# Contributions à l'apprentissage statistique dans les modèles parcimonieux

Pierre Alquier


Soutenance d'HDR
Vendredi 06/12/2013
Université Paris 6

| | | |
|---|---|---|
| 2003-2006 | étudiant en thèse | Université Paris 6 |
| | dir. : Olivier Catoni | (financé par le CREST) |
| 2006-2007 | ATER | Université Paris Dauphine |
| 2007-2012 | maître de conférences | Université Paris Diderot |
| 2012-20XX | *lecturer* | University College Dublin |

# Contexte

On observe, sur un espace $(\Omega, \mathcal{A}, \mathbb{P})$, $n$ paires aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes sous $\mathbb{P}$, $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$. On suppose que

$$Y_i = f(X_i) + W_i$$

pour $f$ mesurable $W_i$ avec $\mathbb{E}(W_i|X_i) = 0$ et $\mathbb{E}(W_i^2|X_i) \leq \sigma^2$.

# Contexte

On observe, sur un espace $(\Omega, \mathcal{A}, \mathbb{P})$, $n$ paires aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes sous $\mathbb{P}$, $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$. On suppose que

$$Y_i = f(X_i) + W_i$$

pour $f$ mesurable $W_i$ avec $\mathbb{E}(W_i|X_i) = 0$ et $\mathbb{E}(W_i^2|X_i) \leq \sigma^2$.

## Prédicteurs linéaires

$$\forall \theta, x \in \mathbb{R}^p, \quad f_\theta(x) := \theta \cdot x.$$

# Contexte

On observe, sur un espace $(\Omega, \mathcal{A}, \mathbb{P})$, $n$ paires aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendantes sous $\mathbb{P}$, $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$. On suppose que

$$Y_i = f(X_i) + W_i$$

pour $f$ mesurable $W_i$ avec $\mathbb{E}(W_i|X_i) = 0$ et $\mathbb{E}(W_i^2|X_i) \leq \sigma^2$.

## Prédicteurs linéaires

$$\forall \theta, x \in \mathbb{R}^p, \quad f_\theta(x) := \theta \cdot x.$$

## Risques

$$r(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f_\theta(X_i)\right)^2, \; R(\theta) := \mathbb{E}\left[r(\theta)\right] \text{ et } \overline{R} = \inf R.$$

# Idée d'inégalité oracle

$$\hat{\theta}_I^{\mathrm{MCO}} = \arg\min\{r(\theta) : \theta_i = 0 \text{ pour } i \notin I\}$$
$$\theta_I = \arg\min\{R(\theta) : \theta_i = 0 \text{ pour } i \notin I\}.$$

# Idée d'inégalité oracle

$$\hat{\theta}_I^{\mathrm{MCO}} = \arg\min\{r(\theta) : \theta_i = 0 \text{ pour } i \notin I\}$$
$$\theta_I = \arg\min\{R(\theta) : \theta_i = 0 \text{ pour } i \notin I\}.$$

A design fixe, $W_i \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{E}\left[R(\hat{\theta}_I^{\mathrm{MCO}}) - \overline{R}\right] \leq R(\theta_I) - \overline{R} + \frac{\sigma^2 \mathrm{card}(I)}{n}$$

# Idée d'inégalité oracle

$$\hat{\theta}_I^{\mathrm{MCO}} = \arg\min\{r(\theta) : \theta_i = 0 \text{ pour } i \notin I\}$$
$$\theta_I = \arg\min\{R(\theta) : \theta_i = 0 \text{ pour } i \notin I\}.$$

A design fixe, $W_i \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{E}\left[R(\hat{\theta}_I^{\mathrm{MCO}}) - \overline{R}\right] \leq R(\theta_I) - \overline{R} + \frac{\sigma^2 \mathrm{card}(I)}{n}$$

Peut-on proposer $\hat{\theta}$ tel que

$$\mathbb{E}\left[R(\hat{\theta}) - \overline{R}\right] \leq \inf_I \left\{ \left[R(\theta_I) - \overline{R}\right] + \frac{\sigma^2 \mathrm{card}(I)}{n} \right\} ?$$

# Idée d'inégalité oracle

$$\hat{\theta}_I^{\mathrm{MCO}} = \arg\min\{r(\theta) : \theta_i = 0 \text{ pour } i \notin I\}$$
$$\theta_I = \arg\min\{R(\theta) : \theta_i = 0 \text{ pour } i \notin I\}.$$

A design fixe, $W_i \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{E}\left[R(\hat{\theta}_I^{\mathrm{MCO}}) - \overline{R}\right] \leq R(\theta_I) - \overline{R} + \frac{\sigma^2 \mathrm{card}(I)}{n}$$

On peut fabriquer $\hat{\theta}$ tel que

$$\mathbb{E}\left[R(\hat{\theta}) - \overline{R}\right] \leq \inf_I \left\{ \left[R(\theta_I) - \overline{R}\right] + C\frac{\sigma^2 \mathrm{card}(I)\log(p)}{n} \right\}$$

et c'est le mieux que l'on puisse faire...

Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007) Aggregation for Gaussian Regression. *The Annals of Statistics* 35(4), pp. 1674–1697.

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | |
|---|---|---|
| Définition | | |
| Inég. oracle ? | | |
| Calcul ? | | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | |
| Inég. oracle ? | | |
| Calcul ? | | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | |
| Inég. oracle ? | oui | |
| Calcul ? | | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | |
| Inég. oracle ? | oui | |
| Calcul ? | $p$ petit | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | |
| Inég. oracle ? | oui | |
| Calcul ? | $p$ petit | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | |
| Calcul ? | $p$ petit | |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | |
| Calcul ? | $p$ petit | $p$ grand |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | hypothèses sur $X_1, \ldots, X_n$ |
| Calcul ? | $p$ petit | $p$ grand |

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | hypothèses sur $X_1, \ldots, X_n$ |
| Calcul ? | $p$ petit | $p$ grand |

Références pour pénalisation $\ell_0$

Barron, A. R., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields* 113(3), pp. 301–413.

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | hypothèses sur $X_1, \ldots, X_n$ |
| Calcul ? | $p$ petit | $p$ grand |

Références pour pénalisation $\ell_0$


Barron, A. R., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields* **113**(3), pp. 301–413.

Références pour pénalisation $\ell_1$


Tibshirani, R. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society : Series B* **58**(1), pp. 267–288.


Bickel, P., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous Analysis of LASSO and Dantzig Selector. *The Annals of Statistics* **37**(4), pp. 1705–1732.

# Comment obtenir des inégalités oracle ?

| | AIC,BIC... | LASSO |
|---|---|---|
| Définition | $\min\{r(\theta) + \lambda\|\theta\|_0\}$ | $\min\{r(\theta) + \lambda\|\theta\|_1\}$ |
| Inég. oracle ? | oui | hypothèses sur $X_1, \ldots, X_n$ |
| Calcul ? | $p$ petit | $p$ grand |

Références pour pénalisation $\ell_0$


Barron, A. R., Birgé, L. and Massart, P. (1999) Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields* **113**(3), pp. 301–413.

Références pour pénalisation $\ell_1$


Tibshirani, R. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society : Series B* **58**(1), pp. 267–288.


Bickel, P., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous Analysis of LASSO and Dantzig Selector. *The Annals of Statistics* **37**(4), pp. 1705–1732.

$\rightarrow$ idée : estimateur agrégé, calculable par Monte-Carlo


Dalalyan, A. and Tsybakov, A. B. (2008) Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity. *Machine Learning* **72**(1-2), pp. 39–61.

# Plan de la présentation

1. Estimation dans un modèle linéaire parcimonieux
   - Introduction
   - Inégalités PAC-Bayésiennes
2. Généralisation à l'autorégression
   - Prédiction de séries temporelles
   - Inégalités PAC-Bayésiennes
3. Au-delà du modèle linéaire
   - Modèle à direction révélatrice parcimonieux
   - Régression de faible rang
   - Modèles à complexité bornée

# Premier estimateur agrégé $\hat{\theta}_\lambda$

$$\pi_I = \frac{\alpha^{\mathrm{card}(I)}}{\binom{p}{\mathrm{card}(I)} \sum_{j=0}^{n} \alpha^j}.$$

# Premier estimateur agrégé $\hat{\theta}_\lambda$

$$\pi_I = \frac{\alpha^{\mathrm{card}(I)}}{\binom{p}{\mathrm{card}(I)} \sum_{j=0}^{n} \alpha^j}.$$

## Définition

$$\hat{\theta}_\lambda = \frac{\sum_{\mathrm{card}(I)\leq n} \pi_I \exp\left[-\lambda\left(r(\hat{\theta}_I^{\mathrm{MCO}}) + \frac{2\sigma^2 \mathrm{card}(I)}{n}\right)\right] \hat{\theta}_I^{\mathrm{MCO}}}{\sum_{\mathrm{card}(I)\leq n} \pi_I \exp\left[-\lambda\left(r(\hat{\theta}_I^{\mathrm{MCO}}) + \frac{2\sigma^2 \mathrm{card}(I)}{n}\right)\right]}.$$

# Premier estimateur agrégé $\hat{\theta}_\lambda$

## Théorème (A & Lounici 2011)

Les $X_i$ déterministes et les $W_i$ iid $\mathcal{N}(0, \sigma^2)$, $\lambda = 1/(4\sigma^2)$,

$$\mathbb{E}\left(R(\hat{\theta}_\lambda) - \overline{R}\right) \leq \inf_{\mathrm{card}(I)\leq n} \left\{ R(\theta_I) - \overline{R} + \frac{\sigma^2 \mathrm{card}(I)}{n}\left(4\log\left(\frac{p\mathrm{e}}{\alpha \mathrm{card}(I)}\right) + 1\right) + \frac{4\sigma^2 \log\left(\frac{1}{1-\alpha}\right)}{n} \right\}.$$

# Premier estimateur agrégé $\hat{\theta}_\lambda$

## Théorème (A & Lounici 2011)

Les $X_i$ déterministes et les $W_i$ iid $\mathcal{N}(0, \sigma^2)$, $\lambda = 1/(4\sigma^2)$,

$$\mathbb{E}\left(R(\hat{\theta}_\lambda) - \overline{R}\right) \leq \inf_{\mathrm{card}(I)\leq n} \left\{ R(\theta_I) - \overline{R} + \frac{\sigma^2 \mathrm{card}(I)}{n}\left(4\log\left(\frac{p\mathrm{e}}{\alpha \mathrm{card}(I)}\right) + 1\right) + \frac{4\sigma^2 \log\left(\frac{1}{1-\alpha}\right)}{n} \right\}.$$

**Preuve :** Lemme de Stein + méthodes PAC-Bayésiennes :


Leung, G. and Barron, A. R. (2006) Information Theory and Mixing Least-Square Regressions. *IEEE Transactions on Information Theory* 52(8), pp. 3396–3410.1.


Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization*. Ecole d'Été de Probabilités de Saint-Flour XXXI - 2001. J. Picard Editor. Springer Lecture Notes in Mathematics 1851.

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

1. on suppose que $R(\overline{\theta}) = \overline{R}$ et $\|\overline{\theta}\|_1 \leq K$, soit $\mathrm{u}_I$ proba. uniforme sur $\{\|\theta\|_1 \leq K+1, \theta_i = 0 \text{ pour } i \notin I\}$ et

$$\pi(\mathrm{d}\theta) = \sum_{\mathrm{card}(I) \leq n} \pi_I \mathrm{u}_I(\mathrm{d}\theta),$$

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

1. on suppose que $R(\overline{\theta}) = \overline{R}$ et $\|\overline{\theta}\|_1 \leq K$, soit $\mathrm{u}_I$ proba. uniforme sur $\{\|\theta\|_1 \leq K+1, \theta_i = 0 \text{ pour } i \notin I\}$ et

$$\pi(\mathrm{d}\theta) = \sum_{\mathrm{card}(I) \leq n} \pi_I \mathrm{u}_I(\mathrm{d}\theta),$$

2. $\|f\|_\infty \leq L$, $\|X_i\|_\infty \leq L'$,

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

1. on suppose que $R(\overline{\theta}) = \overline{R}$ et $\|\overline{\theta}\|_1 \leq K$, soit $\mathrm{u}_I$ proba. uniforme sur $\{\|\theta\|_1 \leq K+1, \theta_i = 0 \text{ pour } i \notin I\}$ et

$$\pi(\mathrm{d}\theta) = \sum_{\mathrm{card}(I) \leq n} \pi_I \mathrm{u}_I(\mathrm{d}\theta),$$

2. $\|f\|_\infty \leq L$, $\|X_i\|_\infty \leq L'$,
3. $W_i$ est sous-exponentiel $(\sigma, \xi)$, $X_i$ peuvent être déterministes ou aléatoires.

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

1. on suppose que $R(\overline{\theta}) = \overline{R}$ et $\|\overline{\theta}\|_1 \leq K$, soit $\mathrm{u}_I$ proba. uniforme sur $\{\|\theta\|_1 \leq K+1, \theta_i = 0 \text{ pour } i \notin I\}$ et

$$\pi(\mathrm{d}\theta) = \sum_{\mathrm{card}(I) \leq n} \pi_I \mathrm{u}_I(\mathrm{d}\theta),$$

2. $\|f\|_\infty \leq L$, $\|X_i\|_\infty \leq L'$,
3. $W_i$ est sous-exponentiel $(\sigma, \xi)$, $X_i$ peuvent être déterministes ou aléatoires.

## Définition

$$\frac{\mathrm{d}\tilde{\rho}_\lambda}{\mathrm{d}\pi}(\theta) = \frac{\exp(-\lambda r(\theta))}{\int_{\Theta_K} \exp(-\lambda r(\theta'))\pi(\mathrm{d}\theta')} \text{ et } \tilde{\theta}_\lambda = \int_{\Theta_K} \theta \tilde{\rho}_\lambda(\mathrm{d}\theta).$$

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

## Théorème (A & Lounici 2011)

Pour $\mathcal{C} = \mathcal{C}(\sigma, \xi, L, L', K)$ connu, $\lambda = n/\mathcal{C}$, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left\{ R(\tilde{\theta}_\lambda) - R(\bar{\theta}) \leq \mathcal{C} \inf_I \left[ R(\theta_I) - R(\bar{\theta}) + \frac{\mathrm{card}(I) \log\left(\frac{\mathrm{e}np(K+1)}{\alpha \mathrm{card}(I)}\right)}{n} + \frac{\log\left(\frac{2}{\varepsilon(1-\alpha)}\right)}{n} \right] \right\} \geq 1 - \varepsilon.$$

# Deuxième estimateur agrégé $\tilde{\theta}_\lambda$

## Théorème (A & Lounici 2011)

Pour $\mathcal{C} = \mathcal{C}(\sigma, \xi, L, L', K)$ connu, $\lambda = n/\mathcal{C}$, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left\{ R(\tilde{\theta}_\lambda) - R(\bar{\theta}) \leq \mathcal{C} \inf_I \left[ R(\theta_I) - R(\bar{\theta}) + \frac{\mathrm{card}(I) \log\left(\frac{\mathrm{e}np(K+1)}{\alpha \mathrm{card}(I)}\right)}{n} + \frac{\log\left(\frac{2}{\varepsilon(1-\alpha)}\right)}{n} \right] \right\} \geq 1 - \varepsilon.$$

**Preuve :** inégalité de Bernstein + PAC-Bayes, suit

Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization*. Ecole d'Été de Probabilités de Saint-Flour XXXI - 2001. J. Picard Editor. Springer Lecture Notes in Mathematics 1851.

# Publications liées à cette thématique



Alquier, P. (2008) Iterative Feature Selection in Least Square Regression Estimation. *Annales de l'Institut Henri Poincarré, Probabilités et Statistiques* 44(1), pp. 47–88.


Alquier, P. (2008) Density Estimation with Quadratic Loss, a Confidence Intervals Method. *ESAIM Probability & Statistics* 12, pp. 438–463.


Alquier, P. (2008) LASSO, Iterative Feature Selection & the Correlation Selector : Oracle Inequalities & Numerical Performances. *Electronic Journal of Statistics* 2, pp. 1129–1152.


Alquier, P. & Lounici, K. (2011) PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics* 5, pp. 127–145.


Alquier, P. & Hebiri, M. (2011) Generalization of L1 Constraints for High Dimensional Regression Problems. *Statistics & Probability Letters* 81(12), pp. 1760–1765.


Alquier, P. & Hebiri, M. (2012) Transductive Versions of the LASSO & the Dantzig Selector. *Journal of Statistical Planning & Inference* 142(9), pp. 2485–2500.

# Menu

1. Estimation dans un modèle linéaire parcimonieux
   - Introduction
   - Inégalités PAC-Bayésiennes
2. Généralisation à l'autorégression
   - Prédiction de séries temporelles
   - Inégalités PAC-Bayésiennes
3. Au-delà du modèle linéaire
   - Modèle à direction révélatrice parcimonieux
   - Régression de faible rang
   - Modèles à complexité bornée

# Contexte

$(X_t)_{t\in\mathbb{Z}}$ série temporelle à valeurs ds $\mathbb{R}^p$. Obs : $X_1, \ldots, X_n$.

# Contexte

$(X_t)_{t\in\mathbb{Z}}$ série temporelle à valeurs ds $\mathbb{R}^p$. Obs : $X_1, \dots, X_n$.

Pour $f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p$ on pose

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \dots, X_{t-k}).$$

# Contexte

$(X_t)_{t\in\mathbb{Z}}$ série temporelle à valeurs ds $\mathbb{R}^p$. Obs : $X_1, \ldots, X_n$.

Pour $f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p$ on pose

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \ldots, X_{t-k}).$$

Hypothèse :

$$\begin{cases} \|f_\theta(x_1, \ldots, x_k) - f_\theta(y_1, \ldots, y_k)\| \leq \sum_{j=1}^k a_j(\theta) \|x_j - y_j\|, \\ \sum_{j=1}^k a_j(\theta) \leq L. \end{cases}$$

# Contexte

$(X_t)_{t\in\mathbb{Z}}$ série temporelle à valeurs ds $\mathbb{R}^p$. Obs : $X_1, \ldots, X_n$.

Pour $f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p$ on pose

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \ldots, X_{t-k}).$$

Hypothèse :

$$\begin{cases} \|f_\theta(x_1, \ldots, x_k) - f_\theta(y_1, \ldots, y_k)\| \leq \sum_{j=1}^k a_j(\theta) \|x_j - y_j\|, \\ \sum_{j=1}^k a_j(\theta) \leq L. \end{cases}$$

## Risques - $\ell$ perte convexe et $K$-Lipshitz

$$R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right] \text{ et } r(\theta) = \frac{1}{n-k}\sum_{i=k+1}^n \ell\left(\hat{X}_i^\theta, X_i\right).$$

# Hypotheses de dépendance/mélange

## $\phi$-mélange pour $(W_i)_{i\in\mathbb{Z}}$ suite de v.a.

$$\varphi^W(r) = \sup_{\substack{A \in \sigma(W_k, W_{k-1}, \dots) \\ B \in \sigma(W_{k+r}) \\ k \in \mathbb{Z}}} \Big\{ |\mathbb{P}(B|A) - \mathbb{P}(B)| \Big\}.$$

Hypothèse $\mathrm{PhiMix}(C)$ : $1 + \sum_{j=1}^{\infty} \sqrt{\varphi^W(j)} \leq C$.

# Hypotheses de dépendance/mélange

## $\phi$-mélange pour $(W_i)_{i\in\mathbb{Z}}$ suite de v.a.

$$\varphi^W(r) = \sup_{\substack{A \in \sigma(W_k, W_{k-1}, \dots) \\ B \in \sigma(W_{k+r}) \\ k \in \mathbb{Z}}} \Big\{ |\mathbb{P}(B|A) - \mathbb{P}(B)| \Big\}.$$

Hypothèse $\mathrm{PhiMix}(C)$ : $1 + \sum_{j=1}^{\infty} \sqrt{\varphi^W(j)} \leq C$.

## Hypothèse $\mathrm{CBS}(c_\xi, (a_j))$

$(\xi_i)_{i\in\mathbb{Z}}$ iid, $\|\xi_i\| \leq c_\xi$,

$$W_i = H(\xi_i, \xi_{i-1}, \xi_{i-2} \dots)$$

avec $\|H(v) - H(v')\| \leq \sum_{j=0}^{\infty} a_j \|v_j - v'_j\|$ et $\sum_{j=0}^{\infty} a_j < +\infty$.

# Hypotheses de dépendance/mélange

## Hypothèse $\mathrm{ThetaDep}(\mathcal{C})$

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S. and Prieur, C. (2007) *Weak Dependance : with Examples and Applications*. Springer Lecture Notes in Statistics 190.

# Hypotheses de dépendance/mélange

## Hypothèse $\mathrm{ThetaDep}(\mathcal{C})$

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S. and Prieur, C. (2007) *Weak Dependance : with Examples and Applications*. Springer Lecture Notes in Statistics 190.

## Proposition

$$\left.\begin{array}{r} \mathrm{PhiMix}(\mathcal{C}) \\ \forall i,\ \|W_i\| \leq c_W \end{array}\right\} \Rightarrow \mathrm{ThetaDep}(c_W\mathcal{C}).$$

# Hypotheses de dépendance/mélange

## Hypothèse $\mathrm{ThetaDep}(\mathcal{C})$

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S. and Prieur, C. (2007) *Weak Dependance : with Examples and Applications*. Springer Lecture Notes in Statistics 190.

## Proposition

$$\left.\begin{array}{r} \mathrm{PhiMix}(\mathcal{C}) \\ \forall i,\ \|W_i\| \leq c_W \end{array}\right\} \Rightarrow \mathrm{ThetaDep}(c_W \mathcal{C}).$$

$$\left.\begin{array}{r} \mathrm{CBS}(c_\xi, (a_j)) \\ \sum_{j=0}^{\infty} j a_j < \infty \end{array}\right\} \Rightarrow \mathrm{ThetaDep}(2 c_\xi \sum_{j=0}^{\infty} j a_j).$$

# Inégalité PAC-Bayésienne

## Estimateur de Gibbs pour une loi *a priori* $\pi$

$$\tilde{\theta}_\lambda = \int_\Theta \theta \tilde{\rho}_\lambda(\mathrm{d}\theta), \text{ avec } \tilde{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r(\theta')}\pi(\mathrm{d}\theta')}.$$

# Inégalité PAC-Bayésienne

## Estimateur de Gibbs pour une loi *a priori* $\pi$

$$\tilde{\theta}_\lambda = \int_\Theta \theta \tilde{\rho}_\lambda(\mathrm{d}\theta), \text{ avec } \tilde{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r(\theta')}\pi(\mathrm{d}\theta')}.$$

## Théorème (A, Li & Wintenberger, 2012)

$(X_t)$ satisfait $\mathrm{ThetaDep}(\mathcal{C})$, $\kappa := K(1+L)(c_X + \mathcal{C})/\sqrt{2}$. Pour $\lambda > 0$, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left( R\left(\tilde{\theta}_\lambda\right) \leq \inf_\rho \left[ \int R(\theta)\rho(\mathrm{d}\theta) + \frac{2\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right] \right) \geq 1-\varepsilon.$$

# Un exemple : ensemble fini de prédicteurs

Cas $\theta \in \Theta$, $\mathrm{card}(\Theta) = M < \infty$. Soit $\lambda = \frac{1-k/n}{\kappa}\sqrt{n \log(M)}$.

# Un exemple : ensemble fini de prédicteurs

Cas $\theta \in \Theta$, $\mathrm{card}(\Theta) = M < \infty$. Soit $\lambda = \frac{1-k/n}{\kappa}\sqrt{n \log(M)}$.

$$R\left(\tilde{\theta}_\lambda\right) \leq \overline{R} + \frac{4\kappa}{\left(1-\frac{k}{n}\right)}\sqrt{\frac{\log\left(M\right)}{n}} + \frac{2\kappa \log\left(\frac{2}{\varepsilon}\right)}{\left(1-\frac{k}{n}\right)\sqrt{n\log(M)}}.$$

# Un exemple : ensemble fini de prédicteurs

Cas $\theta \in \Theta$, $\mathrm{card}(\Theta) = M < \infty$. Soit $\lambda = \frac{1-k/n}{\kappa}\sqrt{n \log(M)}$.

$$R\left(\tilde{\theta}_\lambda\right) \leq \overline{R} + \frac{4\kappa}{\left(1-\frac{k}{n}\right)}\sqrt{\frac{\log\left(M\right)}{n}} + \frac{2\kappa \log\left(\frac{2}{\varepsilon}\right)}{\left(1-\frac{k}{n}\right)\sqrt{n\log(M)}}.$$

- Vitesse optimale

Audibert, J.-Y. (2009) Fast Learning Rates in Statistical Inference Through Aggregation. *The Annals of Statistics* 38(4), pp. 1591–1646.

# Un exemple : ensemble fini de prédicteurs

Cas $\theta \in \Theta$, $\mathrm{card}(\Theta) = M < \infty$. Soit $\lambda = \frac{1-k/n}{\kappa}\sqrt{n\log(M)}$.

$$R\left(\tilde{\theta}_\lambda\right) \leq \overline{R} + \frac{4\kappa}{\left(1-\frac{k}{n}\right)}\sqrt{\frac{\log\left(M\right)}{n}} + \frac{2\kappa\log\left(\frac{2}{\varepsilon}\right)}{\left(1-\frac{k}{n}\right)\sqrt{n\log(M)}}.$$

- Vitesse optimale

Audibert, J.-Y. (2009) Fast Learning Rates in Statistical Inference Through Aggregation. *The Annals of Statistics* 38(4), pp. 1591–1646.

- **Preuve du théorème :** repose sur l'inégalité de Hoeffding pour les variables dépendantes de Rio

Rio, E. (2000) Inégalité de Hoeffding pour les Fonctions Lipshitziennes de Suites Dépendantes. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 330(10), pp. 905–908.

# Un exemple : ensemble fini de prédicteurs

Cas $\theta \in \Theta$, $\mathrm{card}(\Theta) = M < \infty$. Soit $\lambda = \frac{1-k/n}{\kappa}\sqrt{n \log(M)}$.

$$R\left(\tilde{\theta}_\lambda\right) \leq \overline{R} + \frac{4\kappa}{\left(1-\frac{k}{n}\right)}\sqrt{\frac{\log\left(M\right)}{n}} + \frac{2\kappa \log\left(\frac{2}{\varepsilon}\right)}{\left(1-\frac{k}{n}\right)\sqrt{n\log(M)}}.$$

- Vitesse optimale

  

  Audibert, J.-Y. (2009) Fast Learning Rates in Statistical Inference Through Aggregation. *The Annals of Statistics* 38(4), pp. 1591–1646.
- **Preuve du théorème :** repose sur l'inégalité de Hoeffding pour les variables dépendantes de Rio

  

  Rio, E. (2000) Inégalité de Hoeffding pour les Fonctions Lipshitziennes de Suites Dépendantes. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 330(10), pp. 905–908.
- on ne dépassera pas les vitesses en $1/\sqrt{n}$ avec ce théorème...

# Vitesses rapides

$X_t \in \mathbb{R}$, $\theta \in \mathbb{R}^d$, même loi $\pi$ que dans la première partie sur

$$f_\theta(X_{t-1}, \ldots, X_{t-k}) = \sum_{j=1}^{d} \theta_j \phi_j(X_{t-1}, \ldots, X_{t-k}),$$ perte quadratique.

## Théorème (A, Li & Wintenberger, 2012)

$(X_t)$ satisfait $\mathrm{PhiMix}(\mathcal{C})$, $R(\bar{\theta}) = \overline{R}$ et $\|\bar{\theta}\|_1 \leq L$. Pour $\mathcal{C}' = \mathcal{C}'(\mathcal{C}, L, c_X)$ connu $\lambda = \mathcal{C}'/n$, pour $\varepsilon > 0$,

$$\mathbb{P}\left( R(\tilde{\theta}_\lambda) - \overline{R} \leq 4 \inf_J \left\{ R(\bar{\theta}_J) - \overline{R} + \mathcal{C}' \frac{\mathrm{card}(J) \log\left(\frac{(n-k)d}{\mathrm{card}(J)}\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k} \right\} \right) \geq 1 - \varepsilon.$$

# Vitesses rapides

## Théorème (A, Li & Wintenberger, 2012)

$(X_t)$ satisfait $\mathrm{PhiMix}(\mathcal{C})$, $R(\bar{\theta}) = \overline{R}$ et $\|\bar{\theta}\|_1 \leq L$. Pour $\mathcal{C}' = \mathcal{C}'(\mathcal{C}, L, c_X)$ connu $\lambda = \mathcal{C}'/n$, pour $\varepsilon > 0$,

$$\mathbb{P}\left( R(\tilde{\theta}_\lambda) - \overline{R} \leq 4 \inf_J \left\{ R(\bar{\theta}_J) - \overline{R} + \mathcal{C}' \frac{\mathrm{card}(J) \log\left(\frac{(n-k)d}{\mathrm{card}(J)}\right) + \log\left(\frac{2}{\varepsilon}\right)}{n-k} \right\} \right) \geq 1 - \varepsilon.$$

**Preuve du théorème :** repose sur l'inégalité de Bernstein de Samson

Samson, P.-M. (2000) Concentration of Measure Inequalities for Markov Chains and Φ-Mixing Processes. *Annals of Probability* **28**(1), pp. 416–461.

# Publications liées à cette thématique


Alquier, P. & Doukhan, P. (2011) Sparsity Considerations for Dependent Observations. *Electronic Journal of Statistics* 5, pp. 750–774.


Alquier, P. & Wintenberger, O. (2012) Model Selection for Weakly Dependent Time Series Forecasting. *Bernoulli* 18(3), pp. 883–913.


Alquier, P. & Li, X. (2012) Prediction of Quantiles by Statistical Learning & Application to GDP Forecasting. In *Proceedings of the 15th International Conference on Discovery Science (DS'12)*, J.-G. Ganascia, P. Lenca & J.-M. Petit Editors, Springer Lecture Notes in Artificial Intelligence 7569, pp. 22–36.


Alquier, P., Li, X. & Wintenberger, O. (2012) Prediction of Time Series by Statistical Learning : General Losses & Fast Rates. *Preprint arXiv* :1202.4283v1. Currently in minor revision for *Dependence Modeling*.

# Menu

1. Estimation dans un modèle linéaire parcimonieux
   - Introduction
   - Inégalités PAC-Bayésiennes
2. Généralisation à l'autorégression
   - Prédiction de séries temporelles
   - Inégalités PAC-Bayésiennes
3. Au-delà du modèle linéaire
   - Modèle à direction révélatrice parcimonieux
   - Régression de faible rang
   - Modèles à complexité bornée

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\beta \cdot x)$ et $\|\beta\|_0 \ll p$,

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\beta \cdot x)$ et $\|\beta\|_0 \ll p$,
- modèle additif $y = \sum_{j=1}^{p} f_j(x_j)$ et $\mathrm{card}(\{j : f_j \neq 0\}) \ll p$,

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\beta \cdot x)$ et $\|\beta\|_0 \ll p$,
- modèle additif $y = \sum_{j=1}^{p} f_j(x_j)$ et $\mathrm{card}(\{j : f_j \neq 0\}) \ll p$,
- régression de faible rang $y = Mx \in \mathbb{R}^m$, $\mathrm{rg}(M) \ll m \wedge p$,

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\beta \cdot x)$ et $\|\beta\|_0 \ll p$,
- modèle additif $y = \sum_{j=1}^{p} f_j(x_j)$ et $\mathrm{card}(\{j : f_j \neq 0\}) \ll p$,
- régression de faible rang $y = Mx \in \mathbb{R}^m$, $\mathrm{rg}(M) \ll m \wedge p$,
- modèles en physique quantique induisant des contraintes de parcimonie ou de faible rang,

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\beta \cdot x)$ et $\|\beta\|_0 \ll p$,
- modèle additif $y = \sum_{j=1}^{p} f_j(x_j)$ et $\mathrm{card}(\{j : f_j \neq 0\}) \ll p$,
- régression de faible rang $y = Mx \in \mathbb{R}^m$, $\mathrm{rg}(M) \ll m \wedge p$,
- modèles en physique quantique induisant des contraintes de parcimonie ou de faible rang,
- que peut-on dire de façon générale ?

# Autres modèles

Possibles fonctions reliant $x \in \mathbb{R}^p$ et $y$ :

- *single-index* parcimonieux $y = g(\theta^T x)$ et $\|\theta\|_0 \ll p$,
- modèle additif $y = \sum_{j=1}^p f_j(x_j)$ et $\mathrm{card}(\{j : f_j \neq 0\}) \ll p$,
- régression de faible rang $y = Mx \in \mathbb{R}^m$, $\mathrm{rg}(M) \ll m \wedge p$,
- modèles en physique quantique induisant des contraintes de parcimonie ou de faible rang,
- que peut-on dire de façon générale ?

# Modèle à direction révélatrice parcimonieux

$(X_i, Y_i)_{i=1}^n$ i.i.d. sous $\mathbb{P}$, loi commune $P$, $X_i \in [-1,1]^p$,

$$Y_i = \overline{g}(\overline{\beta} \cdot X_i) + W_i, \qquad \overline{\theta} = (\overline{g}, \overline{\beta})$$

$\mathbb{E}(W_i|X_i) = 0$, $\overline{g} : \mathbb{R} \to [-C, C]$, et $\overline{\beta} \in \mathbb{R}^p$, $\|\beta\|_0 \ll p$. Perte quadratique.

# Modèle à direction révélatrice parcimonieux

$(X_i, Y_i)_{i=1}^n$ i.i.d. sous $\mathbb{P}$, loi commune $P$, $X_i \in [-1,1]^p$,

$$Y_i = \overline{g}(\overline{\beta} \cdot X_i) + W_i, \qquad \overline{\theta} = (\overline{g}, \overline{\beta})$$

$\mathbb{E}(W_i|X_i) = 0$, $\overline{g} : \mathbb{R} \to [-C, C]$, et $\overline{\beta} \in \mathbb{R}^p$, $\|\beta\|_0 \ll p$. Perte quadratique.

On peut définir une loi *a priori* dans ce modèle, et un estimateur de Gibbs randomisé $\check{\theta}_\lambda$.

# Inégalité oracle

## Théorème (A & Biau, 2013)

Les $W_i$ sont sous-exponentiels $(\sigma, \xi)$, constante connue $\Xi = \Xi(C, \xi, \sigma) > 0$, $\lambda = n/\Xi$. Pour tout $\varepsilon > 0$,

$$\mathbb{P}\left\{ R(\check{\theta}_\lambda) - \overline{R} \leq \Xi \inf_{I \subset \{1,\dots,p\}} \inf_{1 \leq M \leq n} \left[ R(\overline{\theta}_{I,M}) - \overline{R} + \frac{M \log(n) + \mathrm{card}(I) \log(pn) + \log\left(\frac{2}{\varepsilon}\right)}{n} \right] \right\} \geq 1 - \varepsilon,$$

où $\overline{\theta}_{I,M} = (\overline{g}_M, \overline{\beta}_I)$ est la meilleure approximation $I$-sparse de $\overline{\beta}$ et la meilleure approximation de $\overline{g}$ avec $M$ fonctions d'une base.

# Application : ellipsoïde de Sobolev

Sous quelques hypothèses supplémentaires, dont $\overline{g} \in \mathcal{W}\left(s, \frac{6C^2}{\pi^2}\right)$ pour $s \geq 2$, avec $s$ inconnu :

# Application : ellipsoïde de Sobolev

Sous quelques hypothèses supplémentaires, dont $\overline{g} \in \mathcal{W}\left(s, \frac{6C^2}{\pi^2}\right)$ pour $s \geq 2$, avec $s$ inconnu :

## Corollaire (A & Biau, 2013)

$$\mathbb{P}\left( R(\check{\theta}_\lambda) - \overline{R} \leq \Xi' \left\{ \left(\frac{\log(n)}{n}\right)^{\frac{2s}{2s+1}} + \frac{\|\overline{\beta}\|_0 \log(pn)}{n} + \frac{\log\left(\frac{2}{\varepsilon}\right)}{n} \right\} \right) \geq 1 - \varepsilon.$$

# Modèle de régression de faible rang

## Modèle

$X$ déterministe, $\mathbb{E}(\mathcal{E}) = 0$ et

$$\underbrace{Y}_{n\times m} = \underbrace{X}_{n\times p}\underbrace{B}_{p\times m} + \mathcal{E},\ \mathrm{rg}(B) \ll \min(p, m).$$

# Modèle de régression de faible rang

## Modèle

$X$ déterministe, $\mathbb{E}(\mathcal{E}) = 0$ et

$$\underbrace{Y}_{n\times m} = \underbrace{X}_{n\times p}\underbrace{B}_{p\times m} + \mathcal{E},\ \mathrm{rg}(B) \ll \min(p, m).$$

Il faut définir une loi *a priori* pour approcher les matrices de faibles rang. Utile aussi pour d'autres modèles :

- complétion de matrice,
- régression trace...

# Loi *a priori*

J. Geweke (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* **75** 121–146.

$$k \text{ fixé, } \underbrace{B}_{p\times m} = \underbrace{M}_{p\times k}\underbrace{N^T}_{k\times m}.$$

# Loi *a priori*

J. Geweke (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* **75** 121–146.

$$k \text{ fixé}, \quad \underbrace{B}_{p\times m} = \underbrace{M}_{p\times k}\,\underbrace{N^T}_{k\times m}.$$

Soit $M_{\cdot,\ell}$ la $\ell$-ième colonne de $M$, on a

$$B = \sum_{\ell=1}^{k} M_{\cdot,\ell}(N_{\cdot,\ell})^T \quad \Rightarrow \quad \mathrm{rg}(B) \leq k.$$

# Loi *a priori*

J. Geweke (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* **75** 121–146.

$$k \text{ fixé}, \quad \underbrace{B}_{p\times m} = \underbrace{M}_{p\times k}\,\underbrace{N^T}_{k\times m}.$$

Soit $M_{\cdot,\ell}$ la $\ell$-ième colonne de $M$, on a

$$B = \sum_{\ell=1}^{k} M_{\cdot,\ell}(N_{\cdot,\ell})^T \quad \Rightarrow \quad \mathrm{rg}(B) \leq k.$$

Loi *a priori* $\pi$ : les $M_{\cdot,\ell}$ et $N_{\cdot,\ell}$ sont indépendants $\mathcal{N}(0, \gamma_\ell I)$.

# Loi *a priori*

J. Geweke (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics* **75** 121–146.

$$k \text{ fixé}, \quad \underbrace{B}_{p\times m} = \underbrace{M}_{p\times k}\,\underbrace{N^T}_{k\times m}.$$

Soit $M_{\cdot,\ell}$ la $\ell$-ième colonne de $M$, on a

$$B = \sum_{\ell=1}^{k} M_{\cdot,\ell}(N_{\cdot,\ell})^T \quad \Rightarrow \quad \mathrm{rg}(B) \leq k.$$

Loi *a priori* $\pi$ : les $M_{\cdot,\ell}$ et $N_{\cdot,\ell}$ sont indépendants $\mathcal{N}(0, \gamma_\ell I)$.

R. Salakhutdinov & A. Mnih (2008). Bayesian probabilistic matrix factorization using MCMC. *Proceedings of ICML'08*.

Fixer $k = \min(m, p)$, les $\gamma_\ell$ sont i.i.d. $1/\gamma_\ell \sim \Gamma(a, b)$, calibrée de telle façon que $\gamma_\ell$ proche de 0.

# Inégalité oracle

$$\hat{\rho}_\lambda(\mathrm{d}b) \propto \exp[-\lambda\|Xb - Y\|_F^2]\pi(\mathrm{d}b) \text{ et } \hat{B}_\lambda = \int b\hat{\rho}_\lambda(\mathrm{d}b).$$

Si $\mathcal{E}_{i,j}$ i.i.d. $\mathcal{N}(0,\sigma^2)$, $\lambda = \frac{1}{4\sigma^2}$, $a = 1$, $b = \frac{\sigma^2}{npk^2(m^2+p^2)}$.

## Théorème (A, 2013)

$\forall$ $r$, $M$ et $N$ tq $M_{\cdot,\ell} = N_{\cdot,\ell} = 0$ si $\ell > r$, $|M_{i,j}|, |N_{i,j}| \leq c'$,

$$\mathbb{E}\left(\|X\hat{B} - XB\|_F^2\right) \leq \|X(MN^T - B)\|_F^2 + C(c,c')\left\{\sigma^2 r(m+p)\log\left(\frac{nmp}{\sigma^2}\right) + \sigma^4\right\}.$$

# Inégalité oracle

$$\hat{\rho}_\lambda(\mathrm{d}b) \propto \exp[-\lambda\|Xb - Y\|_F^2]\pi(\mathrm{d}b) \text{ et } \hat{B}_\lambda = \int b\hat{\rho}_\lambda(\mathrm{d}b).$$

Si $\mathcal{E}_{i,j}$ i.i.d. $\mathcal{N}(0,\sigma^2)$, $\lambda = \frac{1}{4\sigma^2}$, $a = 1$, $b = \frac{\sigma^2}{npk^2(m^2+p^2)}$.

## Théorème (A, 2013)

$\forall\ r$, $M$ et $N$ tq $M_{\cdot,\ell} = N_{\cdot,\ell} = 0$ si $\ell > r$, $|M_{i,j}|, |N_{i,j}| \leq c'$,

$$\mathbb{E}\left(\|X\hat{B} - XB\|_F^2\right) \leq \|X(MN^T - B)\|_F^2 + C(c,c')\left\{\sigma^2 r(m+p)\log\left(\frac{nmp}{\sigma^2}\right) + \sigma^4\right\}.$$

F. Bunea, Y. She & M. Wegkamp (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics* **39** 1282–1309.

A. Rohde & A. Tsybakov (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39** 887–930.

# Contexte

$n$ paires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, $(X_i, Y_i) \sim P$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$.

# Contexte

$n$ paires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, $(X_i, Y_i) \sim P$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$.

$$\left\{ f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \bigcup_{i \in I} \Theta_i = \Theta \right\}$$

# Contexte

$n$ paires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, $(X_i, Y_i) \sim P$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$.

$$\left\{ f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \bigcup_{i \in I} \Theta_i = \Theta \right\}$$

Fonction de perte $\ell : \mathcal{Y}^2 \to \mathcal{R}_+$, $\mathbb{P}(\ell(f_\theta(X), Y) \leq C) = 1$.

# Contexte

$n$ paires i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$, $(X_i, Y_i) \sim P$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$.

$$\left\{ f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \bigcup_{i \in I} \Theta_i = \Theta \right\}$$

Fonction de perte $\ell : \mathcal{Y}^2 \to \mathcal{R}_+$, $\mathbb{P}(\ell(f_\theta(X), Y) \leq C) = 1$.

$$R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)], \text{ et } r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i).$$

# Contexte

$n$ paires i.i.d. $(X_1, Y_1), \dots, (X_n, Y_n)$, $(X_i, Y_i) \sim P$, $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$.

$$\left\{ f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \bigcup_{i \in I} \Theta_i = \Theta \right\}$$

Fonction de perte $\ell : \mathcal{Y}^2 \to \mathcal{R}_+$, $\mathbb{P}(\ell(f_\theta(X), Y) \leq C) = 1$.

$$R(\theta) = \mathbb{E}_{(X,Y)\sim P}[\ell(f_\theta(X), Y)], \text{ et } r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(X_i), Y_i).$$

$$\overline{\theta} \in \arg\min_{\Theta} R, \ \overline{\theta}_i \in \arg\min_{\Theta_i} R.$$

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,
- $\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta))\pi_i(\mathrm{d}\theta)\mathbf{1}(\theta \in \Theta_i)$,

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,
- $\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta)) \pi_i(\mathrm{d}\theta) \mathbf{1}(\theta \in \Theta_i)$,
- $\Lambda = \{2^0, 2^1, \ldots, 2^{\left\lfloor \frac{\log(n)}{\log(2)} \right\rfloor}\}$,

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,
- $\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta))\pi_i(\mathrm{d}\theta)\mathbf{1}(\theta \in \Theta_i)$,
- $\Lambda = \{2^0, 2^1, \ldots, 2^{\lfloor \frac{\log(n)}{\log(2)} \rfloor}\}$,
- $\check{\theta}_{i,\lambda} \sim \tilde{\rho}_{i,\lambda}$,

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,
- $\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta))\pi_i(\mathrm{d}\theta)\mathbf{1}(\theta \in \Theta_i)$,
- $\Lambda = \{2^0, 2^1, \ldots, 2^{\lfloor \frac{\log(n)}{\log(2)} \rfloor}\}$,
- $\check{\theta}_{i,\lambda} \sim \tilde{\rho}_{i,\lambda}$,
- un niveau de confiance $\varepsilon$.

# Un estimateur de Gibbs dans chaque modèle...

- loi *a priori* $\pi_i$ dans chaque $\Theta_i$,
- poids $\mu_i \geq 0$ tels que $\sum_{i \in I} \mu_i = 1$,
- $\tilde{\rho}_{i,\lambda}(\mathrm{d}\theta) \propto \exp(-\lambda r(\theta))\pi_i(\mathrm{d}\theta)\mathbf{1}(\theta \in \Theta_i)$,
- $\Lambda = \{2^0, 2^1, \ldots, 2^{\lfloor \frac{\log(n)}{\log(2)} \rfloor}\}$,
- $\check{\theta}_{i,\lambda} \sim \tilde{\rho}_{i,\lambda}$,
- un niveau de confiance $\varepsilon$.

Il existe $\tilde{B}_\varepsilon((i,\lambda),(i',\lambda'))$, qui ne dépend pas de $P$, t.q.

$$\mathbb{P}\left(\forall i, i', \lambda, \lambda', R(\check{\theta}_{i,\lambda}) - R(\check{\theta}_{i',\lambda'}) \leq \tilde{B}_\varepsilon((i,\lambda),(i',\lambda'))\right) \geq 1 - \varepsilon$$

$$\tilde{B}_\varepsilon((i,\lambda),(i',\lambda')) \leq \tilde{B}_\varepsilon((i,\lambda),(i'',\lambda'')) + \tilde{B}_\varepsilon((i'',\lambda''),(i',\lambda')),$$

et il existe une mesure de complexité $\mathcal{C}_\varepsilon((i,\lambda))$.

# Procédure de Lepski

On organise les paires $(i, \lambda)$ par complexité croissante : $\{t_j, j \in \{1, \dots, M\}\} = I \times \Lambda$ et $\mathcal{C}_\varepsilon(t_1) \leq \cdots \leq \mathcal{C}_\varepsilon(t_M)$.

# Procédure de Lepski

On organise les paires $(i, \lambda)$ par complexité croissante : $\{t_j, j \in \{1, \dots, M\}\} = I \times \Lambda$ et $\mathcal{C}_\varepsilon(t_1) \leq \cdots \leq \mathcal{C}_\varepsilon(t_M)$.

$$s(k) = \inf\{j \in \{1, \dots, M\},\ \tilde{B}_\varepsilon(t_k, t_j) > 0\},$$

par convention $s(k) = 0$ si pour tout $j$, $\tilde{B}_\varepsilon(t_k, t_j) \leq 0$.

# Procédure de Lepski

On organise les paires $(i, \lambda)$ par complexité croissante : $\{t_j, j \in \{1, \dots, M\}\} = I \times \Lambda$ et $\mathcal{C}_\varepsilon(t_1) \leq \cdots \leq \mathcal{C}_\varepsilon(t_M)$.

$$s(k) = \inf\{j \in \{1, \dots, M\},\ \tilde{B}_\varepsilon(t_k, t_j) > 0\},$$

par convention $s(k) = 0$ si pour tout $j$, $\tilde{B}_\varepsilon(t_k, t_j) \leq 0$. Enfin

$$\hat{k} = \min(\arg\max s)$$

et $(\hat{i}, \hat{\lambda})$ est la paire telle que $(\hat{i}, \hat{\lambda}) = t_{\hat{k}}$, finalement :

$$\check{\theta} = \check{\theta}_{\hat{i},\hat{\lambda}}.$$

# Hypothèses

## Hypothèse de marge pour $(\kappa, c) \in [1, +\infty[ \times \mathbb{R}_+^*$

$$V(\theta, \theta') = \mathbb{E}_{(X,Y)\sim P}\left\{ [\ell(f_\theta(X), Y) - \ell(f_{\theta'}(X), Y)]^2 \right\}$$

$$V(\theta, \bar{\theta}) \leq c \left[ R(\theta) - R(\bar{\theta}) \right]^{\frac{1}{\kappa}} .$$

Mammen, E. and Tsybakov, A. B. (1999) Smooth Discrimination Analysis. *The Annals of Statistics* **27**(6), pp. 1808–1829.

# Hypothèses

## Hypothèse de marge pour $(\kappa, c) \in [1, +\infty[ \times \mathbb{R}_+^*$

$$V(\theta, \theta') = \mathbb{E}_{(X,Y)\sim P}\left\{[\ell(f_\theta(X), Y) - \ell(f_{\theta'}(X), Y)]^2\right\}$$
$$V(\theta, \bar{\theta}) \leq c\left[R(\theta) - R(\bar{\theta})\right]^{\frac{1}{\kappa}}.$$

Mammen, E. and Tsybakov, A. B. (1999) Smooth Discrimination Analysis. *The Annals of Statistics* 27(6), pp. 1808–1829.

## Dimension

$$\sup_{\xi\in\mathbb{R}}\left\{\xi\left[\int_{\Theta_i} R(\theta)\pi^i_{\exp(-\xi R)}(\mathrm{d}\theta) - R\left(\bar{\theta}_i\right)\right]\right\} \leq d_i, \text{ où}$$

$$\pi^i_{\exp(-\xi R)}(\mathrm{d}\theta) \propto \exp[-\xi R(\theta)]\pi_i(\mathrm{d}\theta)\mathbf{1}(\theta \in \Theta_i).$$

Catoni, O. (2007) *PAC-Bayesian Supervised Classification : The Thermodnamics of Statistical Learning*. Institute of Mathematical Statistics (IMS) Lecture notes - monograph series 56.

# Inégalité oracle

## Théorème (A, 2008)

Sous l'hypothèse de marge et l'hypothèse de dimension, il existe $\mathcal{C} = \mathcal{C}(\kappa, c, C)$ t.q.

$$\mathbb{P}\Bigg\{ R(\check{\theta}) - R(\overline{\theta}) \leq \inf_{i \in I} \Bigg[ R\big(\overline{\theta}_i\big) - R(\overline{\theta}) + \mathcal{C} \max\Bigg\{ \Bigg( \frac{\big[R\left(\overline{\theta}_i\right) - R\left(\overline{\theta}\right)\big]^{\frac{1}{\kappa}} \left( d_i + \log \frac{1+\log_2(n)}{\varepsilon \mu_i} \right)}{n} \Bigg)^{\frac{1}{2}}, \Bigg( \frac{d_i + \log \frac{1+\log_2(n)}{\varepsilon \mu(i)}}{n} \Bigg)^{\frac{\kappa}{2\kappa-1}} \Bigg\} \Bigg] \Bigg\} \geq 1 - \varepsilon.$$

# Publications liées à cette chapitre



Alquier, P. (2008) PAC-Bayesian Bounds for R&omized Empirical Risk Minimizers. *Mathematical Methods of Statistics* **17**(4), pp. 279–304.


Alquier, P. (2010) An Algorithm for Iterative Selection of Blocks of Features. In *Proceedings of the 21th International Conference on Algorithmic Learning Theory (ALT'10)*, M. Hutter, F. Stephan, V. Vovk & T. Zeugmann Editors, Springer Lecture Notes in Artificial Intelligence 6331, pp. 35–49.


Guedj, B. & Alquier, P. (2013) PAC-Bayesian Estimation & Prediction in Sparse Additive Models. *Electronic Journal of Statistics* **7**, pp. 264–291.


Alquier, P. & Biau, G. (2013) Sparse Single-Index Models. *Journal of Machine Learning Research* **14**, pp. 243–280.


Alquier, P., Meziani, K. & Peyré, G. (2013) Adaptive Estimation of the Density Matrix in Quantum Homodyne Tomography with Noisy Data. *Inverse Problems* **29**(7), 075017.


Alquier, P., Butucea, C., Hebiri, M., Meziani, K. & Morimae, T. (2013) Rank Penalized Estimation of a Quantum System. *Physical Review A* **88**(3), 032113.


Alquier, P. (2013) Bayesian Methods for Low-rank Matrix Estimation : Short Survey & Theoretical Study. In *Proceedings of the 24th International Conference on Algorithmic Learning Theory (ALT'13)*, S. Jain, R. Munos, F. Stephan & T. Zeugmann Editors, Springer Lecture Notes in Artificial Intelligence 8139, pp. 309–323.