Conditional density estimation by penalized maximum likelihood model selection

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

Valparaiso
24/03/2012

informatiques mathématiques

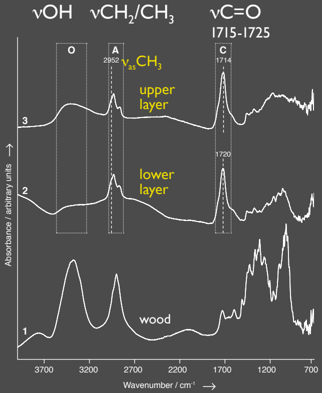
Outline

1. Hyperspectral image segmentation (preview of CLAPEM talk...)
2. Conditional density estimation by a penalized maximum likelihood approach
3. Abstract model selection theorem and related tools
4. Application to partition based conditional density estimation

A. Stradivari (1644 - 1737)

Provigny (1716)

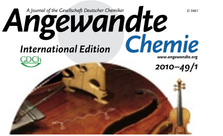
A. Giordan © Cité de la Musique



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.

Hyperspectral image segmentation with GMM

- *Classical* stochastic model of spectrum $\mathcal{S}$:
  - K spectrum classes,
  - with proportion π_k for each class ($\sum_{k=1}^{K} \pi_k = 1$),
  - Gaussian law $\mathcal{N}(\mu_k, \Sigma_k)$ within each class (strong assumption!)
- Heuristic: true density s_0 of $\mathcal{S}$ close from

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

- Goal: estimate all parameters (K, π_k, μ_k and Σ_k) from the data.
- Why: yields a classification/segmentation by a maximum likelihood principle

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

- Typical result in term of density estimation and not classification...

Gaussian Mixture Model

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

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

- Gaussian Mixture Model $S_m = \{s_m\}$ specified by
  - a number of classes K,
  - a structure for the means μ_k and the covariance matrices $\Sigma_k = L_k D_k A_k D_k'$
- Structure $[\mu\, L\, D\, A]^K$: structural constraints (know, common or free values...) on the means μ_k , the volumes L_k, the diagonalization basis D_k and the rescaled eigenvalues A_k plus compactness and condition number assumptions.
- GMM S_m: parametric model of dimension $(K-1) + \dim([\mu\, L\, D\, A]^K)$.
- Maximum likelihood estimation by EM algorithm of:
  - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D_k'$ for each class
  - and the mixing proportions π_k

Model selection

- How to choose the *good* model S_m:
  - the number of classes K,
  - the structure model $[\mu\, L\, D\, A]^K$?
- Penalized model selection principle:
  - Choice of a collection of models $S_m = \{s_m\}$ with $m \in \mathcal{S}$,
  - Maximum likelihood estimation of a density $\hat{s}_m$ for each model S_m,
  - Selection of a model $\widehat{m}$ by

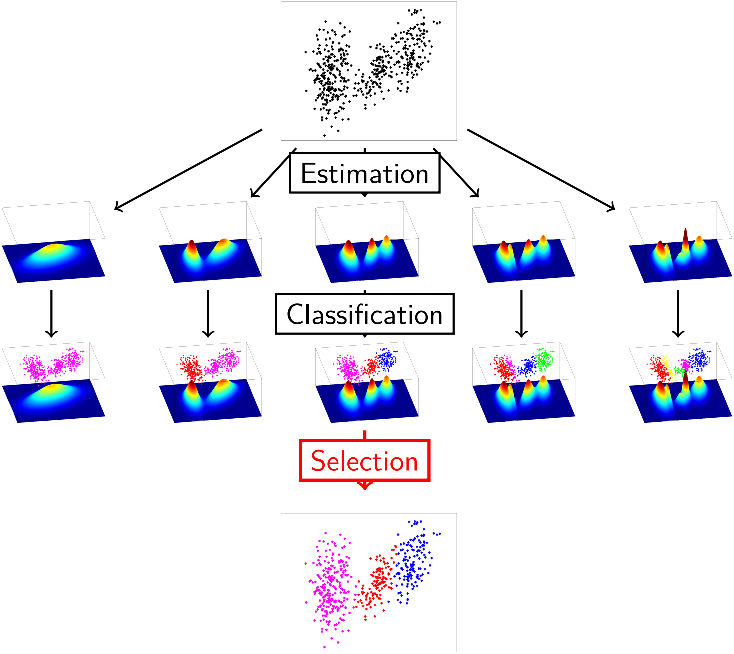
$$\widehat{m} = \operatorname{argmin} - \ln(\widehat{s}_m) + \operatorname{pen}(m).$$

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

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

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

Methodology



Segmentation and Spatialized GMM

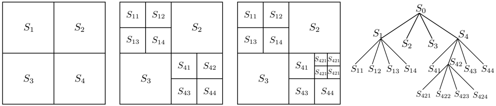
- Initial goal: segmentation $\neq$ clustering.
- Idea of Kolaczyk et al (cf Bigot): take into account the spatial position x of the spectrum in the mixing proportions .
- Conditional density model:

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

- Estimation from the data:
  - the mean μ_k and the covariance matrix $\Sigma_k = L_k D_k A_k D_k'$ for each class
  - and the mixing proportion functions $\pi_k(x)$.
- Non parametric model ($\pi_k(x)$ function): regularization required!
- Model selection principle...

Spat. GMM and hierarchical partition

- How to choose the *right* model S_m ?:
  - the number of classes K,
  - the structure model $[\mu\, L\, D\, A]^K$,
  - the structure of the mixing proportion functions $\pi_k(x)$.
- Simple structure for $\pi_k(x)$: $\pi_k(x) = \sum_{\mathcal{R}\in\mathcal{P}} \pi_k[\mathcal{R}]\chi_{\{x\in\mathcal{R}\}} = \pi_k[\mathcal{R}(x)]$
  - piecewise constant on a *hierarchical* partition,
  - efficient optimization algorithm,
  - good approximation properties.



- $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
- Penalty $\mathrm{pen}(m) = \kappa \ln(n) \dim(S_m)$ allows
  - a numerical optimization scheme (EM + dynamic programing)
  - a theoretical control: for κ large enough

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

Conditional density and selection

- General framework: observation of (X_i, Y_i) with X_i independent and Y_i cond. independent of law of density $s_0(y|X_i)$.
- Goal: estimation of $s_0(y|x)$.
- Penalized model selection principle:
  - choice of a collection of cond. dens. models $S_m = \{s_m(y|x)\}$ with $m \in \mathcal{S}$,
  - Maximum likelihood estimation of a cond. density $\hat{s}_m$ for each model S_m:
  $$\hat{s}_m = \operatorname*{argmin}_{s_m \in S_m} - \sum_{i=1}^{n} \ln s_m(Y_i|X_i)$$
  - Selection of a model $\widehat{m}$ by
  $$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{S}} - \sum_{i=1}^{n} \ln \widehat{s}_m(Y_i|X_i) + \mathrm{pen}(m).$$
  with $\mathrm{pen}(m)$ *well chosen*.
- Conditional density estimation result of type:
$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$
- Short biblio: Rosenblatt, Fan et al., de Gooijer and Zerom, Efromovitch, Brunel, Comte, Lacour... / Plugin, direct estimation, L^2, minimax, censure...

Theorem

Assumption (H): For every model S_m in the collection $\mathcal{S}$, there is a non-decreasing function $\phi_m(\delta)$ such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0,+\infty)$ and for every $\sigma \in \mathbb{R}^+$ and every $s_m \in S_m$

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

Assumption (K): There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative number such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty$$

Theorem

Assume we observe (X_i, Y_i) with unknown conditional s_0. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ a at most countable collection of conditional density sets. Assume Assumptions (H), (K) and (S) hold.
Let $\widehat{s}_m$ be a δ -log-likelihood minimizer in S_m:

$$\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^n -\ln(s_m(Y_i|X_i))\right) + \delta$$

Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there are two constants κ_0 and C_2 depending only on ρ and C_1 such that,
as soon as for every index $m \in \mathcal{M}$ $\mathrm{pen}(m) \geq \kappa\left(\mathfrak{D}_m + x_m\right)$ with $\kappa > \kappa_0$
where $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$,
the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ defined by

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

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

Simplified Theorem...

- Oracle inequality:

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

  as soon as

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

  where $\mathfrak{D}_m$ measure the complexity of the model S_m (entropy term) and x_m the coding cost within the collection.
- *Distances* used $KL^{\otimes_n}$ and $JKL_\rho^{\otimes_n}$: *tensorized* Kullback divergence and *Jensen-Kullback* divergence.
- $\mathfrak{D}_m$ linked to the *bracketing entropy* of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.
- Often $\mathfrak{D}_m \propto (\log n)\dim(S_m)$...

Kullback, Hellinger and extensions

- Model selection oracle inequality of type

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

- Density: Hellinger $d^2(s,s')$ (or affinity) (Kolaczyk, Barron, Bigot) on the left...
- Refinement with a bounded convexification of KL: $JKL(s,s') = 2KL(s,(s'+s)/2)$ (Massart, van de Geer)
- Jensen-Kullback-Leibler: generalization to $JKL_\rho(s,s') = \frac{1}{\rho}KL(s,\rho s'+(1-\rho)s)$.
- **Prop.:** For all probability measures $sd\lambda$ and $td\lambda$ and all $\rho\in(0,1)$

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

  with $C_\rho = \frac{1}{\rho}\min(\frac{1-\rho}{\rho},1)\left(\ln\left(1+\frac{\rho}{1-\rho}\right)-\rho\right)$.
- $C_\rho \simeq 1/5$ if $\rho\simeq 1/2$.

Tensorized divergences

- Need to adapt to conditional density design:
  - Divergence on the product density conditioned on the design (Kolaczyk, Bigot).
  - *Tensorization* principle and expectation on the design: design:

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

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

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

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

- Classical density estimation theorem if $s(\cdot|X_i) = s(\cdot)$.

Penalty and complexities

- Oracle inequality:

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

- A good $\mathrm{pen}(m)$ should be of order $\mathbb{E}\left[\left|KL^{\otimes_n}(s_0, \widehat{s}_m) - \mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_m)\right]\right|\right]$ (variance term).
- Control in expectation requires a larger $\mathrm{pen}(m)$:
  - with an intrinsic term corresponding to the complexity of the model (upper bound of the variance/deviation bound),
  - and with a term corresponding to the complexity of the collection (simultaneous control on all the collection/union bound)
- Complexity used here:
  - Model (entropy): $\mathfrak{D}_m$ defined from the *bracketing entropy* $H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)$ of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.
  - Collection (coding): Kraft type inequality $\sum_{m \in \mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- Classical constraint on the penalty

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

Bracketing entropy and complexity

- Bracketing entropy: $H_{[\cdot],d^{\otimes_n}}(\epsilon, S)$ = logarithm of the minimum number of brackets $[t_i^-, t_i^+]$ such that
  - $\forall i, d^{\otimes_n}(t_i^-, t_i^+) \leq \epsilon$
  - $\forall s \in S, \exists i, t_i^- \leq s \leq t_i^+$

  where $d^{\otimes_n} = \sqrt{d^{2\otimes_n}} = \sqrt{\mathbb{E}\left[\frac{1}{n}\sum d^2(s(\cdot|X_i), s'(\cdot|X_i))\right]}$ is the tensorized Hellinger distance.
- Model $S_m \Rightarrow$ Local model $S_m(s_m, \sigma) = S_m \cap \{s, d^{\otimes_n}(s_m, s) \leq \sigma\}$.
- Assumption (H): for all model S_m, there is a non decreasing $\phi_m(\delta)$ such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non increasing $(0, +\infty)$ and such that for all $\sigma \in \mathbb{R}^+$ and all $s_m \in S_m$

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

- Complexity $\mathfrak{D}_m$ def. as $n\sigma_m^2$ with σ_m unique root of $\phi_m(\sigma) = \sqrt{n}\sigma^2$.
- *Key*: Dudley type integral and optimization of a deviation bound.
- Typically, $n\sigma_m^2 \propto (\ln n)\dim(S_m)$...

Sketch of proof

- Close from Theorem 7.11 of Massart's book.
- For all function $g(x,y)$, let $P_n^{\otimes_n}(g)$ be its empirical process

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

$P^{\otimes_n}(g)$ the expectation of this process

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

with (X_i', Y_i') a phantom sample of same law than (X_i, Y_i) but independent and $\nu_n^{\otimes_n}(g) = P_n^{\otimes_n}(g) - P^{\otimes_n}(g)$ the recentred process.

- By definition,

$$KL^{\otimes_n}(s_0, t) = P^{\otimes_n}\left(-\ln\left(\frac{t}{s_0}\right)\right)$$

$$JKL_\rho^{\otimes_n}(s_0, t) = P^{\otimes_n}\left(-\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho t}{s_0}\right)\right)$$

Best(s) model(s)

- Define
  - $\widehat{s}_m = \mathrm{argmin}_{s_m \in S_m} P_n^{\otimes_n}(-\ln s_m) = \mathrm{argmin}_{s_m \in S_m} P_n^{\otimes_n}\left(-\ln \frac{s_m}{s_0}\right)$
  - $\overline{s}_m = \mathrm{argmin}_{s_m \in S_m} P^{\otimes_n}\left(-\ln \frac{s_m}{s_0}\right) = \mathrm{argmin}_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m)$.
- Let

$$kl(\overline{s}_m) = -\ln\left(\frac{\overline{s}_m}{s_0}\right)$$

$$kl(\widehat{s}_m) = -\ln\left(\frac{\widehat{s}_m}{s_0}\right)$$

$$jkl(\widehat{s}_m) = -\frac{1}{\rho}\ln\left(\frac{(1-\rho)s_0 + \rho\widehat{s}_m}{s_0}\right)$$

- By convexity, $jkl(\widehat{s}_m) = -\frac{1}{\rho}\ln\frac{\rho\widehat{s}_m + (1-\rho)s_0}{s_0} \leq -\ln\frac{\widehat{s}_m}{s_0} = kl(\widehat{s}_m)$

Log-likelihood majorization

- Let $m \in \mathcal{S}$, for all m' such that

$$P_n^{\otimes_n}(kl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \leq P_n^{\otimes_n}(kl(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n} :$$

$$\begin{aligned} P_n^{\otimes_n}(jkl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} &\leq P_n^{\otimes_n}(kl(\widehat{s}_{m'})) + \frac{\mathrm{pen}(m')}{n} \\ &\leq P_n^{\otimes_n}(kl(\widehat{s}_m)) + \frac{\mathrm{pen}(m)}{n} \\ &\leq P_n^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} \end{aligned}$$

- This implies

$$\begin{aligned} &P^{\otimes_n}(jkl(\widehat{s}_{m'})) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \\ &\qquad \leq P^{\otimes_n}(kl(\overline{s}_m)) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n} \end{aligned}$$

Oracle inequality up to deviation

- The previous inequality can be rewritten

$$JKL_{\rho}^{\otimes_n}(s_0, \hat{s}_{m'}) - \nu_n^{\otimes_n}(kl(\overline{s}_m)) \leq KL^{\otimes_n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n} - \nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n}$$

- Appear
  - the integrated loss of the estimate in the model m': $JKL_{\rho}^{\otimes_n}(s_0, \hat{s}_{m'})$
  - a simple and centered process: $-\nu_n^{\otimes_n}(kl(\overline{s}_m))$,
  - the oracle $KL^{\otimes_n}(s_0, \overline{s}_m) + \frac{\mathrm{pen}(m)}{n}$
  - a random *remainder* $-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \frac{\mathrm{pen}(m')}{n}$
- It turns out that $\mathbb{E}\left[-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))\right]$ can be essentially bounded by $\epsilon JKL_{\rho}^{\otimes_n}(s_0, \hat{s}_{m'}) + \frac{\mathrm{pen}(m')}{n}$ as soon as $\mathrm{pen}(m') \geq \kappa(n\sigma_{m'}^2 + x_m)$...

Deviation lemma

- **Lemma:** $\exists \kappa'_0 > 4$, κ'_1 and κ'_2 such that, under assumption (H), for all $m \in \mathcal{M}$, and all $x > 0$, for every $y_m > \sigma_m$

$$\mathbb{P}\left\{ \frac{-\nu_n^{\otimes_n}(jkl(\widehat{s}_m))}{y_m^2 + \kappa'_0 d^{2\otimes_n}(s_0, \widehat{s}_m)} > \frac{\kappa'_1 \sigma_m}{y_m} + \kappa'_2 \sqrt{\frac{x_m + x}{n y_{m'}^2}} + \frac{18}{\rho} \frac{x_m + x}{n y_m^2} \right\} \leq 2e^{-x_m - x}$$

- Using $y_{m'} = \kappa_1 \sqrt{n\sigma_{m'}^2 + x_{m'} + x} / \sqrt{n}$, we obtain, thanks to the Kraft inequality, simultaneously on all model with proba $2\Sigma e^{-x}$:

$$\frac{-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))}{\kappa_1 \left(n\sigma_{m'}^2 + x_{m'} + x\right)/n + \kappa'_0 d^{2\otimes_n}(s_0, \widehat{s}_{m'})} \leq \kappa_1^{-1}(\kappa'_1 + \kappa'_2) + \frac{18}{\rho}\kappa_1^{-2} = \kappa''_0$$

- That is with proba $2\Sigma e^{-x}$:

$$\sup_{m'} -\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'})) - \underbrace{\kappa_1^2 \kappa''_0 \frac{n\sigma_{m'}^2 + x_{m'}}{n}}_{\sim \mathrm{pen}(m')/n} - \underbrace{\kappa'_0 \kappa''_0 d^{2\otimes_n}(s_0, \widehat{s}_{m'})}_{\sim \epsilon JKL_\rho^{\otimes n}(s_0, \hat{s}_{m'})} \leq \kappa_1^2 \kappa''_0 \frac{x}{n}$$

- End of the proof: Choice of κ_1 such that $\kappa''_0 \kappa' 0 = \epsilon C_\rho$ and integration...

Deviation of $-\nu_n^{\otimes n}(jkl(\hat{s}_{m'}))$

- Control of

$$-\nu_n^{\otimes n}(jkl(\widehat{s}_{m'})) = -\left(P_n^{\otimes n}(jkl(\widehat{s}_{m'})) - P^{\otimes n}(jkl(\widehat{s}_{m'}))\right)$$

  with

$$jkl(\widehat{s}_{m'}) = \frac{1}{\rho} \ln \frac{\rho \widehat{s}_{m'} + (1-\rho) s_0}{s_0}$$

- Two main difficulties:
  - Empirical processes,
  - Functions $\widehat{s}_{m'}$ are random!
- Strategy and tools:
  - $-jkl(\widehat{s}_{m'}) = -jkl(\widetilde{s}_{m'}) + (-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))$ with $\widetilde{s}_{m'}$ non random.
  - Concentration of the first term around its mean using Bernstein
  - Control of a weighted supremum

$$\frac{\nu_n^{\otimes n}(-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))}{\epsilon JKL_\rho^{\otimes n}(s_0, \hat{s}_{m'}) + \frac{\mathrm{pen}(m')}{n}} \leq \sup_{s_{m'} \in S_{m'}} \frac{\nu_n^{\otimes n}(-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))}{\epsilon JKL_\rho^{\otimes n}(s_0, s_{m'}) + \frac{\mathrm{pen}(m')}{n}}$$

  by maximal inequality, chaining and pealing.

Chernoff and Bernstein

- X_i independents: study of $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$.
- **Chernoff**: $\forall \lambda \geq 0, \mathbb{P}\{S > x\} \leq \dfrac{\mathbb{E}\left[e^{\lambda S}\right]}{e^{\lambda x}} = e^{-\left(\lambda x - \mathbb{E}\left[e^{\lambda S}\right]\right)}$
- Let $\psi_S(\lambda) = \ln \mathbb{E}\left[e^{\lambda S}\right]$, $\psi_S^*(x) = \sup_{\lambda \in \mathbb{R}^+}(\lambda x - \psi_S(\lambda))$ and ψ_S^{*-} its generalized inverse, we deduce

$$\mathbb{P}\{S > x\} \leq e^{-\psi_S^*(x)} \Leftrightarrow \mathbb{P}\left\{S > \psi_S^{*-}(t)\right\} \leq e^{-t}$$

- **Bernstein**: If

$$\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] \leq V \qquad \text{and} \qquad \forall k \geq 3, \sum_{i=1}^n \mathbb{E}\left[(X_i)_+^k\right] \leq \frac{k!}{2} V b^{k-2}$$

then $\quad \psi_S(\lambda) \leq \dfrac{V\lambda^2}{2(1-b\lambda)}, \quad \psi_S^*(x) \geq \dfrac{v}{b^2}\left(1 + \dfrac{bx}{v} - \sqrt{1 + 2\dfrac{bx}{v}}\right)$

and $\quad \psi_S^{*-}(t) \leq \sqrt{2Vt} + bt.$

Bernstein and JKL

- Berstein revisited: if

$$P^{\otimes_n}\left(f^2\right) \leq V \quad \text{and} \quad \forall k \geq 3, P^{\otimes_n}\left((f)_+^k\right) \leq \frac{k!}{2} V b^{k-2}$$

$$\text{then} \qquad \mathbb{P}\left\{\nu_n^{\otimes_n}(f) \geq \sqrt{\frac{2V}{n}} + b\frac{t}{n}\right\} \leq e^{-t}.$$

- Useful with $-jkl(\widetilde{s}_m) = -\frac{1}{\rho} \ln \frac{s_0}{\rho\widetilde{s}_m + (1-\rho)s_0}$ with $\widetilde{s}_m$ non random?
- **Lemma of van de Geer:** For all positive functions t, u and all integer $k \geq 2$

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

- Apparition of Jensen-Kullback-Leibler:

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

- Bernstein possible for $-jkl(\widetilde{s}_m)$ with $V = \frac{9d^{2\otimes_n}(s_0,\widetilde{s}_m)}{8\rho(1-\rho)}$ and $b = \frac{2}{\rho}$!

Controf of the supremum

- Simple case: $\sup f$ with $f \in \mathcal{F}$ finite and $\forall f \in \mathcal{F}, \psi_f(\lambda) \leq \psi_{\mathcal{F}}(\lambda)$.
- **Control by union bound**:

$$\mathbb{P}\left\{\sup_{f \in \mathcal{F}} f > x\right\} \leq \sum_{f \in \mathcal{F}} \mathbb{P}\{f > x\} \leq |\mathcal{F}| e^{-\psi^*_{\mathcal{F}}(x)}$$

- **Control by *conditioning***:
  - Prop:

$$\forall A, \mathbb{E}^A[Z] = \frac{\mathbb{E}\left[Z\chi_{\{A\}}\right]}{\mathbb{P}\{A\}} \leq \Psi\left(\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right) \Rightarrow \mathbb{P}\{Z > \Psi(x)\} \leq e^{-x}$$

  - Application to recover the union bound:

$$\mathbb{E}^A\left[\sup_{f \in \mathcal{F}} f\right] = \frac{1}{\lambda}\ln\left(e^{\lambda \mathbb{E}^A\left[\sup_{f \in \mathcal{F}} f\right]}\right) \leq \frac{1}{\lambda}\ln\left(E^A\left[e^{\lambda \sup_{f \in \mathcal{F}} f}\right]\right) \leq \frac{1}{\lambda}\ln\left(\sum_{f \in \mathcal{F}} E^A\left[e^{\lambda f}\right]\right)$$

$$\leq \frac{1}{\lambda}\ln\left(\frac{|\mathcal{F}|\psi_{\mathcal{F}}(\lambda)}{\mathbb{P}\{A\}}\right) \leq \psi^{\star -}_{\mathcal{F}}\left(\ln\left(\frac{|\mathcal{F}|}{\mathbb{P}\{A\}}\right)\right)$$

$$\Rightarrow \mathbb{P}\left\{\sup_{f \in \mathcal{F}} f > \psi^{\star -}_{\mathcal{F}}\left(\ln|\mathcal{F}| + x\right)\right\} \leq e^{-x}$$

  - Much more versatile tool...

Countable family and bracketing entropy

- Using chaining technique, extension possible to countable family (much more technical...)
- **Theorem:** Let $\mathcal{F}$ be a countable family of functions. Assume it exists V and b such that for all $f \in \mathcal{F}$ and all integer $k \geq 2$

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

and for all $\delta > 0$, it exists a bracket covering of $\mathcal{F}$ with brackets $[g^-, g^+]$ such that for all integer $k \geq 2$

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

and let $e^{H(\delta)}$ be the cardinality of this covering. It exists an absolute constant $\kappa \leq 27$ such that for $\epsilon \in (0,1]$ and all measurable set A with $\mathbb{P}\{A\} > 0$,

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

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

Jensen-KL and bracketing entropy

- Control of $\mathbb{E}^A\left[\sup \nu_n^{\otimes_n}(f)\right]$ for $f \in \mathcal{F}$ under two assumptions
  - Berstein type assumption: $\exists V$ and b such that for all $f \in \mathcal{F}$ and all integer $k \geq 2$, $\quad P^{\otimes_n}(|f|^k) \leq \frac{k!}{2} V b^{k-2}$.
  - Bracketing entropy assumption on $\mathcal{F}$: For all $\delta > 0$, it exists a bracketing covering of cardinality $H(\delta)$ such that for all bracket $[g^-, g^+]$ and all integer $k \geq 2$, $P^{\otimes_n}(|g^+ - g^-|^k) \leq \frac{k!}{2} \sigma^2 b^{k-2}$.
- **Lemma of van de Geer:** (importance of JKL)

$$P^{\otimes_n}\left(\left|-jkl(s_{m'}) + jkl(\widetilde{s}_{m'})\right|^k\right) \leq \frac{k!}{2}\left(\frac{9d^{2\otimes_n}(s_{m'}, \widetilde{s}_{m'})}{8\rho(1-\rho)}\right)\left(\frac{2}{\rho}\right)^{k-2}$$

- *Natural* choice for $\mathcal{F}$:

$$\left\{-jkl(s_{m'}) + jkl(\widetilde{s}_{m'}) \middle| s_{m'} \in S_{m'}(\widetilde{s}_{m'}, \sigma) = S_{m'} \cap \{s, d^{\otimes_n}(s, \widetilde{s}_{m'}) \leq \sigma\}\right\}.$$

- Bracketing entropy assumptions on $\mathcal{F} \Rightarrow$ Bracketing entropy assumptions on $S_{m'}(\widetilde{s}_{m'}, \sigma)$ with respect to $d^{\otimes_n}$.

Assumption (H) and $\sigma_{m'}$

- Let $W_{m'}(\sigma) = \sup_{s_{m'} \in S_{m'}(\widetilde{s}_{m'},\sigma)} \left(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'})\right)$
- Theorem yields with $\epsilon = 2\sqrt{2\rho(1-\rho)}/3$:

$$\mathbb{E}^A\left[W_{m'}(\sigma)\right] \leq E + \frac{(1+6\epsilon_\rho)3\sigma}{2\sqrt{2\rho(1-\rho)}\sqrt{n}}\sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n}\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

with

$$E = \frac{\kappa}{\epsilon_\rho}\frac{1}{\sqrt{n}}\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}\left(u, S_{m'}(\widetilde{s}_{m'},\sigma)\right) \wedge n}\,\mathrm{d}u + \frac{2(\frac{2}{\rho} + \frac{3\sigma}{2\sqrt{2\rho(1-\rho)}})}{n} H_{[\cdot],d^{\otimes n}}\left(\sigma, S_{m'}(\widetilde{s}_{m'},\sigma)\right)$$

- Assumption $(H) \Rightarrow \int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes n}}\left(\epsilon, S_m(\widetilde{s}_{m'},\sigma)\right)}\,d\epsilon \leq \phi_{m'}(\sigma)$
- Implication: $E \leq \left(\frac{\kappa}{\epsilon} + 2\left(\frac{2}{\rho} + \frac{3}{2\sqrt{\rho(1-\rho)}}\right)\frac{\phi_{m'}(\sigma)}{\sqrt{n}\sigma^2}\right)\frac{\phi_{m'}(\sigma)}{\sqrt{n}}$
- Def. of $\sigma_{m'}$ and monotony prop. of $\phi_{m'}$: $\frac{\phi_{m'}(\sigma_{m'})}{\sqrt{n}\sigma_{m'}^2} = 1$ and $\forall \sigma \geq \sigma_{m'}$

$$\mathbb{E}^A\left[W_{m'}(\sigma)\right] \leq \kappa_1''\frac{\phi_{m'}(\sigma)}{\sqrt{n}} + \frac{\kappa_2''\sigma}{\sqrt{n}}\sqrt{\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)} + \frac{4}{\rho n}\ln\left(\frac{1}{\mathbb{P}\{A\}}\right)$$

$$\leq \Psi_{m'}\left(\sigma, \ln\left(\frac{1}{\mathbb{P}\{A\}}\right)\right)$$

Pealing

- $S_{m'}(\widetilde{s}_{m'}, \sigma)$: $\forall \sigma \geq \sigma_{m'}$

$$\mathbb{E}^A \left[\sup_{s_{m'} \in S_m(\widetilde{s}_{m'}, \sigma)} \left(\nu_n^{\otimes n}(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'})) \right) \right] \leq \Psi_{m'} \left(\sigma, \ln \left(\frac{1}{\mathbb{P}\{A\}} \right) \right)$$

- Choice of $\widetilde{s}_m$: $\widetilde{s}_m = \operatorname{argmin}_{s \in S_m} d^{2\otimes n}(s_0, s_m)$.
- As $\sigma \mapsto \Psi_{m'}(\sigma, \cdot)/\sigma$ is decreasing, the *pealing lemma* applies and implies

$$\forall y_{m'} \geq \sigma_{m'},\ \mathbb{E}^A \left[\sup_{s_{m'} \in S_{m'}} \frac{\nu_n^{\otimes n}(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'}))}{y_{m'}^2 + d^{2\otimes n}(\widetilde{s}_{m'}, s_{m'})} \right] \leq 4 \frac{\Psi_{m'} \left(y_{m'}, \ln \left(\frac{1}{\mathbb{P}\{A\}} \right) \right)}{y_{m'}^2}$$

$$\leq \kappa_1' \frac{\sigma_{m'}}{y_{m'}} + 4\kappa_2'' \sqrt{\frac{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)}{n y_{m'}}} + \frac{16}{\rho} \frac{\ln \left(\frac{1}{\mathbb{P}\{A\}} \right)}{n y_{m'}}$$

- For the deviations: $\forall y_{m'} \geq \sigma_{m'}$,

$$\mathbb{P} \left\{ \sup_{s_{m'} \in S_{m'}} \frac{\nu_n^{\otimes n}(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'}))}{y_{m'}^2 + d^{2\otimes n}(\widetilde{s}_{m'}, s_{m'})} > \kappa_1' \frac{\sigma_{m'}}{y_{m'}} + 4\kappa_2'' \sqrt{\frac{x}{n y_{m'}^2}} + \frac{16}{\rho} \frac{x}{n y_{m'}^2} \right\} \leq e^{-x}$$

Bound summary

- $-\nu_n^{\otimes n}(jkl(\widehat{s}_{m'})) = - - \nu_n^{\otimes n}(jkl(\widetilde{s}_{m'})) + \nu_n^{\otimes n}(-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))$
- For the first term, $-\nu_n^{\otimes n}(jkl(\widetilde{s}_{m'}))$:
  - Bernstein with $V = \frac{9d^{2\otimes n}(s_0,\widetilde{s}_{m'})}{8\rho(1-\rho)}$ and $b = \frac{2}{\rho}$

$$\mathbb{P}\left\{-\nu_n^{\otimes n}(jkl(\widetilde{s}_{m'})) > \sqrt{\frac{9d^{2\otimes n}(s_0,\widetilde{s}_{m'})}{4\rho(1-\rho)}}\sqrt{\frac{x}{n}} + \frac{2}{\rho}\frac{x}{n}\right\} \le e^{-x}$$

  - Renormalization by $y_{m'}^2 + \kappa_0' d^{2\otimes n}(s_0,\widetilde{s}_{m'}) \ge 2y_{m'}\sqrt{\kappa_0'}\sqrt{d^{2\otimes n}(s_0,\widetilde{s}_m)}$:

$$\mathbb{P}\left\{\frac{-\nu_n^{\otimes n}(jkl(\widetilde{s}_{m'}))}{y_m^2 + \kappa_0' d^{2\otimes n}(s_0,\widetilde{s}_{m'})} > \sqrt{\frac{9}{16\rho(1-\rho)\kappa_0'}}\sqrt{\frac{x}{ny_{m'}^2}} + \frac{2}{\rho}\frac{x}{ny_{m'}^2}\right\} \le e^{-x}.$$

- For the second term, $\nu_n^{\otimes n}(-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))$:
  - For the deviations: $\forall y_{m'} \ge \sigma_{m'}$,

$$\mathbb{P}\left\{\sup_{s_{m'}\in S_{m'}} \frac{\nu_n^{\otimes n}(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'}))}{y_{m'}^2 + d^{2\otimes n}(\widetilde{s}_{m'}, s_{m'})} > \kappa_1' \frac{\sigma_{m'}}{y_{m'}} + 4\kappa_2''\sqrt{\frac{x}{ny_{m'}^2}} + \frac{16}{\rho}\frac{x}{ny_{m'}^2}\right\} \le e^{-x}$$

Recombination

- Previous bounds:
  - For the first term, $-\nu_n^{\otimes_n}(jkl(\widetilde{s}_{m'}))$:

$$\mathbb{P}\left\{\frac{-\nu_n^{\otimes_n}(jkl(\widetilde{s}_{m'}))}{y_m^2 + \kappa_0' d^{2\otimes_n}(s_0, \widetilde{s}_{m'})} > \sqrt{\frac{9}{16\rho(1-\rho)\kappa_0'}}\sqrt{\frac{x}{ny_{m'}^2}} + \frac{2}{\rho}\frac{x}{ny_{m'}^2}\right\} \leq e^{-x}.$$

  - For the second term, $\nu_n^{\otimes_n}(-jkl(\widehat{s}_{m'}) + jkl(\widetilde{s}_{m'}))$: $\forall y_{m'} \geq \sigma_{m'}$,

$$\mathbb{P}\left\{\sup_{s_{m'}\in S_{m'}} \frac{\nu_n^{\otimes_n}(-jkl(s_{m'}) + jkl(\widetilde{s}_{m'}))}{y_{m'}^2 + d^{2\otimes_n}(\widetilde{s}_{m'}, s_{m'})} > \kappa_1' \frac{\sigma_{m'}}{y_{m'}} + 4\kappa_2''\sqrt{\frac{x}{ny_{m'}^2}} + \frac{16}{\rho}\frac{x}{ny_{m'}^2}\right\} \leq e^{-x}$$

- $\forall s_{m'} \in S_{m'}$, $d^{2\otimes_n}(s_0, \widetilde{s}_{m'}) \leq d^{2\otimes_n}(s_0, s_{m'})$ and for $\kappa_0' > 4$, $d^{2\otimes_n}(\widetilde{s}_{m'}, s_{m'}) \leq \kappa_0' d^{2\otimes_n}(s_0, s_{m'})$.
- Simple union bounds yields

$$\mathbb{P}\left\{\sup_{s_{m'}\in S_{m'}} \frac{-\nu_n^{\otimes_n}(jkl(s_{m'}))}{y_{m'}^2 + \kappa_0' d^{2\otimes_n}(s_0, s_{m'})} > \kappa_1' \frac{\sigma_{m'}}{y_{m'}} + \kappa_2'\sqrt{\frac{x}{ny_{m'}^2}} + \frac{18}{\rho}\frac{x}{ny_{m'}^2}\right\} \leq 2e^{-x}$$

- Bound valid for $-\nu_n^{\otimes_n}(jkl(\widehat{s}_{m'}))$ i.e. the announced lemma...

Back to the spatialized GMM

- Computation of an upper bound of $H_{[\cdot],d^{\otimes n}}(\epsilon, S_m(s_m,\sigma))$ for the spatialized GMM (cf Maugis and Michel):
  - Bound on an upper bound of the entropy: $H_{[\cdot],d^{\sup}}(\epsilon, S_m)$ where $d^{\sup} = \sqrt{d^{2\,\sup}} = \sqrt{\sup_x d^2(s(\cdot|x), s'(\cdot|x))}$,
  - Result valid for every structure ($[\mu\, L\, D\, A]^K$) and every partition:
  $$H_{[\cdot],d^{\sup}}(\epsilon, S_m) \leq \dim(S_m)(C + \ln\frac{1}{\epsilon})$$
  with an (almost) explicit common C (use of a lemma from Szarek for the entropy of $SO(n)$ without explicit constant) and $\dim(S_m) = |\mathcal{P}|(K-1) + \dim([\mu\, L\, D\, A]^K)$.
  - Consequence: $\mathfrak{D}_m \leq \kappa' \left(C' + \frac{1}{2}\left(\ln\left(\frac{n}{C'\dim(S_m)}\right)\right)_+\right)\dim(S_m)$.
- Collection coding with $x_m \leq \kappa''|\mathcal{P}| \leq \frac{\kappa''}{K-1}\dim(S_m)$.
- Condition on the penalty:

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

Conditional density estimators

- Much work for only one example of model collection: Spatialized GMM!
- Generality of Theorem (luckily) allows more cases!
- Conditional density estimators already analyzed:
  - Covariate Partition based (piecewise constant with respect to X) estimators with density conditioned to X modeled by
    - a GMM (spat. GMM),
    - a piecewise polynomial density.
- Extension to non constant cases:
  - piecewise logistic weights GMM (L. Montuelle),
  - piecewise polynomial on both X and Y.
- For all cases, $\mathrm{pen}(m) \propto (\ln n) \dim(S_m)$.
- Non partition based approach possible theoretically but numerical issues.

Conclusion

- Framework:
  - Unsupervised segmentation problem and Spatialized GMM.
  - Penalized maximum likelihood conditional density estimation.
  - Partition based conditional density estimator.
- Results:
  - Theoretical guaranty for the conditional density estimation problem.
  - Applicable to Spatialized GMM (and unsupervised segmentation...)
  - Efficient minimization algorithm.
- Proof tools:
  - Convexification of KL which allows Bernstein type bound,
  - Supremum of empirical processes and pealing.
- Perspectives:
  - Formal link between conditional density estimation and unsupervised segmentation.
  - Penalty calibration by slope heuristic.
  - Dimension reduction adapted to unsupervised segmentation/classification.
  - Enh. Spatialized GMM with piecewise logistic weights (L. Montuelle).

¿Estan perdidos?

- Pueden:
  - Pregunctarme en Valparaiso,
  - Ir a Viña y ver la misma historia pero con mas aplicaciones y menos detalles matemáticos,
  - Ir a Quintay...

Bibliography

- Model selection and (bracketing) entropy:
  - 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.* ESAIM : P&S (2010) (and erratum)
  - S. Cohen and E. Le Pennec (2011)
- Segmentation by penalization:
  - 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), 216–237
- Segmentation and model selection:
  - 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
- MDL based model selection:
  - 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.
- Entropy of $SO(n)$:
  - J. Szarek (1998). *Metric entropy of homogeneous spaces.* Banach Center Pub., 43, 395–410
- Model selection and conditional density:
  - E. Brunel, C. Lacour and F. Comte (2007). *Adaptive estimation of the conditional density in presence of censoring.* Sankhya, 69(4), 734–763