Segmentation d'images hyperspectrales par mélange de gaussiennes et sélection de modèles

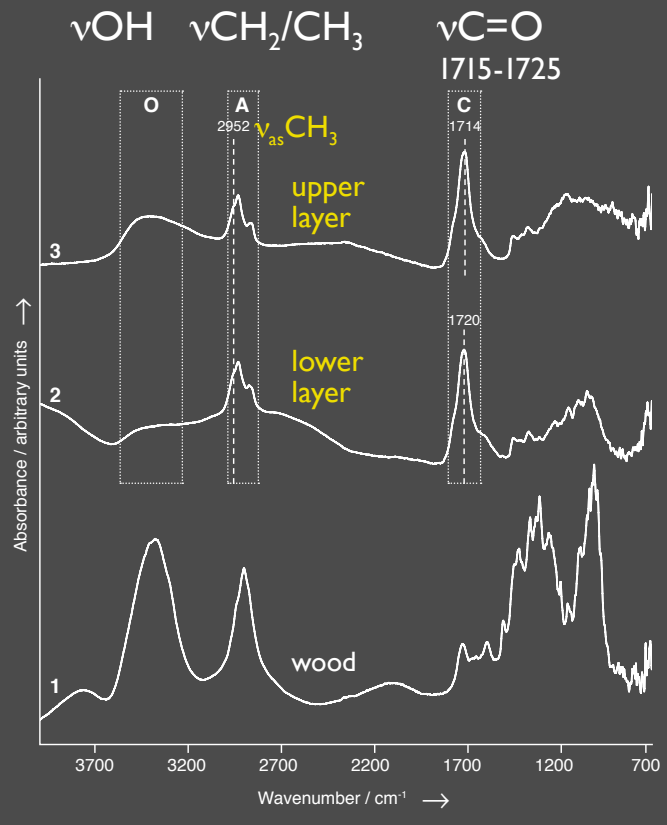
E. Le Pennec
(SELECT - INRIA Saclay / Université Paris Sud)
et
S. Cohen (IPANEMA - Soleil Saclay)
avec
G. Celeux et P. Massart (SELECT - INRIA Saclay / Université Paris Sud)
et C. Maugis (INSA Toulouse)

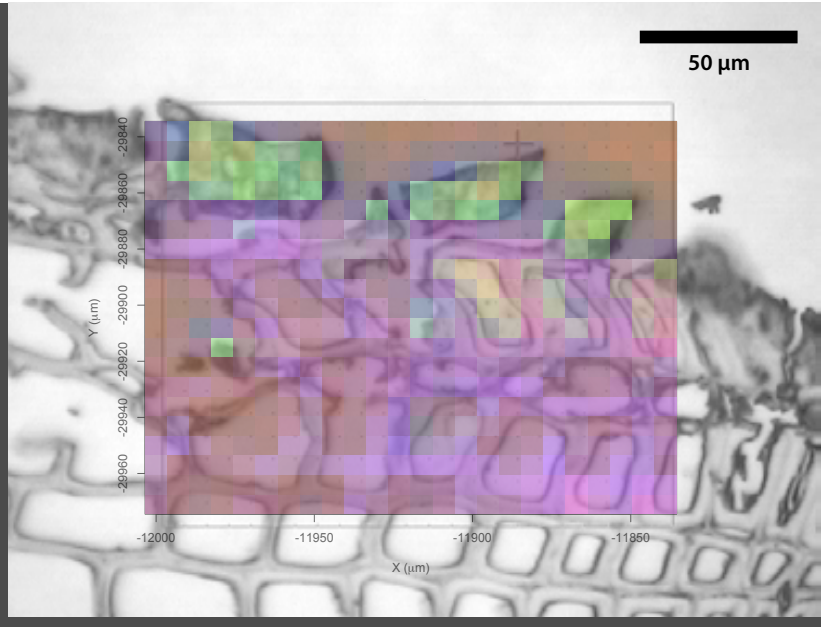
INRIA

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...

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 :

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

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

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.

Sélection de modèles

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

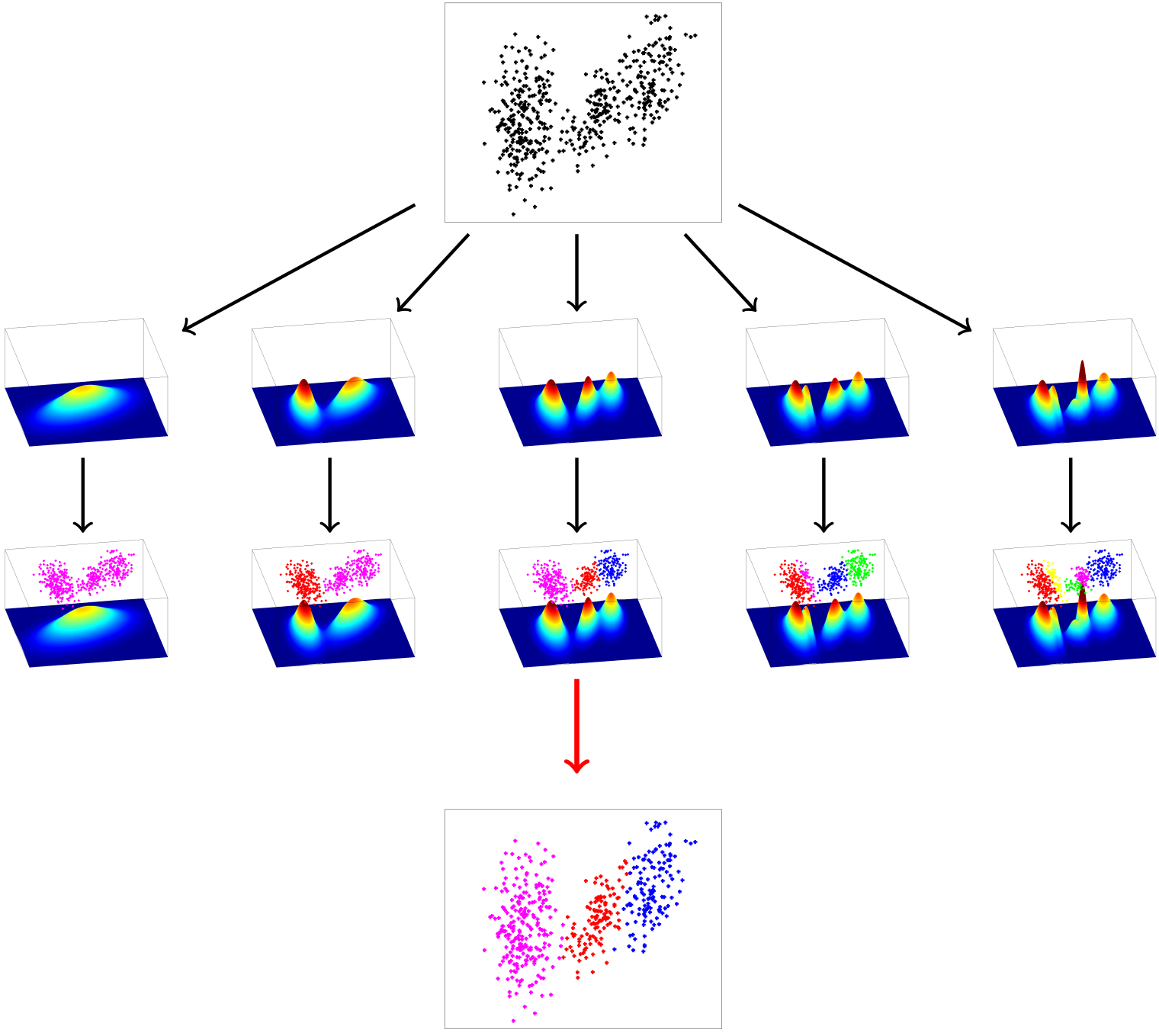
$$\widehat{m} = \operatorname{argmin} - \ln(\widehat{s}_m) + \operatorname{pen}(m).$$

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

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

  - Consistance de la classification si $\ln\ln(n)$ dans la pénalité...

Méthodologie



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 pour $\pi_k(x)$:
  - 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
  - l'optimisation numérique (EM + programmation dynamique),
  - le contrôle théorique : pour κ assez grand,

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

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 $m \in \mathcal{M}$,
  - estimation par maximum de vraisemblance d'une densité $\hat{s}_m$ pour chaque modèle S_m :

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

  - sélection d'un modèle $\widehat{m}$ par

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

  avec $\operatorname{pen}(m)$ à définir...
- Résultat d'estimation de densité du type

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

Theorem

Under Assumption (H_{β_ϕ}) : the existence, for a given $\beta_\phi \geq 1$, of a non-decreasing function $\phi_m(\delta, \beta_\phi)$ such that $\delta \mapsto \frac{1}{\beta_\phi \delta}\phi_m(\delta, \beta_\phi)$ is non-increasing on $(0, +\infty)$ and for every $\sigma \in \mathbb{R}^+$ and every $s_m \in S_m$

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

This assumption implies the following theorem (up to a technical measurability condition) :

Theorem

Assume we observe (X_i, Y_i) with unknown conditional s_0. Let $(S_m)_{m\in\mathcal{M}}$ a at most countable model collection.

Assume that there is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative number such that $\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty \qquad (K)$

and, under assumption (H), let σ_m be the unique root of $\frac{1}{\sigma}\phi_m(\sigma, \beta_\phi) = \frac{1}{\beta_\sigma}\sqrt{n}\sigma. \qquad (H_\sigma)$

and let $\widehat{s_m}$ be a ρ maximum likelihood minimizer in S_m : $\sum_{i=1}^n -\ln(\widehat{s_m}(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^n -\ln(s_m(Y_i|X_i))\right) + \rho$

For any $\rho \in (0,1)$, any $\beta_x > 0$ and any $C_1 > 1$, there are two absolute constants κ_0 and C_2 such as soon as for every model $m \in \mathcal{M}$

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

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 $\mathbb{E}\left[\mathfrak{d}_\rho^{2\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n}\right) + C_2\frac{\Sigma}{n} + \frac{\rho}{n}.$

Théorème

- Inégalité oracle

$$\mathbb{E}\left[\mathfrak{d}_{\rho}^{2\otimes_n}(s_0,\widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0,s_m)+\frac{\mathrm{pen}(m)}{n}\right)+C_2\frac{\Sigma}{n}+\frac{\rho}{n}$$

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 $\mathfrak{d}_{\rho}^{2\otimes_n}$: distance de Kullback "tensorisée" et relaxation.
- $n\sigma_m$ 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

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

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

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

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

De plus, si $\forall\omega \in \Omega, s(\omega) = 0 \implies t(\omega) = 0$

$$d^2_\lambda(s, t) \leq KL_\lambda(s, t) \leq \left(2 + \ln\left\|\frac{s}{t}\right\|_\infty\right) d^2_\lambda(s, t).$$

Densités conditionnelles

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

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

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

- 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

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

- On retrouve exactement le théorème classique si $s(\cdot|X_i) = s(\cdot)$.
- Bon "scaling" de $\mathfrak{d}_\rho^{2\otimes_n}(s_{,0}\, \widehat{s}_{\widehat{m}})$ et $KL^{\otimes_n}(s_0, s_m)$ avec n.
- Pb dans Bigot et al avec Hellinger : $\frac{1}{n} d^2(s_0, \widehat{s}_{\widehat{m}}) \leq \frac{2}{n}$!
- Pas ce soucis avec Bhattacharyya-Renyi de Kolaczyk et Barron mais pas de distance "intégrée"...

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 lié à la distance de Hellinger tensorisée $(d^{\otimes n} = \sqrt{d^{2\otimes n}} = \sqrt{\mathbb{E}\left[\frac{1}{n}\sum d^2(s(X_i,\cdot), s'(X_i,\cdot))\right]})$.
  - 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 $\displaystyle\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty$
- Contrainte (classique) sur la pénalité

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

Esquisse de preuve

- Preuve très proche de celle du théorème 7.11 du livre « Concentration Inequalities and Model Selection » de P. Massart.
- Pour toute fonction $g(x, y)$, on note $P_n^{\otimes n}(g)$ son processus empirique

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

et $P^{\otimes n}(g)$ l'espérance de ce processus

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

et $\nu_n^{\otimes n}(g) = P_n^{\otimes n}(g) - \nu_n^{\otimes n}(g)$ le processus recentré.

- On note
  - $\widehat{s}_m = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n}(-\log s_m) = \operatorname{argmin}_{s_m \in S_m} P_n^{\otimes n}\left(-\log \frac{s_m}{s_0}\right)$
  - $\overline{s}_m = \operatorname{argmin}_{s_m \in S_m} P^{\otimes n}\left(-\log \frac{s_m}{s_0}\right) = \operatorname{argmin}_{s_m \in S_m} KL^{\otimes n}(s_0, s_m)$.
- On pose

$$\widehat{g}_m = -\ln\left(\frac{\widehat{s}_m}{s_0}\right) \qquad \overline{g}_m = -\ln\left(\frac{\overline{s}_m}{s_0}\right).$$

Majoration des log-vraisemblances

- Soit $m \in \mathcal{M}$, pour tout m' tel que

$$P_n^{\otimes n}(\widehat{g}_{m'}) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes n}(\widehat{g}_m) + \frac{\mathrm{pen}(m)}{n}$$

on a

$$P_n^{\otimes n}(\widehat{g}_{m'}) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes n}(\overline{g}_m) + \frac{\mathrm{pen}(m)}{n}$$

ou encore

$$\begin{aligned} P^{\otimes n}(\widehat{g}_{m'}) + (P_n^{\otimes n}(\widehat{g}_{m'}) - P^{\otimes n}(\widehat{g}_{m'})) + \frac{\mathrm{pen}(m')}{n} & \\ \leq P^{\otimes n}(\overline{g}_m) & \\ + (P_n^{\otimes n}(\overline{g}_m) - P^{\otimes n}(\overline{g}_m)) + \frac{\mathrm{pen}(m)}{n} & \end{aligned}$$

- Soit

$$\begin{aligned} P^{\otimes n}(\widehat{g}_{m'}) - \nu_n^{\otimes n}(\overline{g}_m) & \\ \leq P^{\otimes n}(\overline{g}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes n}(\widehat{g}_{m'}) - \frac{\mathrm{pen}(m')}{n} & \end{aligned}$$

Inégalité oracle à déviation près

- L'inégalité précédente s'écrit

$$KL^{\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(\overline{g}_m)$$
$$\leq KL^{\otimes n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n}$$
$$- \nu_n^{\otimes n}(\widehat{g}_{m'}) - \frac{\mathrm{pen}(m')}{n}$$

- On a fait apparaître
  - l'erreur intégré de l'estimateur dans le modèle m' : $KL^{\otimes n}(s_0, \hat{s}_{m'})$
  - un processus centré et simple : $-\nu_n^{\otimes n}(\overline{g}_m)$,
  - l'oracle $KL^{\otimes n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n}$
  - un « reste » aléatoire $-\nu_n^{\otimes n}(\widehat{g}_{m'}) - \frac{\mathrm{pen}(m')}{n}$
- Ne reste plus qu'à « contrôler les déviations » de $-\nu_n^{\otimes n}(\widehat{g}_{m'})$ par $\epsilon KL^{\otimes n}(s_0, \hat{s}_{m'}) + \frac{\mathrm{pen}(m')}{n}$...

Déviation et log-vraisemblance

- On doit contrôler

$$-\nu_n^{\otimes n}(\widehat{g}_{m'}) = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{s_0(Y_i|X_i)}{\widehat{s}_{m'}(Y_i|X_i)} - \mathbb{E}\ln\frac{s_0(Y_i|X_i)}{\widehat{s}_{m'}(Y_i|X_i)}\right)$$

- Première idée :

$$\frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{s_0(Y_i|X_i)}{\widehat{s}_{m'}(Y_i|X_i)} - \mathbb{E}\ln\frac{s_0(Y_i|X_i)}{\widehat{s}_{m'}(Y_i|X_i)}\right)$$
$$\leq \sup_{s_{m'}\in S_{m'}}\frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{s_0(Y_i|X_i)}{s_{m'}(Y_i|X_i)} - \mathbb{E}\ln\frac{s_0(Y_i|X_i)}{s_{m'}(Y_i|X_i)}\right)$$

- Beaucoup trop grossier...
- Localisation pour mettre un peu de $KL(s_0, s_{m'})$.

Lemme d'épluchage

- Soit S un ensemble dénombrable, $u \in S$ et $a : S \to \mathbb{R}^+$ tel que $a(u) = \inf_{t \in S} a(t)$. Soit Z un processus indexé par S et soit $B(\sigma) = \{t \in S, a(t) \leq \sigma\}$, supposons que pour tout σ positif la variable aléatoire $\sup_{t \in B(\sigma)} (Z(t) - Z(u))$ a une espérance finie. Alors, pour toute fonction ψ sur $\mathbb{R}^+$ tel que $\psi(x)/x$ soit décroissante sur $\mathbb{R}^+$ et

$$\mathbb{E}\left[\sup_{t \in B(\sigma)} (Z(t) - Z(u))\right] \leq \psi(\sigma), \quad \text{pour tout } \sigma \geq \sigma_\star \geq 0,$$

on a $\forall x \geq \sigma_\star \quad \mathbb{E}\left[\sup_{t \in S} \frac{Z(t) - Z(u)}{x^2 + a^2(t)}\right] \leq 4x^{-2}\psi(x).$

- Applicable pour $S_{m'}$, $Z(t) = \nu_n^{\otimes n}\left(\ln \frac{s_0(Y_i|X_i)}{t_{m'}(Y_i|X_i)}\right)$, $u = \overline{s}_{m'} = \operatorname{argmin}_{s_{m'} \in S_{m'}} KL^{\otimes n}(s_0, s_{m'})$ et $a(t) = \sqrt{KL^{\otimes n}}(u, t)$?

Déviation locale

- Contrôle de

$$\mathbb{E}\Bigg[\sup_{t\in B(\sigma)} \nu_n^{\otimes n}\left(\ln\frac{s_0(Y_i|X_i)}{t(Y_i|X_i)}\right) - \nu_n^{\otimes n}\left(\ln\frac{s_0(Y_i|X_i)}{\overline{s}_{m'}(Y_i|X_i)}\right)\Bigg] = \mathbb{E}\Bigg[\sup_{t\in B(\sigma)} \nu_n^{\otimes n}\left(\ln\frac{\overline{s}_{m'}(Y_i|X_i)}{t(Y_i|X_i)}\right)\Bigg]$$

  avec $B(\sigma) = \{s_{m'} \in S_{m'}, KL(s_0, s_{m'}) \leq \sigma^2\}$.
- Outil de contrôle des déviations : Théorème 6.8 du livre de P. Massart...

Déviations et entropie

- **Théorème :** Soit $\mathcal{F}$ une famille dénombrable de fonctions à valeur réelles. Si ils existent des nombres positifs σ et b tels que pour tout $f \in \mathcal{F}$ et tout entier $k \geq 2$

$$P^{\otimes_n}(|f|^k) \leq \frac{k!}{2}\sigma^2 b^{k-2}$$

et que de plus pour tout nombre positif δ il existe un ensemble $B(\delta)$ de crochets recouvrant $\mathcal{F}$ tel que pour tout crochet $[g^-, g^+] \in B(\delta)$ et tout entier $k \geq 2$

$$P^{\otimes_n}(|g^+ - g^-|^k) \leq \frac{k!}{2}\sigma^2 b^{k-2}$$

Soit $e^{H(\delta)}$ le cardinal de ce recouvrement Il existe un constante absolue κ telle pour tout $\epsilon \in (0, 1]$ et tout ensemble mesurable A A avec $P[A] > 0$, on a

$$\mathbb{E}^A\left[\sup_{f\in\mathcal{F}} \nu_n^{\otimes_n}(f)\right] \leq E + \frac{(1+6\epsilon)\sigma}{\sqrt{n}}\sqrt{2\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{2b}{n}\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

où

$$E = \frac{\kappa}{\epsilon}\frac{1}{\sqrt{n}}\int_0^{\epsilon\sigma}\sqrt{H(u)\wedge n}\mathrm{d}u + \frac{2(b+\sigma)}{n}H(\sigma)$$

avec $\kappa \leq 27$.

Hypothèse à vérifier

- Hypothèse nécessaire (c.f. Bernstein) : $\exists \sigma$ et b tels que pour tout

  $f \in \mathcal{F}$ et tout entier $k \geq 2$, $\quad P^{\otimes n}(|f|^k) \leq \frac{k!}{2}\sigma^2 b^{k-2}$.

- Pour $f = \ln \frac{\overline{s}_{m'}}{t}$ et $t \in B(\sigma) = \{s_{m'} \in S_{m'}, KL(s_0, s_{m'}) \leq \sigma^2\}$, impossible à obtenir (caractère non bornée par exemple...).
- **Lemme de van de Geer :** Pour toutes fonctions positives t, u et tout entier $k \geq 2$

$$P\left(\left|\ln\left(\frac{s_0+t}{s_0+u}\right)\right|^k\right) \leq \frac{k!}{2}\left(\frac{9\|\sqrt{t}-\sqrt{u}\|^2_{\lambda,2}}{8}\right)2^{k-2}.$$

- Ici

$$P^{\otimes n}\left(\left|\frac{1}{\rho}\ln\left(\frac{s_0+\frac{\rho}{1-\rho}t}{s_0+\frac{\rho}{1-\rho}u}\right)\right|^k\right) \leq \frac{k!}{2}\left(\frac{9d^{2\otimes n}(t,u)}{8\rho(1-\rho)}\right)\left(\frac{2}{\rho}\right)^{k-2}$$

$$P^{\otimes n}\left(\left|\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0+\rho t}{(1-\rho)s_0+\rho u}\right)\right|^k\right) \leq \frac{k!}{2}\left(\frac{9d^{2\otimes n}(t,u)}{8\rho(1-\rho)}\right)\left(\frac{2}{\rho}\right)^{k-2}$$

Jensen-Shannon

- Inégalité de base

$$P^{\otimes_n}\left(\left|\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0+\rho t}{(1-\rho)s_0+\rho u}\right)\right|^k\right)\leq\frac{k!}{2}\left(\frac{9d^{2\otimes_n}(t,u)}{8\rho(1-\rho)}\right)\left(\frac{2}{\rho}\right)^{k-2}$$

- Lien immédiat avec la distance de Jensen-Shannon tensorisée

$$\mathfrak{d}_\rho^{2\otimes_n}(s_0,t)=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\frac{1}{\rho}KL(s_0(\cdot|X_i),\rho t(\cdot|X_i)+(1-\rho)s_0(\cdot|X_i))\right]$$

et l'entropie à crochet par rapport à la distance de Hellinger tensorisée

$$d^{2\otimes_n}(s_0,t)=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n d^2(s_0(\cdot|X_i),t(\cdot|X_i))\right].$$

- Retour au début de la preuve...

Majoration des « log-vraisemblances »

- On pose $\widehat{f}_m = -\frac{1}{\rho} \ln \frac{\rho \widehat{s}_m + (1-\rho) s_0}{s_0} \leq -\ln \frac{\widehat{s}_m}{s_0} = \widehat{g}_m$
- Soit $m \in \mathcal{M}$, pour tout m' tel que

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

- Soit

$$\begin{aligned} &P^{\otimes n}(\widehat{f}_{m'}) - \nu_n^{\otimes n}(\overline{g}_m) \\ &\qquad \leq P^{\otimes n}(\overline{g}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes n}(\widehat{f}_{m'}) - \frac{\mathrm{pen}(m')}{n} \end{aligned}$$

Inégalité oracle à déviation près

- L'inégalité précédente s'écrit

$$\mathfrak{d}_\rho^{2\otimes n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes n}(\overline{g}_m) \leq KL^{\otimes n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes n}(\widehat{f}_{m'}) - \frac{\mathrm{pen}(m')}{n}$$

- On a fait apparaître
  - l'erreur intégré de l'estimateur dans le modèle m' : $\mathfrak{d}_\rho^{2\otimes n}(s_0, \hat{s}_{m'})$
  - un processus centré et simple : $-\nu_n^{\otimes n}(\overline{g}_m)$,
  - l'oracle $KL^{\otimes n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n}$
  - un « reste » aléatoire $-\nu_n^{\otimes n}(\widehat{f}_{m'}) - \frac{\mathrm{pen}(m')}{n}$
- On peut alors « contrôler les déviations de » $-\nu_n^{\otimes n}(\widehat{f}_{m'})$ par $\epsilon\mathfrak{d}_\rho^{2\otimes n}(s_0, \hat{s}_{m'}) + \frac{\mathrm{pen}(m')}{n}$ dès que $\mathrm{pen}(m') \geq \kappa(n\sigma_{m'}^2 + x_m)$...

Déviation de $\nu_n^{\otimes n}(\widehat{f}_{m'})$

- On pose $\widetilde{s}_{m'} = \operatorname{argmin}_{s \in S_{m'}} d^{2\otimes n}(s_0, s_{m'})$ et $\widetilde{f}_{m'} = \frac{1}{\rho} \ln \frac{s_0}{\rho \widetilde{s}_{m'} + (1-\rho) s_0}$
- $-\nu_n^{\otimes n}(\widehat{f}_{m'}) = \left(-\nu_n^{\otimes n}(\widehat{f}_{m'}) + \nu_n^{\otimes n}(\widetilde{f}_{m'})\right) - \nu_n^{\otimes n}(\widetilde{f}_{m'})$
- $\nu_n^{\otimes n}(\widetilde{f}_{m'})$ facile à contrôler avec Bernstein :

$$\mathbb{P}\left\{ -\nu_n^{\otimes n}(\widetilde{f}_m) > \frac{3}{2\sqrt{\rho(1-\rho)}} \sqrt{d^{2\otimes n}(s_0, \widetilde{s}_m)} \sqrt{\frac{x_m + x}{n}} + \frac{2}{\rho} \frac{x_m + x}{n} \right\}$$
$$\leq e^{-x_m - x}$$

qui implique, en utilisant la définition de $\widetilde{s}_m$,

$$\mathbb{P}\left\{ \nu_n^{\otimes n} \left(\frac{-\widetilde{f}_m}{y_m^2 + 4 d^{2\otimes n}(s_0, \widehat{s}_m)} \right) > \frac{3}{4\sqrt{\rho(1-\rho)} \kappa'_d y_m} \sqrt{\frac{x_m + x}{n}} + \frac{2}{\rho} \frac{x_m + x}{n y_m^2} \right\}$$
$$\leq e^{-x_m - x}.$$

Contrôle par un sup

- On pose
  $W(s) = \nu_n^{\otimes n}\left(\frac{1}{\rho}\ln(\rho s + (1-\rho)s_0) - \frac{1}{\rho}\ln(\rho\widetilde{s}_{m'} + (1-\rho)s_0)\right)$.
- $W(\widehat{s}_{m'}) = -\nu_n^{\otimes n}(\widehat{f}_{m'}) + \nu_n^{\otimes n}(\widetilde{f}_{m'})$
- Contrôle direct de $\sup_{s_{m'}\in S_{m'}} W(s)$ insuffisant.
- Théorème de contrôle de la déviation avec entropie à crochet : pour tout A tel que $P(A) > 0$

$$\mathbb{E}^A\left[\sup_{s_{m'}\in S_{m'}, d^{\otimes n}(\tilde{s}_{m'}, s_{m'})\leq\sigma} W(s)\right] \leq E + \psi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right)$$

avec

$$\psi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right) = \frac{(1+6\epsilon)3\sigma}{2\sqrt{2\rho(1-\rho)}\sqrt{n}}\sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n}\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

$$E \leq \frac{81}{\beta_\phi\sqrt{2}}\frac{1}{\sqrt{n}}\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}\left(u, S_{m'}(\widetilde{s}_{m'}, \sigma)\right) \wedge n}\mathrm{d}u$$

$$+ \frac{\frac{8}{\rho} + \frac{3}{\sqrt{\rho(1-\rho)}}}{n} H_{[\cdot],d^{\otimes n}}\left(\sigma, S_{m'}(\widetilde{s}_{m'}, \sigma)\right).$$

Déviation et épluchage

- De $\phi_{m'}(\sigma_{m'}) = \sqrt{n}\sigma$ on déduit

$$\forall \sigma > \sigma_{m'}, \quad E \leq \left(\frac{81}{\sqrt{2}} + \frac{8}{\rho} + \frac{3}{\sqrt{\rho(1-\rho)}} \right) \frac{\phi_{m'}(\sigma)}{\sqrt{n}}.$$

- Le lemme d'épluchage donne alors $\forall y_m \geq \sigma_m$

$$\mathbb{E}^A \left[\nu_n^{\otimes n} \left(\frac{\widetilde{f}\, m - \widehat{f}\, m}{y_{m'}^2 + 4d^{2\otimes n}(s_0, \widehat{s}_{m'})} \right) \right] \leq \frac{4\kappa_1'' \sigma_{m'}}{y_{m'}} + \frac{4\kappa_2''}{\sqrt{n} y_{m'}} \sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}} \right)} + \frac{16}{n\rho y_{m'}^2} \ln\left(\frac{1}{\mathbb{P}\{A\}} \right).$$

- En utilisant si $\forall A, P(A) > 0,\ \mathbb{E}^A[Z] \leq \phi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}} \right) \right)$ alors $\forall x > 0, \mathbb{P}[Z \geq \phi(x)] \leq e^{-x}$, on obtient

$$\mathbb{P}\left\{ \nu_n^{\otimes n} \left(\frac{-\widehat{f}_{m'} + \widetilde{f}_{m'}}{y_{m'}^2 + 4d^{2\otimes n}(s_0, \widehat{s}_{m'})} \right) > \frac{4\kappa_1'' \sigma_{m'}}{y_{m'}} + 4\kappa_2'' \sqrt{\frac{x_{m'} + x}{n y_{m'}}} + \frac{16}{\rho} \frac{x_{m'} + x}{n y_{m'}} \right.$$

Contrôle sur tous les modèles

- En combinant avec Bernstein, on obtient

$$\mathbb{P}\left\{\nu_n^{\otimes n}\left(\frac{-\widehat{f}_{m'}}{y_{m'}^2+\kappa_0' d^{2\otimes n}(s_0,\widehat{s}_{m'})}\right) > \frac{\kappa_1'\sigma_{m'}}{y_{m'}} + \kappa_2'\sqrt{\frac{x_{m'}+x}{ny_{m'}^2}} + \frac{18}{\rho}\frac{x_{m'}+x}{ny_{m'}^2}\right\}$$
$$\leq 2e^{-x_{m'}-x}$$

- Par une borne d'union et Kraft, on en déduit

$$\mathbb{P}\left\{\forall m\in\mathcal{M},\quad \nu_n^{\otimes n}\left(\frac{-\widehat{f}_{m'}}{y_{m'}^2+\kappa_0' d^{2\otimes n}(s_0,\widehat{s}_{m'})}\right) > \frac{\kappa_1'\sigma_{m'}}{y_{m'}} + \kappa_2'\sqrt{\frac{x_{m'}+x}{ny_{m'}^2}} + \frac{18}{\rho}\frac{x_{m'}+x}{ny_{m'}^2}\right\} \leq 2\Sigma e^{-x}$$

- La preuve se finit alors de manière calculatoire en posant $y_{m'} = \kappa^{-1}\sqrt{\sigma_{m'}^2 + \frac{x_{m'}+x}{n}}$ et en utilisant $C_\rho d^{2\otimes n}(s_0,\widehat{s}_{m'}) \leq \mathfrak{d}_\rho^{2\otimes n}(s_0,\widehat{s}_{m'})$...

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 $d^{\sup} = \sqrt{d^{2\sup}} = \sqrt{\sup_x d^2(s(x,\cdot), s'(x,\cdot))}$,
  - 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...).
  - implication : $n\phi_m^2(\sigma_m) \leq \kappa' \ln(n) \dim(S_m)$.
- Codage de la collection avec $x_m \leq \kappa''|\mathcal{P}| \leq \frac{\kappa''}{K-1}\dim(S_m)$.
- Condition sur la pénalité :

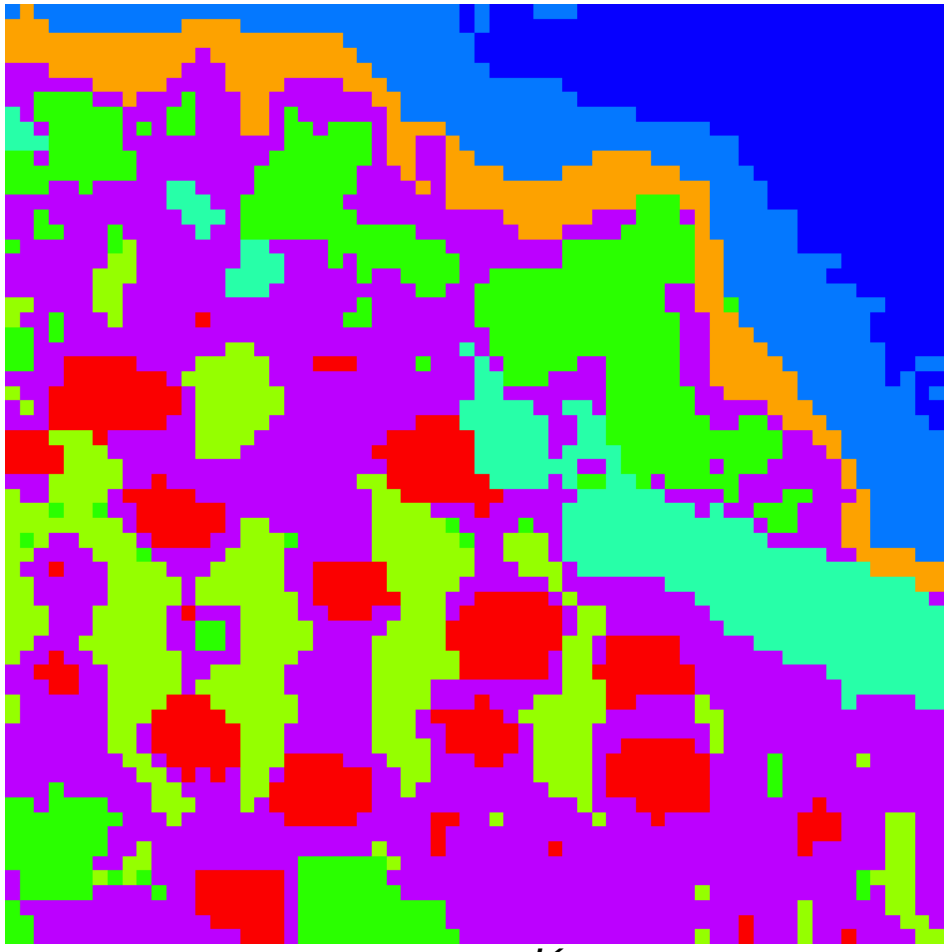
$$\mathrm{pen}(m) \geq \left(\kappa' \ln(n) + \frac{\kappa''}{K-1}\right)\dim(S_m).$$

Segmentation automatique

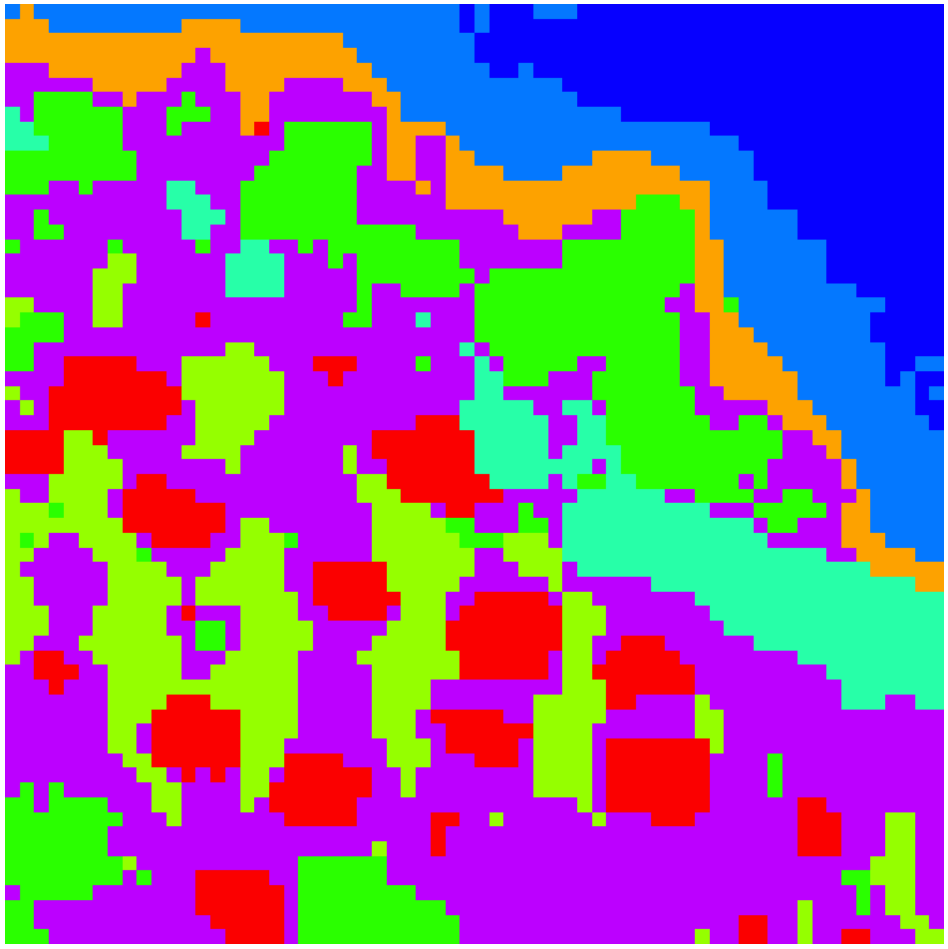
- Résultat numérique :

Prise en compte du caractère spatial

Sans



Avec



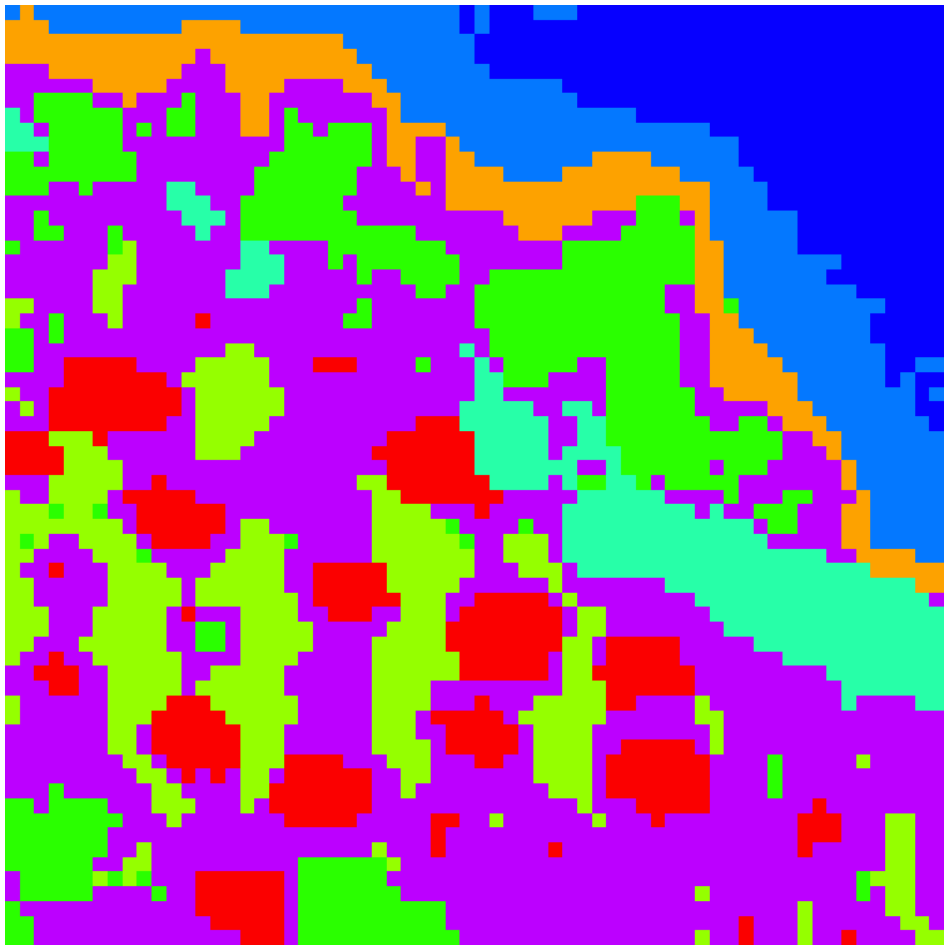
- $K = 8$, $[L_k DA_k]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente...

Segmentation automatique

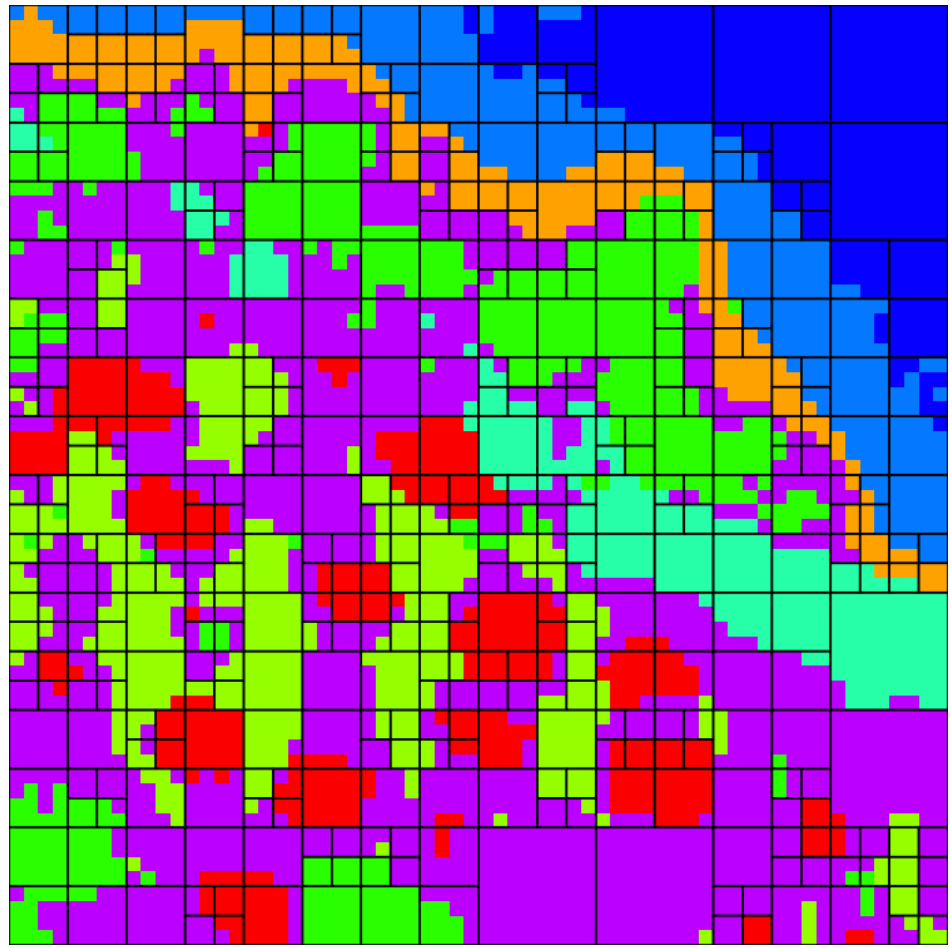
- Résultat numérique :

Prise en compte du caractère spatial

Sans

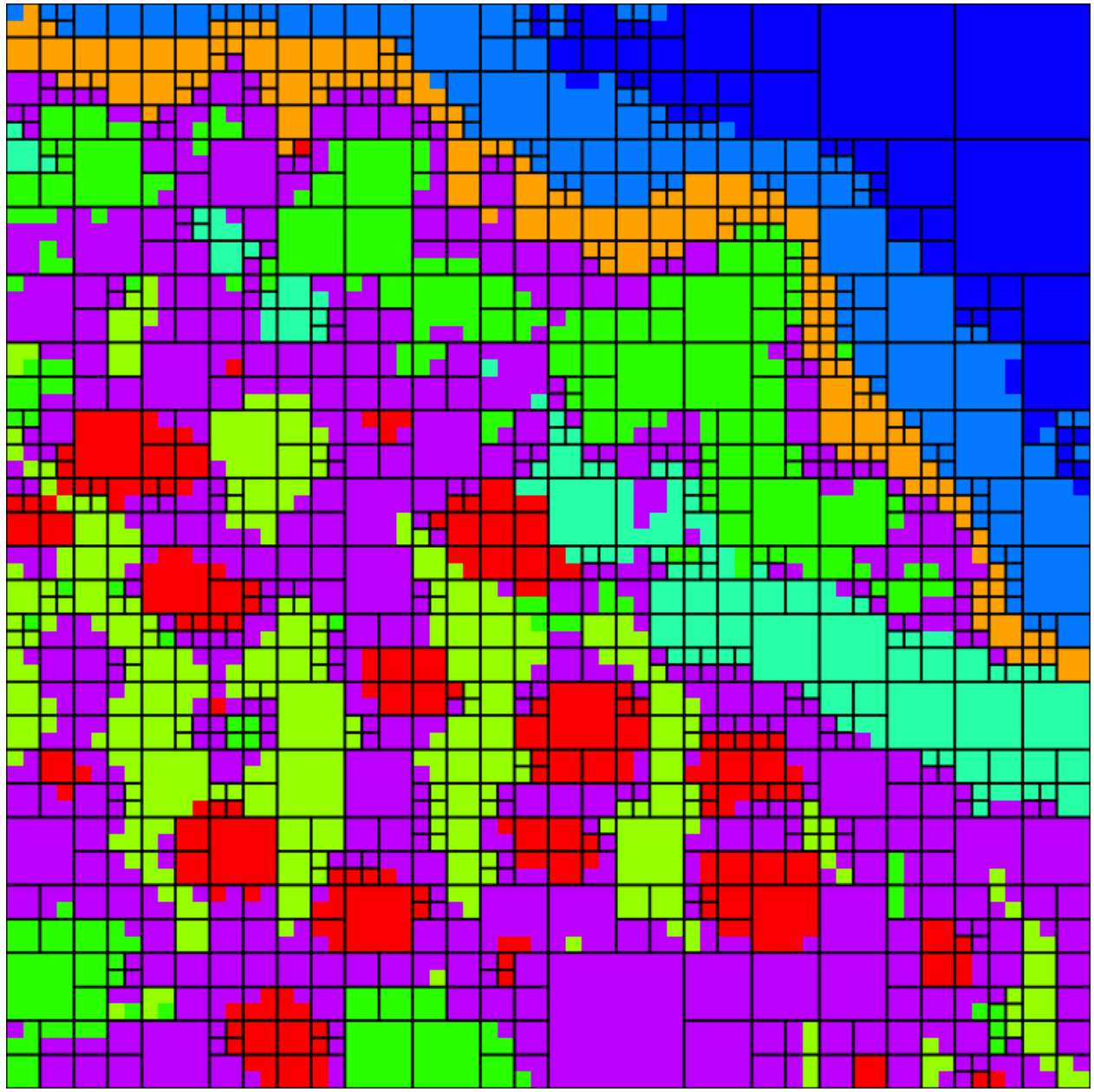


Avec

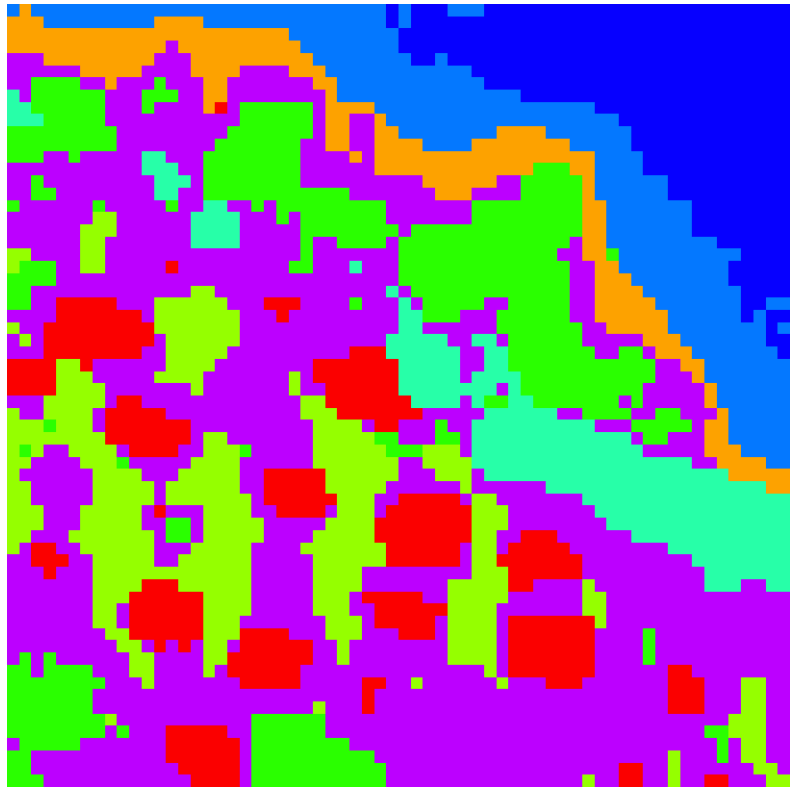


- $K = 8$, $[L_k DA_k]^K$ et partition optimale.
- Calibration de la pénalité par heuristique de pente...

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 !

Bibliographie

- Papier de départ présentant une solution complète :
  - A. Antoniadis, J. Bigot and R. von Sachs (2008) A multiscale approach for statistical characterization of functional images, Journal of Computational and Graphical Statistics, 18(1), pp. 216-237
  - Discussion paper qui lui correspond.
- Segmentation et sélection de modèles :
  - E.D. Kolaczyk, J. Ju and S. Gopal (2005). Multiscale, multigranular statistical image segmentation. Journal of the American Statistical Association, 100, 1358-1369.
  - E.D. Kolaczyk and R.D. Nowak (2004). Multiscale likelihood analysis and complexity penalized estimation. Annals of Statistics, 32, 500-527
- Sélection de modèles par MDL :
  - A.R. Barron, C. Huang, J. Q. Li and Xi Luo (2008). MDL Principle, Penalized Likelihood, and Statistical Risk. In Feschrift for Jorma Rissanen. Presented to Rissanen Nov. 2007.
- Sélection de modèles et entropie (à crochet) :
  - P. Massart (2003). Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII.
  - C. Maugis and B. Michel. (2009). A non asymptotic penalized criterion for Gaussian mixture model selection. Accepté à ESAIM : P&S (et l'erratum)
  - S. Cohen and E. Le Pennec (201 ?)
- Entropie de $SO(n)$:
  - J. Szarek (1998). Metric entropy of homogeneous spaces. Banach Center Pub., 43, 395–410