Agrégation d'estimateurs pour la régression hétéroscédastique

Arnak Dalalyan [1] et Joseph Salmon [2]

[1]École des Ponts ParisTech

[2]Université Paris Diderot - Paris 7

Septembre 2010 -Journées MAS
Bordeaux

Plan

Introduction

Motivations

- Théorique : inégalités oracles
- Grande dimension — sparsité
  Applications : traitement d'images, génétique, internet, finance
- Problèmes inverses, adaptation

Notations et modèle

- Données : $\mathcal{D}_n = \{(x_1, Y_n), \cdots, (x_n, Y_n)\}$, $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$
- Modèle hétéroscédasctique gaussien :
  $\{x_i\}$ déterministes, fonction f,

$$Y_i = f(x_i) + \sigma_i \varepsilon_i, \quad i = 1, \cdots, n \qquad (\star)$$

$$\varepsilon_i \quad \text{i.i.d} \quad \mathcal{N}(0,1) \quad \text{et} \quad \Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_n)$$

  Rem : Σ connue (ou seulement $\max(\sigma_i)$)
- Risque : pour tout estimateur $\hat{f}$

$$r = \mathbb{E}\left(\left\|f - \hat{f}_n\right\|_n^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2\right)$$

- Estimateur sans biais du risque $\hat{r}_n$: $\mathbb{E}(\hat{r}_n) = r$

Lien problème inverse/hétéroscédasticité

T : opérateur ***connu*** sur un Hilbert $(\mathcal{H}, \langle \cdot | \cdot \rangle_{\mathcal{H}})$ (penser matrice...)
Y : processus aléatoire indexé par $g \in \mathcal{H}$, pour tout $h \in \mathcal{H}$

$$Y = Th + \varepsilon\xi \quad \Longleftrightarrow \quad Y(g) = \langle Th | g \rangle_{\mathcal{H}} + \varepsilon\xi(g), \quad \forall g \in \mathcal{H},$$

T^* : l'adjoint de T ; si $T^* T$ est compact, par SVD

$$T\phi_k = b_k \psi_k, \quad T^*\psi_k = b_k \phi_k, \qquad k \in \mathbb{N},$$

b_k : valeurs singulières,$\{\phi_k\}$: base orthonormale de $\mathcal{H}$, $\{\psi_k\}$: base orthonormale de $\mathrm{Im}(T) \subset \mathcal{H}$. Le modèle se ré-écrit

$$Y(\psi_k) = \langle h | \phi_k \rangle_{\mathcal{H}} \, b_k + \varepsilon\xi(\psi_k), \qquad k \in \mathbb{N}.$$

Si $b_k \neq 0$ le modèle est équivalent à ($\star$), avec $f(x_i) = \langle h | \phi_i \rangle_{\mathcal{H}}$ et $\sigma_i = \varepsilon b_i^{-1}$
Exemples de problèmes inverses : estimation de dérivée, déconvolution avec un noyau connu, tomographie, etc.

Agrégation d'estimateurs

Famille de « pré-estimateurs » : $\mathcal{F}_\Lambda = (\hat{f}_\lambda)_{\lambda \in \Lambda} \in \mathbb{R}^n$ avec $\Lambda \subset \mathbb{R}^M$.
$\hat{r}_\lambda$: estimateur sans biais du risque, $\mathbb{E}(\hat{r}_\lambda) = \mathbb{E}(\|f - \hat{f}_\lambda\|_n^2)$

Méthode par pénalisation : $\hat{f}_\lambda = f_\lambda = X\lambda$

$$\hat{f}^{\mathrm{Pen}} = \hat{f}_{\hat{\lambda}}, \quad \text{où} \quad \hat{\lambda} = \arg\min_{\lambda \in \Lambda} \Big(\underbrace{\hat{r}_\lambda}_{\text{adéquation}} + \underbrace{\mathrm{Pen}(\lambda)}_{\text{régularisation}} \Big)$$

- $\mathrm{Pen}(\lambda) = c\|\lambda\|_0$: pénalité BIC
- $\mathrm{Pen}(\lambda) = c\|\lambda\|_2^2$: pénalité Ridge (ou filtre de Wiener)
- $\mathrm{Pen}(\lambda) = c\|\lambda\|_1$: pénalité LASSO
- Versions par blocs, etc.

Agrégation à poids exponentiels (EWA)

- Extension : élargir l'espace de recherche, changer de pénalité
- Espace de recherche :
  $\mathcal{P} = \{p$: probabilité tq $\mathbb{E} \int_\Lambda \|\hat{f}_\lambda\|_n^2 p(d\lambda) < \infty\}$
- Pénalisation étendue : $\hat{f}^{\mathrm{Pen}} = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\mathrm{Pen}}(d\lambda)$ avec

$$\hat{\pi}^{\mathrm{Pen}} = \arg\min_{p \in \mathcal{P}} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \int_\Lambda \mathrm{Pen}(\lambda) p(d\lambda) \right)$$

- Pénalisation KL : π *a priori* sur Λ, l'EWA est

$$\hat{f}^{\mathrm{Ewa}} = \int_\Lambda \hat{f}_\lambda \hat{\pi}^{\mathrm{Ewa}}(d\lambda)$$

$$\hat{\pi}^{\mathrm{Ewa}} = \arg\min_{p \in \mathcal{P}} \left(\int_\Lambda \hat{r}_\lambda p(d\lambda) + \frac{\beta}{n} \mathcal{K}(p, \pi) \right)$$

- Solution explicite : $\boxed{\hat{\pi}^{\mathrm{Ewa}}(d\lambda) \propto \exp(-n\hat{r}_\lambda/\beta)\pi(d\lambda)}$

Pré-estimateurs Affines

Forme des pré-estimateurs : $\hat{f}_\lambda = A_\lambda \boldsymbol{Y} + b_\lambda$

Risque associé : $r_\lambda = \mathbb{E}[\|\hat{f}_\lambda - f\|_n^2]$

Formule de Stein (cas hétéroscédasctique)

Pour le modèle $(\star)$, si $\hat{f}$ est différentiable presque partout en Y et que $\partial_{y_i}\hat{f}_i$ est intégrable alors

$$\hat{r} = \| \boldsymbol{Y} - \hat{f}\|_n^2 + \frac{2}{n}\sum_{i=1}^{n} \sigma_i^2 \partial_{y_i}\hat{f}_i - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \,,$$

est un estimateur sans biais du risque r

Conclusion : $\hat{r}_\lambda = \| \boldsymbol{Y} - \hat{f}_\lambda\|_n^2 + \frac{2}{n}\mathrm{Tr}(\Sigma A_\lambda) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$

Cas constant $A_\lambda = 0$

Pré-estimateurs déterministes : $\hat{f}_\lambda = f_\lambda$

Exemples

- cadre linéaire usuel : $f_\lambda = X\lambda$
- $\mathcal{F}_\Lambda = (f_\lambda)_{\lambda \in \Lambda}$ est un « dictionnaire » fini
- $\mathcal{F}_\Lambda = (\hat{f}_\lambda)_{\lambda \in \Lambda}$ est obtenue par découpage (splitting) de l'échantillon :
  Première partie des données : création des pré-estimateurs
  Deuxième partie des données : agrégation des pré-estimateurs

Cas linéaire : $b_\lambda = 0$ (1)

Moindres carrés ordinaires $\hat{f}_\lambda = A_\lambda \boldsymbol{Y}$

$\{\mathcal{S}_\lambda : \lambda \in \Lambda\}$ ensemble de sous-espaces de $\mathbb{R}^n$

A_λ : projecteurs orthogonaux sur $\mathcal{S}_\lambda$, Leung et Barron (2006)

Pré-estimateurs diagonaux

$$\hat{f} = A\,\boldsymbol{Y} \quad \text{avec} \quad A = \operatorname{diag}(a_1, \ldots, a_n)$$

- Projections ordonnées : $a_k = \mathbb{1}_{(k \leq \lambda)}$ pour λ entier, ie. $\Lambda = \{1, \ldots, n\}$
- Projections par blocs : $a_k = \mathbb{1}_{(k \leq w_1)} + \sum_{j=1}^{m-1} \lambda_j \mathbb{1}_{(w_j \leq k \leq w_{j+1})}$ avec $\lambda_j \in \{0, 1\}$. $\Lambda = \{0, 1\}^{m-1}$
- Filtre de Tikhonov-Philipps : $a_k = \frac{1}{1 + (k/w)^\alpha}$, où $w, \alpha > 0$. $\Lambda = (\mathbb{R}_+^*)^2$.
- Filtre de Pinsker : $a_k = \left(1 - \frac{k^\alpha}{w}\right)_+$, où $x_+ = \max(x, 0)$ et $w, \alpha > 0$. $\Lambda = (\mathbb{R}_+^*)^2$.

Cas linéaire : $b_\lambda = 0$ (2)

Kernel ridge regression

Noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, f est dans l'espace de de Hilbert à noyau associé $(\mathcal{H}_k, \| \cdot \|_{H_k})$. L'estimateur est défini par

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}_k} \left(\| \boldsymbol{Y} - f \|_n^2 + \lambda \| f \|_{H_k}^2 \right)$$

K : matrice $n \times n$ du noyau, $K_{i,j} = k(x_i, x_j)$
Solution : $\hat{f}_\lambda = A_\lambda \boldsymbol{Y}$, avec $A_\lambda = K(K + n\lambda I_{n\times n})^{-1}$

Multiple Kernel ridge regression

$k_1, \ldots, k_M$: M noyaux, $K_1, \ldots, K_M$ les matrices associées. Si $\lambda = (\lambda_1, \ldots, \lambda_M) \in \Lambda = \mathbb{R}_+^M$, alors $\hat{f}_\lambda = A_\lambda \boldsymbol{Y}$ avec

$$A_\lambda = \left(\sum_{j=1}^{M} \lambda_j K_j \right) \left(\sum_{j=1}^{M} \lambda_j K_j + n I_{n\times n} \right)^{-1} .$$

Rem : forme liée au group Lasso, Arlot et Bach (2009)

Conditions du Théorème Principal

Condition $\mathbf{C}_1$

Matrices A_λ : projections orthogonales ($A_\lambda^2 = A_\lambda^\top = A_\lambda$)
Vecteurs b_λ : $A_\lambda b_\lambda = 0$

Exemple : A_λ projections des sous espaces de $\mathbb{R}^M$ Leung et Barron (2006)

Condition $\mathbf{C}_2$

Matrices A_λ : symétriques, semi-définies positives et
$A_\lambda A_{\lambda'} = A_{\lambda'} A_\lambda$, $A_\lambda \Sigma = \Sigma A_\lambda \ \forall \lambda, \lambda' \in \Lambda$.
Vecteurs b_λ : $b_\lambda = 0$.

Exemple : les A_λ sont des estimateurs par seuillage dans la base canonique Leung (2004)

Énoncé Théorème Principal

Borne PAC-Bayesienne

Si $\mathbf{C}_1$ ou $\mathbf{C}_2$ est vérifiée, alors l'agrégat à poids exponentiel $\hat{f}_{\text{EWA}}$ vérifie pour tout choix d' *a priori* π :

$$\mathbb{E}(\|\hat{f}_{\text{EWA}} - f\|_n^2) \leq \inf_{p \in \mathcal{P}_\Lambda} \left(\int_\Lambda \mathbb{E}\|\hat{f}_\lambda - f\|_n^2 \, p(d\lambda) + \frac{\beta}{n} \, \mathcal{K}(p, \pi) \right)$$

pour $\beta \geq \alpha \max_{i=1,\dots,n} \sigma_i^2$. $\alpha = 4$ sous $\mathbf{C}_1$ et $\alpha = 8$ sous $\mathbf{C}_2$.

Rem : $A_\lambda = 0$, inégalité donnée dans le cas Λ continu, Dalalyan et Tsybakov (2007)

Corollaire : cas discret

Inégalité Oracle : $\Lambda = [\![1, M]\!]$, π uniforme

Si $\mathbf{C}_1$ ou $\mathbf{C}_2$ est vérifiée, et si π est uniforme sur $[\![1, M]\!]$, alors

$$\mathbb{E}(\|\hat{f}_{\mathsf{EWA}} - f\|_n^2) \leq \inf_{\lambda \in [\![1,M]\!]} \left(\mathbb{E}\|\hat{f}_\lambda - f\|_n^2 + \frac{\beta \log(M)}{n} \right)$$

pour $\beta \geq \alpha \max_{i=1,\ldots,n} \sigma_i^2$. $\alpha = 4$ sous $\mathbf{C}_1$ et $\alpha = 8$ sous $\mathbf{C}_2$.

- Pour $b_\lambda = 0$, résultat de Leung et Barron (2006)
- Pour $A_\lambda = 0$, et si pour tout i, $\sigma_i = \sigma$: l'inégalité est optimale Tsybakov (2003)

Inégalité Oracle Sparse

Scénario sparse : il existe un vecteur sparse $\lambda^* \in \Lambda = \mathbb{R}^M$ tq $\hat{f}_{\lambda^*} \approx f$. Choix d'un *a priori* favorisant la sparsité

$$\pi(d\lambda) \propto \prod_{j=1}^{M} \frac{1}{(1+|\lambda_j/\tau|^2)^2} \mathbb{1}_\Lambda(\lambda),$$

$\tau > 0$: paramètre de concentration.

Inégalité Oracle

Prenons π définit ci-dessus, supposons que $\lambda \mapsto r_\lambda$ est $\mathcal{C}^1$, et qu'il existe une matrice $\mathcal{M}$ de taille $M \times M$ tq :

$$r_\lambda - r_{\lambda'} - \nabla r_{\lambda'}^\top(\lambda - \lambda') \leq (\lambda - \lambda')^\top \mathcal{M}(\lambda - \lambda'), \quad \forall \lambda, \lambda' \in \Lambda.$$

Si $\mathbf{C}_1$ ou $\mathbf{C}_2$ est vérifiée

$$\mathbb{E}(\|\hat{f}_{\mathsf{EWA}} - f\|_n^2) \leq \inf_{\lambda \in \mathbb{R}^M} \left\{ \mathbb{E}\|\hat{f}_\lambda - f\|_n^2 + \frac{4\beta}{n} \sum_{j=1}^{M} \log\left(1 + \frac{|\lambda_j|}{\tau}\right) \right\} + \mathrm{Tr}(\mathcal{M})\tau^2$$

pour $\beta \geq \alpha \max_{i=1,\dots,n} \sigma_i^2$. $\alpha = 4$ sous $\mathbf{C}_1$ et $\alpha = 8$ sous $\mathbf{C}_2$.

Point de vue minimax (cas homoscédastique)

$\theta_k(f) = \langle f | \varphi_k \rangle_n$: coefficients de la transformée (orthogonale) Fourier discrète de f, notée $\mathcal{D}f$

Ellipsoïde de Sobolev : $\mathcal{F}(\alpha, R) = \{f \in \mathbb{R}^n : \sum_{k=1}^n k^{2\alpha} \theta_k(f)^2 \leq R\}$

Théorème de Pinsker

$$\begin{aligned} \inf_{\hat{f}} \sup_{f \in \mathcal{F}(\alpha,R)} \mathbb{E}(\|\hat{f} - f\|_n^2) &\sim \inf_{A} \sup_{f \in \mathcal{F}(\alpha,R)} \mathbb{E}(\| A\,\boldsymbol{Y} - f\|_n^2) \\ &\sim \inf_{w>0} \sup_{f \in \mathcal{F}(\alpha,R)} \mathbb{E}(\| A_{\alpha,w}\,\boldsymbol{Y} - f\|_n^2) \end{aligned}$$

$\inf$ est sur tous les estimateurs $\hat{f}$ possibles et
$A_{\alpha,w} = \mathcal{D}^\top \mathrm{diag}\left((1 - k^\alpha / w)_+ ; k = 1, \ldots, n\right) \mathcal{D}$: Filtre de Pinsker

Morale : Estimateurs linéaires minimax sur les ellipsoïdes

EWA pour l'adaptation (cas homoscédastique)

Estimateur adaptatif : ne dépend pas de (α, R)
Exemple : Estimateur de James-Stein par blocs

Adaptation

Pour $\Lambda = (\mathbb{R}_+^*)^2$, *a priori* $\pi(d\lambda) = \frac{2}{w^3} e^{-\alpha} \mathbb{1}_{(0,\infty)\times(1,\infty)}(\alpha, w)$, où $\lambda = (\alpha, w)$, l'estimateur $\hat{f}_{\text{EWA}}$ avec la température $\beta = 8\sigma^2$ et les pré-estimateurs : $\hat{f}_\lambda = \hat{f}_{\alpha,w} = A_{\alpha,w}\boldsymbol{Y}$ ($A_{\alpha,w}$: filtre de Pinsker) est adaptatif au sens minimax exacte sur la famille $\{\mathcal{F}(\alpha, R) : \alpha > 0, R > 0\}$

Rem : pour implémenter l'EWA, l'intrégrale est seulement dans $\mathbb{R}^2$

Conclusion

Améliorations

- Extension au cadre hétéroscédasctique
- Famille de pré-estimateurs plus large

Limites

- Variance ou borne sur la variance supposée connue
- Restriction importante sur la forme des pré-estimateurs

Travail en cours

- Simulations
- ...

Références

- S. Arlot and F. Bach.
  Data-driven calibration of linear estimators with minimal penalties.
  In *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.
- A. S. Dalalyan and A. B. Tsybakov.
  Aggregation by exponential weighting, sharp oracle inequalities and sparsity.
  In *COLT*, pages 97–111, 2007.
- G. Leung and A. R. Barron.
  Information theory and mixing least-squares regressions.
  IEEE Trans. Inf. Theory, 52(8) :3396–3410, 2006.
- G. Leung.
  Information Theory and Mixing Least Squares Regression.
  PhD thesis, Yale University, 2004.
- A. B. Tsybakov.
  Optimal rates of aggregation.
  In *COLT*, pages 303–313, 2003.