Autour de deux problèmes d'optimisation en estimation statistique

E. Le Pennec
(SELECT - Inria Saclay / Université Paris Sud)

S. Cohen (IPANEMA - CNRS / Soleil)
K. Bertin (U. Valparaiso, Chile) et V. Rivoirard (U. Paris Dauphine)

Avignon
17 novembre 2011

informatiques mathématiques
Inria

Estimation statistique et optimisation

- Observations : $Z_1, \ldots, Z_n$.
- Modèlisation statistique vrai : modèle aléatoire $\mathcal{M}_0$ (partiellement) inconnu ayant générer les données.
- Modèle statistique utilisé : modèle aléatoire $\mathcal{M}_\theta$ dépendant d'un *paramètre* θ permettant de générer des données $Z'_1, \ldots, Z'_n$.
- Estimation statistique : estimer un paramètre $\widehat{\theta}$ à partir des données de sorte que $\mathcal{M}_{\widehat{\theta}}$ soit proche du *vrai* modèle $\mathcal{M}_0$.
- Estimation et optimisation : bien souvent

$$\widehat{\theta} = \operatorname*{argmin}_{\theta} F(\theta, Z_1, \ldots, Z_n)$$

Plan

1. Estimation et optimisation
2. Densité, maximum de vraisemblance, mélange de Gaussienne et algorithme EM
3. Densité conditionnelle, maximum de vraisemblance pénalisé, mélange de Gaussienne spatialisé, algorithme EM et programmation dynamique
4. Densité, moindre carré, approche dictionnaire et pénalisation ℓ_1

Plan

Régression linéaire

- Observation $Z_i = (X_i, Y_i)$.
- Modèlisation statistique vrai $\mathcal{M}_0$: $X_i = s_0(Y_i) + \epsilon W_i$ avec $W_i \sim \mathcal{N}(0,1))$ i.i.d.
- Modèle statistique utilisé $\mathcal{M}_\theta$: $X_i = \theta_1 Y_i + \theta_2 + \epsilon W_i$ avec $\theta \in \mathbb{R}^2$.
- Méthode des moindres carrés :

$$(\widehat{\theta}_1, \widehat{\theta}_2) = \underset{(\theta_1,\theta_2)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (X_i - (\theta_1 Y_i + \theta_2))^2$$

- Optimisation sans soucis...
- Modèle optimal $\mathcal{M}_{\theta_\star}$: $\theta_\star = \operatorname{argmin}_\theta \mathbb{E}\left[(X_1 - (\theta_1 Y_1 + \theta_2))^2\right]$
- Résultat de consistance ($\widehat{\theta}$ tend vers $\theta_\star$ quand n tend vers $+\infty$), vitesse ($\mathbb{E}[\|\theta_\star - \widehat{\theta}\|] = O(1/\sqrt{n})$),...

Estimation de densité

- Observations venant du modèle vrai $\mathcal{M}_0$: $X_1, \ldots, X_n$ i.i.d. de loi de densité $s_0(x)$ par rapport à la mesure de Lebesgue.
- Deux propriétés importantes :
  - $\forall\ g$ mesurable, $\mathbb{E}\left[g(X_1)\right] = \int g(x)s_0(x)dx$ (déf. de densité).
  - $\forall\ g$ intégrable, $\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathbb{E}\left[g(X_1)\right]$ (loi des grands nombres).
- Modèle statistique utilisé $\mathcal{M}_{\mathcal{F}}$: X_i i.i.d. de densite $s(x)$ avec $s \in \mathcal{F}$.
- Comment estimer s_0 à partir de $X_1, \ldots, X_n$?

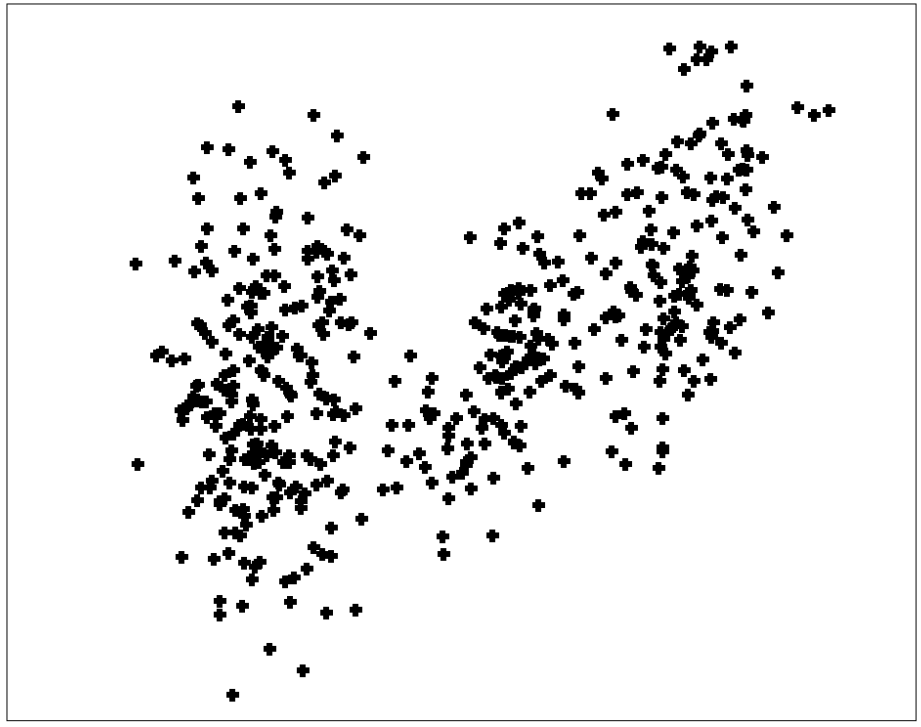
Méthode de contraste

- Construction de critère (convexe) $\mathcal{L}(s, s_0)$ avec min. unique en s_0 :
  - Moindre carré : $\|s - s_0\|^2$
  - Maximum de vraisemblance : $KL(s_0, s) = \mathbb{E}\left[\ln \frac{s_0(x)}{s(x)}\right]$
- Utilisation d'une version *empirique* $L(s, s_0)$ telle que $\mathbb{E}L = \mathcal{L}$:
  - Moindre carré : $-\frac{2}{n}\sum_{i=1}^{n} s(X_i) + \|s\|^2 \left(+\|s_0\|^2\right)$
  - Maximum de vraisemblance : $\frac{1}{n}\sum_{i=1}^{n} -\ln s(X_i) \left(+\frac{1}{n}\sum_{i=1}^{n} \ln s_0(X_i)\right)$
- Estimation par minimisation de la version empirique
  - Moindre carré : $\widehat{s} = \underset{s\in\mathcal{F}}{\operatorname{argmin}} -\frac{2}{n}\sum_{i=1}^{n} s(X_i) + \|s\|^2$
  - Maximum de vraisemblance $\widehat{s} = \underset{s\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} -\ln s(X_i)$

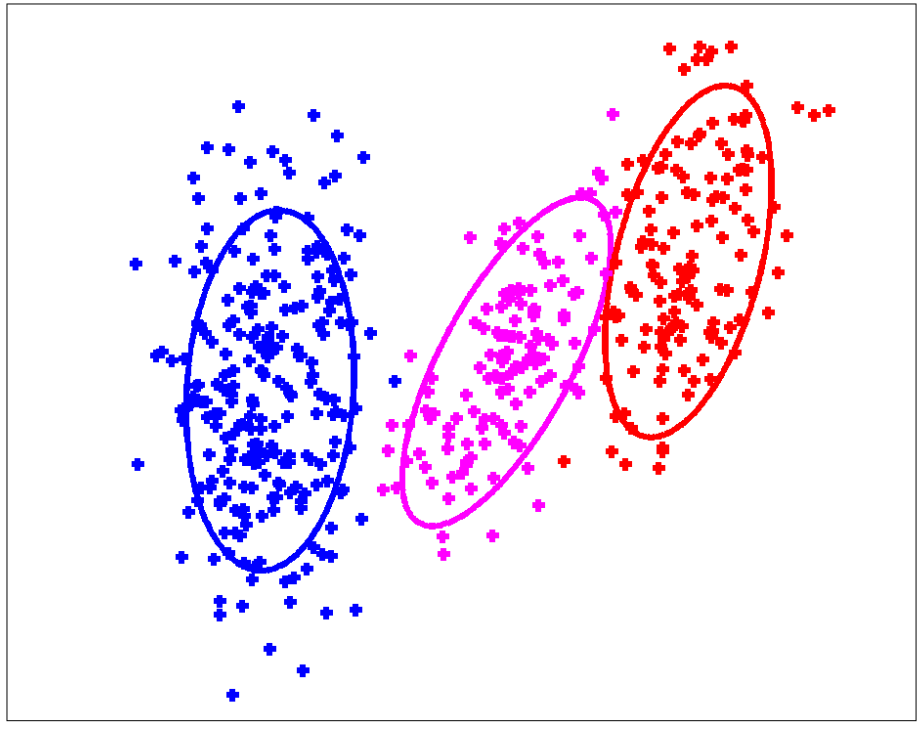
Plan

Estimation de densité et mélange de Gaussienne

Estimation de densité et mélange de Gaussienne

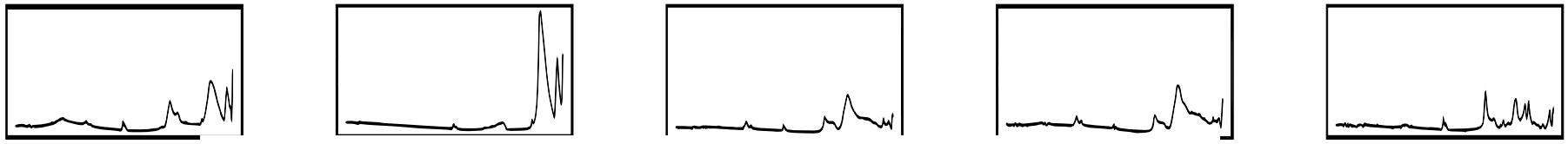


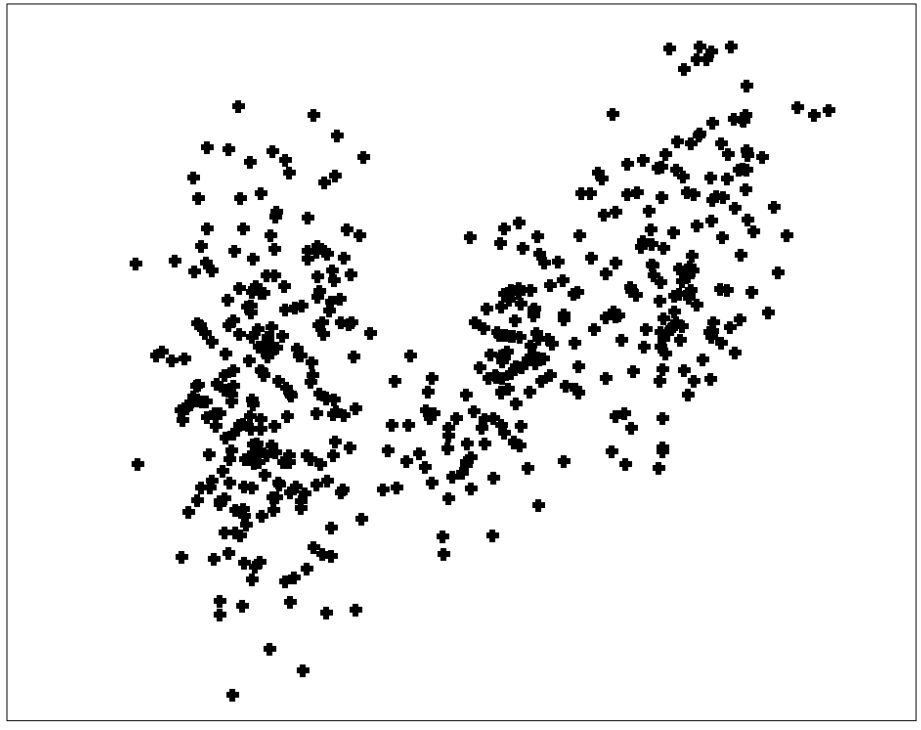
Estimation de densité et mélange de Gaussienne



- Mélange de Gaussiennes

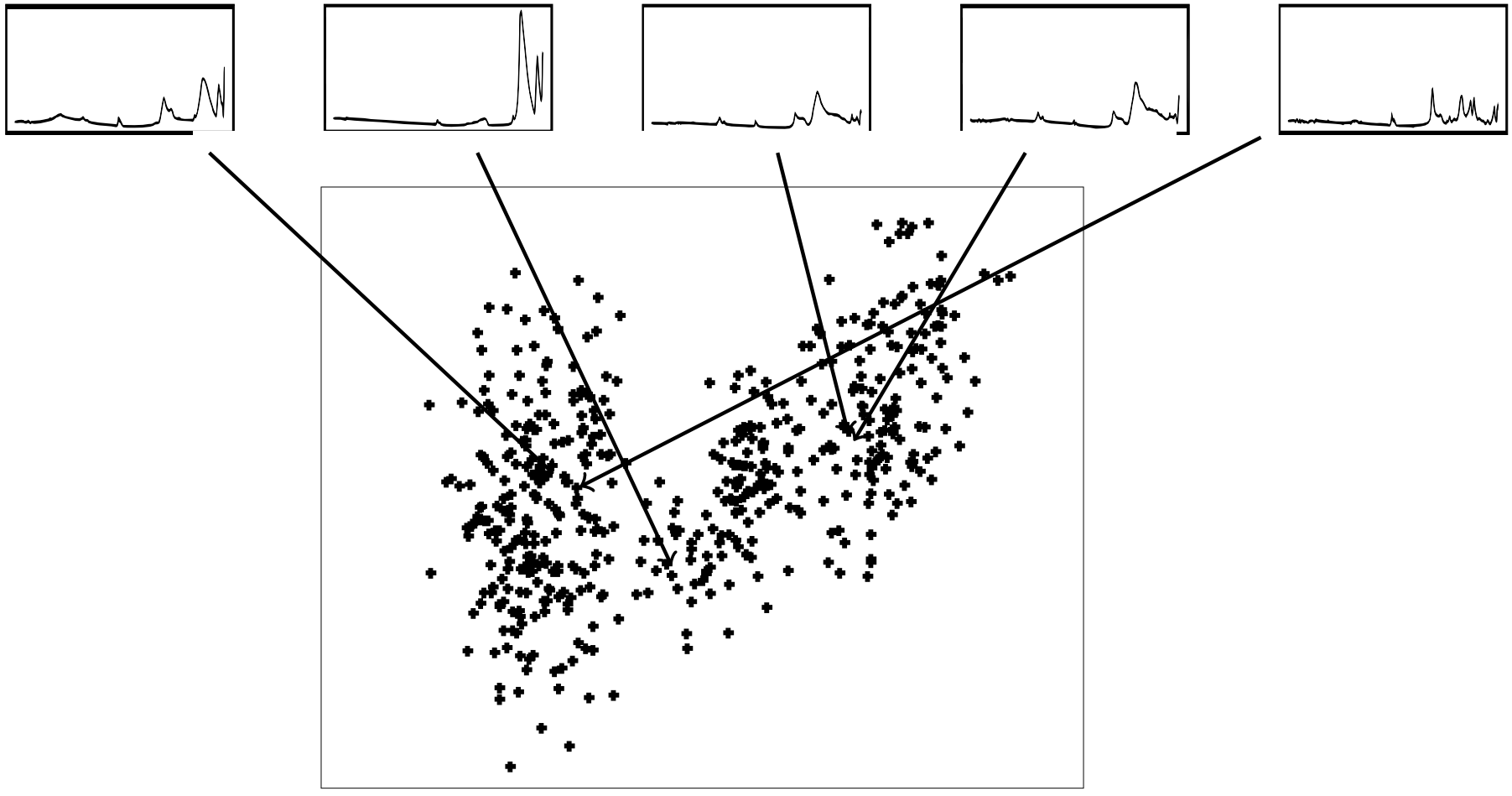
Estimation de densité et mélange de Gaussienne





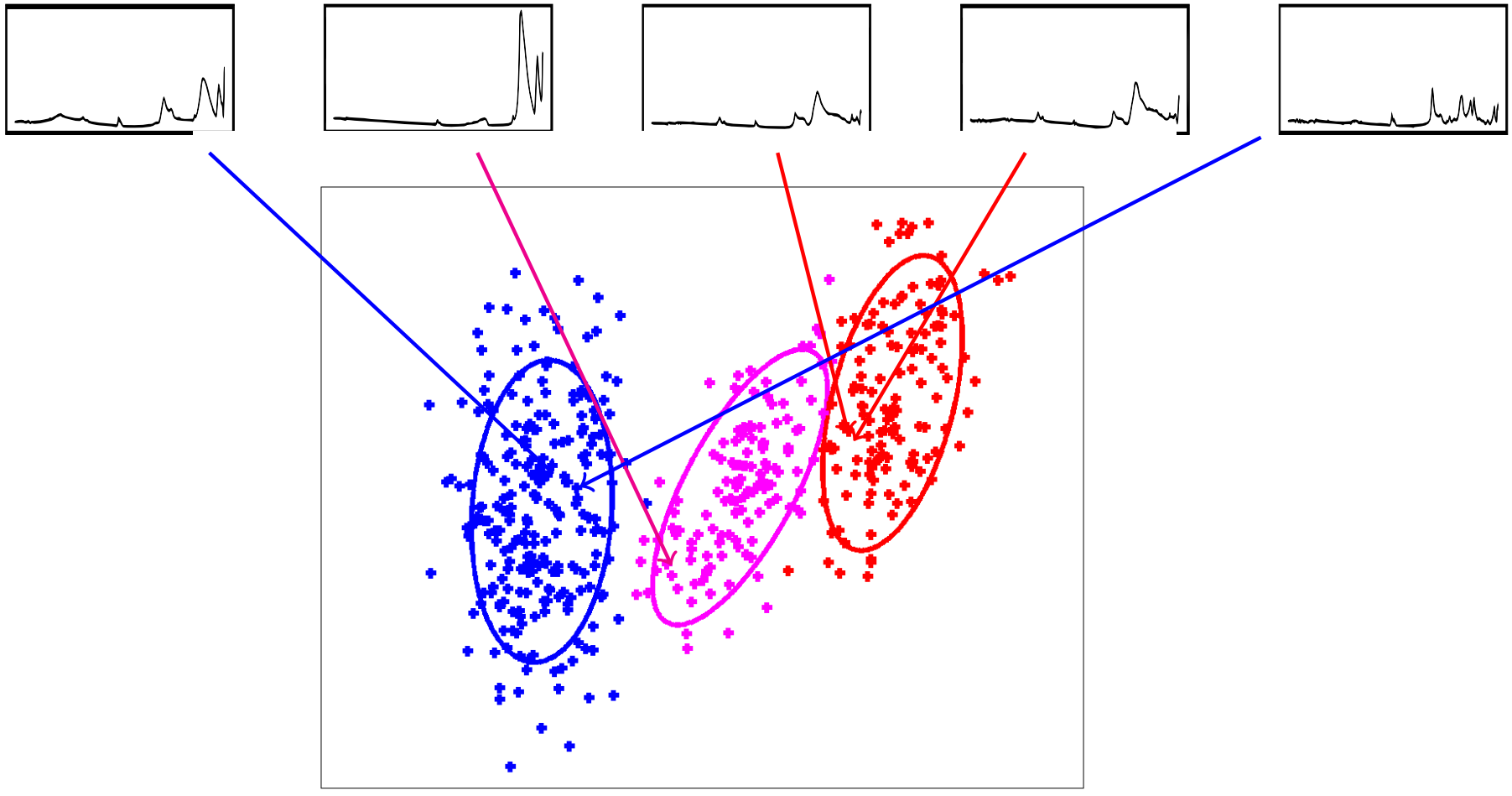
- Mélange de Gaussiennes
- Modèle très structuré lié à une idée de classification.

Estimation de densité et mélange de Gaussienne



- Mélange de Gaussiennes
- Modèle très structuré lié à une idée de classification.

Estimation de densité et mélange de Gaussienne



- Mélange de Gaussiennes
- Modèle très structuré lié à une idée de classification.

Modélisation par un mélange de gaussiennes

- Modèle statistique vrai $\mathcal{M}_0$: $X_1, \ldots, X_n$ i.i.d. de loi de densité $s_0(x)$ par rapport à la mesure de Lebesgue.
- Modéle statistique utilisé "$\mathcal{M}_\theta$" :
  - existence de K classes,
  - 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 !)
- Heuristique : densité s_0 de X_i proche de

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

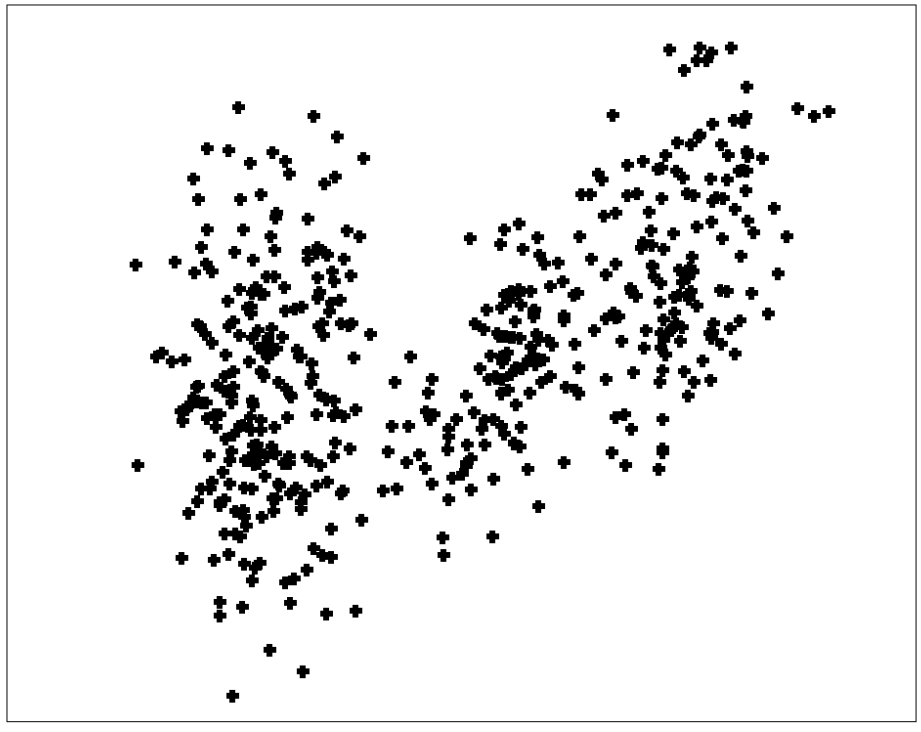
- Objectif : estimer les paramètres K, π_k, μ_k, Σ_k à partir des données.
- Pourquoi ? : possibilité d'assigner ensuite une classe à une observation par maximum de vraisemblance

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

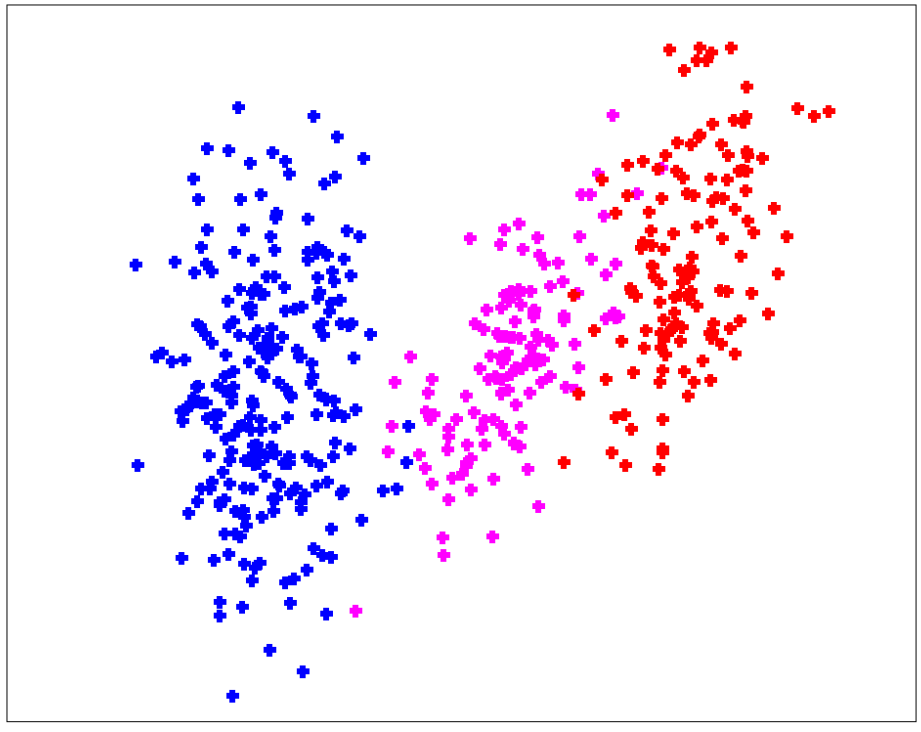
- Résultat en terme d'estimation de densité...

Modélisation “stochastique”

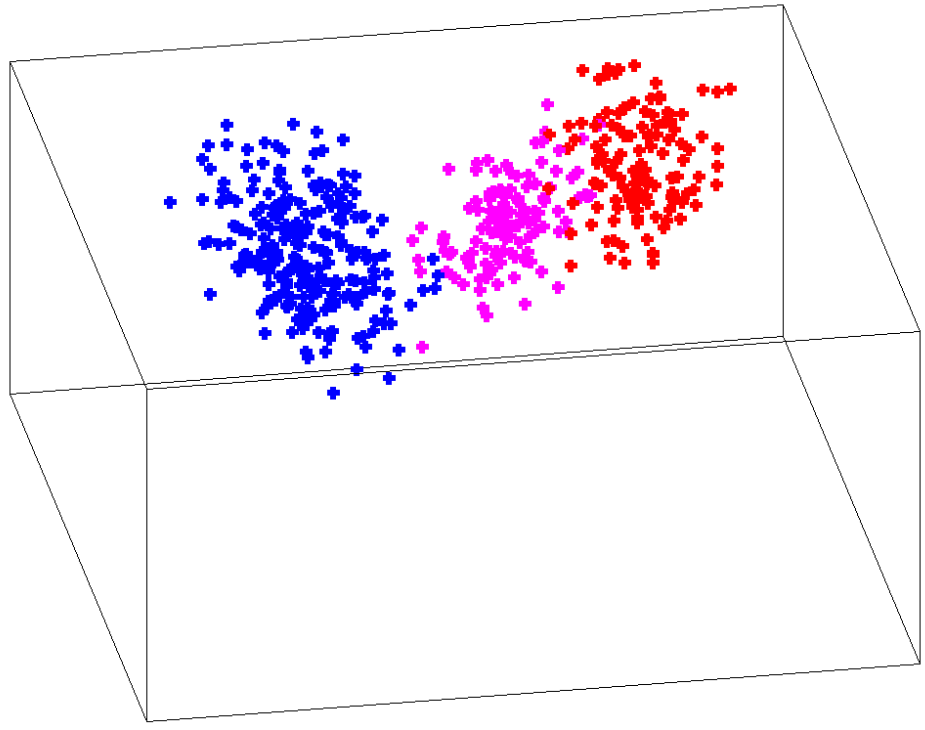
Modélisation “stochastique”



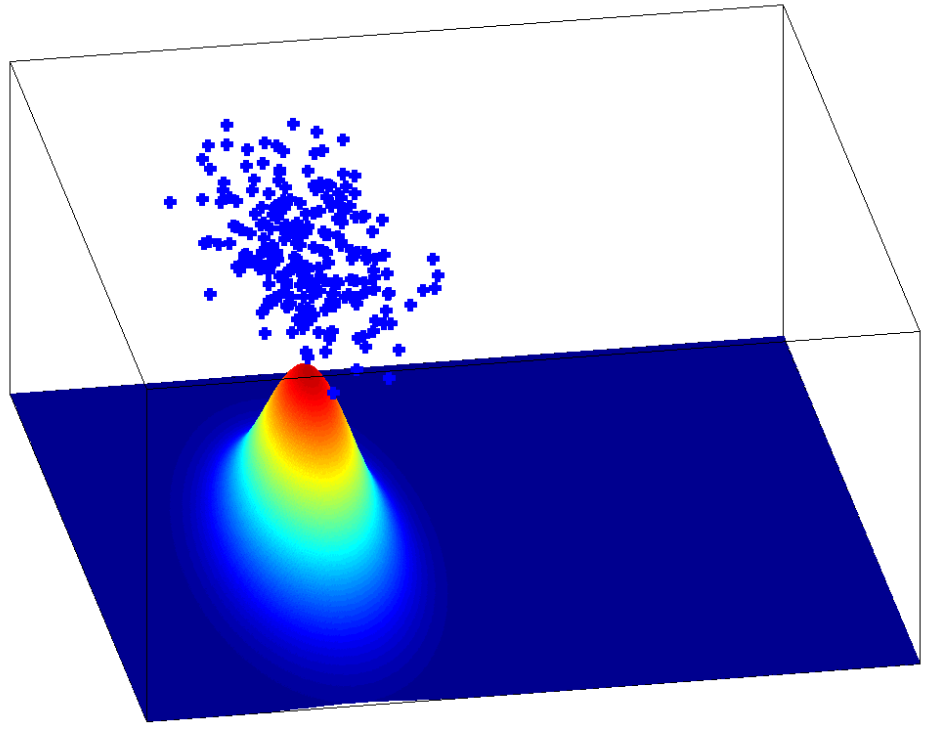
Modélisation “stochastique”



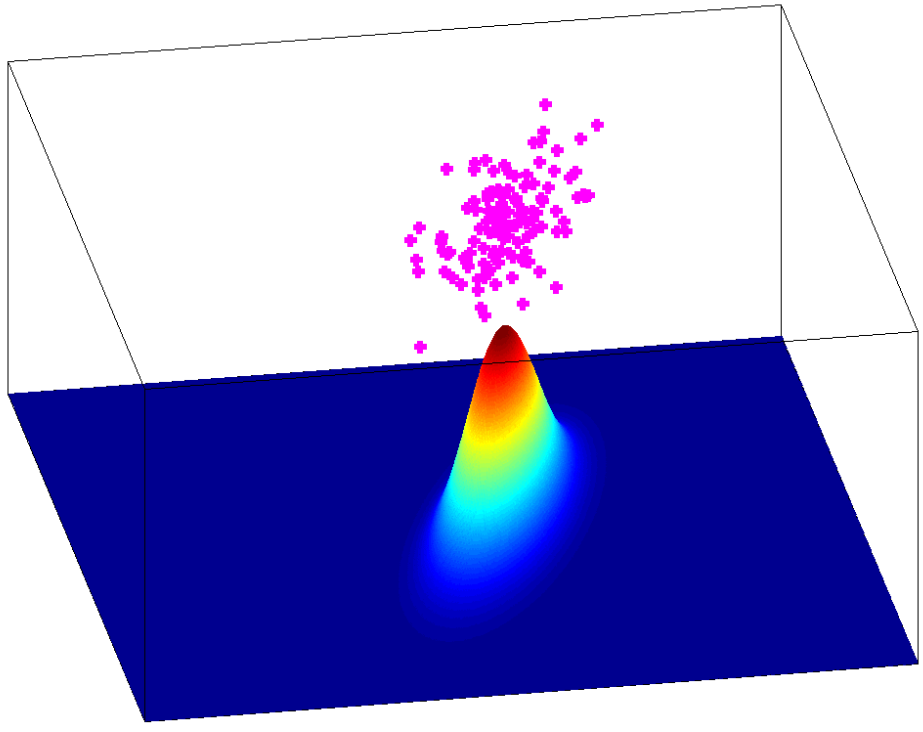
Modélisation “stochastique”



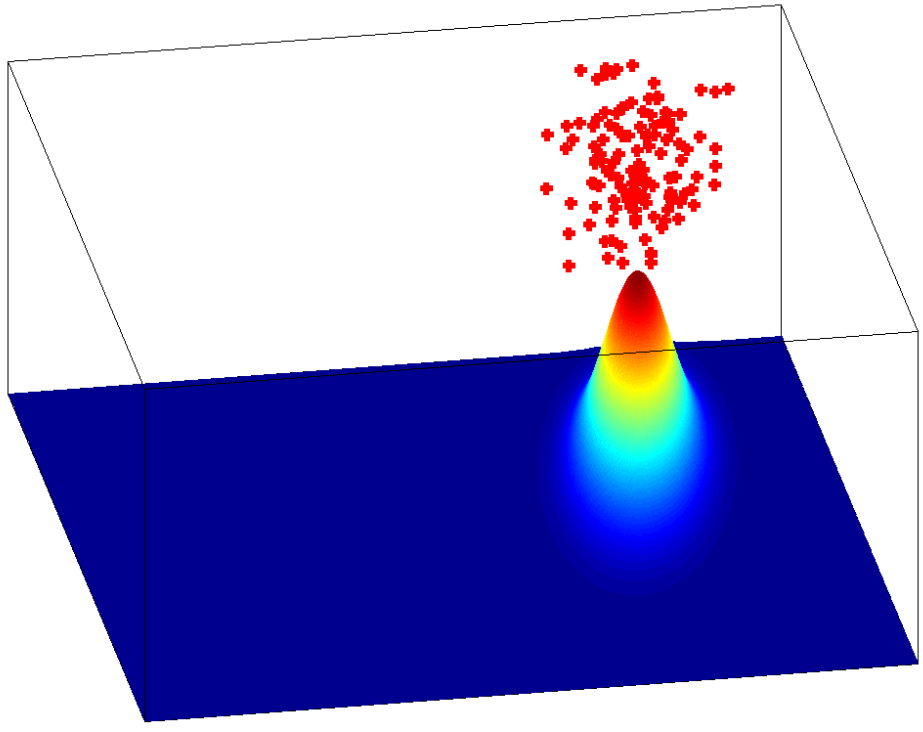
Modélisation “stochastique”



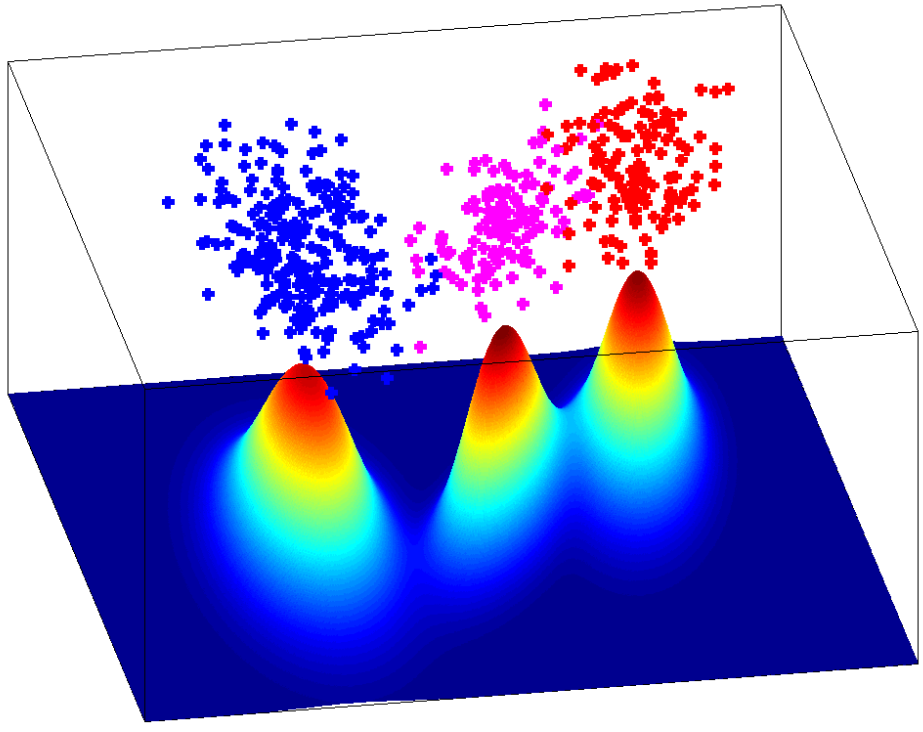
Modélisation “stochastique”



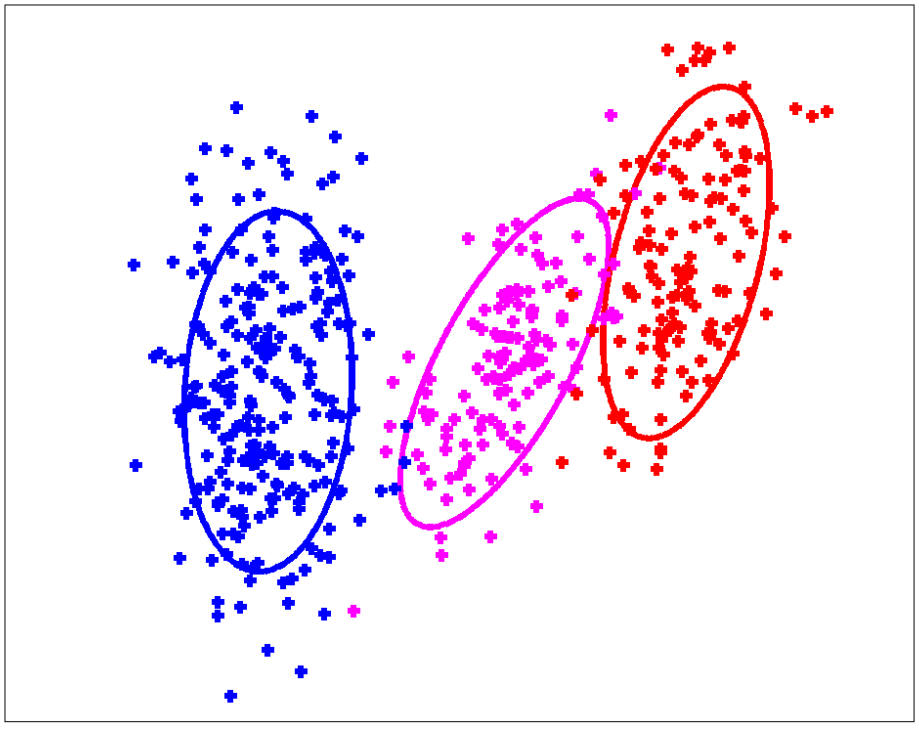
Modélisation “stochastique”



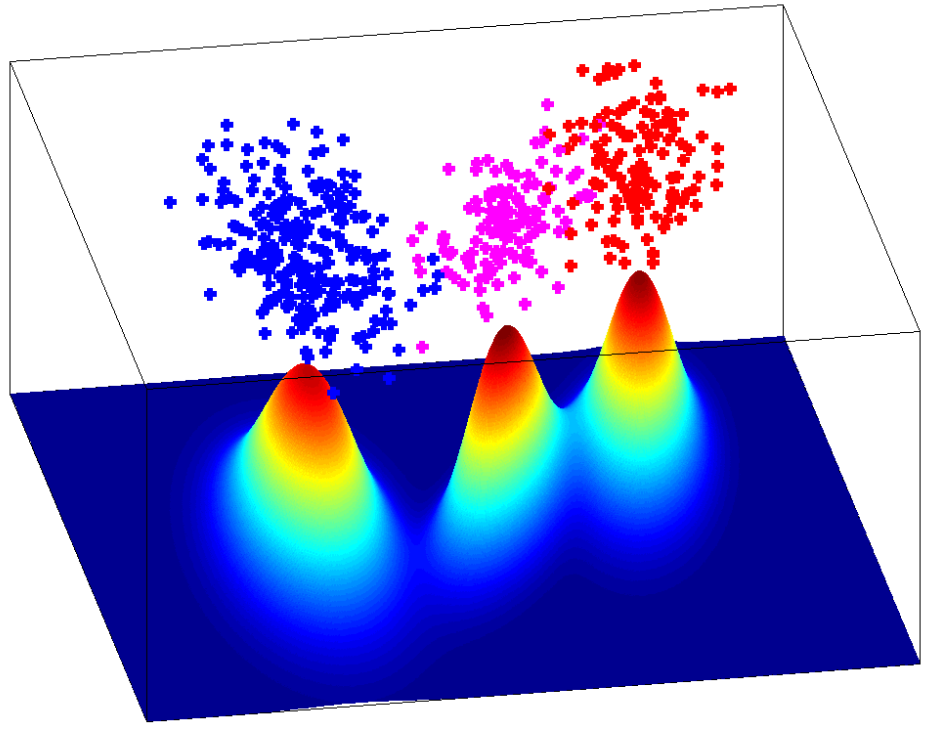
Modélisation “stochastique”



Modélisation “stochastique”



Modélisation “stochastique”

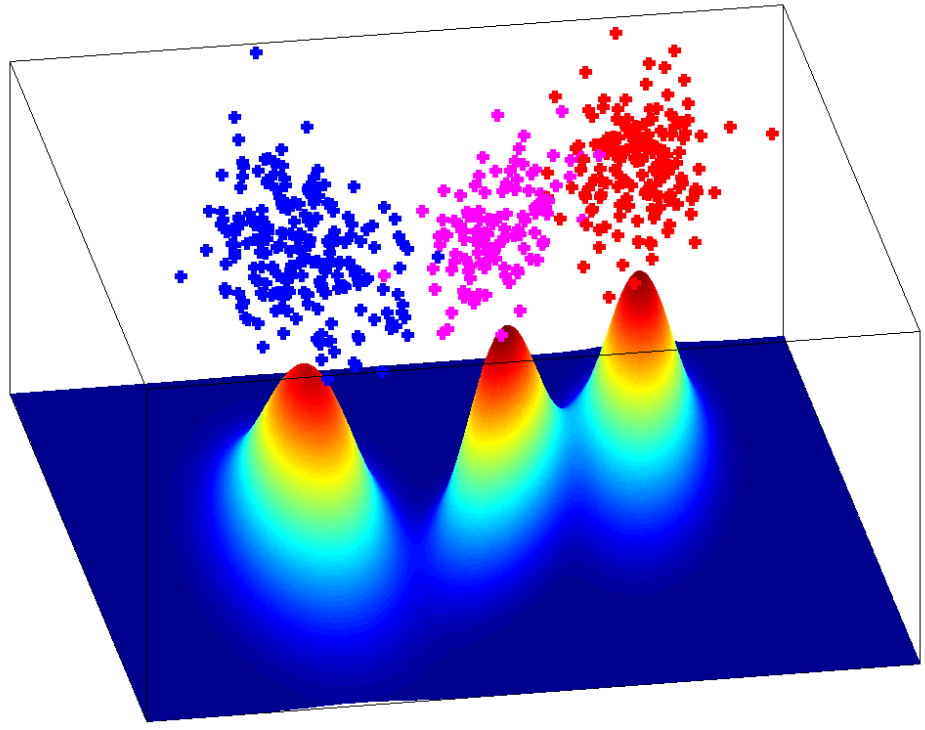


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”

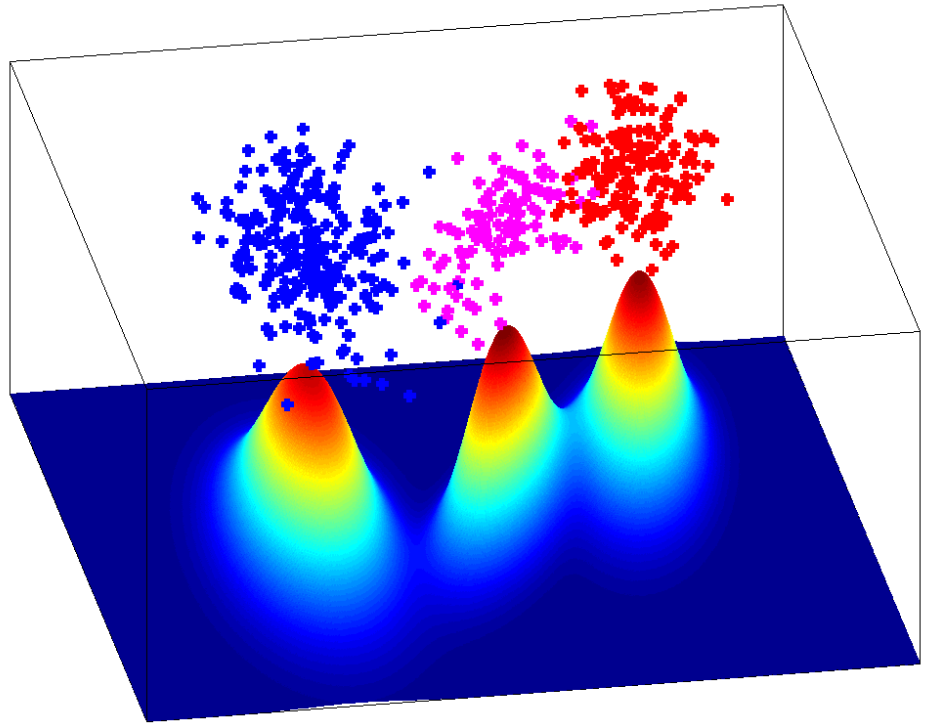


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”

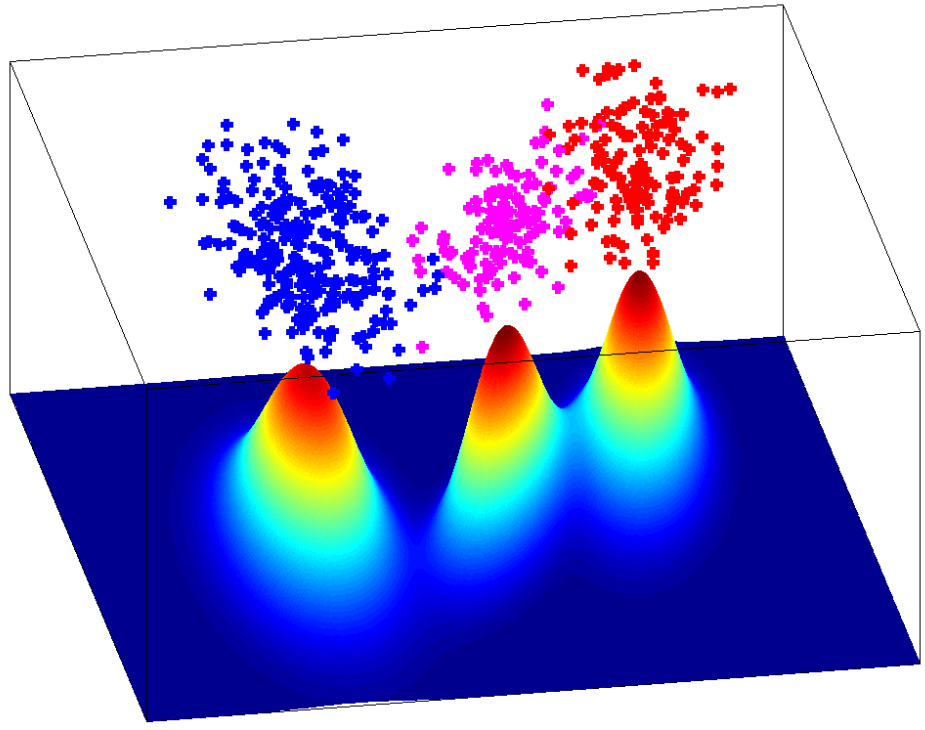


- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

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

Modélisation “stochastique”



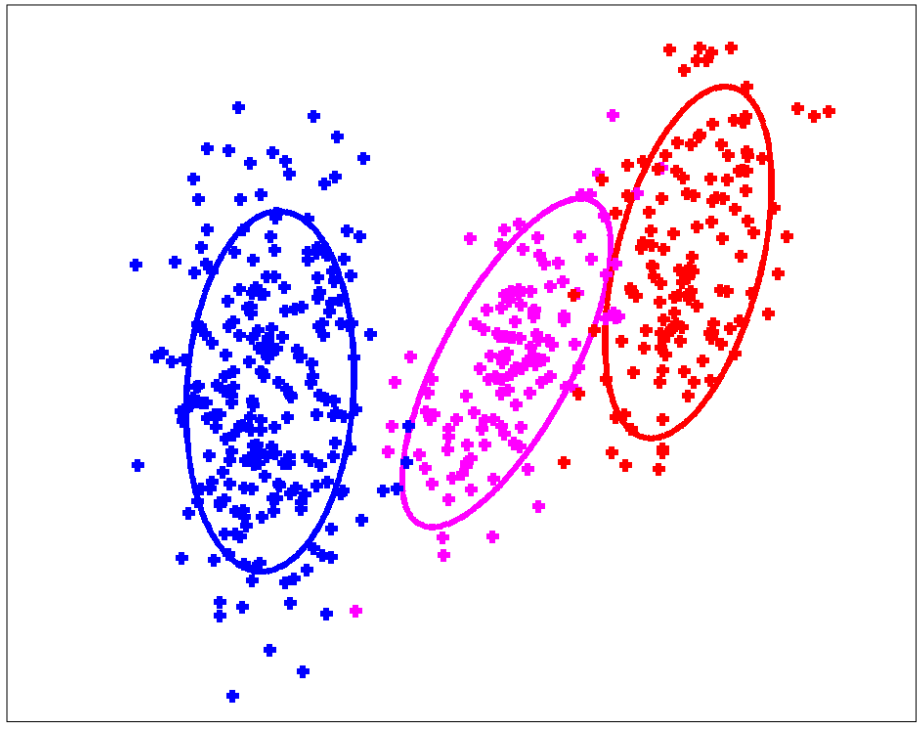
- Modèle : mélange de gaussiennes à K classes.
- Densité du mélange :

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

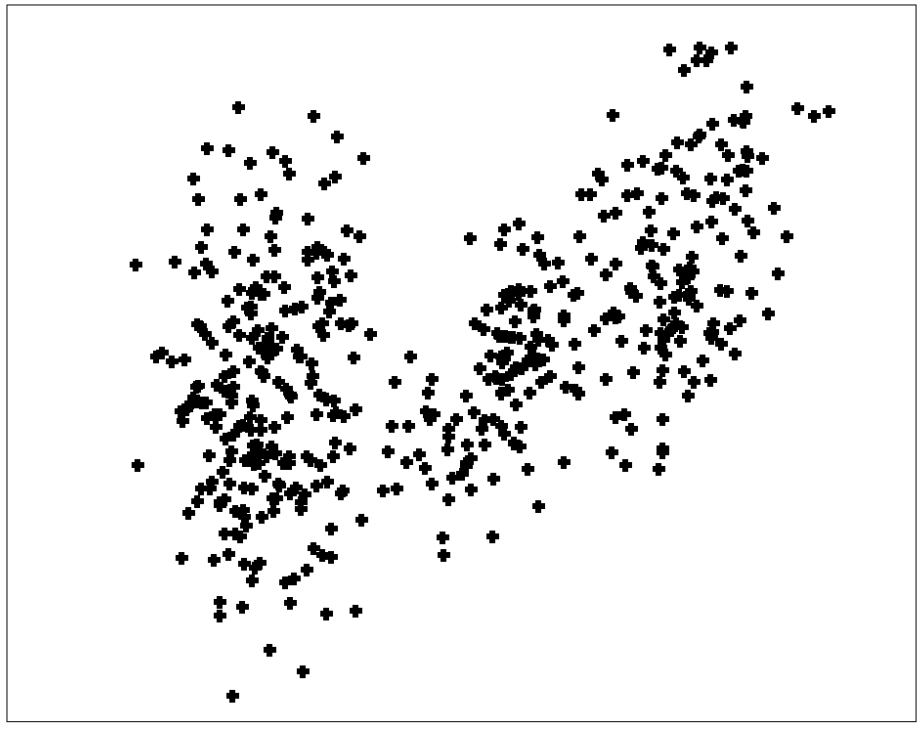
$$= \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k,\Sigma_k}(X)$$

Estimation “statistique”

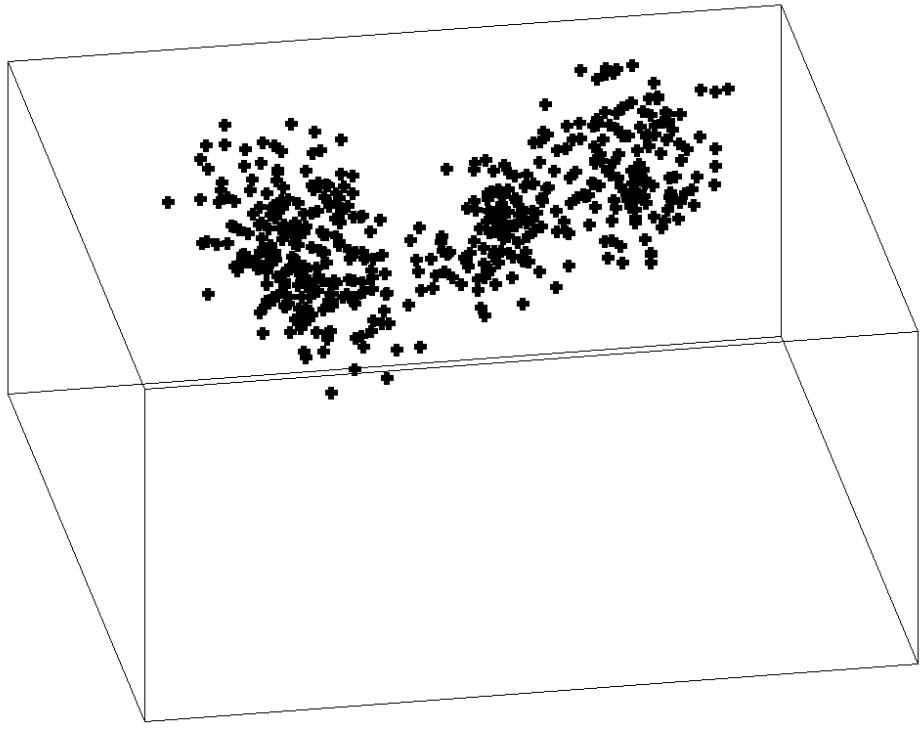
Estimation “statistique”



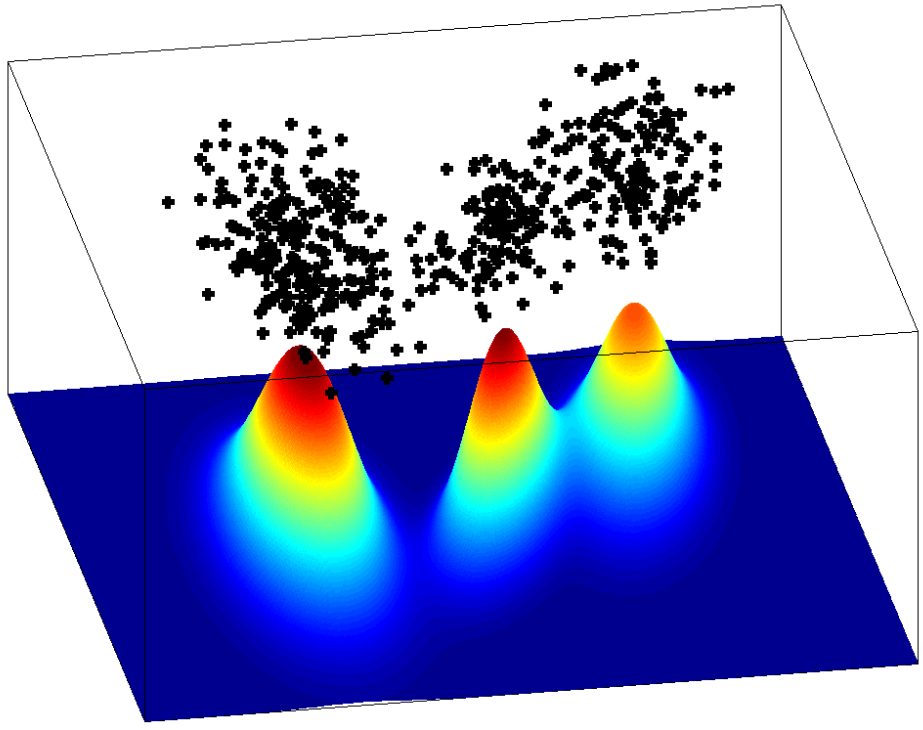
Estimation “statistique”



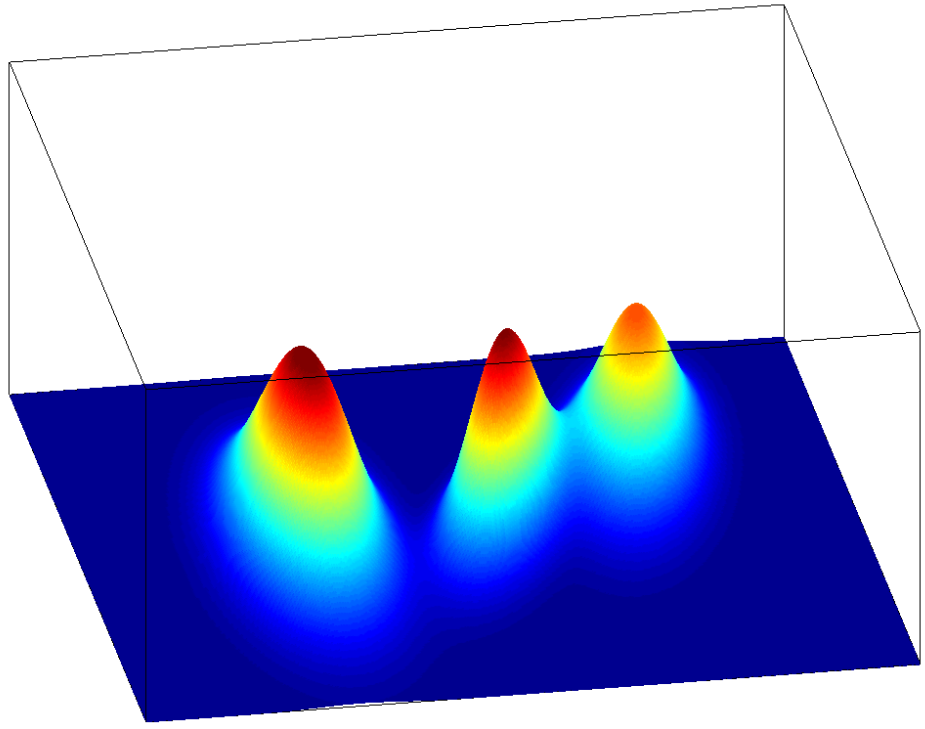
Estimation “statistique”



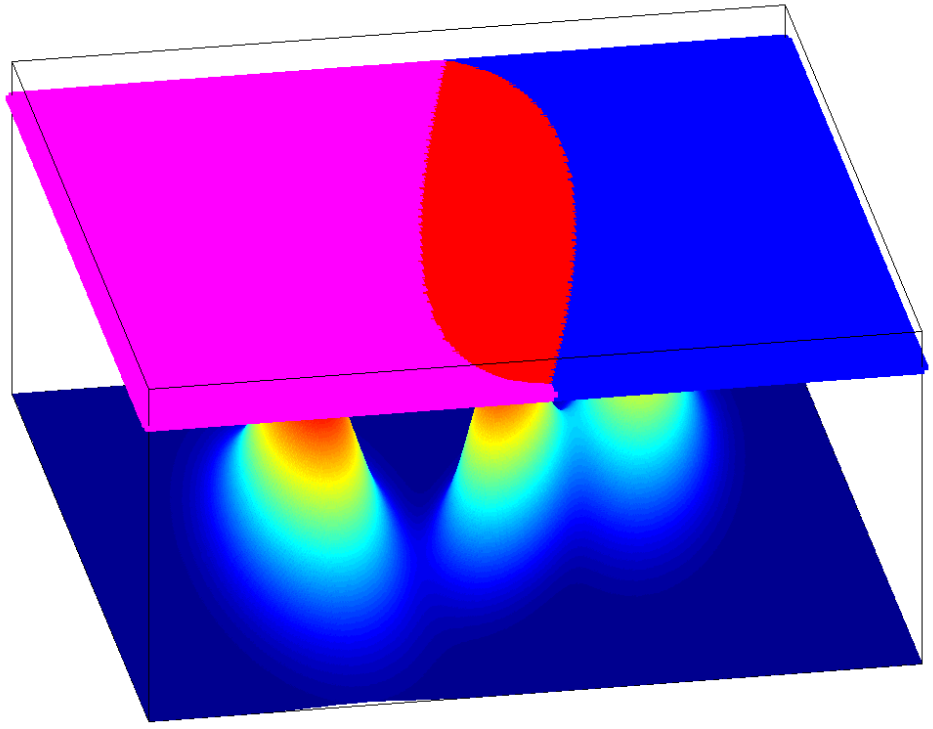
Estimation “statistique”



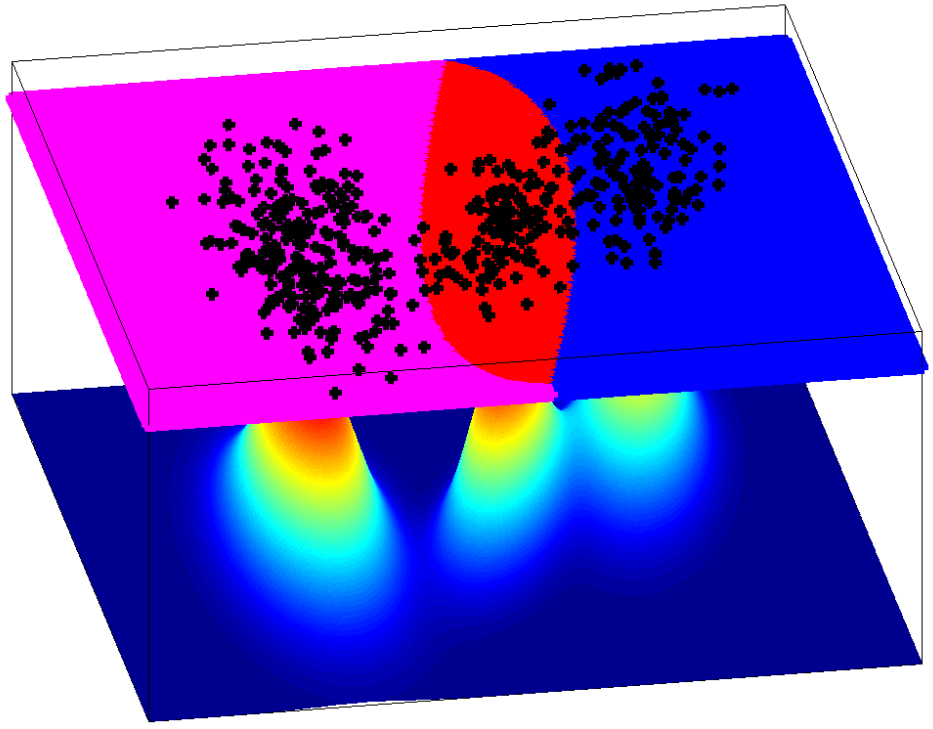
Estimation “statistique”



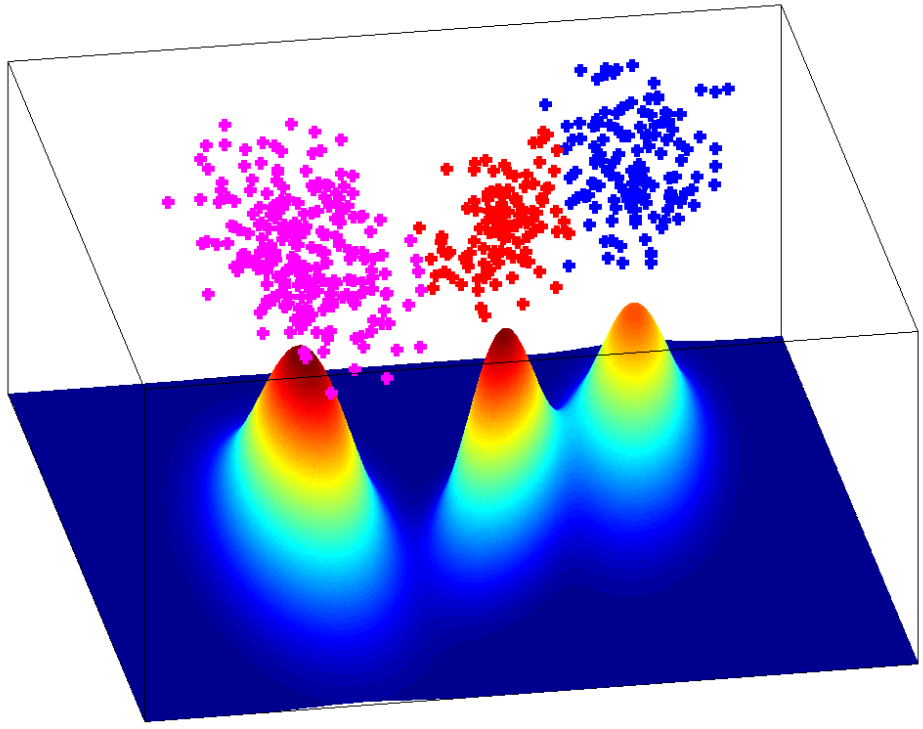
Estimation “statistique”



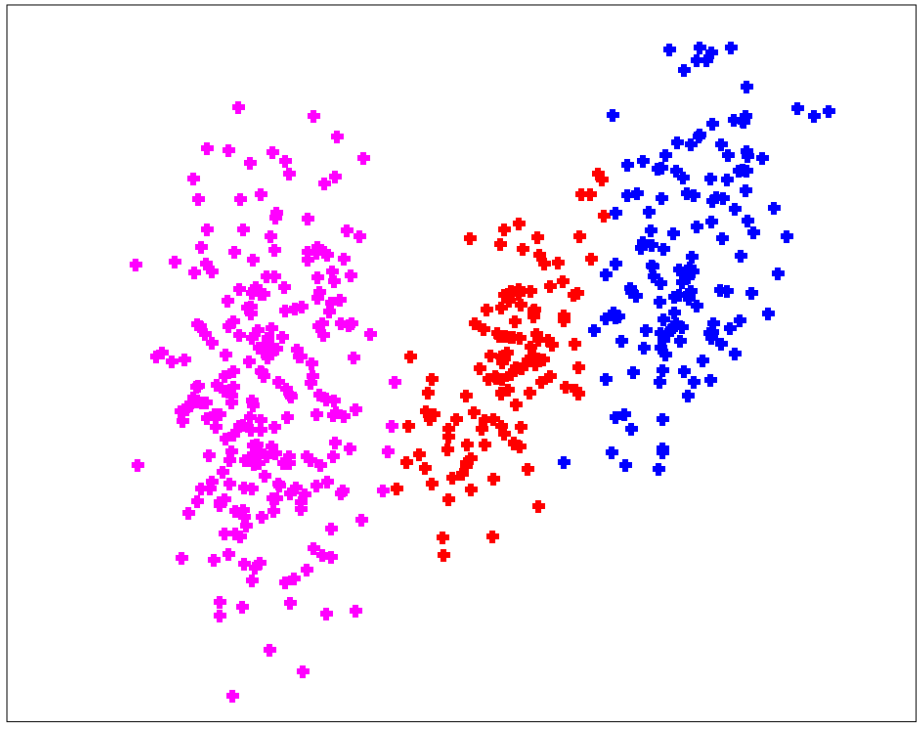
Estimation “statistique”



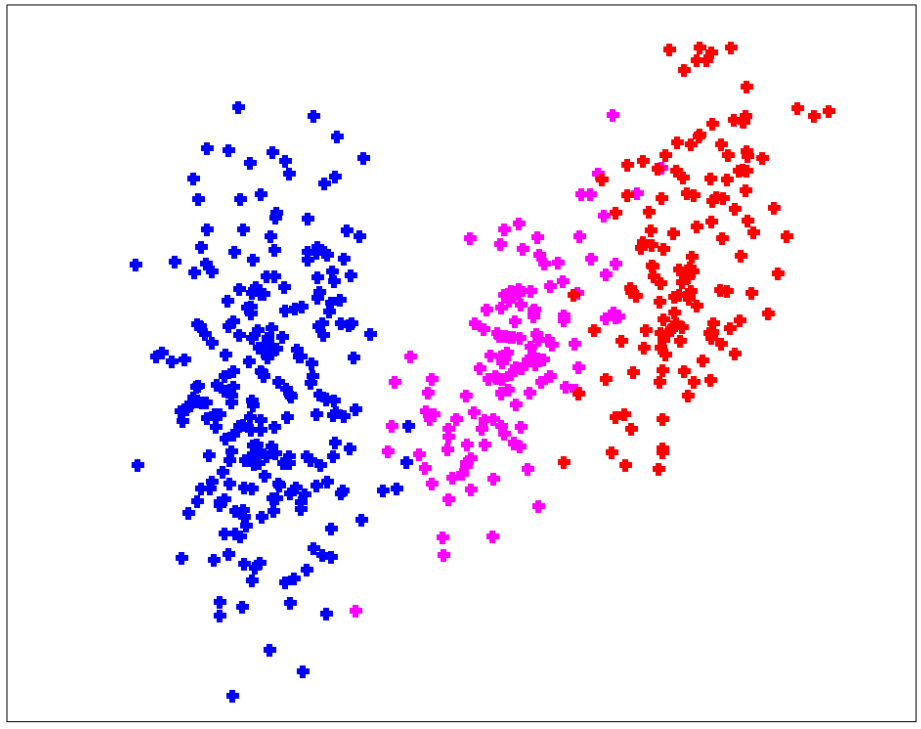
Estimation “statistique”



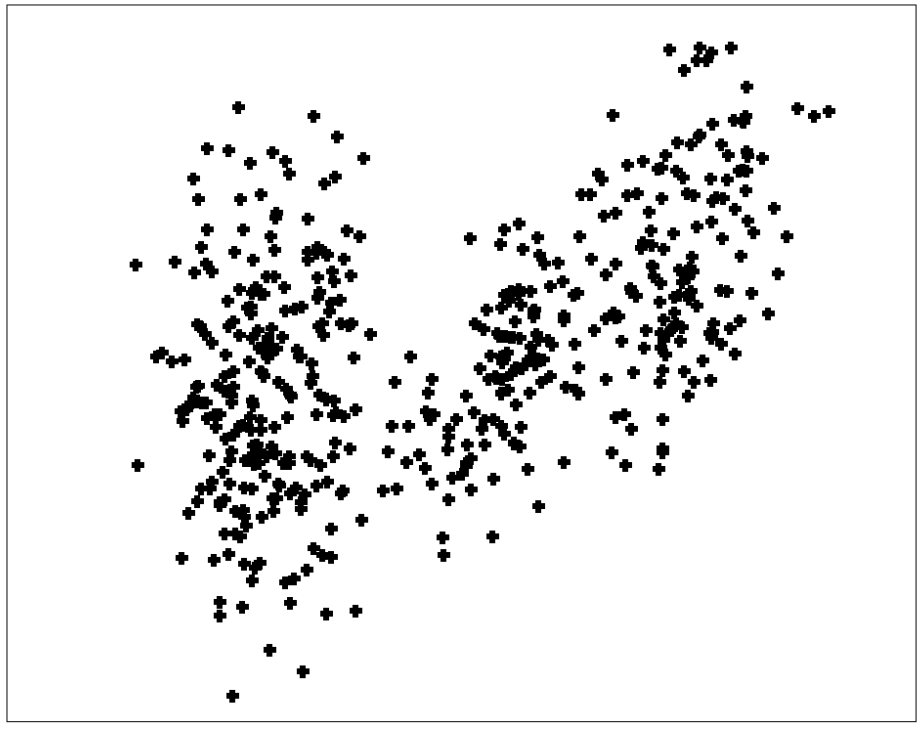
Estimation “statistique”



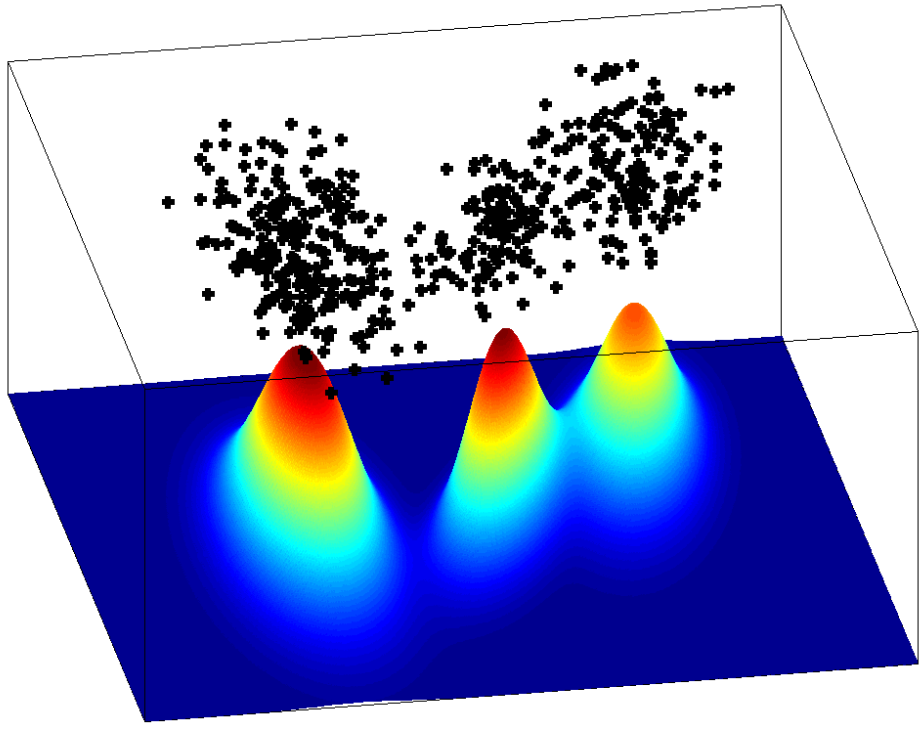
Estimation “statistique”



Estimation “statistique”



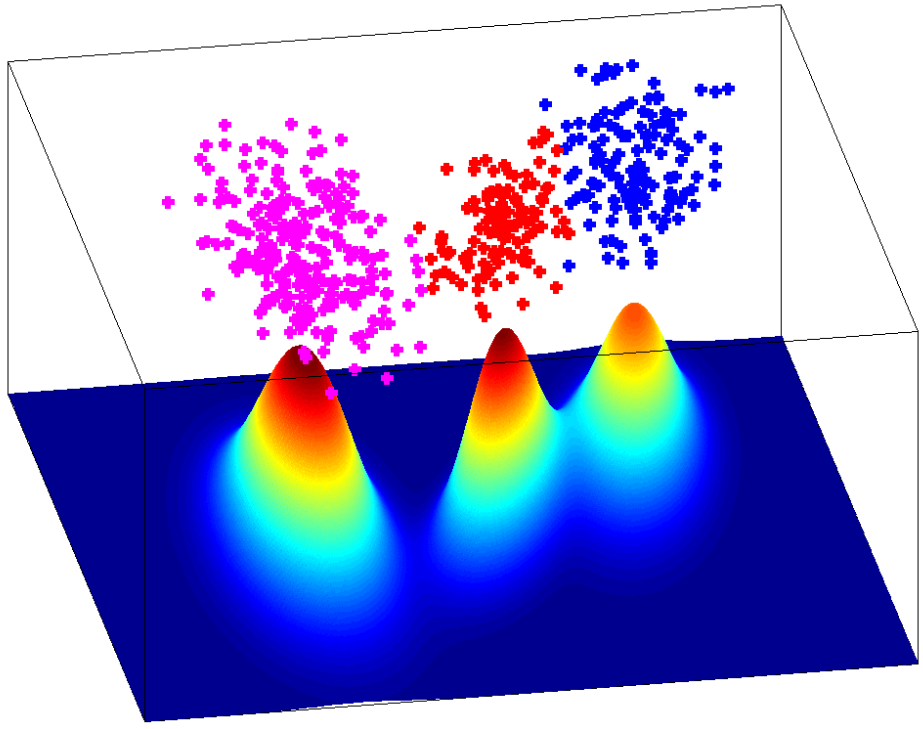
Estimation “statistique”



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

$$(\widehat{\pi_k}, \widehat{\mu_k}, \widehat{\Sigma_k}) = \operatorname{argmax} \sum_{i=1}^{n} \log s_{K,(\pi_k,\mu_k,\Sigma_k)}(X_i)$$

Estimation “statistique”



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

$$(\widehat{\pi_k}, \widehat{\mu_k}, \widehat{\Sigma_k}) = \operatorname{argmax} \sum_{i=1}^{n} \log s_{K,(\pi_k,\mu_k,\Sigma_k)}(X_i)$$

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

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

Modèle de mélange de gaussiennes

- Densité s_0 de X proche de $s_m(X) = \sum_{k=1}^{K} \pi_k \mathcal{N}_{\mu_k, \Sigma_k}(X)$.
- Modèle $\mathcal{M}_m = \{s_m\}$:
  - 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 $\mathcal{M}_m$: 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 de tous les paramètres.
- Technique classique avec algorithme (EM) efficace disponible.
- Résultat théorique :

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_m)\right] \leq (1+\epsilon)\left(\inf_{s_m \in \mathcal{M}_m} KL^{\otimes n}(s_0, s_m) + \kappa_0(\epsilon)\frac{\dim(\mathcal{M}_m)}{n}\right)$$

Max. de vraisemblance et MM

- "Maximum" de vraisemblance à K fixé :

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

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

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

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

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

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

Max. de vraisemblance et EM

- Retour vers L :

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

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

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

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

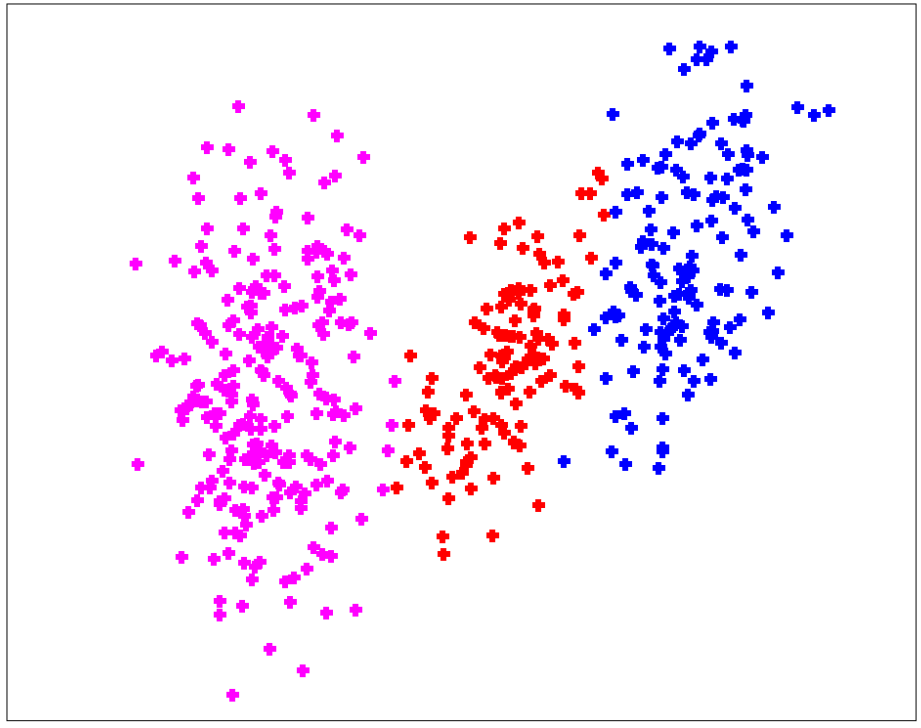
  - Kullback : $L^i \leq L^{i,(j)} + \text{Cst}^{i,(j)}$ avec égalité en $(\pi^{(j)}, \mu^{(j)}, \Sigma^{(j})$.
- Bonus :
  - Séparabilité de $L^{i,(j)}$ en π et (μ, Σ) :

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

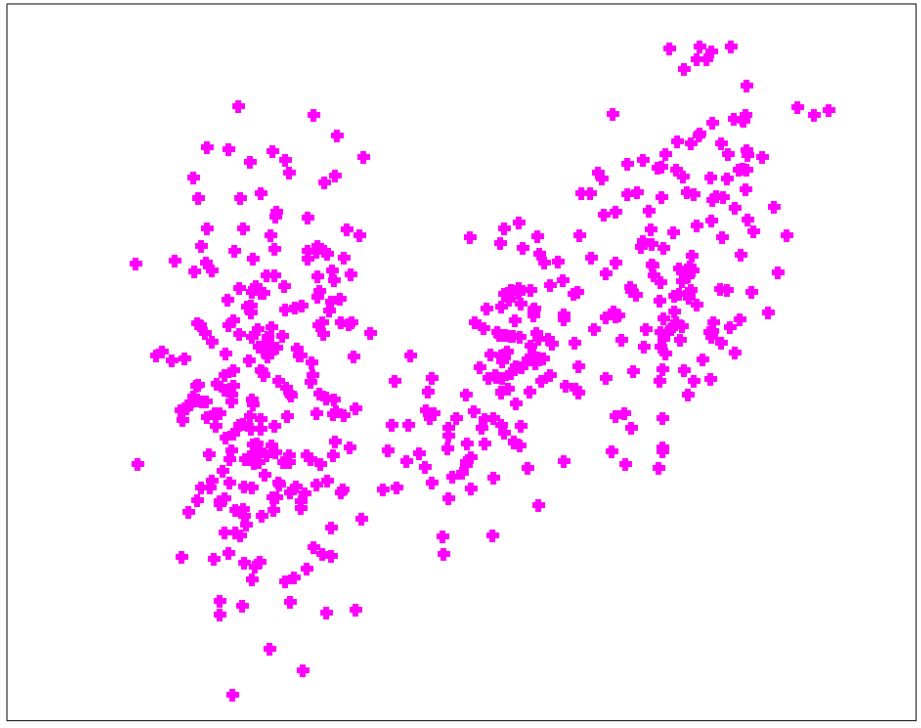
  - Formules closes pour la Minimisation de $L^{(j)}$ en π et (μ, Σ) !

Combien de classes ?

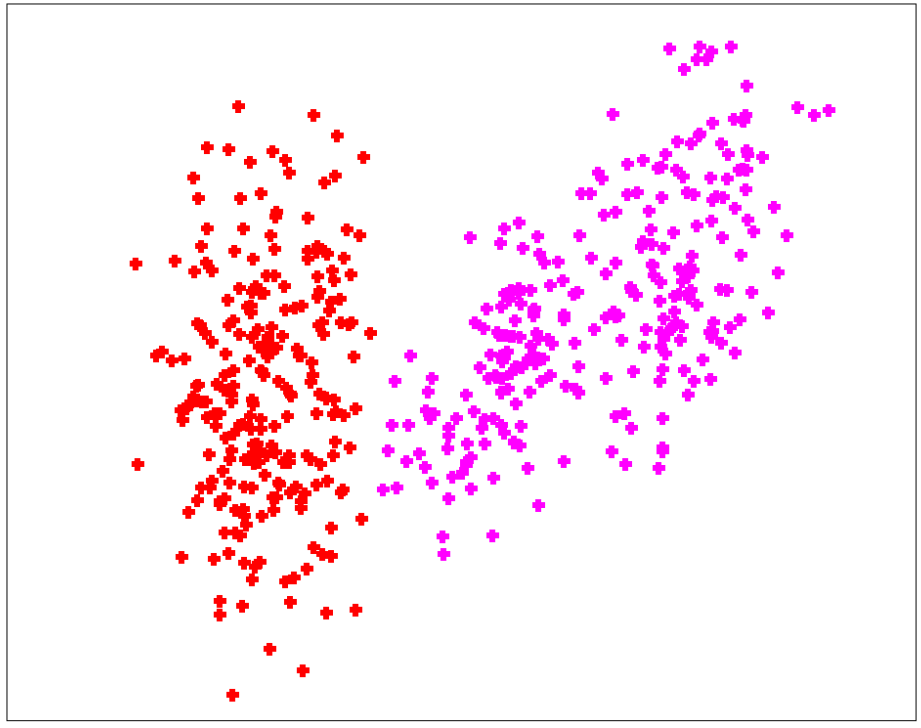
Combien de classes ?



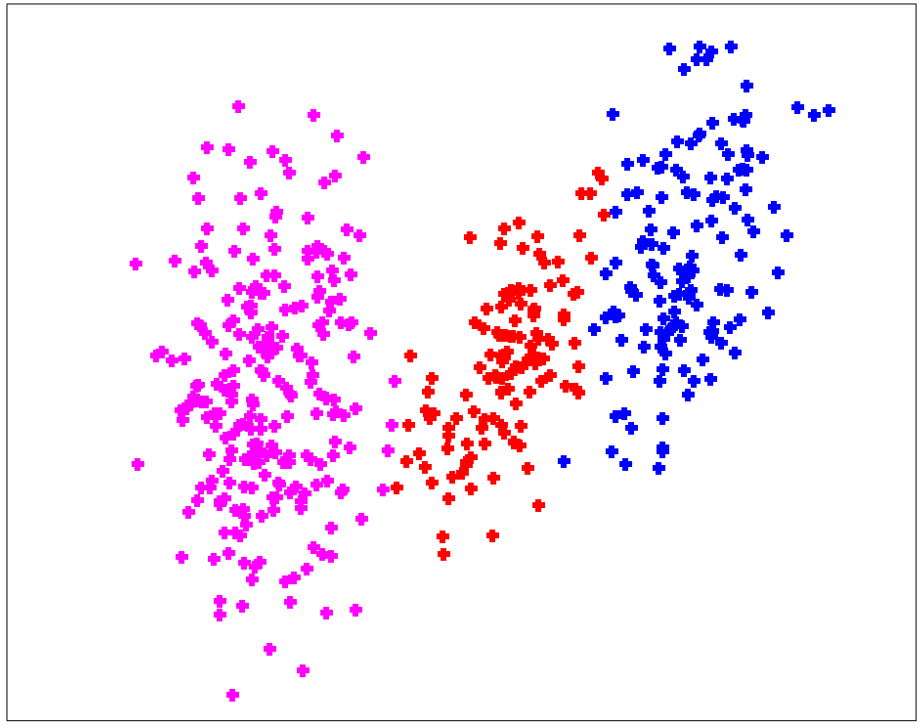
Combien de classes ?



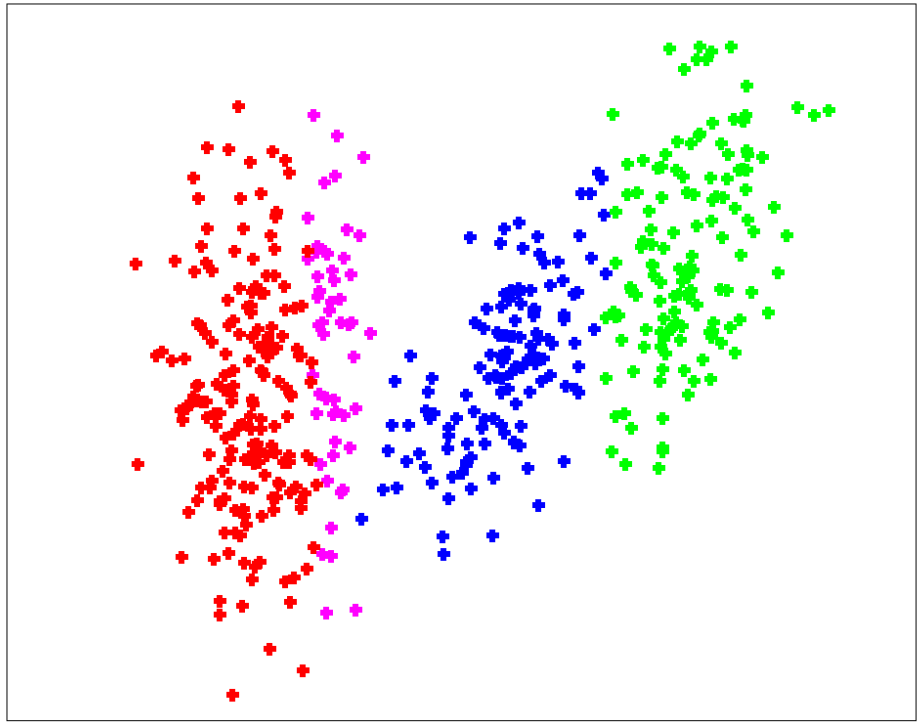
Combien de classes ?



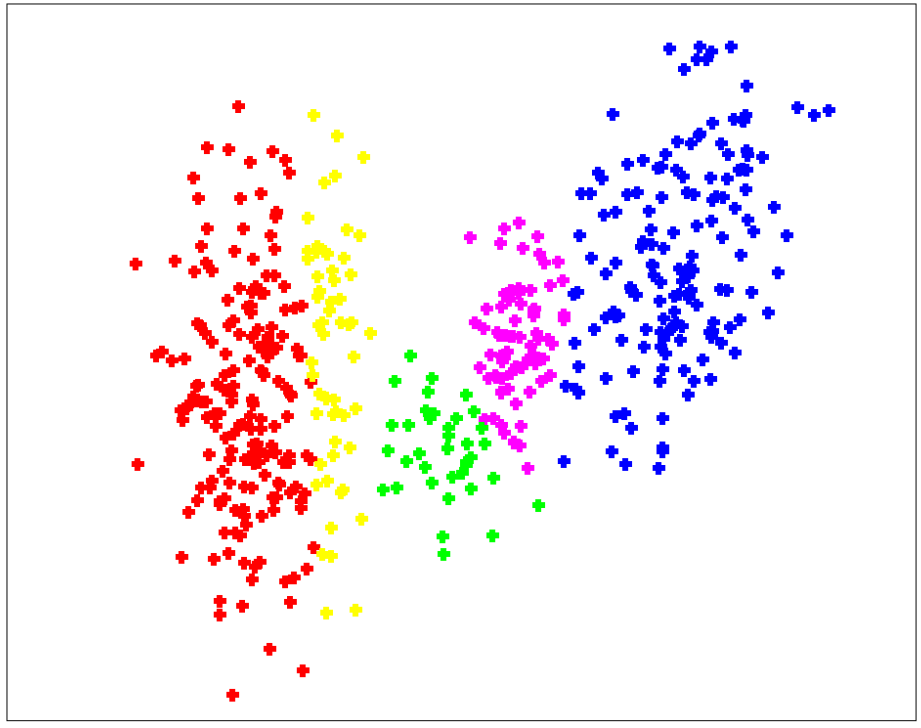
Combien de classes ?



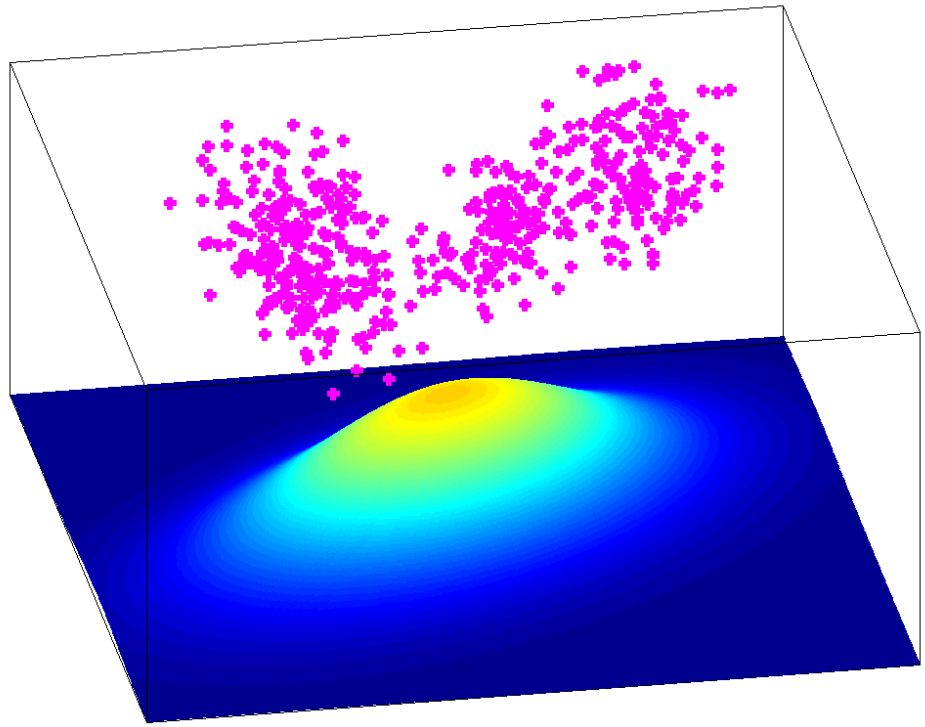
Combien de classes ?



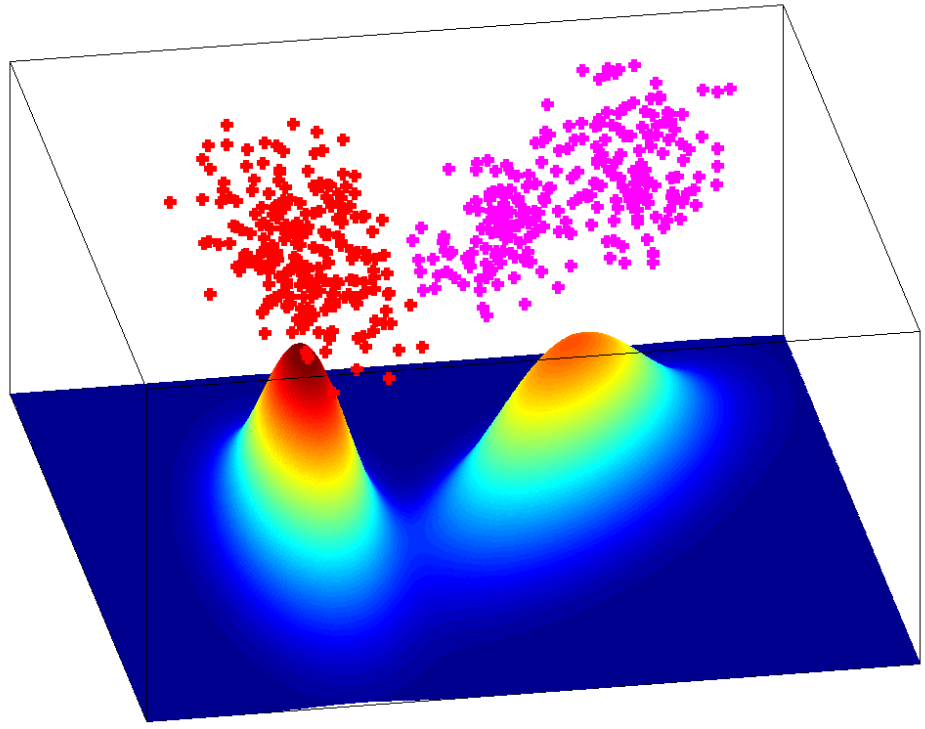
Combien de classes ?



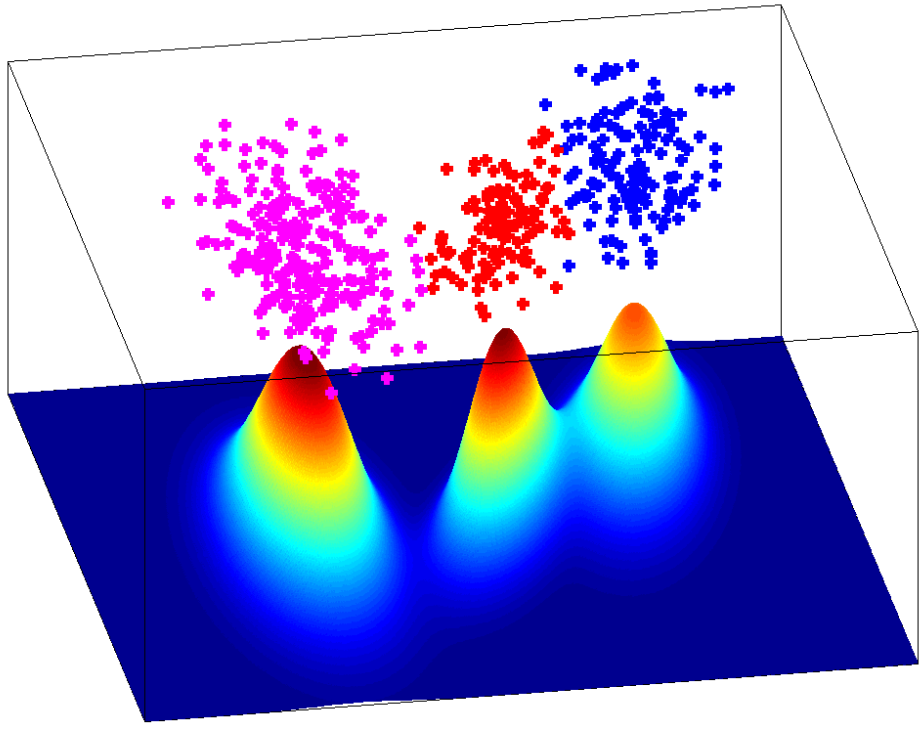
Combien de classes ?



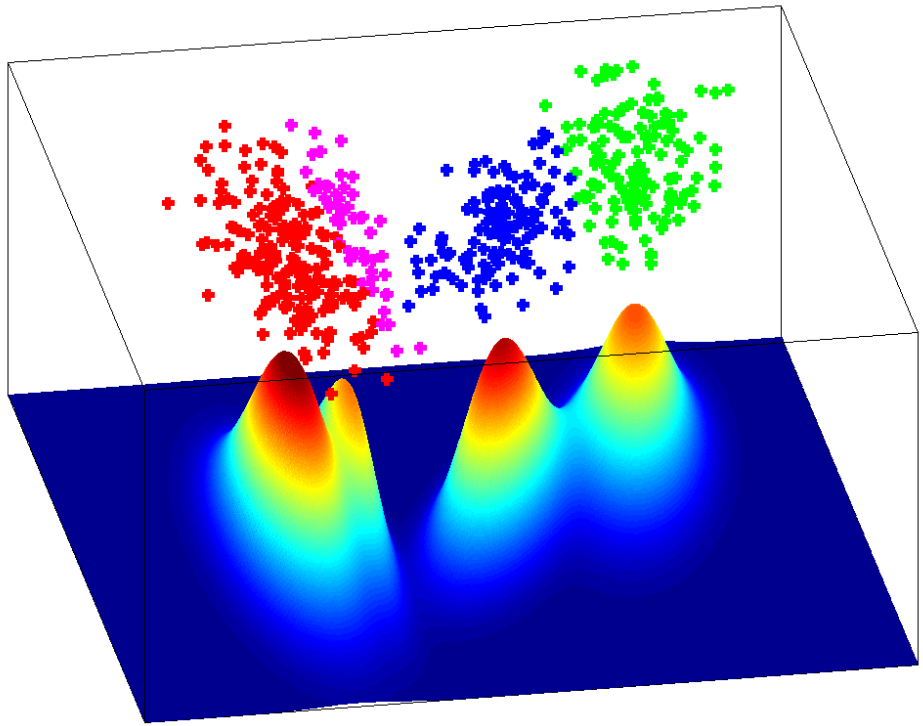
Combien de classes ?



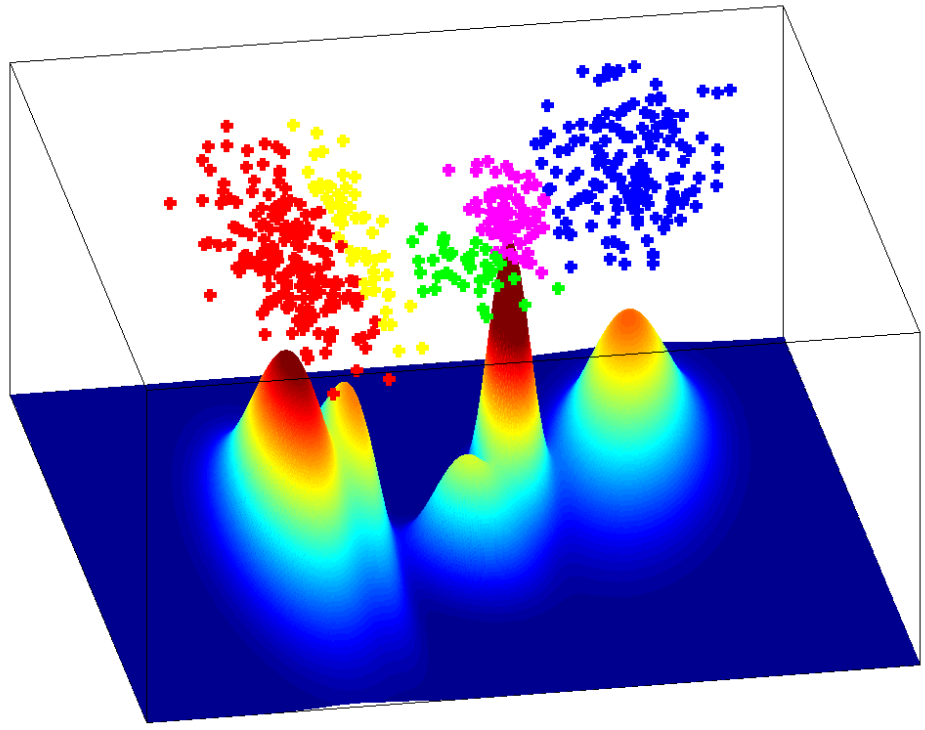
Combien de classes ?



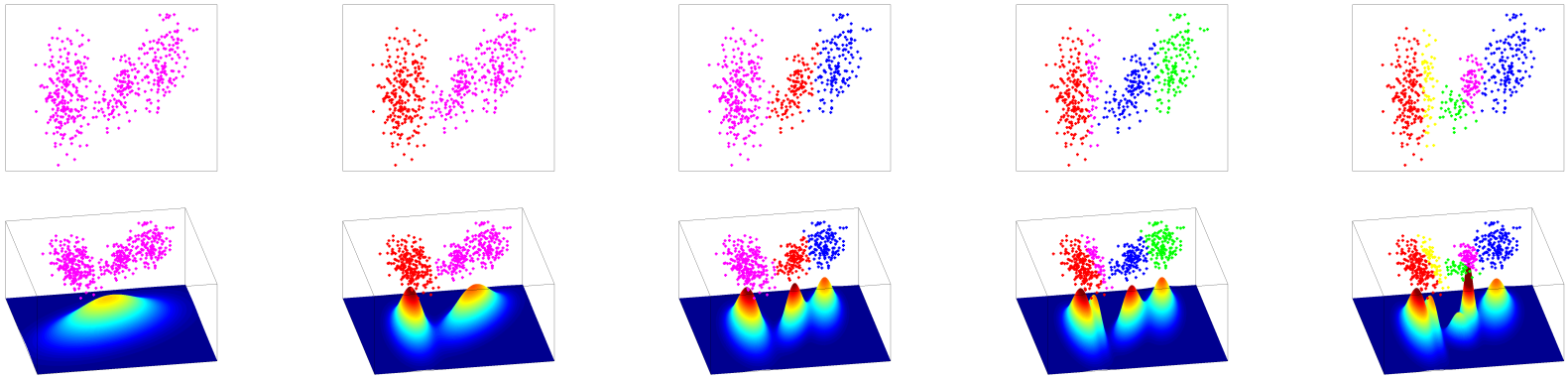
Combien de classes ?



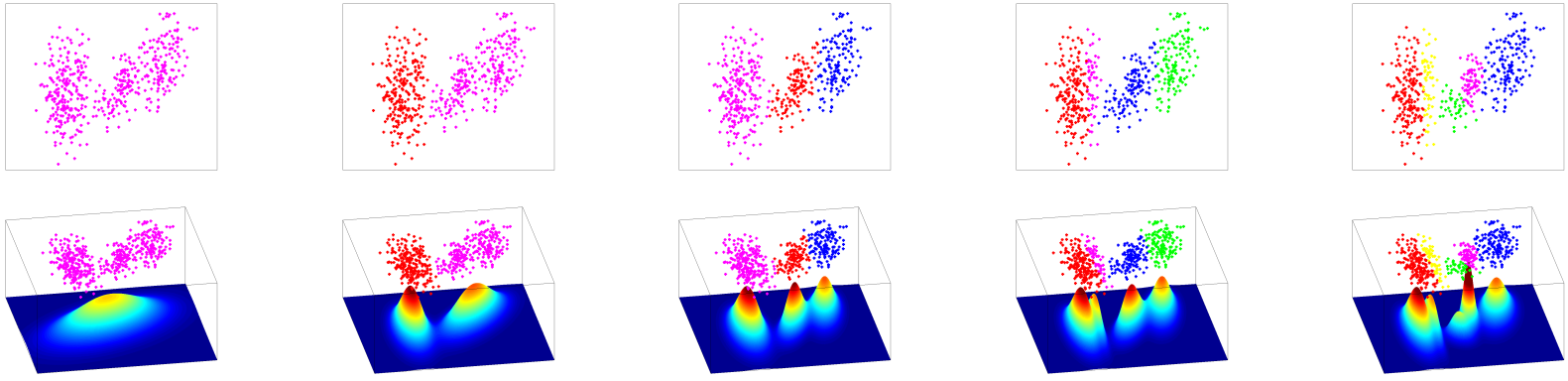
Combien de classes ?



Combien de classes ?

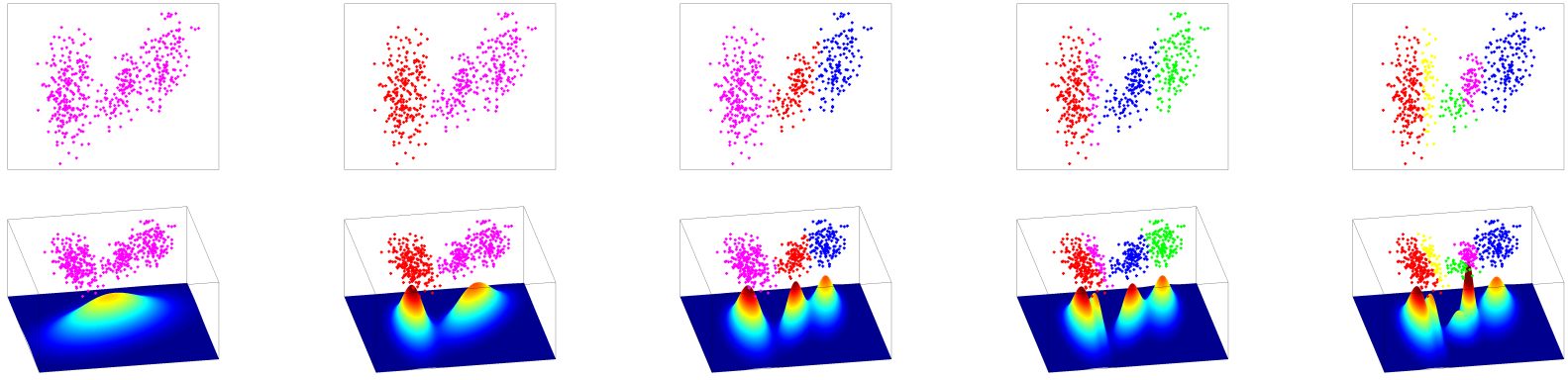


Combien de classes ?



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

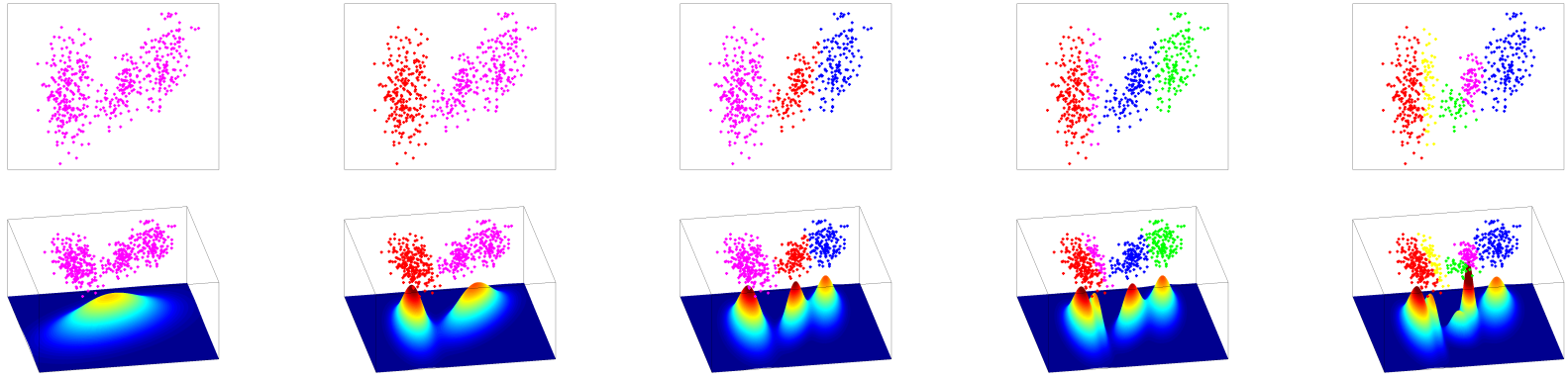
Combien de classes ?



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

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

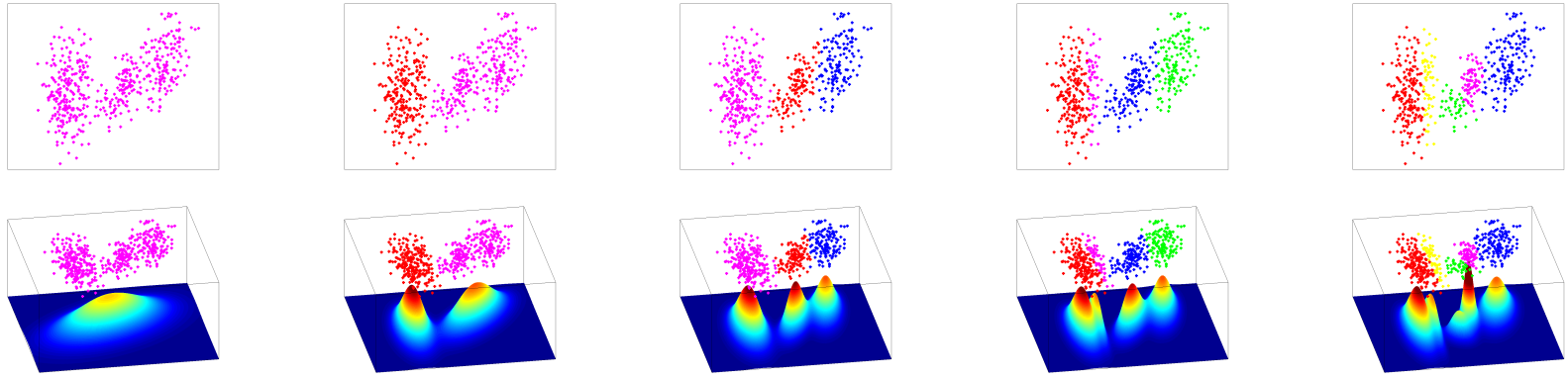
Combien de classes ?



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)

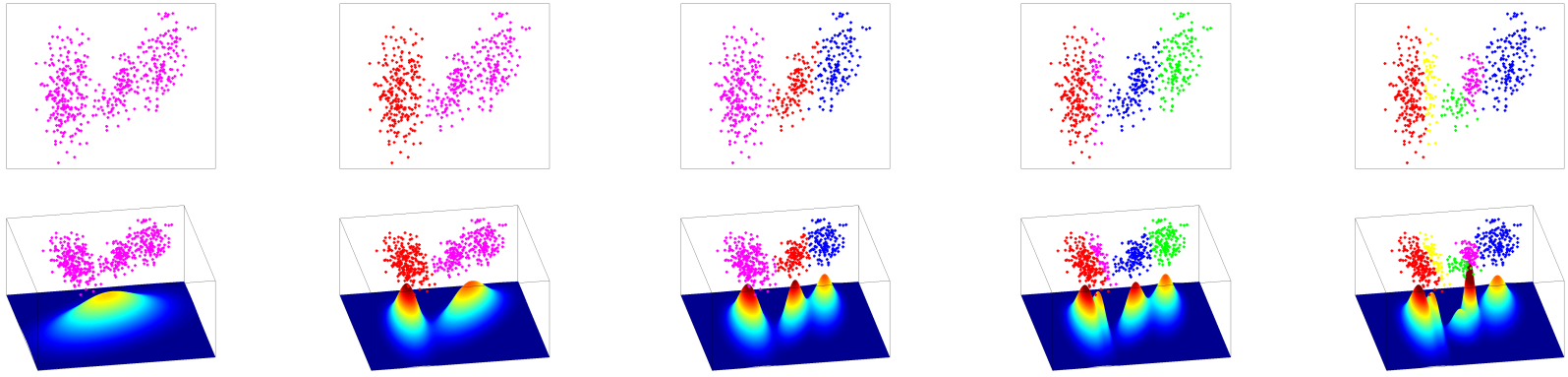
Le rasoir d'Ockham

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

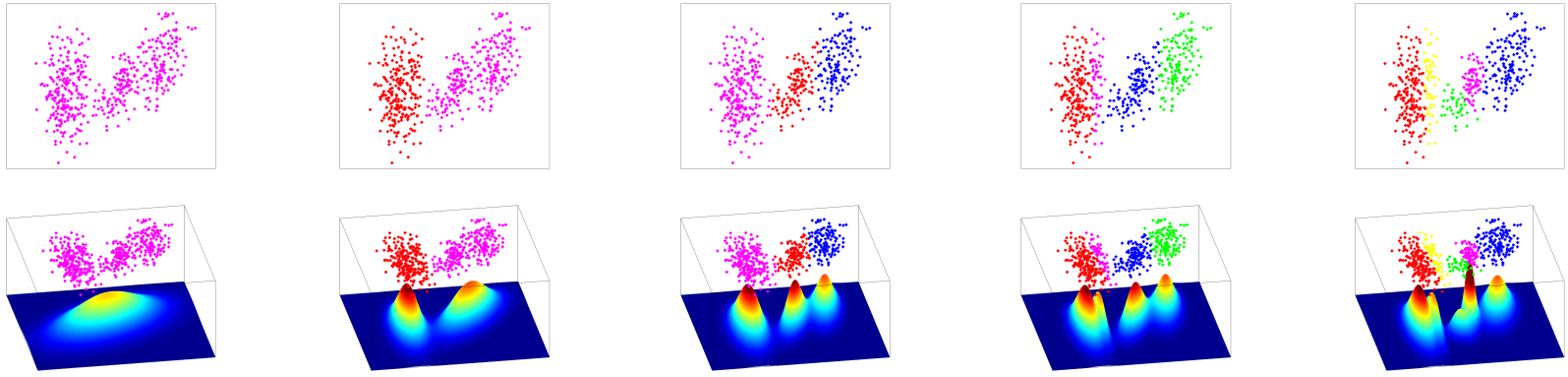
- Rasoir d'Ockham (principe de simplicité) : il ne faut pas ajouter des hypothèses, si celles utilisées suffisent déjà !
- Compromis entre pouvoir d'explication et simplicité.

Sélection par pénalisation

Sélection par pénalisation

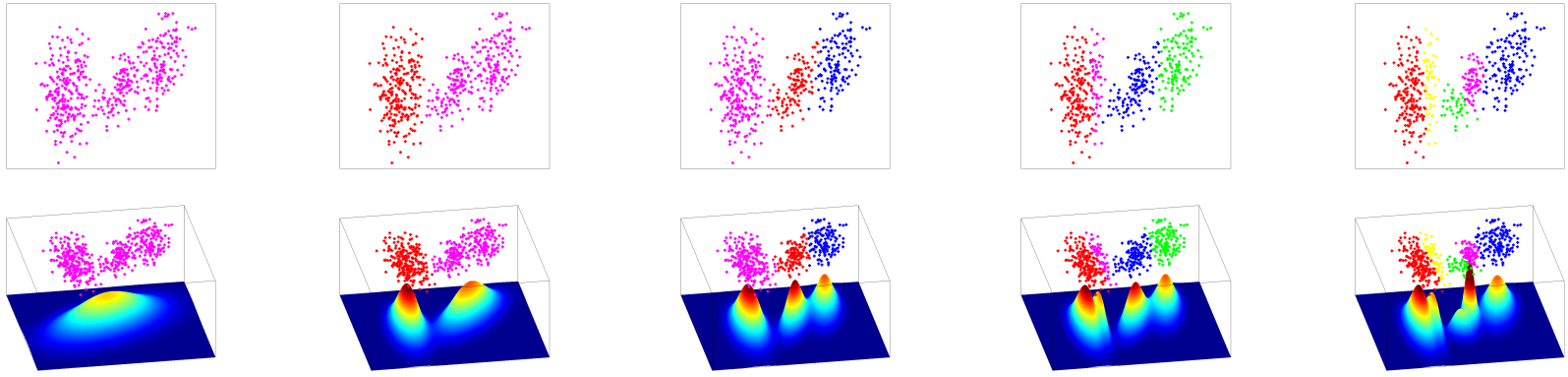


Sélection par pénalisation



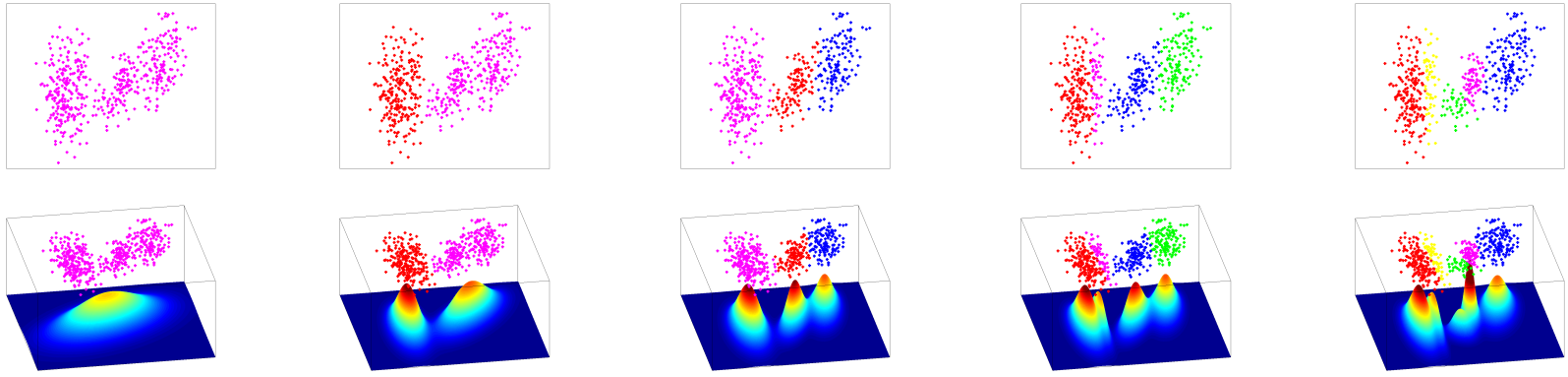
Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
Simplicité	$+++++$	$+++$	$++$	$+$	$-$

Sélection par pénalisation



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

Sélection par pénalisation

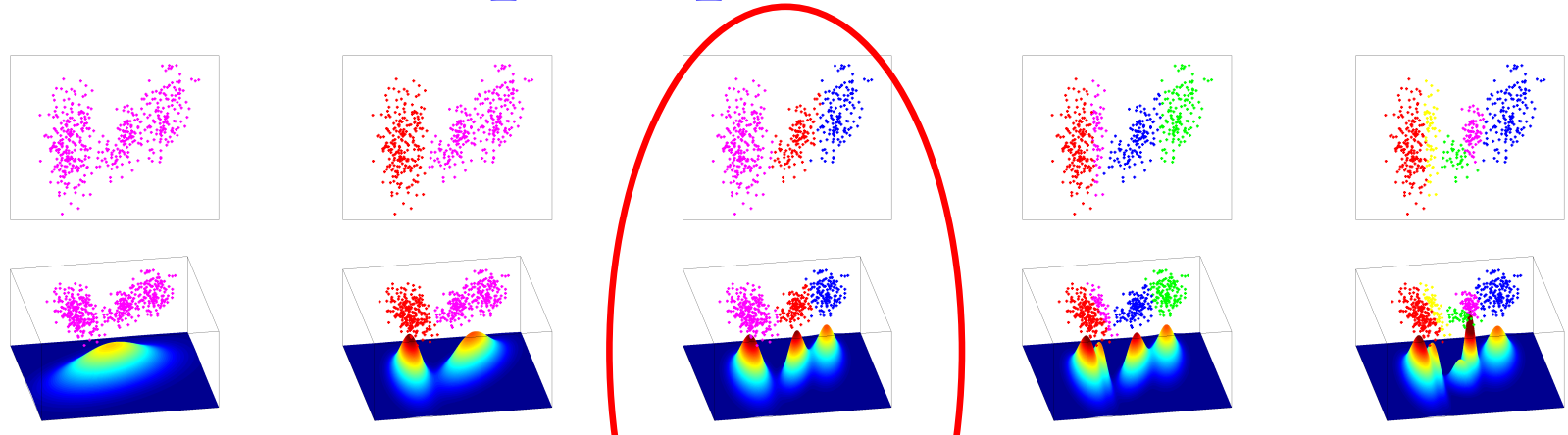


Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicité	$+++++$	$+++$	$++$	$+$	$-$
$=$ Compromis	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

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

$$\operatorname{argmax} \underbrace{\sum_{i=1}^{n} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \dim(SM_K)}_{\text{Pénalité}}$$

Sélection par pénalisation

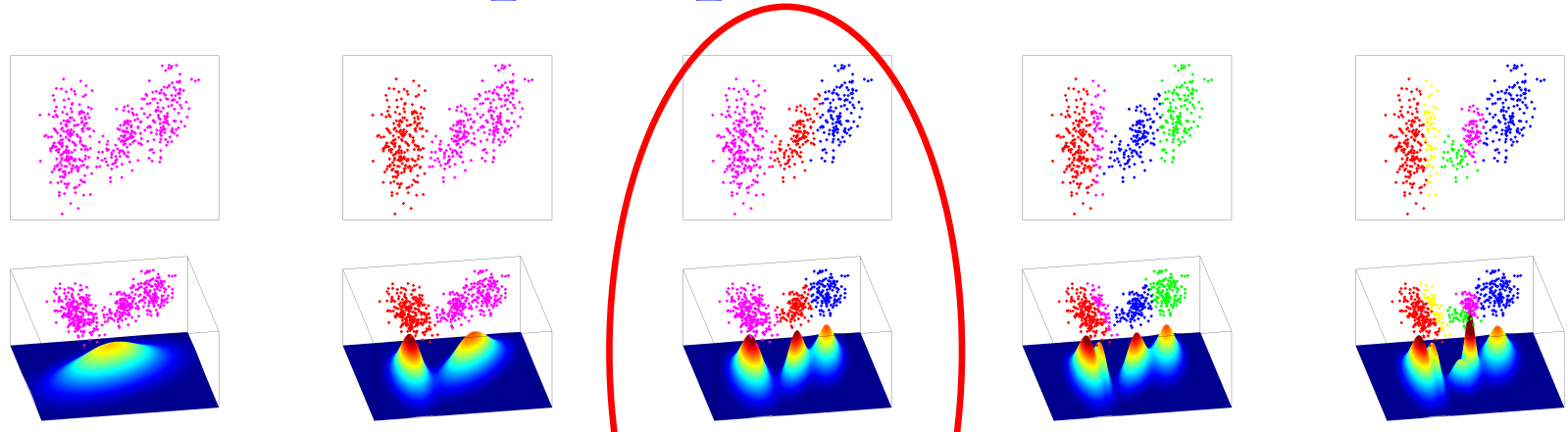


Vraisemblance	$--$	$+$	$+++$	$+++\cdot$	$++++$
$+$ Simplicité	$+++++$	$+++$	$++$	$+$	$-$
$=$ Compromis	$++$	$++++$	$+++++$	$++++\cdot$	$+++$

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

$$\operatorname{argmax} \underbrace{\sum_{i=1}^{n} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \dim(SM_K)}_{\text{Pénalité}}$$

Sélection par pénalisation



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

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

$$\operatorname{argmax} \underbrace{\sum_{i=1}^{n} \log \hat{s}_K(X_i)}_{\text{Vraisemblance}} \underbrace{- \lambda \dim(SM_K)}_{\text{Pénalité}}$$

- Optimisation en K par exploration exhaustive !

Sélection de modèles

- Comment choisir le "modèle" $\mathcal{M}_m$:
  - 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 $\mathcal{M}_m = \{s_m\}$ avec $m \in \mathcal{S}$,
  - estimation par maximum de vraisemblance d'une densité $\hat{s}_m$ pour chaque modèle $\mathcal{M}_m$,
  - sélection d'un modèle $\widehat{m}$ par

$$\widehat{m} = \operatorname{argmin} - \sum_{i=1}^{n} \ln \widehat{s}_m(X_i) + \operatorname{pen}(m).$$

    avec $\operatorname{pen}(m) = \kappa(\ln(n))\dim(\mathcal{M}_m)$ (dimension intrinsèque de $\mathcal{M}_m$) par exploration exhaustive.
- Résultats (Birgé, Massart, Celeux, Maugis, Michel...) :
  - théorique d'estimation du mélange : pour $\kappa > \kappa_0(\epsilon)$,

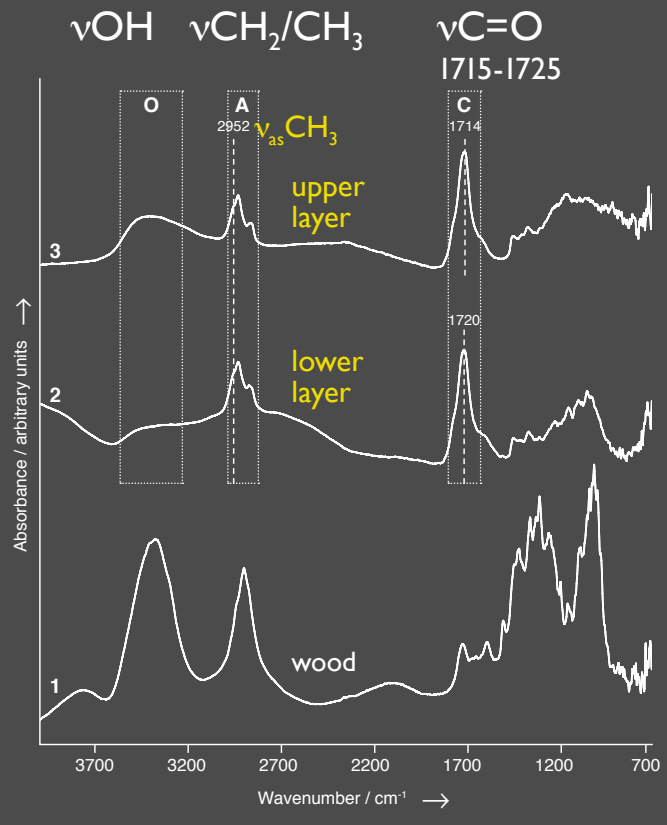
$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq (1+\epsilon) \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in \mathcal{M}_m} KL(s_0, s_m) + \frac{\operatorname{pen}(m)}{n} \right) + \frac{C'}{n}.$$

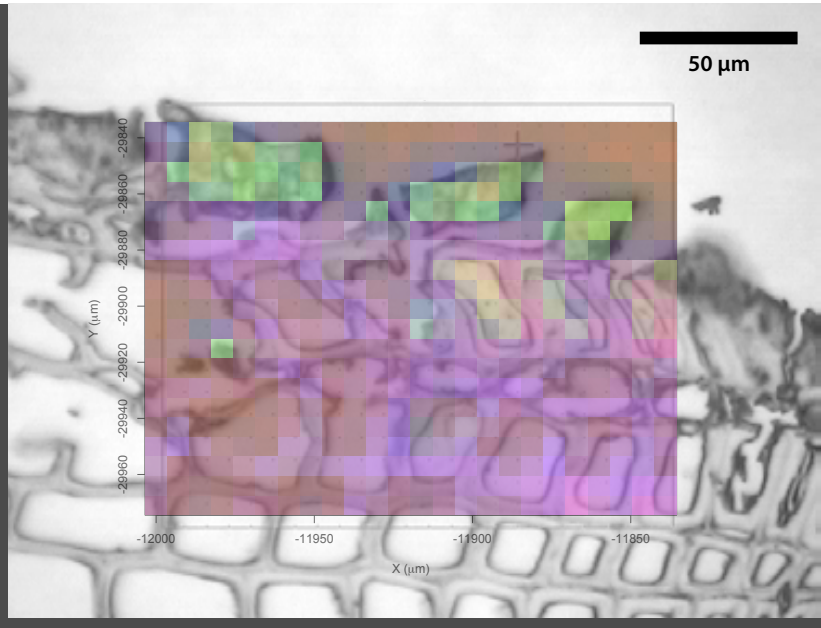
  - pratique de classification non supervisée.

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

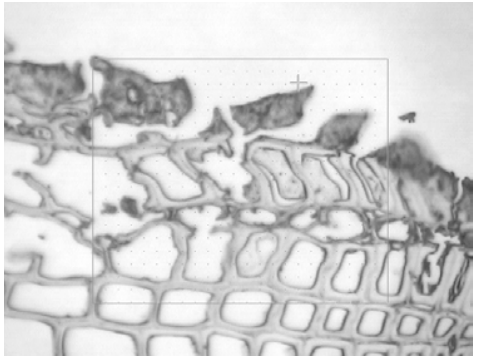
CRCC

CENTRE DE RECHERCHE ET DE RESTAURATION DES MUSÉES DE FRANCE

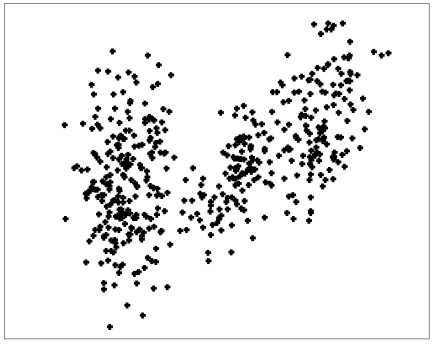
Laboratoire de Dynamique, Interactions et Réactivité

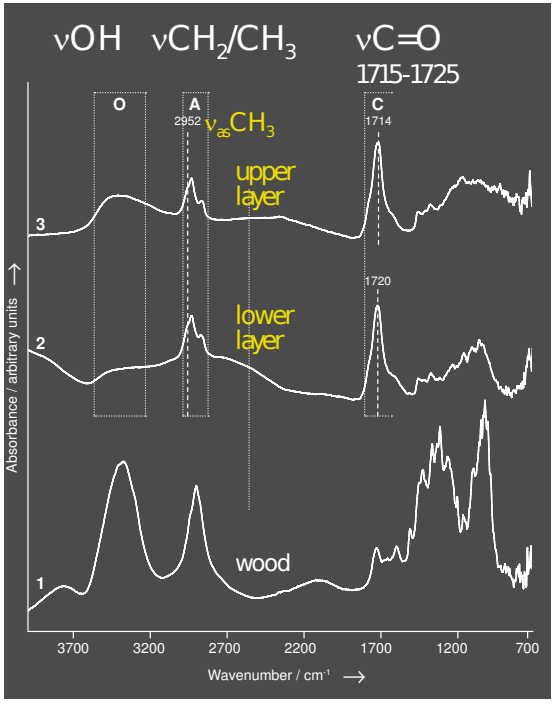
J.-P. Echard, L. Bertrand, A. von Bohlen, A.-S. Le Hô, C. Paris, L. Bellot-Gurlet, B. Soulier, A. Lattuati-Derieux, S. Thao, L. Robinet, B. Lavédrine, and S. Vaiedelich. *Angew. Chem. Int. Ed.*, 49(1), 197-201, 2010.

Application



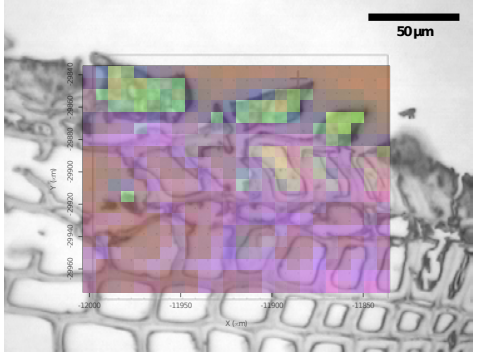
Représentation

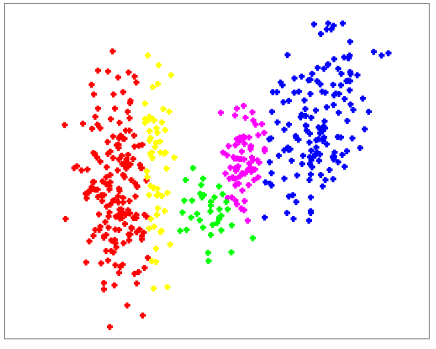




Segmentation

Classification





Info. Spatiale

Plan

Segmentation et mélange de gaussiennes

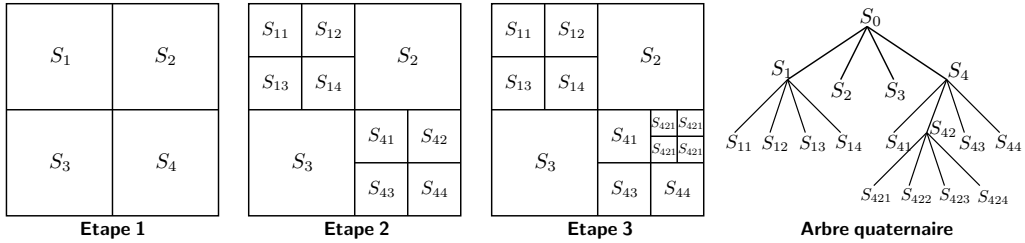
- Objectif initial : segmentation $\neq$ classification non supervisée.
- Prise en compte de la position spatiale Y du spectre à travers les proportions du mélange (Kolaczyk et al) : modèle de densités conditionnelles

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

- 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 $\pi_k(y)$.
- $\pi_k(y)$ fonction : régularisation nécessaire.
- Principe de sélection de modèles...

Mélange de gaussiennes et partition hiérarchique

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



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

Densités conditionnelles

- Modèle statistique vrai $\mathcal{M}_0$: observation de (X_i, Y_i) avec Y_i indépendants et X_i indépendants de loi de densité $s_0(x|Y_i)$.
- Objectif : estimation de $s_0(x|y)$.
- Principe de sélection de modèles par pénalisation :
  - choix d'une collection de modèles $\mathcal{M}_m = \{s_m(x|y)\}$ avec $m \in \mathcal{S}$,
  - estim. par max. de vraisemblance d'une dens. $\hat{s}_m$ pour chaque modèle S_m :

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

  - avec $\mathrm{pen}(m)$ bien choisie, sélection d'un modèle $\widehat{m}$ par

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

- Résultat d'estimation de densité : conditions sur $\mathrm{pen}(m)$ telles

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq (1+\epsilon) \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

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 $s_m \in S_m$

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

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.

- « 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 $d^{2\otimes n}$.

Kullback, Hellinger et extensions

- Inégalité oracle en sélection de modèles de la forme

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

- Densité : Hellinger $d^2(s, s')$ (ou affinité) (Kolaczyk, Barron, Bigot).
- Raff. avec $JKL(s, s') = 2KL(s, (s' + s)/2)$ (Massart, van de Geer).
- Jensen-Kullback-Leibler : généralisation à $JKL_\rho(s, s') = \frac{1}{\rho} KL(s, \rho s' + (1 - \rho)s)$.
- **Prop. :** Pour toutes mesures de proba $sd\lambda$ et $td\lambda$ et tout $\rho \in (0, 1)$

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

avec $C_\rho = \frac{1}{\rho} \min(\frac{1-\rho}{\rho}, 1)\left(\ln\left(1 + \frac{\rho}{1-\rho}\right) - \rho\right)$.

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

Densités conditionnelles

- Nécessité de s'adapter pour les densités conditionnelles :
  - Divergence sur la densité produit conditionnée au design (Kolaczyk, Bigot).
  - Principe de tensorisation et de passage à l'espérance sur le design :

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

- Approche similaire sauf pour Hellinger et la possibilité du passage à l'espérance sur le design dans l'inégalité oracle.
- Inégalité oracle de la forme

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

- On retrouve exactement le théorème classique si $s(\cdot|X_i) = s(\cdot)$.

Pénalité et complexité

- Pénalité liée à la complexité du modèle et de la collection.
- Complexité du modèle S_m (entropie) :
  - $H_{[\cdot],d^{\otimes n}}(\epsilon, S_m)$ entropie à crochet liée à la distance de Hellinger tensorisée $(d^{\otimes n} = \sqrt{d^{2\otimes n}} = \sqrt{\mathbb{E}\left[\frac{1}{n}\sum d^2(s(\cdot|X_i), s'(\cdot|X_i))\right]})$.
  - Hypothèse (H) : pour tout modèle S_m, il existe une fonction croissante $\phi_m(\delta)$ telle que $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ soit décroissante sur $(0, +\infty)$ et telle que pour tout $\sigma \in \mathbb{R}^+$ et tout $s_m \in S_m$

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

  - Complexité mesurée par $n\sigma_m^2$ avec σ_m l'unique racine de $\dfrac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$.
- Complexité de la collection (codage) :
  - complexité donnée par x_m satisfaisant Kraft $\displaystyle\sum_{m\in\mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- Contrainte (classique) sur la pénalité

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

Retour vers les modèles de mélanges spatiaux

- Contrôle de $H_{[\cdot],d^{\otimes n}}(\epsilon, S_m(s_m,\sigma))$ pour les modèles de mélanges spatiaux (cf Maugis et Michel) :
  - contrôle d'un majorant de l'entropie : $H_{[\cdot],d^{\sup}}(\epsilon, S_m)$ où $d^{\sup} = \sqrt{d^{2\sup}} = \sqrt{\sup_x d^2(s(\cdot|x), s'(\cdot|x))}$,
  - résultat valide pour toutes les classes de mélanges ($[\mu\, L\, D\, A]^K$) et toutes les partitions :
    $$H_{[\cdot],d^{\sup}}(\epsilon, S_m) \leq \dim(S_m)(C + \ln\frac{1}{\epsilon})$$
    avec C presque explicite (utilisation d'un lemme de Szarek sur l'entropie de $SO(n)$ sans constante explicite...) et $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
  - implication : $n\sigma_m^2 \leq \kappa'\left(C' + \frac{1}{2}\left(\ln\left(\frac{n}{C'\dim(S_m)}\right)\right)_+\right)\dim(S_m)$.
- Codage de la collection avec $x_m \leq \kappa''|\mathcal{P}| \leq \frac{\kappa''}{K-1}\dim(S_m)$.
- Condition sur la pénalité :

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

Optimisation numérique

pour les mélanges de Gaussiennes spatialisés

- Sélection de modèle par pénalisation :

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

- 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 :

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

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

Algorithme EM

- Étape E : avec $P_k^{i,(j)} = P(k_i = k | Y_i, X_i, \mathcal{P}^{(j)}, \pi^{(j)}, \mu^{(j)}, \Sigma^{(j)})$

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

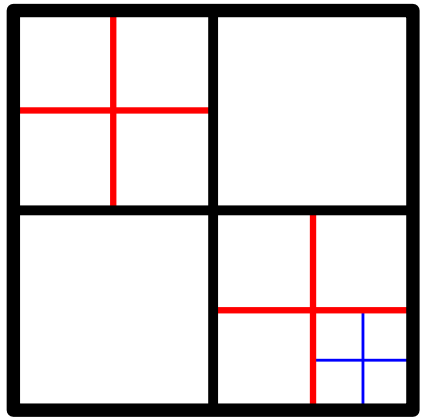
$$\leq -\sum_{i=1}^{n} \sum_{k=1}^{K} P_k^{i,(j)} \ln \left(\pi_k [\mathcal{R}(Y_i)] \right) + \lambda_{0,n} |\mathcal{P}| (K-1)$$

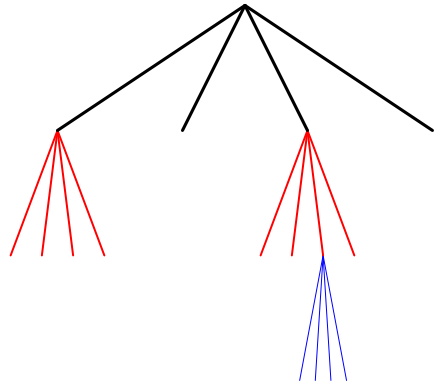
$$+ \left(-\sum_{i=1}^{n} \sum_{k=1}^{K} P_k^{i,(j)} \ln \left(\mathcal{N}_{\mu_k, \Sigma_k}(X_i) \right) \right) + \text{Cst}^{(j)}$$

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

- 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





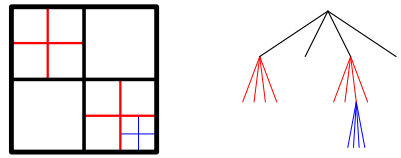
- Optimisation en $(\mathcal{P}, \pi)$ de

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

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

- 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).

Optimisation CART



- Pb : déterminer efficacement $\operatorname{argmin}_{\mathcal{P}} \sum_{\mathcal{R}\in\mathcal{P}} C[\mathcal{R}]$ où $\mathcal{P}$ appartient à l'ensemble des partitions récursives dyadiques (associées à des arbres dyadiques) de profondeur limitée connue.
- Observation : la partition optimale $\widehat{\mathcal{P}}[\mathcal{R}]$ d'un carré dyadique est
  - soit ce carré, $\widehat{\mathcal{P}}[\mathcal{R}] = \{\mathcal{R}\}$
  - soit l'union des partitions opt. de ses enfants, $\widehat{\mathcal{P}}[\mathcal{R}] = \cup_{\mathcal{R}'\in \mathsf{Enfant}[\mathcal{R})} \widehat{\mathcal{P}}[\mathcal{R}']$

  avec pour critère de décision

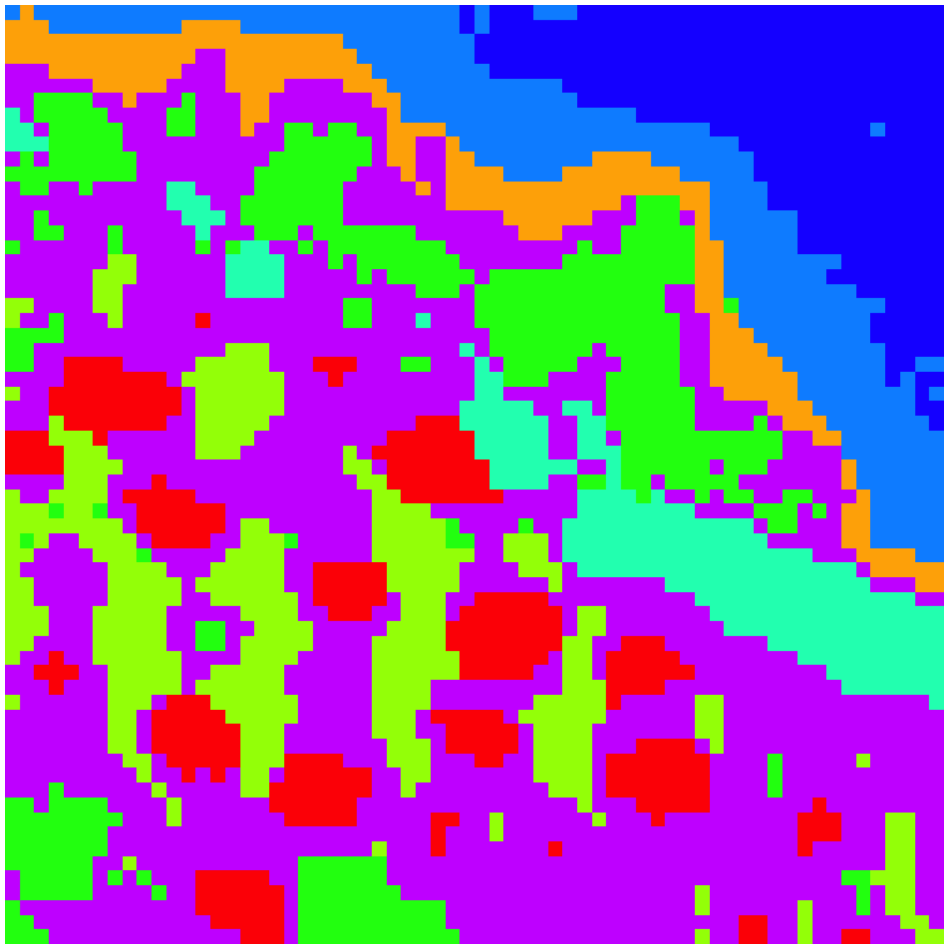
$$C[\mathcal{R}] \leq \sum_{\mathcal{R}'\in \mathsf{Enfant}(\mathcal{R})} \sum_{\mathcal{R}''\in\widehat{\mathcal{P}}[\mathcal{R}']} C[\mathcal{R}'']$$

- Algorithme : Précalcul des $C[\mathcal{R}]$ puis détermination récursive de $\widehat{\mathcal{P}}[\mathcal{R}]$ et de $\widehat{C}[\mathcal{R}] = \sum_{\mathcal{R}''\in\widehat{\mathcal{P}}} C[\mathcal{R}'']$ (soit $C[\mathcal{R}]$ soit la somme des $\widehat{C}$ des enfants) avec arrêt de la récursion dès qu'il n'y a plus d'enfants.
- Version non récursive possible.

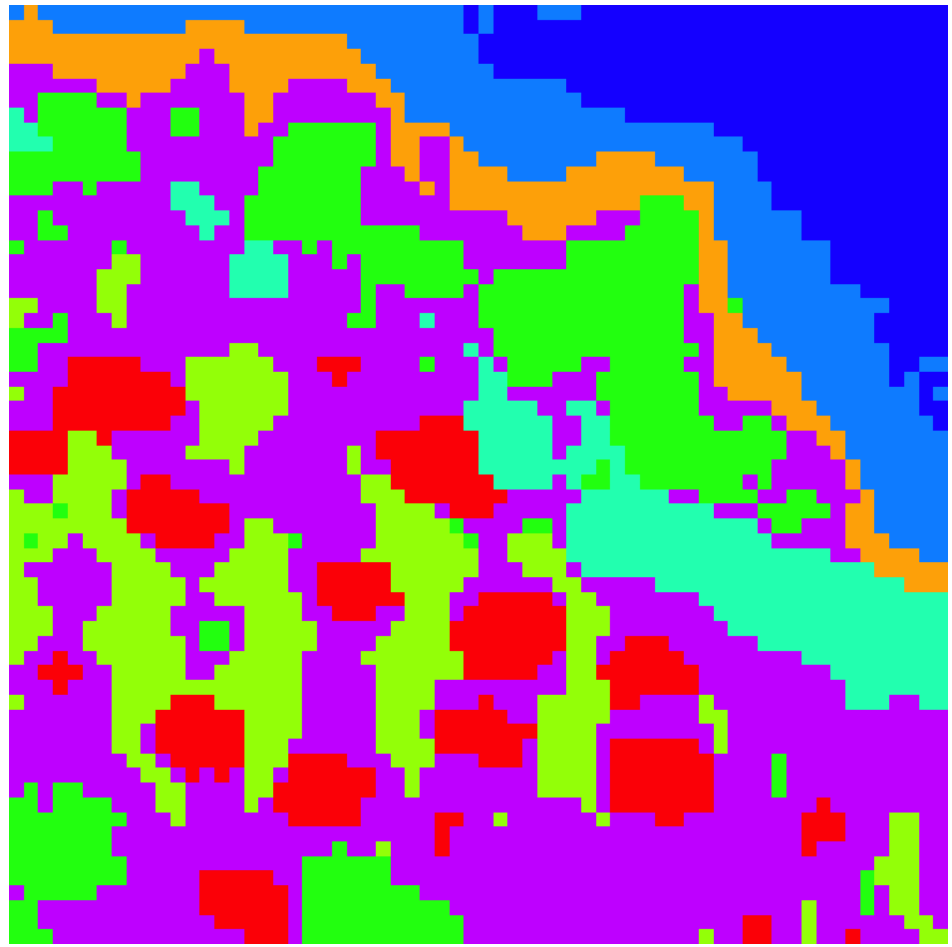
Segmentation automatique

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

Sans



Avec

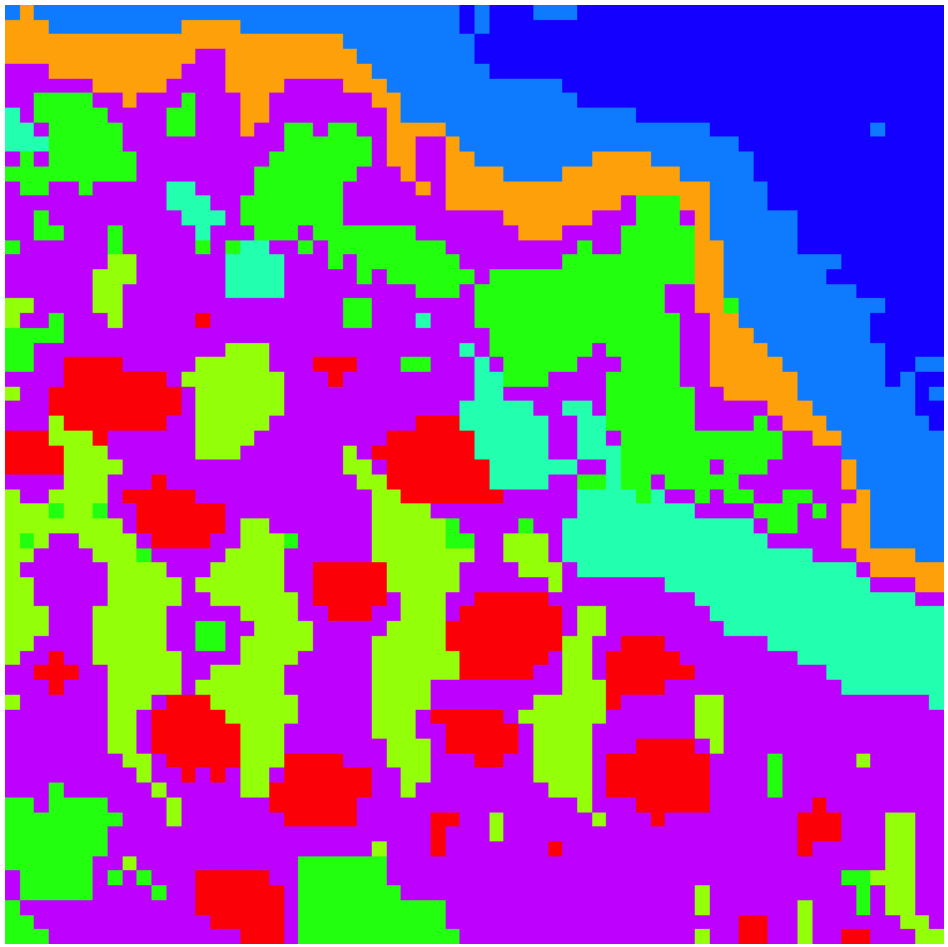


- $K = 8$, $[L_k\, D\, B]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente.
- Réduction de dimension par (simple) ACP...

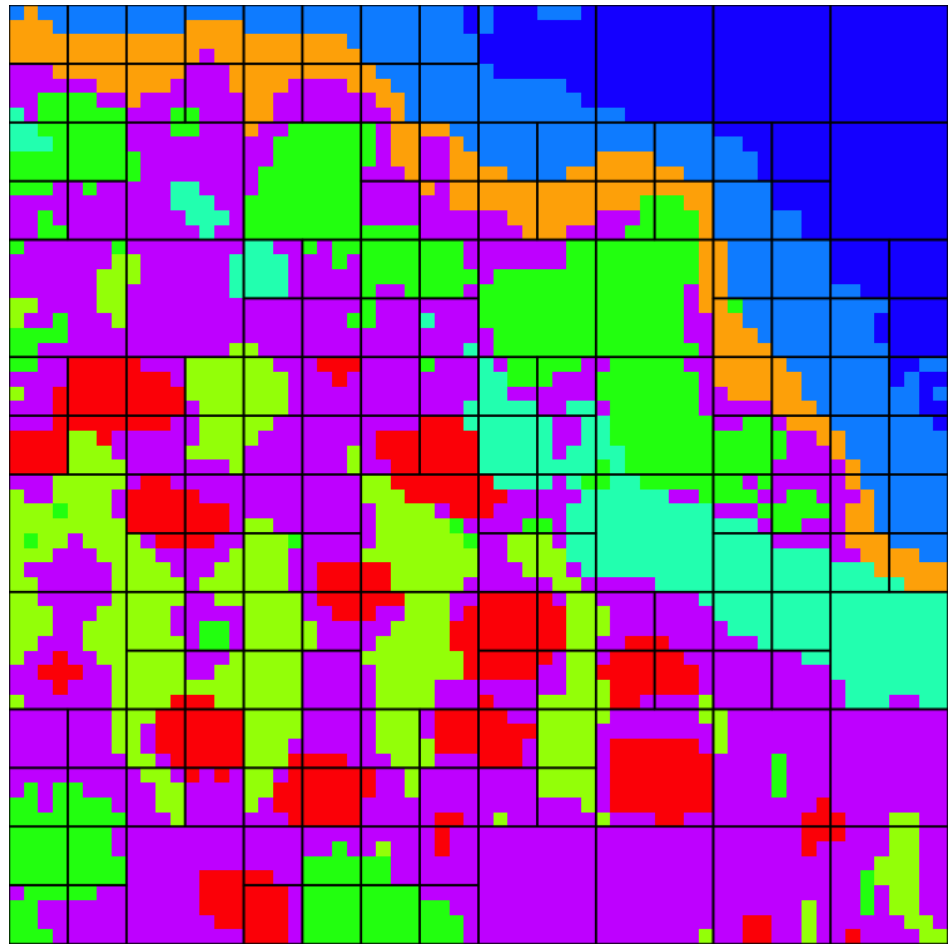
Segmentation automatique

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

Sans

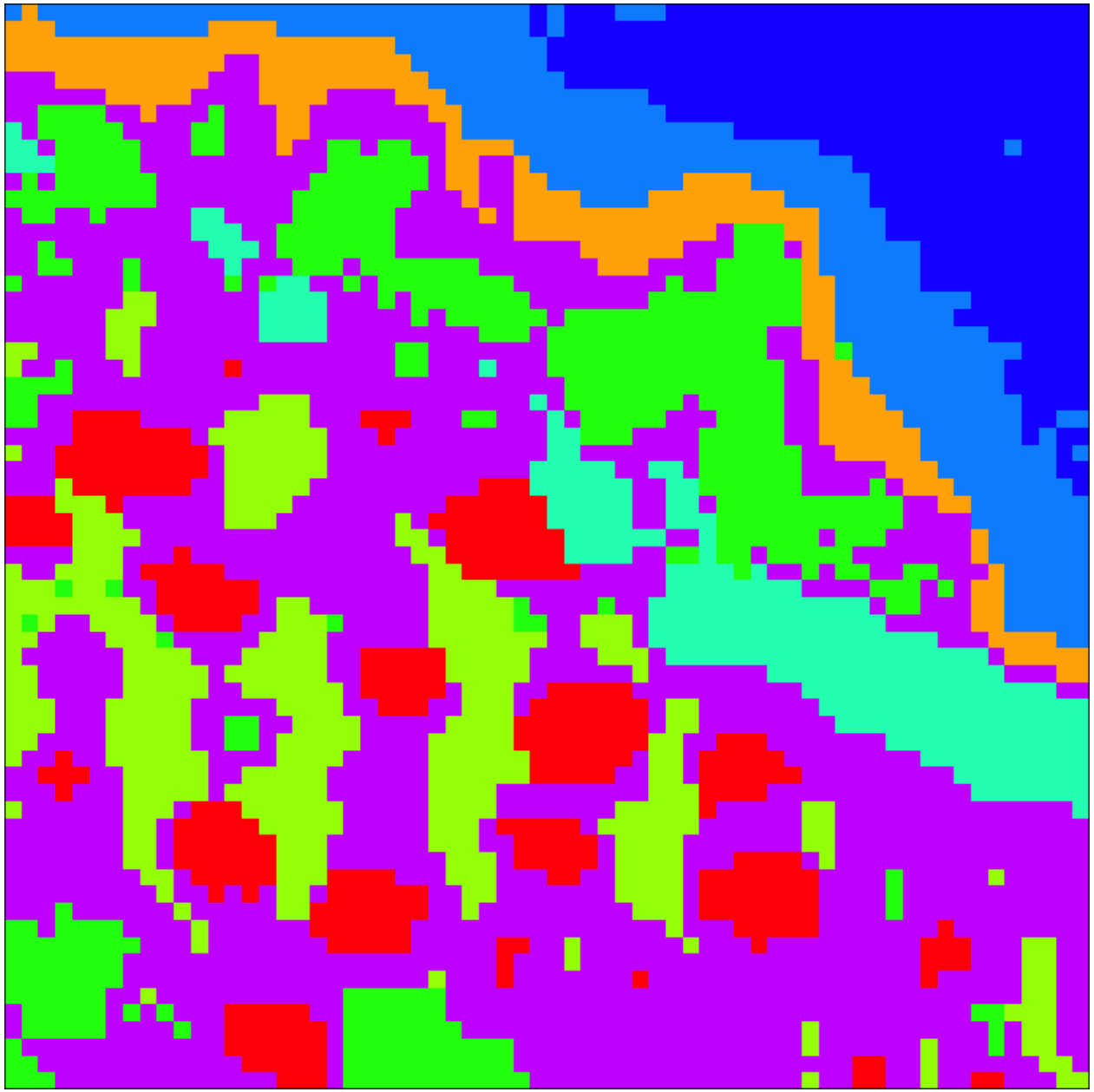


Avec

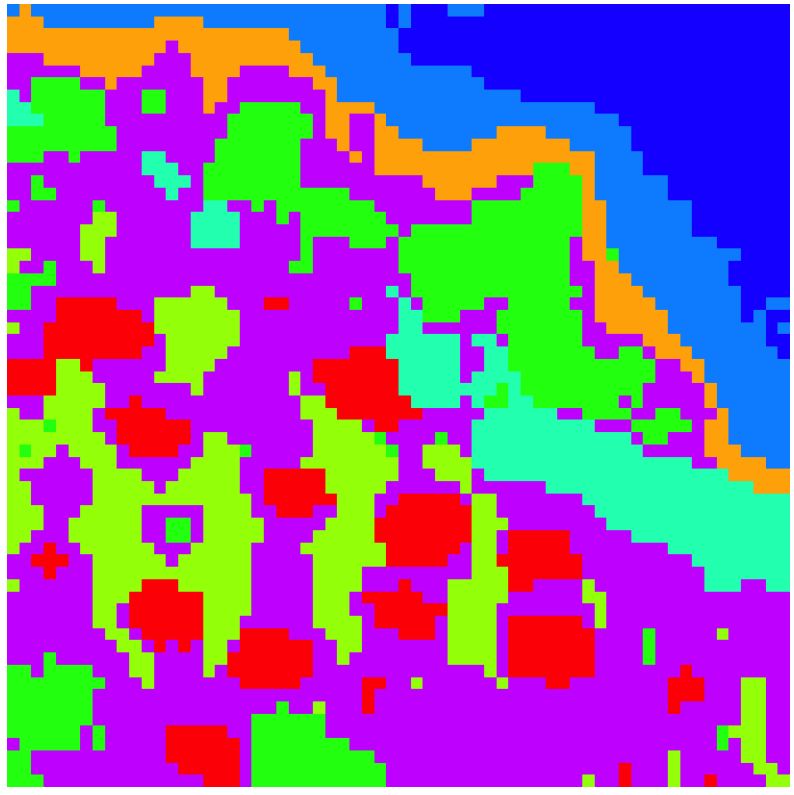


- $K = 8$, $[L_k\, D\, B]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente.
- Réduction de dimension par (simple) ACP...

Segmentations



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 !

Plan

Approche dictionnaire

- Modèle statistique vrai $\mathcal{M}_0$: $X_1, \ldots, X_n$ i.i.d de densité $s_0(x)$.
- Modèle stat. $\mathcal{M}_{\mathcal{F}}$ engendré par un dictionnaire $\mathcal{D} = (\phi_k)_{1 \leq k \leq K}$.
- Estimation de s_0 par $s_\lambda = \sum_{k=1}^{K} \lambda_k \phi_k$
- Moindres carrés :

$$\begin{aligned}
\widehat{\lambda} &= \operatorname*{argmin}_{\lambda} -\frac{2}{n} \sum_{i=1}^{n} s_\lambda(X_i) + \|s_\lambda\|^2 \\
&= \operatorname*{argmin}_{\lambda} -\frac{2}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \lambda_k \phi_k(X_i) + \|s_\lambda\|^2 \\
&= \operatorname*{argmin}_{\lambda} -2\Phi'\lambda + \lambda' G \lambda
\end{aligned}$$

  avec $\Phi_k = \frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i)$ et $G_{k,k'} = \langle \phi_k, \phi_{k'} \rangle$.
- Si G inversible : $\widehat{\lambda} = G^{-1}\Phi$ et

$$\mathbb{E}\left[\|s_0 - s_{\widehat{\lambda}}\|^2\right] \leq \inf_{\lambda} \|s_0 - s_\lambda\|^2 + \kappa \frac{K}{n}$$

Pénalisation

- Comment optimiser le dictionnaire dans

$$\mathbb{E}\left[\|s_0 - s_{\widehat{\lambda}}\|^2\right] \leq \inf_{\lambda} \|s_0 - s_\lambda\|^2 + \kappa \frac{K}{n} ?$$

- $\mathcal{D}$ doit permettre d'approcher efficacement s_0 (terme de bais) mais ne pas être trop grand (terme de variance).
- Sélection de variable : choix de $I \subset \{1, \ldots, K\}$ tel que $\mathcal{D} = (\phi_k)_{k \in I}$ donne un bon compromis.
- Principe de parcimonie : pénalisation ℓ_0 (par la dimension)

$$\widehat{I} = \operatorname*{argmin}_{I \subset \{1,\ldots,K\}} \min_{\lambda, \lambda_{\bar{I}}=0} -2\Phi'\lambda + \lambda' G \lambda + \kappa \log K \frac{|I|}{n}$$

- Résultat théorique

$$\mathbb{E}\left[\|s_0 - s_{\widehat{\lambda}_{\widehat{I}}}\|^2\right] \leq (1+\epsilon) \inf_{I \subset \{1,\ldots,K\}} \inf_{\lambda, \lambda_{\bar{I}}=0} \|s_0 - s_{\lambda_I}\|^2 + \kappa \log K \frac{|I|}{n}$$

Relaxation convexe

- Pénalisation ℓ_0 (par la dimension) non convexe

$$\widehat{I} = \underset{I \subset \{1,\dots,K\}}{\operatorname{argmin}} \min_{\lambda, \lambda_{\bar{I}}=0} -2\Phi'\lambda + \lambda' G \lambda + \kappa \log K \frac{|I|}{n}.$$

- Optimisation difficile si K est grand...
- Lasso : relaxation de la norme ℓ_0 par la norme ℓ_1

$$\widehat{\lambda}^L = \underset{\lambda}{\operatorname{argmin}} -2\Phi'\lambda + \lambda' G \lambda + \gamma \|\lambda\|_1.$$

- Dantzig : condition de premier ordre du Lasso + minisation ℓ_1 pour faire

$$\widehat{\lambda}^D = \underset{\lambda}{\operatorname{argmin}} \|\lambda\|_1 \quad \text{sous} \quad \|\Phi'\lambda - G\lambda\|_\infty \leq \frac{\gamma}{2}$$

- Interp. probab. du Dantzig : concentration de moyennes empiriques permettant une calibration fine en faisant dépendre γ de k

$$\forall k \in \{1,\dots,K\}, \left| \frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i) - \mathbb{E}\phi_k(X) - \langle \phi_k, s_\lambda - s_0 \rangle \right| \leq \frac{\gamma_k}{2}$$

Lasso et Dantzig

- Lasso :

$$\widehat{\lambda}^L = \underset{\lambda}{\operatorname{argmin}} -2\Phi'\lambda + \lambda' G\lambda + \sum_{k=1}^{K} \gamma_k |\lambda_k|.$$

- Dantzig :

$$\widehat{\lambda}^D = \underset{\lambda}{\operatorname{argmin}} \|\lambda\|_1 \quad \text{sous} \quad \forall k, |(\Phi'\lambda - G\lambda)_k| \leq \frac{\gamma_k}{2}$$

- Bon choix : $\gamma_k = \gamma c_k \sqrt{\frac{\log K}{n}}$ avec γ à bien choisir.
- Résultat théorique (sous des hypothèses fortes sur $\mathcal{D}$!) : avec grande probabilité,

$$\|s_0 - s_{\widehat{\lambda}}\|^2 \leq \inf_{I \subset \{1,\dots,K\}} \inf_{\lambda, \lambda_{\bar{I}}=0} \|s_0 - s_{\lambda_I}\|^2 + \kappa(I) \log K \frac{|I|}{n}$$

Optimisation convexe

- Lasso :

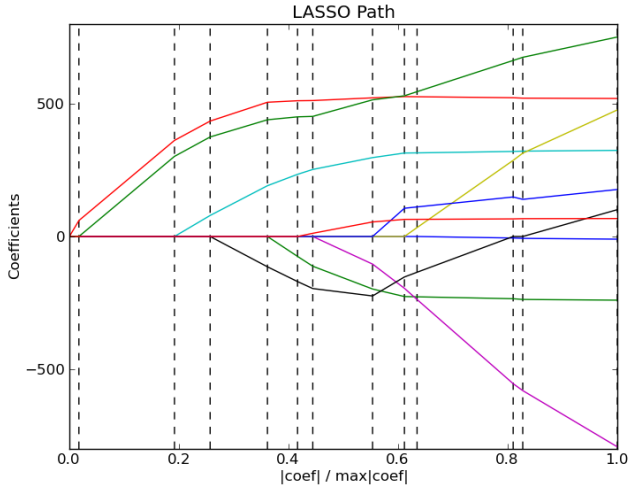
$$\widehat{\lambda}^L = \underset{\lambda}{\operatorname{argmin}} -2\Phi'\lambda + \lambda' G\lambda + \sum_{k=1}^{K} \gamma_k |\lambda_k|.$$

- Dantzig :

$$\widehat{\lambda}^D = \underset{\lambda}{\operatorname{argmin}} \|\lambda\|_1 \quad \text{sous} \quad \forall k, |(\Phi'\lambda - G\lambda)_k| \leq \frac{\gamma_k}{2}$$

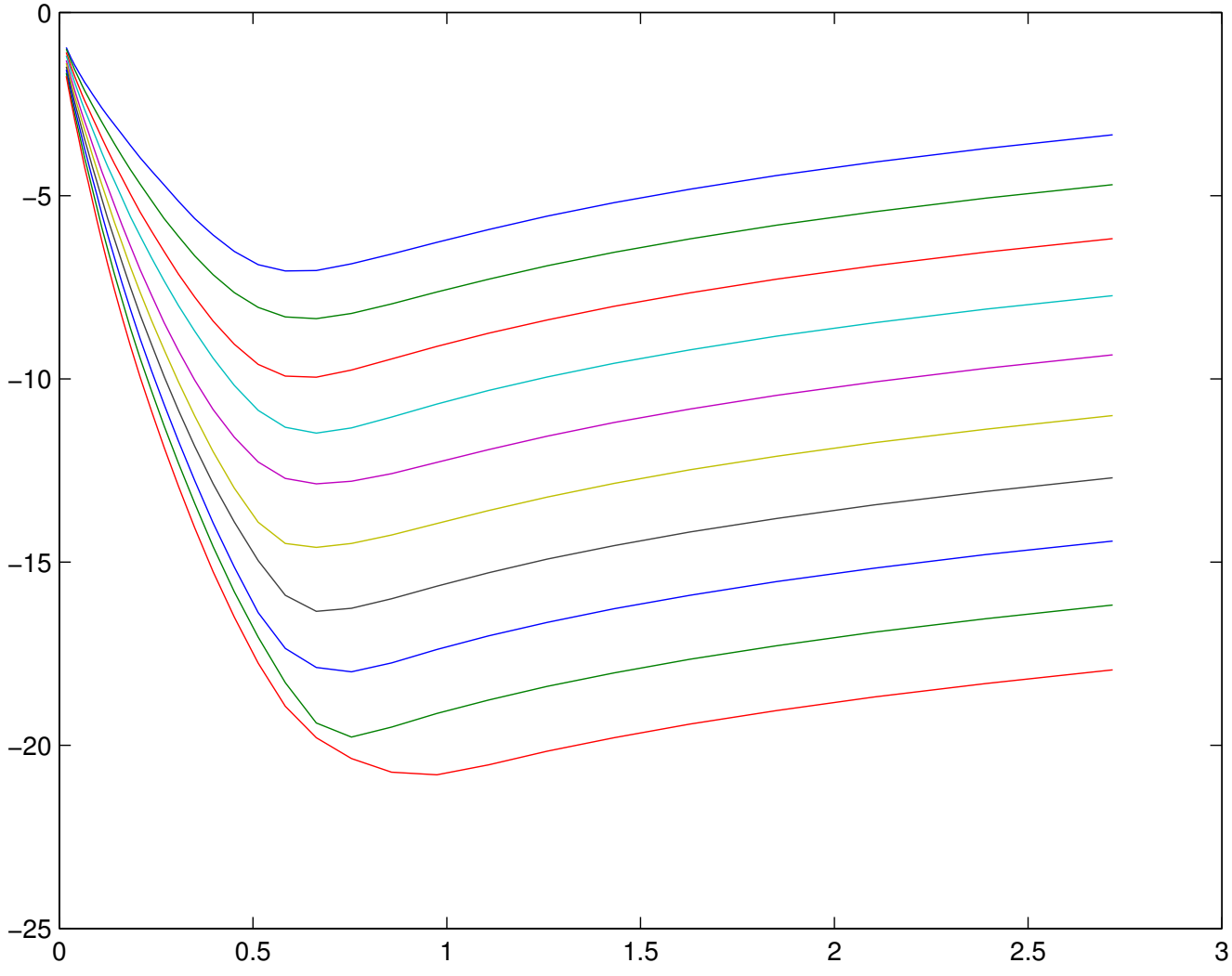
- Deux problèmes proches de minimisation convexe !
- Nombreuses approches utilisées :
  - Points intérieurs,
  - Méthodes proximales (itératives),
  - Méthodes homotopiques...
- Ici $\gamma_k = \gamma c_k \sqrt{\frac{\log K}{n}}$ avec γ à bien choisir $\Longrightarrow$ méthodes homotopiques permettant de calculer la solution simultanément pour tout γ !

Homotopie



- Trajectoire de λ est linéaire par morceaux.
- Deux premières observations :
  - Quand $\gamma \mapsto +\infty$, $\lambda = 0$.
  - Premier coefficient non nul correspond au ϕ_k le mieux corrélé.
- Les contraintes d'optimalité des problèmes primal et dual permettent de déterminer :
  - à support fixé, la direction de linéarité jusqu'à rupture des conditions,
  - la modification du support à effectuer pour satisfaire à nouveau ces contraintes.
- Algorithme d'homotopie utilise ces observations pour construire le chemin de régularisation.

Calibration de γ



- Choix fin des c_k dans $\gamma_k = \gamma \frac{c_k}{\sqrt{n}} \implies \gamma \simeq 1$ bon choix !

Conclusion

- Exemples de problèmes d'optimisation en estimation statistiques :
  - Optimisation MM, programmation dynamique et optimisation convexe...
  - Trois types d'optimisation assez générales mais autres techniques possibles...
  - Algorithme déterministe $\rightarrow$ algorithme stochastique...
- Rôle de la partie optimisation :
  - Crucial pour la partie numérique,
  - Prise en compte de plus en plus fréquente de la faisabilité numérique dans la théorie...
- Perspectives :
  - Convergence !
  - Recherche de meilleurs outils existants ou à construire...
  - Problème des architectures matérielles ?