



Agrégation incomplète des estimateurs à seuillage dur

Yuri Golubev

► **To cite this version:**

Yuri Golubev. Agrégation incomplète des estimateurs à seuillage dur. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386563>

HAL Id: inria-00386563

<https://hal.inria.fr/inria-00386563>

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AGRÉGATION INCOMPLÈTE DES ESTIMATEURS À SEUILLAGE DUR

Yuri Golubev

Université de Provence, CMI, 22 rue F. Joliot-Curie, 13453 Marseille

Résumé. Cet exposé traite du problème de la reconstruction des vecteurs creux inconnus à partir des données bruitées. Plus précisément, il se focalise sur la méthode du seuillage dur et le choix optimal de la pénalité dans la méthode de minimisation du risque empirique. Dans le cas où les seuils utilisés appartiennent à une grille fine, nous démontrons que la pénalité optimale se trouve dans le voisinage de la pénalité minimale admissible.

This talk deals with the reconstruction of sparse vectors from noisy data. More precisely, it focuses on the hard thresholding technique and the optimal choice of the penalty in the method of the empirical risk minimization. In the case where the thresholds in use belong to a fine grid, we show that the optimal penalty is in the vicinity of the minimal admissible penalty.

Mots-clés: *seuillage dur, vecteur creux, minimisation du risque empirique, pénalité.*

Cet exposé traite de l'estimation d'un vecteur inconnu $\theta \in \mathbb{R}^n$ à partir des données bruitées

$$Y_i = \theta_i + \sigma \xi_i, \quad i = 1, \dots, n,$$

où ξ_i sont i.i.d. $\mathcal{N}(0, 1)$. Afin d'estimer θ , on utilise les estimateurs à seuillage dur

$$\hat{\theta}_i(t) = Y_i \mathbf{1}\{|Y_i| \geq t\}$$

et le but est de choisir le seuil t en se basant sur les observations. L'idée du seuillage dur est liée à l'hypothèse que le vecteur à estimer est creux. De tels vecteurs apparaissent typiquement dans les méthodes statistiques se basant sur les ondelettes, voir, par exemple, Donoho, D., Johnstone, I., Kerkycharian, G. et Picard, D. (1995).

Ensuite, les seuils sont choisis à l'aide du principe de minimisation du risque empirique

$$\hat{t} = \arg \min_{t \in \mathcal{T}} \left\{ \|Y - \hat{\theta}(t)\|^2 + Pen[\|\hat{\theta}(t)\|_0, t] \right\},$$

où \mathcal{T} est un sous-ensemble dans \mathbb{R}^+ , $\|\cdot\|$ est la norme euclidienne,

$$\|x\|_0 = \sum_{i=1}^n \mathbf{1}\{|x_i| > 0\},$$

et $Pen[\cdot, \cdot]$ est une pénalité. En général, le but principal du concept de choix adaptatif de t est de trouver la pénalité qui minimise le risque de l'estimateur $\hat{\theta}(\hat{t})$

$$R(\theta, Pen) = \mathbf{E} \sum_{i=1}^n [\theta_i - \hat{\theta}_i(\hat{t})]^2,$$

uniformément $\theta \in \mathbb{R}^n$. Pour illustrer cette idée, nous commençons par le cas, où $\mathcal{T} = \mathbb{R}^+$. Dans la littérature statistique, ce cas est connu sous le nom "sélection de modèle"s (voir, par exemple, Birgé, L. et Massart, P. (2007) pour une brève histoire de cette méthode).

Pour borner le risque de $\hat{\theta}(\hat{t})$, on introduit quelques notations supplémentaires. Pour un vecteur $x \in \mathbb{R}^n$ on va noter par $x_{(\cdot)}$ la permutation décroissante de ces composantes $x_{(1)} \geq x_{(2)} \geq \dots \geq x_{(n)}$. Soit

$$pen_\alpha(x) \stackrel{\text{def}}{=} (2x + 1) \log \frac{2Q_\alpha n}{2x + 1}, \quad \text{avec } Q_\alpha = \left(1 + \frac{1}{\alpha}\right) \exp(2 + \epsilon).$$

On commence par un résultat dans l'esprit de Birgé, L. et Massart, P. (2007).

Théorème 1. *Soit $\mathcal{T} = \mathbb{R}^+$ et $Pen_\alpha(x) = (1 + \alpha)\sigma^2 pen_\alpha(x)$, alors pour tout $\epsilon > 0$, uniformément en $\theta \in \mathbb{R}^n$*

$$R(\theta, Pen_\alpha) \leq \left(1 + \frac{1}{\alpha}\right) \mathcal{R}(\theta, Pen_\alpha) + \frac{C\sigma^2(1 + \alpha)^2}{\alpha\sqrt{\epsilon}},$$

où

$$\mathcal{R}(\theta, Pen_\alpha) = \min_k \left\{ \sum_{i=1}^k \theta_{(n-k)}^2 + Pen_\alpha(k) \right\}$$

et $C > 0$ est une constante générique.

Ce théorème prédit que: la méthode de minimisation du risque empirique s'explode lorsque $\alpha \rightarrow 0$, la pénalité presque optimale correspond à $\alpha \approx 1$, le risque de cette méthode (avec la pénalité optimale) sur l'ensemble de vecteurs creux $\mathbb{S}_{\gamma_n} = \{\theta : \|\theta\|_0 \leq \gamma_n n\}$ est asymptotiquement quatre fois supérieure au risque minimax (voir Abramovich, F., Benjamini, Y., Donoho, D. et Johstone, I. (2006)). Malheureusement, ces conclusions ne s'accordent pas bien avec la pratique statistique car elles se basent sur la borne supérieure qui est grossière.

Le résultat suivant montre que la borne supérieure du théorème 1 peut être améliorée. Définissons l'ensemble des vecteurs creux

$$\mathcal{Q}^n = \left\{ \theta \in \mathbb{R}^n : \#\{i : |\theta_i| \leq \sigma_n\} \geq qn \right\},$$

où $q > 0$ et $\sigma_n = \sigma/\sqrt{2 \log(n)}$. On va supposer que

- $\mathcal{T} = \{t_1, t_2, \dots, t_M\}$, où $M = n/\lceil \exp(1)\sqrt{\log(n)} \rceil$ et t_k sont définis par

$$n \exp[-t_k^2/(2\sigma^2)] = 1 + (k-1)\sqrt{\log(n)}.$$

- $Pen_\alpha(x, t) = (1 + \alpha)\sigma^2 pen_{\alpha/2}(k) + \alpha n(t^2/2 + \sigma^2) \exp[-t^2/(2\sigma^2)]$ avec $\alpha > 0$.

Théorème 2. Uniformément en $\theta \in \mathbb{Q}^n$ et $t \in \mathcal{T}$

$$\begin{aligned} R(\theta, Pen_\alpha) \leq & \mathcal{R}_0(\theta, t) + \sigma^2 \#\{i : |\theta_i| \geq \sigma_n\} \\ & + \frac{C\mathcal{R}_1(\theta, t)}{\alpha^{3/2}\epsilon^{1/4}} \left[1 + \log_+ \frac{C\sqrt{\epsilon}\alpha^3 n\sigma^2}{\mathcal{R}_1(\theta, t)} \right]^{-1/4}, \end{aligned}$$

où

$$\mathcal{R}_0(\theta, t) = \sum_{i=1}^n \theta_i^2 \mathbf{P}\{|Y_i| \leq t\} + Pen_\alpha \left[\sum_{i=1}^n \mathbf{P}\{|Y_i| > t\}, t \right]$$

et

$$\mathcal{R}_1(\theta, t) = \mathcal{R}_0(\theta, t) + t^2 \sqrt{\#\{i : |\theta_i| \geq \sigma_n\}}.$$

Le sens statistique de ce théorème est assez transparent. Si le vecteur à estimer est creux, alors $\mathcal{R}_1(\theta, t) \asymp \mathcal{R}_0(\theta, t)$ et $\mathcal{R}_0(\theta, t) \ll n\sigma^2$ pour tout t raisonnable. Cela signifie que $\mathcal{R}_0(\theta, t)$ est le terme principal dans la borne du théorème 2 et son minimum par rapport à α est atteint pour $\alpha \approx 0$. Notons également que cette borne supérieure s'explode, mais contrairement au théorème 1, c'est le terme de deuxième ordre qui s'explode. Cette observation permet de démontrer en particulier que l'estimateur $\hat{\theta}(\hat{t})$ est asymptotiquement minimax sur \mathbb{S}_{γ_n} .

Bibliographie

- [1] Abramovich, F., Benjamini, Y., Donoho, D. et Johnstone, I. (2006) Adapting to unknown sparsity by controlling false discovery rate. *Ann. Statist.*, 34, 584–653.
- [2] Birgé, L. et Massart, P. (2007) Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* 138, 33–73.
- [3] Donoho, D., Johnstone, I., Kerkycharian, G. et Picard, D. (1995) Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57, 301–369.