PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights

Joint work with Karim Lounici, University of Cambridge

Pierre Alquier

Université Paris 7, Laboratoire de Probabilités et Modèles Aléatoires
& CREST, Laboratoire de Statistiques

Workshop on Foundations and New Trends of PAC Bayesian Learning
March 23, 2010, London

High-dimensional regression estimation

Regression model

We observe n independent pairs (X_1, Y_1), ..., (X_n, Y_n) in $\mathcal{X} \times \mathbb{R}$ with

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

and $\mathbb{E}(W_i) = 0$, $\mathbb{E}(W_i^2) \leq \sigma^2$.

Objective: to approximate $f(.)$ by $f_\theta(.) = \sum_{j=1}^{p} \theta_j \phi_j(.)$ where $(\phi_j(.))_{j=1}^{p}$ is some dictionary of functions.

Problem: $p > n$.

Measures of the risk

Empirical norm: $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(X_i)^2$.

Empirical risk: $r(\theta) = \frac{1}{n}\sum_{i=1}^n \left[Y_i - f_\theta(X_i)\right]^2 = \|Y - f_\theta\|_n^2$.

Prevision risk: $R(\theta) = \mathbb{E}\left[r(\theta)\right]$.

Sparse regression estimation

Assumption: there is a $p_0 \ll n$ such that $\exists \overline{\theta} \in \arg\min R(.)$ with at most p_0 non-zero coordinates: "sparse" regression.

If these coordinates were known, we can build the LSE $\hat{\theta}_n^0$ and obtain, at least in the fixed design case

$$\mathbb{E}\left[R(\hat{\theta}_n^0) - R(\overline{\theta})\right] \leq \mathrm{cst}.\frac{\sigma^2 p_0}{n}.$$

Problem: Usually, these coordinates and even p_0 are unknown.

ℓ_0-type penalization

ℓ_0-type penalization

Define the estimator

$$\arg\min_{\theta\in\mathbb{R}^p}\left\{ r(\theta)+\lambda_{n,p}\|\theta\|_0 \right\}$$

where $\|\theta\|_0$ is the number of non-zero coordinates in θ.

Examples: C_p (Mallows, 1973), AIC (Akaike, 1973), BIC (Schwarz, 1978)...

Results with ℓ_0-type penalization

Good theoretical properties. For example:

Theorem (Bunea *et al.*, 2007)

In the fixed design case,

$$\mathbb{E}\left[R(\hat{\theta}_n^{\mathrm{BIC}}) - R(\overline{\theta})\right] \leq \mathrm{cst.}\frac{\sigma^2 p_0 \log(p)}{n}.$$

Results with ℓ_0-type penalization

Good theoretical properties. For example:

Theorem (Bunea *et al.*, 2007)

In the fixed design case,

$$\mathbb{E}\left[R(\hat{\theta}_n^{\mathrm{BIC}}) - R(\bar{\theta})\right] \leq \mathrm{cst.}\frac{\sigma^2 p_0 \log(p)}{n}.$$

Problem: 2^p possible submodels. In practice, $\hat{\theta}_n^{\mathrm{BIC}}$ can be computed for p at most a few tens!!

ℓ_1-type penalization

ℓ_1-type penalization - the LASSO (Tibshirani, 1996)

Define the estimator

$$\arg\min_{\theta\in\mathbb{R}^p} \left\{ r(\theta) + \lambda_{n,p}\|\theta\|_1 \right\}.$$

Can be computed for very large p, using for example the very popular LARS algorithm (Efron, Hastie, Johnstone & Tibshirani, 2004).

Variants: bridge regression (Frank & Friedman, 1993), nonnegative garrote (Breiman, 1995), basis pursuit (Chen, Donoho, Saunders, 2001), Dantzig selector (Candès & Tao, 2007), LOL (Kerkyacharian, Mougeot, Picard & Tribouley, 2010)...

Problem: restrictive assumption on the design are required to prove sparsity oracle inequalities:

- mutual coherence assumption (Bunea, Tsybakov & Wegkamp 2007),
- restricted eigenvalue condition (Koltchinskii, *to appear*, Bickel, Ritov & Tsybakov, *to appear*),
- ...

Bayesian statistics

Possible idea: bayesian estimator with a prior distribution $\pi(d\theta)$ that gives large probability to sparse parameters θ (George 2000 good review, Casella & Moreno 2006, Cui & George 2008 ...).

Monte Carlo methods usually allow to compute the estimators.

No theoretical results like sparsity oracle inequalities.

PAC-Bayesian approach

References: everybody in this room!

Dalalyan & Tsybakov 2008: use tools from Catoni (2007) to build an estimator

1. that can be approximated by Monte Carlo methods;
2. that satisfies a spartisy oracle inequality.

But:

1. **fixed design only**;
2. $\theta \in \mathbb{R}^p$ with $\|\theta\|_2 \leq C$ only.

Overview of the talk

The submodels

For any $J \subset \{1, ..., p\}$ and $K > 0$, we put

$$\Theta_K = \{\theta \in \mathbb{R}^p : \quad \|\theta\|_1 \leq K\},$$

$$\Theta(J) = \{\theta \in \mathbb{R}^p : \quad \theta_j \neq 0 \Leftrightarrow j \in J\},$$

$$\Theta_K(J) = \Theta_K \cap \Theta(J),$$

$$u_{\Theta_K(J)}(d\theta) = \text{ the uniform proba. measure on } \Theta_K(J).$$

For any $\theta \in \mathbb{R}^p$, **only one** $J(\theta)$ such that $\theta \in J(\theta)$.

Definition of $\hat{\theta}_n$

For the sake of simplicity, $\|\phi_j\|_n = 1$.
For any $J \subset \{1, ..., p\}$ let $\hat{\theta}_J \in \arg\min_{\theta \in \Theta(J)} r(\theta)$.

Definition

Let us choose $\lambda > 0$, we define the prior $\pi_J = 2^{-|J|-1} \binom{p}{|J|}^{-1}$ and:

$$\hat{\theta}_n = \frac{\sum_{|J| \leq n} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2 |J|}{n}\right)} \hat{\theta}_J}{\sum_{|J| \leq n} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2 |J|}{n}\right)}}.$$

Theoretical result for $\hat{\theta}_n$

We assume that there is a θ^* such that $f = f_{\theta^*}$.

Theorem

Let us assume that X_1, ..., X_n are deterministic. Let us assume W_1, ..., W_n i.i.d. $\mathcal{N}(0,\sigma^2)$, let us choose $\lambda = \frac{n}{4\sigma^2}$, then:

$$\mathbb{E}\left(\left\|f_{\hat{\theta}_n} - f\right\|_n^2\right) \leq \frac{4\sigma^2|J(\theta^*)|}{n}\log\left(\frac{7p}{|J(\theta^*)|}\right)$$

Theoretical result for $\hat{\theta}_n$

Theorem

Let us assume that X_1, ..., X_n are deterministic. Let us assume W_1, ..., W_n i.i.d. $\mathcal{N}(0,\sigma^2)$, let us choose $\lambda = \frac{n}{4\sigma^2}$, then:

$$\mathbb{E}\left(\left\|f_{\hat{\theta}_n} - f\right\|_n^2\right) \leq \min_{\theta\in\mathbb{R}^p}\left\{\left\|f_\theta - f\right\|_n^2 + \frac{4\sigma^2|J(\theta)|}{n}\log\left(\frac{7p}{|J(\theta)|}\right)\right\}$$

Definition of $\tilde{\theta}_n$

We put $\mathrm{m}(d\theta) = \sum_J 2^{-|J|-1} \binom{p}{|J|}^{-1} u_{\Theta_{K+\frac{1}{n}}(J)}(d\theta)$ for a given $K > 0$.

Definition

Let us choose $\lambda > 0$, we put

$$\tilde{\theta}_n = \frac{\displaystyle\int \theta e^{-\lambda r(\theta)} \mathrm{m}(d\theta)}{\displaystyle\int e^{-\lambda r(\theta)} \mathrm{m}(d\theta)}.$$

Motivation for definition of $\tilde{\theta}_n$

Variant of a result by Catoni (2001).

PAC-Bayesian inequality

For any $0 < \lambda < n/w$, $\theta \in \Theta_{K+c}$ and $\theta' \in \Theta_K$, $\varepsilon \in]0;1[$, with prob. at least $1-\varepsilon$,

$$R(\tilde{\theta}_\lambda) - R(\theta') \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta_{K+c})} \frac{\int r d\rho - r(\theta') + \frac{1}{\lambda}\left[\mathcal{K}(\rho, \mathrm{m}) + \log\frac{1}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}$$

where $C 8\sigma^2 + (2\|f\|_\infty + L(2K+1/n))^2$,
$w = 8[\xi + 2(\|f\|_\infty + L(K+1/(2n)))]L(2K+1/n)$.

Theoretical result for $\tilde{\theta}_n$

Theorem

Random or deterministic design. Known $\sigma > 0$ and $\xi > 0$ with $\mathbb{E}(W_i^2) \leq \sigma^2$ and $\mathbb{E}(|W_i|^k) \leq \sigma^2 k! \xi^{k-2}$ (sub-gaussian). Then, with probability at least $1-\varepsilon$, for $\lambda = \frac{n}{2\mathcal{C}_1}$,

$$R(\tilde{\theta}_n) \leq \min_{\theta \in \Theta_K} \Bigg\{ R(\theta) + \frac{3\mathcal{C}_2}{n} + \frac{8\mathcal{C}_1}{n} \left[|J(\theta)| \log \frac{np2e(K+1)}{|J(\theta)|} + \log \frac{2}{\varepsilon} \right] \Bigg\}$$

where $\mathcal{C}_1 = \mathcal{C}_1(\sigma, \xi, \|\phi_1\|_\infty, ..., \|\phi_p\|_\infty, \|f\|_\infty)$ and $\mathcal{C}_2 = \mathcal{C}_2(...)$ are known constants.

MCMC methods for the computation of the estimators

Hastings-Metropolis algorithm for $\hat{\theta}_n$ (1/2)

$$\hat{\theta}_n = \sum_{|J| \leq n} w_J \hat{\theta}_J.$$

We simulate a Markov Chain $J^{(0)}$, ..., $J^{(N)}$ with invariant distribution $(w_J)_{|J|\leq n}$.
Hastings-Metropolis:

- draw $I^{(t)}$ from $k(J^{(t)}, \cdot)$;
- take

$$J^{(t+1)} = \begin{cases} I^{(t)} & \text{with proba.} \quad \alpha(J^{(t)}, I^{(t)}) \\ & \qquad = \min\left(1, \frac{w_{I^{(t)}} k(I^{(t)}, J^{(t)})}{w_{J^{(t)}} k(J^{(t)}, I^{(t)})}\right), \\ J^{(t)} & \text{with proba.} \quad 1 - \alpha(J^{(t)}, I^{(t)}). \end{cases}$$

Hastings-Metropolis algorithm for $\hat{\theta}_n$ (2/2)

$$k(J,\cdot) = k_+(J,\cdot)\mathbb{1}_{\{|J|=0\}} + \frac{k_+(J,\cdot)+k_-(J,\cdot)}{2}\mathbb{1}_{\{0<|J|<n\}} + k_-(J,\cdot)\mathbb{1}_{\{|J|=n\}}$$

where, for $j \notin J$,

$$k_+(J, J\cup\{j\}) = \frac{e^{\zeta\left|\frac{1}{n}\sum_{i=1}^{n}[Y_i - f_{\hat{\theta}_J}(X_i)]\phi_j(X_i)\right|}}{\sum_{h\notin J} e^{\zeta\left|\frac{1}{n}\sum_{i=1}^{n}[Y_i - f_{\hat{\theta}_J}(X_i)]\phi_h(X_i)\right|}}$$

and, for $j \in J$,

$$k_-(J, J\setminus\{j\}) = \frac{e^{-\zeta\left|(\hat{\theta}_J)_j\right|}}{\sum_{h\in J} e^{-\zeta\left|(\hat{\theta}_J)_h\right|}}.$$

Reversible Jump MCMC algorithm for $\tilde{\theta}_n$

For $\tilde{\theta}_n$, we have to simulate $\theta^{(1)}$, ..., $\theta^{(N)}$ from

$$\frac{e^{-\lambda r(\theta)}\mathrm{m}(d\theta)}{\int_{\Theta_K} e^{-\lambda r(t)}\mathrm{m}(dt)}.$$

Rmk: Hastings-Metropolis with a measure $\mathrm{m}(.)$ on several subspaces known as "Reversible Jump" MCMC (Green 1995, Green & Richardson 1997).

Empirical remarks

On a small set of experiments:

1. we are able to compute $\hat{\theta}_n$ and $\tilde{\theta}_n$ for $p = 1000$,
2. better than the LASSO when $p \nearrow$ with σ fixed,
3. the LASSO is better when $\sigma \nearrow$ with p fixed,
4. computation time depends heavily on $|J(\theta^*)|$.