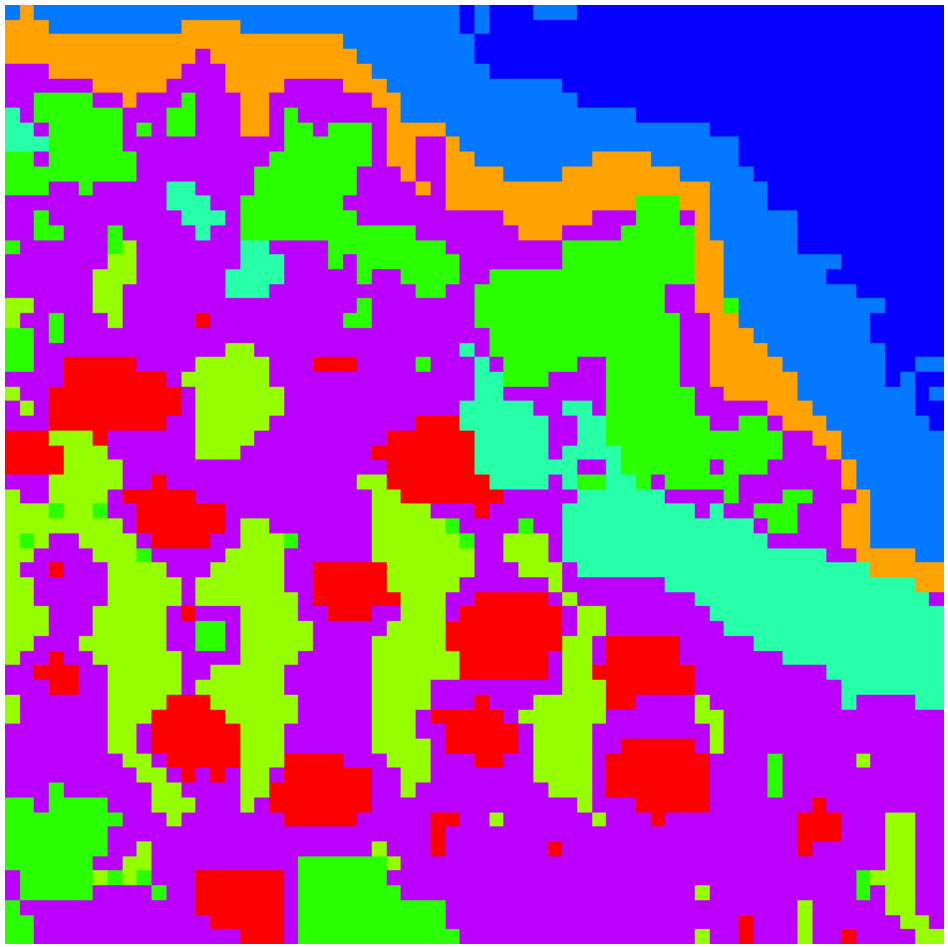
$$= -\sum_{\mathcal{R}\in\mathcal{P}} \left(\sum_{i|x_i\in\mathcal{R}} \sum_{k=1}^{K} P_k^{i,(n)} \ln\left(\pi_k[\mathcal{R}(x_i)]\right) + \lambda_{0,n}(K-1) \right)$$

- Deux propriétés clés :
  - Pour chaque $\mathcal{R}$, optimisation simple de $\pi_k[\mathcal{R}]$.
  - Structure de coût additive en $\mathcal{R}$...
- $\Rightarrow$ Algorithme d'optimisation rapide de type CART (Programmation dynamique sur la structure d'arbre).

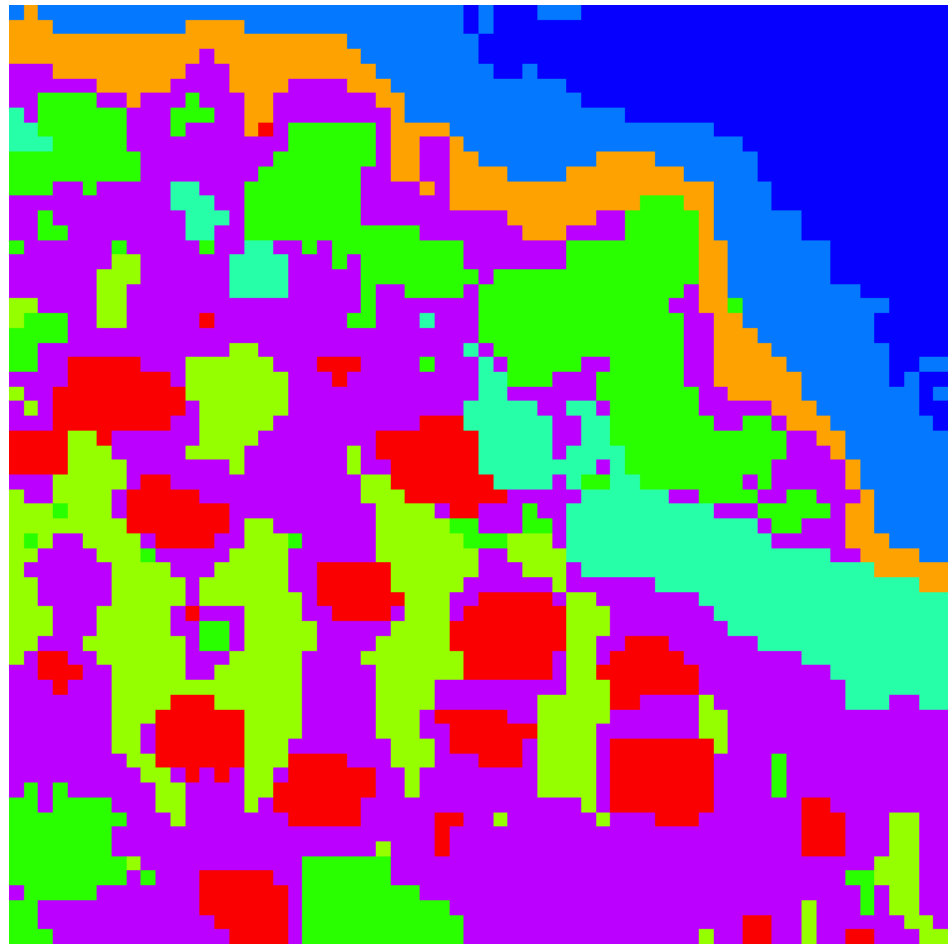
Segmentation automatique

- Résultat numérique selon la prise en compte du caractère spatial :

Sans



Avec

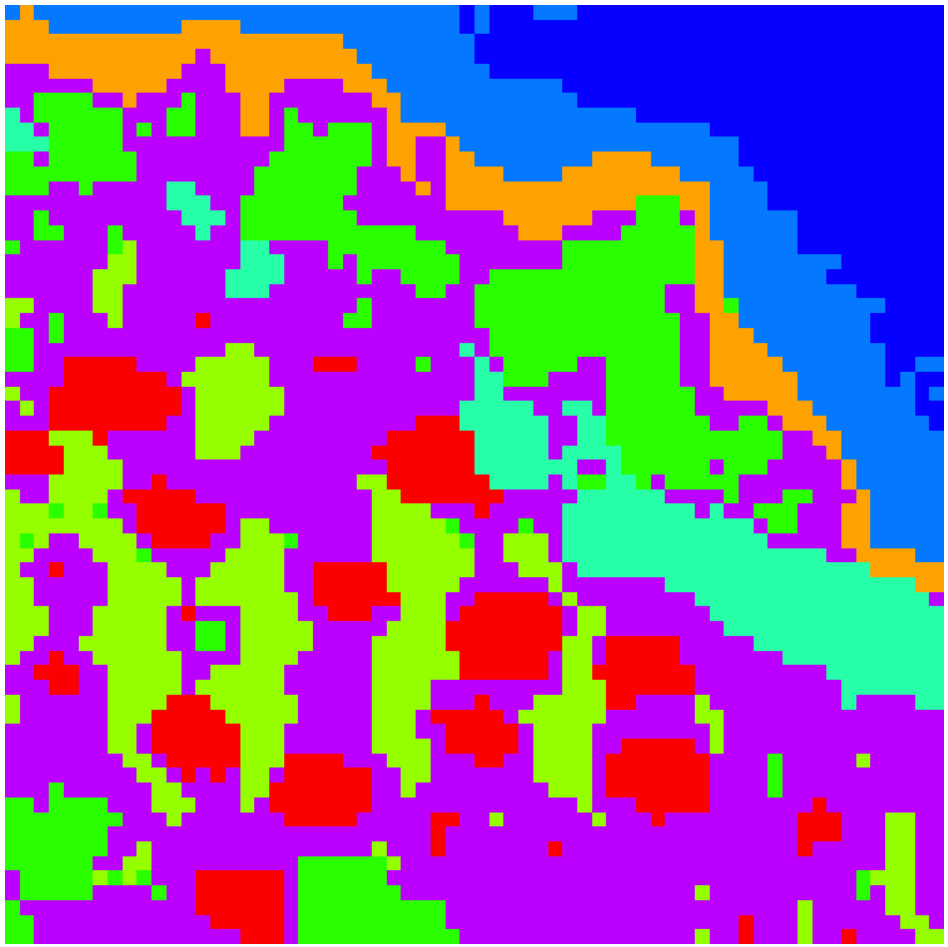


- $K = 8$, $[L_k\, D\, A_k]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente.
- Réduction de dimension par (simple) ACP...

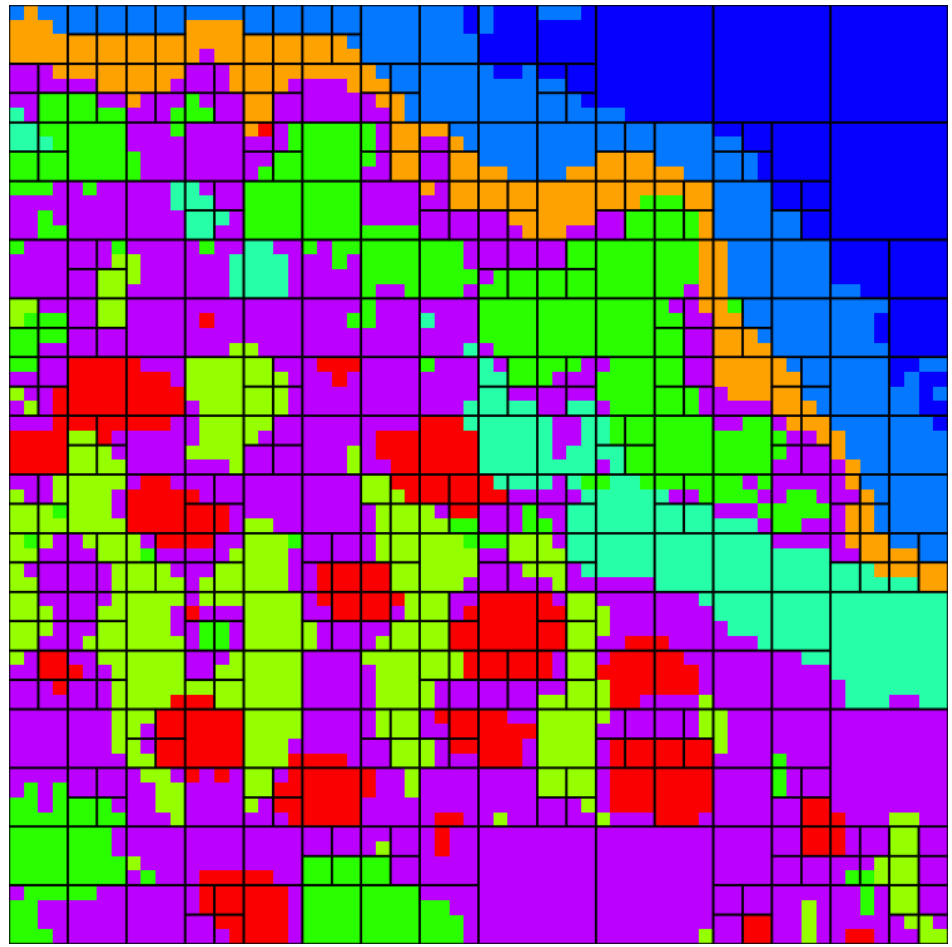
Segmentation automatique

- Résultat numérique selon la prise en compte du caractère spatial :

Sans

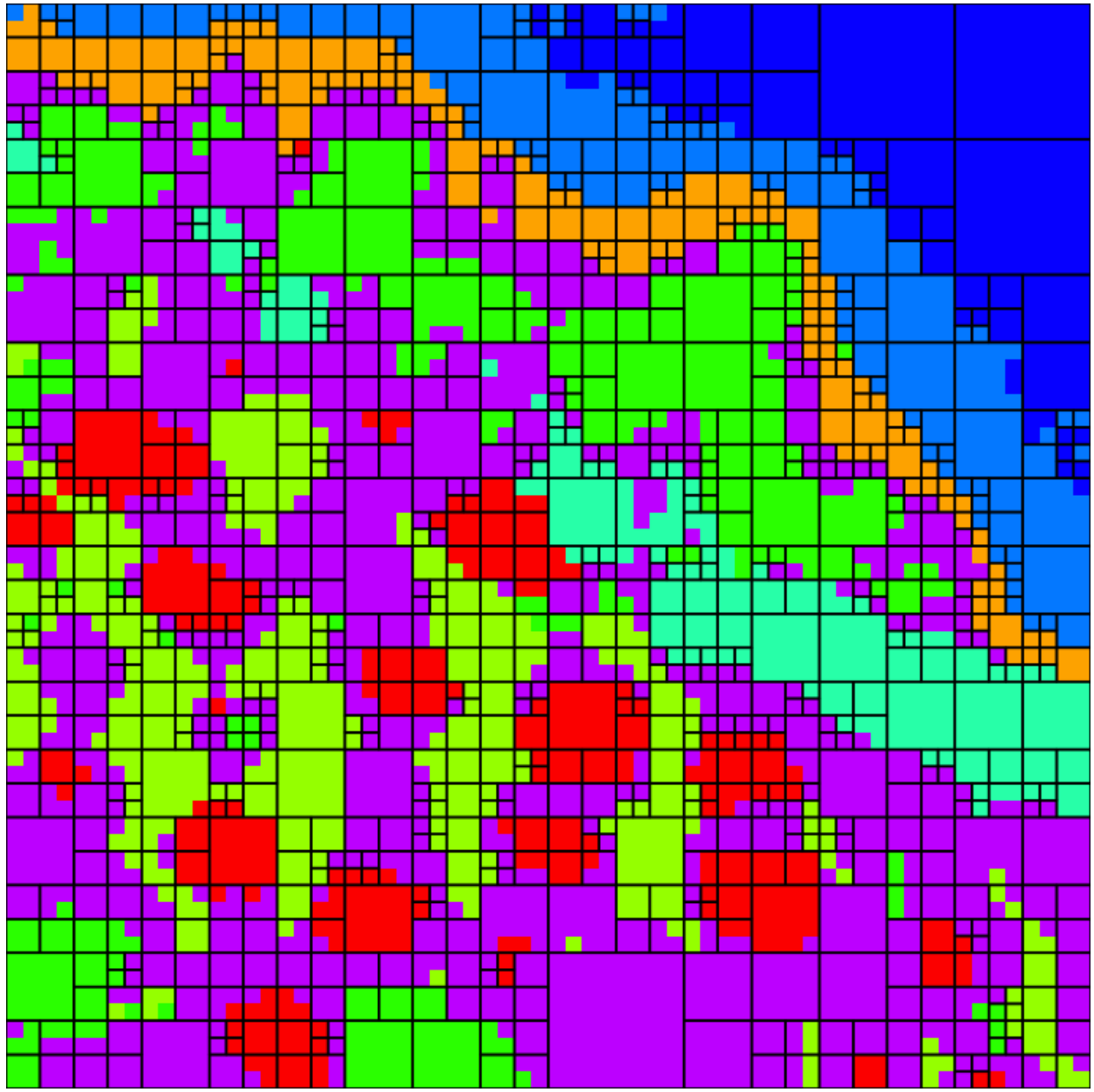


Avec

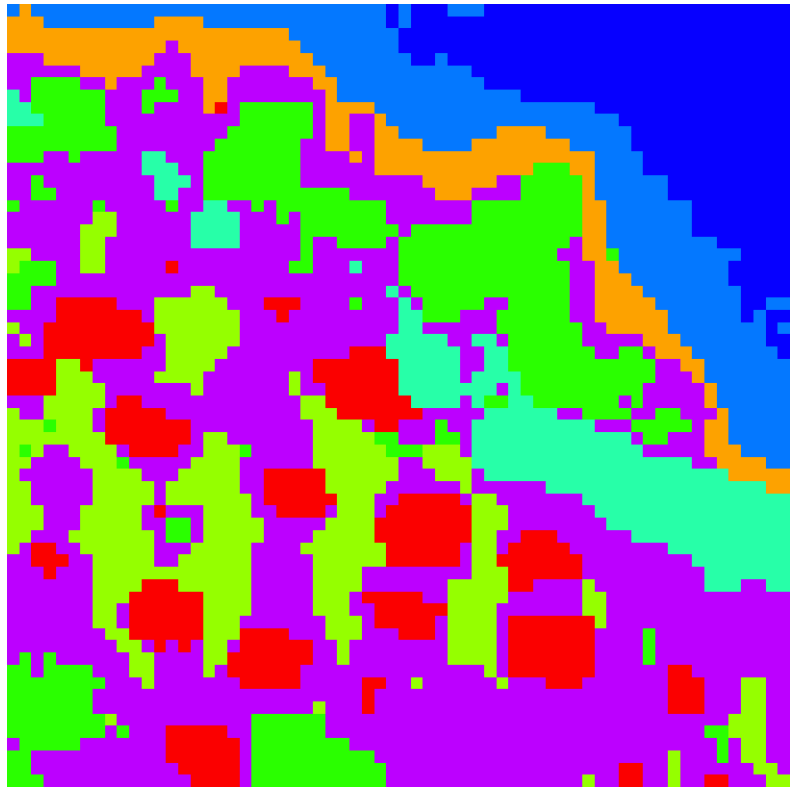


- $K = 8$, $[L_k\, D\, A_k]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente.
- Réduction de dimension par (simple) ACP...

Segmentation et classification



Le secret de Stradivarius



- Deux couches fines de vernis :
  - une première couche d'huile simple, similaire à celle des peintres, pénétrant légèrement le bois,
  - une seconde d'un mélange huile, résine de pin, pigments donnant cette couleur rouge caractéristique.
- Technique classique pour l'époque.
- Le secret de Stradivarius n'est pas dans le vernis !

Conclusion

- Cadre :
  - Problème de segmentation non supervisée.
  - Estimateur de densités conditionnelles par maximum de vraisemblance et pénalisation.
- Résultats :
  - Garantie théorique pour l'estimation de densités avec des distances *tensorisées*.
  - Applicable au problème de segmentation.
  - Algorithme efficace de minimisation.
  - Algorithme de segmentation intermédiaire ente les méthodes *spectrales* et les méthodes *spatiales*.
- Perspectives :
  - Lien entre l'estimation de densités conditionnelles et les performances de segmentation.
  - Calibration par heuristique de pente des deux problèmes.
  - Réduction de dimension adaptée à la classification non supervisée...

Conclusion

- Cadre :
  - Problème de segmentation non supervisée.
  - Estimateur de densités conditionnelles par maximum de vraisemblance et pénalisation.
- Résultats :
  - Garantie théorique pour l'estimation de densités avec des distances *tensorisées*.
  - Applicable au problème de segmentation.
  - Algorithme efficace de minimisation.
  - Algorithme de segmentation intermédiaire ente les méthodes *spectrales* et les méthodes *spatiales*.
- Perspectives :
  - Lien entre l'estimation de densités conditionnelles et les performances de segmentation.
  - Calibration par heuristique de pente des deux problèmes.
  - Réduction de dimension adaptée à la classification non supervisée...