

Estimation de densité par pénalisation ℓ_1 minimale

Erwan LE PENNEC

SELECT, INRIA Saclay - Univ. Paris Sud

en collaboration avec Karine BERTIN (Univ. de Valparaiso, Chili)
et Vincent RIVOIRARD (Univ. Paris Dauphine)

GdR ISIS - 10 novembre 2010

Cadre

Estimation de densité

- $X_1, \dots, X_n$ i.i.d de loi $f_0(x)dx$.
- Hyp : $f_0 \in L^2$.
- Pas d'hypothèse $f_0 \in L^\infty$ et surtout pas $\|f_0\|_\infty$ connue.

Approche dictionnaire

- Dictionnaire arbitraire $\mathcal{D} = \{\phi_k\}_{1 \leq k \leq p}$.
- Hyp : $\phi_k \in L^2 \cap L^\infty$ et $\|\phi_k\|_\infty$ connue.
- Notation : $f_\lambda = \sum_{k=1}^p \lambda_k \phi_k$.

Objectifs

- Estimation de f_0 par une combinaison linéaire f_λ du dictionnaire $\mathcal{D}$.
- Mesure de l'erreur en norme L^2 .
- Cas $p \gg n$ possible...

Estimateur de Dantzig

Produits scalaires

- Produits scalaires et espérances : $\beta_k = \langle f_0, \phi_k \rangle = \mathbb{E}(\phi_k(X))$.
- Produits scalaires empiriques : $\hat{\beta}_k = \frac{1}{n} \sum_{i=1}^n \phi_k(X_i) \simeq \beta_k$.
- G : matrice de Gram du dictionnaire $\mathcal{D}$ ($G_{k,k'} = \langle \phi_k, \phi_{k'} \rangle$).
- Pour $f_\lambda = \sum_{k=1}^p \lambda_k \phi_k$, $\langle f_\lambda, \phi_k \rangle = (G\lambda)_k$

Contraintes et régularisation

- Recherche de λ tel que $G\lambda \simeq \hat{\beta} \simeq \beta$.
- Contraintes de Dantzig : $\forall k \in \{1, \dots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k$.
- + Régularisation par la norme ℓ_1 de λ .
- Estimateur de Dantzig :

$$\hat{\lambda}^D = \operatorname{argmin} \|\lambda\|_{\ell_1} \quad \text{sous} \quad \forall k \in \{1, \dots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k.$$

Estimateur Lasso / SPADES

Norme quadratique

- Erreur L^2 : $\|f - f_0\|_2^2 = \|f\|_2^2 - 2\mathbb{E}(f(X)) + \|f_0\|_2^2$.
- Contraste L^2 : $\gamma(f) = -2\frac{1}{n} \sum_{i=1}^n f(X_i) + \|f\|_2^2 \simeq \|f - f_0\|_2^2 - \|f_0\|_2^2$.
- Pour $f_\lambda = \sum_{k=1}^p \lambda_k \phi_k$, $\gamma(f_\lambda) = -2\lambda^* \hat{\beta} + \lambda^* G \lambda$.

Contraste et pénalisation

- Minimisation du contraste empirique pénalisé par $2 \sum_k \eta_k |\lambda_k|$.
- SPADES/Estimateur Lasso :
$$\hat{\lambda}^L = \operatorname{argmin} \frac{1}{2} \left(-2\lambda^* \hat{\beta} + \lambda^* G \lambda \right) + \sum_{k=1}^p \eta_k |\lambda_k|.$$

Lien Dantzig/Lasso

- Condition d'optimalité de premier ordre du Lasso
= Contraintes de Dantzig, $\forall k \in \{1, \dots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k$
- Lien fort entre les deux estimateurs...

Bibliographie parcimonieuse

Pénalisation ℓ_1 en statistique

- Frank et Friedman (91–93) Bridge.
- Tibshirani et al. (94–96) Lasso.
- Donoho et al. (95–01) Basis Pursuit.

Parcimonie et système sous déterminé

- Donoho (2004) Parcimonie et matrice gaussienne.
- Candes, Tao et Romberg (2004) Parcimonie et matrice de Fourier.

Parcimonie et ℓ_1 en statistique

- Candes et Tao (2007) Dantzig.
- Lasso/Dantzig, SPADES : Tsybakov, Bickel, Ritov, Bunea, Wegkamp
- Énorme littérature : Plan, Temlyakov, Lounici, van de Geer,...

Inégalités de concentration

Choix de η_k

- Comment choisir η_k pour que $\mathbb{P}(|\beta_k - \hat{\beta}_k| > \eta_k) \leq 2e^{-u^2/2}$?
- Concentration d'un processus empirique simple autour de sa moyenne.

Hoeffding

- $\eta_k = u \frac{\|\phi_k\|_\infty}{\sqrt{n}} \implies \mathbb{P}(|\beta_k - \hat{\beta}_k| > \eta_k) \leq 2e^{-u^2/2}.$
- Souvent utilisé mais majoration grossière n'utilisant pas f_0 et limitation sur la norme des ϕ_k .

Bernstein

- Utilisation de la variance : $\sigma_k^2 = \text{Var}(\phi_k(X))$.
- $\eta_k = u \frac{\sigma_k}{\sqrt{n}} + \frac{u^2}{3} \frac{\|\phi_k\|_\infty}{n} \implies \mathbb{P}(|\beta_k - \hat{\beta}_k| > \eta_k) \leq 2e^{-u^2/2}.$
- Prise en compte fine de f_0 et limitation beaucoup moins forte sur la norme de ϕ_k .

Inégalités de concentration

Bernstein (suite)

- Pb : Comme f_0 , la variance est inconnue...
- Majoration possible mais perte du caractère fin :
 - $\sigma_k \leq \|\phi_k\|_\infty$ ($\simeq$ Hoeffding)
 - $\sigma_k \leq \|f\|_\infty^{1/2} \|\phi_k\|_2$ (Pb hypothèse sur $\|f\|_\infty$)

Adaptation/Auto renormalisation

- Variance empirique : $\hat{\sigma}_k^2 = \frac{1}{n-1} \sum_{i=1}^n (\phi_k(X_i) - \hat{\beta}_k)^2$
- $\eta_k = u \frac{\hat{\sigma}_k}{\sqrt{n}} + 7 \frac{u^2}{3} \frac{\|\phi_k\|_\infty}{n} \implies \mathbb{P}(|\beta_k - \hat{\beta}_k| > \eta_k) \leq 2e^{-u^2/2}.$
- Prise en compte adaptative de la variabilité de f_0 .
- De plus, $\mathbb{P}\left(|\hat{\sigma}_k - \sigma_k| > 2u \frac{\|\phi_k\|_\infty}{n}\right) \leq 2e^{-u^2/2}.$

Contrôle sur les coefficients avec grande probabilité

Contrôle sur les coefficients

- Borne d'union avec $u = \sqrt{2\gamma \log p}$.

- $$\eta_k^\gamma = \sqrt{2\gamma \log p} \frac{\hat{\sigma}_k}{\sqrt{n}} + \frac{14}{3} \gamma \log p \frac{\|\phi_k\|_\infty}{n}$$

L'évènement Ω_{γ_1}

- Pour tout $\gamma_1 > 0$, évènement Ω_{γ_1} de proba $\geq 1 - 2 \left(\frac{1}{p}\right)^{\gamma_1 - 1}$:

$$\Omega_{\gamma_1} = \{\forall k \in \{1, \dots, p\}, |\hat{\beta}_k - \beta_k| \leq \eta_k^{\gamma_1}\}$$

- De plus sous Ω_{γ_1} : $\forall k \in \{1, \dots, p\}, |\hat{\sigma}_k - \sigma_k| \leq 2\sqrt{2\gamma_1 \log p} \frac{\|\phi_k\|_\infty}{\sqrt{n}}$

$$\begin{aligned} \sup_k \eta_k^\gamma &= \|\eta^\gamma\|_\infty \leq \sqrt{2\gamma \log p} \frac{\sigma_k}{\sqrt{n}} \\ &\quad + 4(\sqrt{\gamma \gamma_1} + \frac{7}{6} \gamma) \log p \frac{\|\phi_k\|_\infty}{n} \end{aligned}$$

Estimateurs

Contrainte de Dantzig

- Pour $\gamma_0 > 0$, $\forall k \in \{1, \dots, p\}$

$$|(G\lambda)_k - \hat{\beta}_k| \leq \eta_k^{\gamma_0} = \sqrt{2\gamma_0 \log p} \frac{\hat{\sigma}_k}{\sqrt{n}} + \frac{14}{3} \gamma_0 \log p \frac{\|\phi_k\|_\infty}{n}.$$

Estimateurs pénalisés adaptatifs

- Dantzig :

$$\hat{\lambda}^{D, \gamma_0} = \operatorname{argmin} \|\lambda\|_{\ell_1} \text{ sous } \forall k \in \{1, \dots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k^{\gamma_0}.$$

- Lasso : $\hat{\lambda}^{L, \gamma_0} = \operatorname{argmin} \frac{1}{2} \left(-2\lambda^* \hat{\beta} + \lambda^* G\lambda \right) + \sum_{k=1}^p \eta_k^{\gamma_0} |\lambda_k|.$

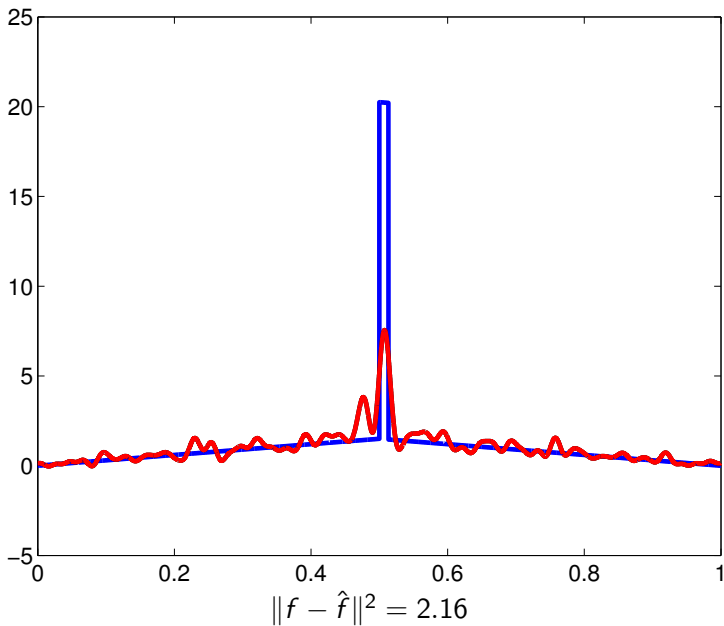
Variantes

- Estimateurs non adaptatifs en remplaçant $\eta_k^{\gamma_0}$ par $\tilde{\eta}_k^{\gamma_0}$:

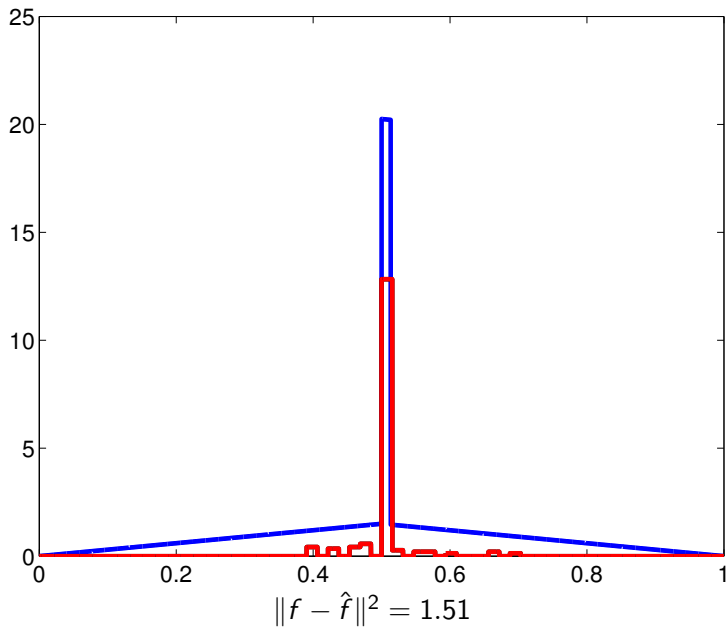
$$\tilde{\eta}_k^{\gamma_0} = \min \left(\sqrt{2\gamma_0 \log p} \frac{\|\phi_k\|_\infty}{\sqrt{n}}, \sqrt{2\gamma_0 \log p} \frac{\|f_0\|_\infty^{1/2} \|\phi_k\|_2}{\sqrt{n}} + \frac{2}{3} \gamma_0 \log p \frac{\|\phi_k\|_\infty}{n} \right)$$

- Amélioration possible avec une seconde étape de type moindre carré sur le support de $\hat{\lambda}$ (à la Gauss-Dantzig).

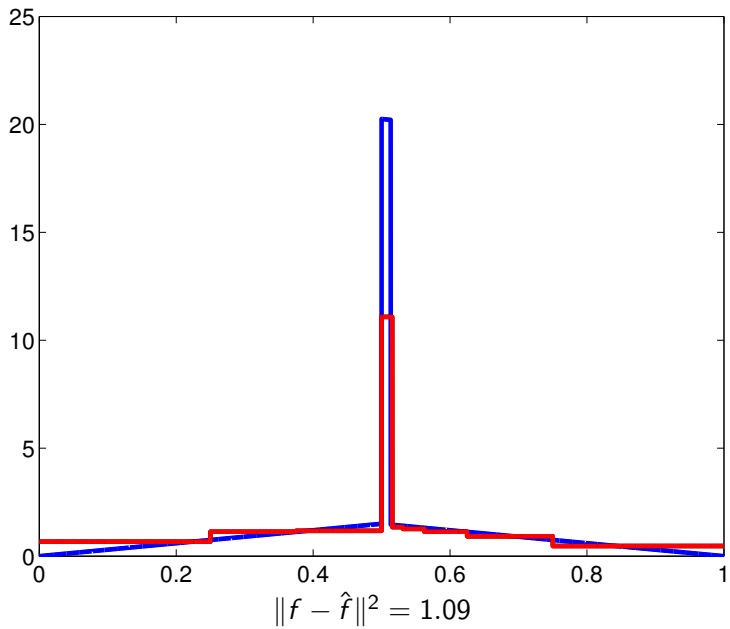
Fourier



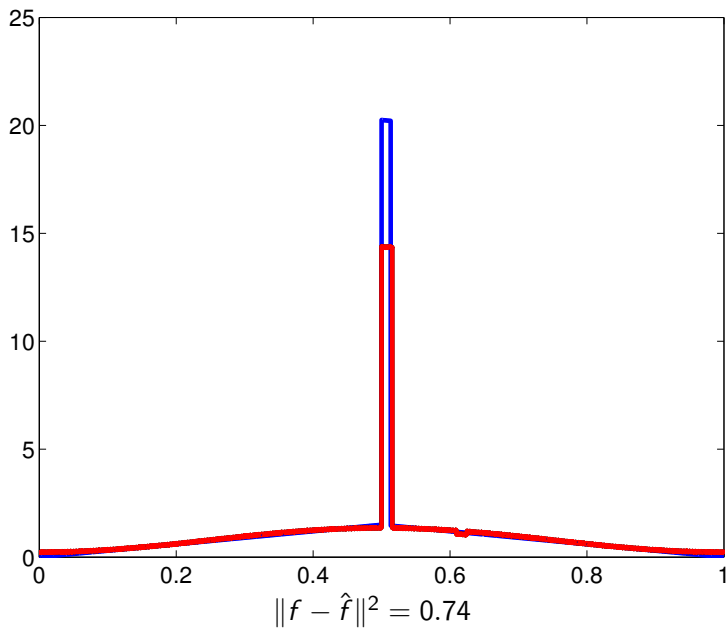
Boxes



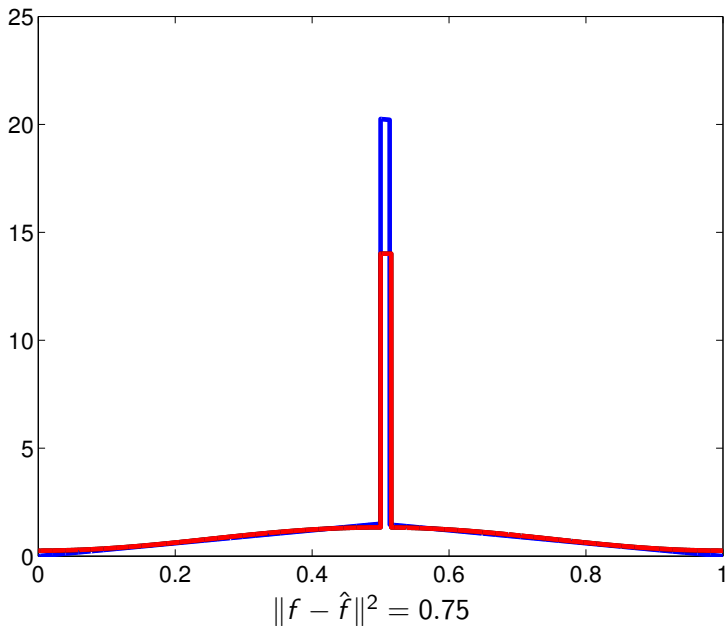
Haar



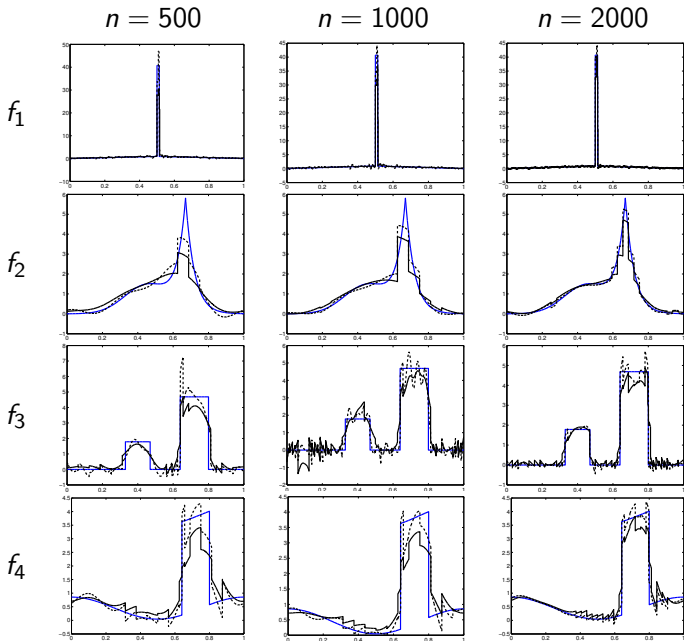
Mix = Fourier + Boxes



Mix2 = Fourier + Boxes + Haar

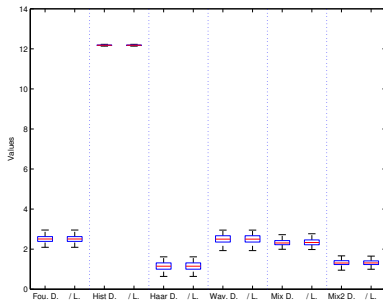
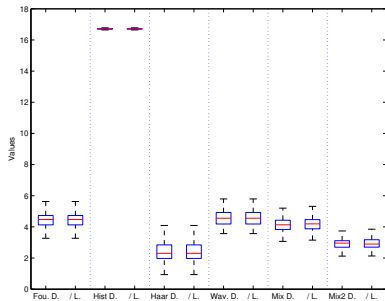


Dantzig et Dantzig+LS

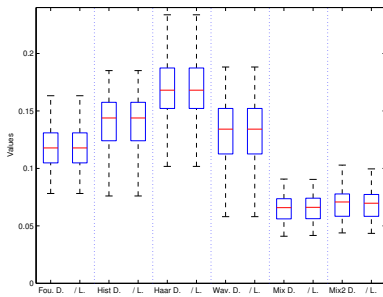
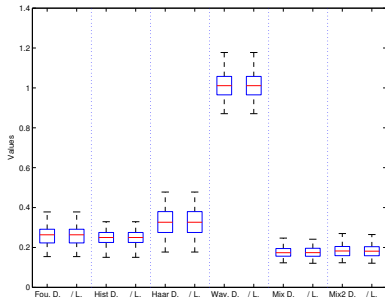


Dantzig / Lasso f_1/f_2

f_1



f_2

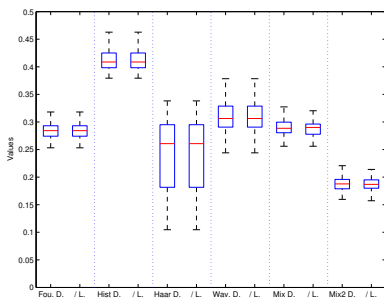
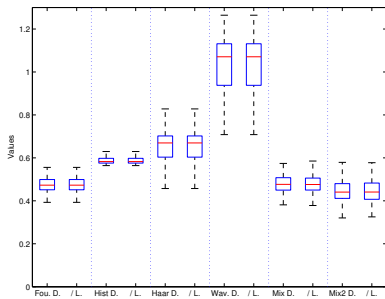


$n = 500$

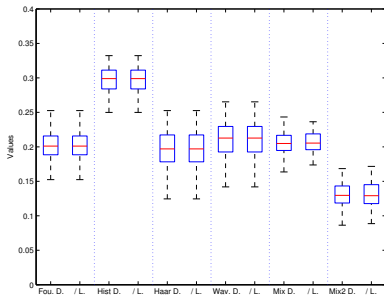
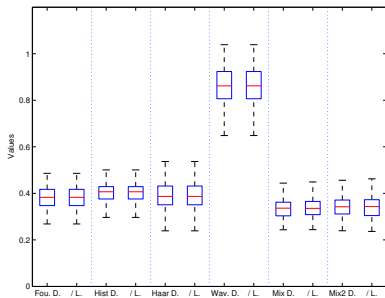
$n = 2000$

Dantzig / Lasso f_3/f_4

f_3



f_4

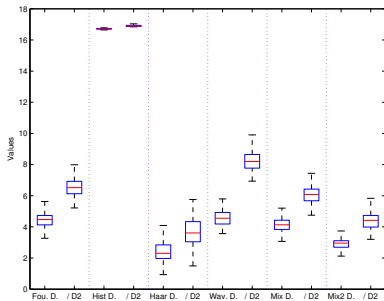


$n = 500$

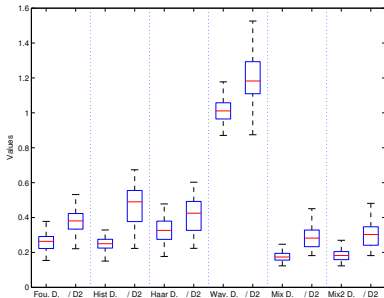
$n = 2000$

Dantzig / Non adaptive D. f_1/f_2

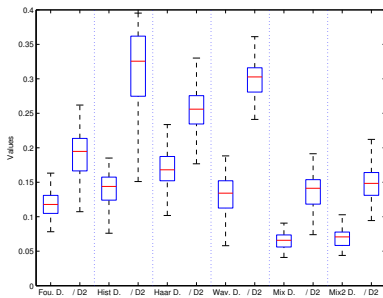
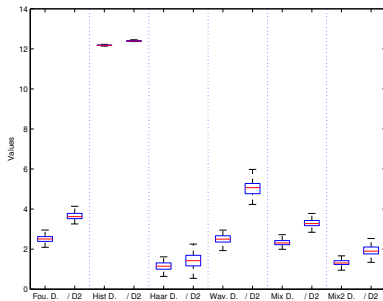
f_1



f_2



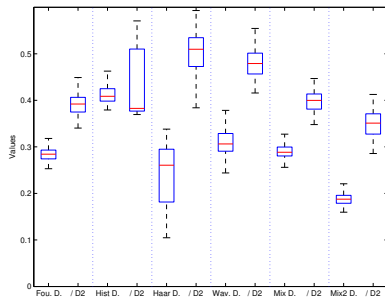
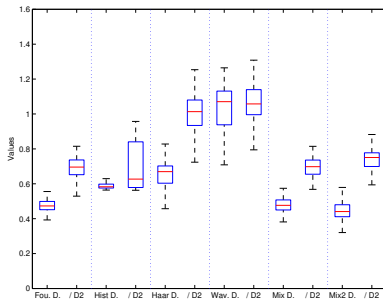
$n = 500$



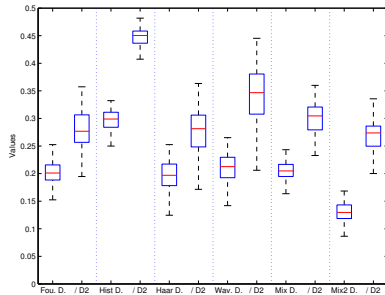
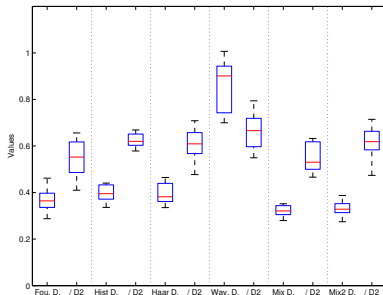
$n = 2000$

Dantzig / Non adaptive D. f_3/f_4

f_3



f_4

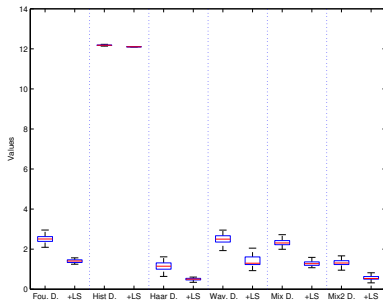
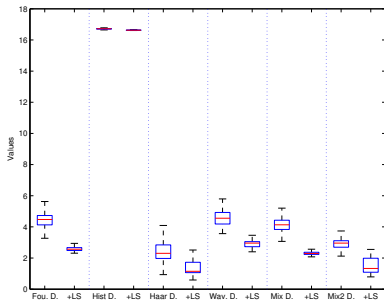


$n = 500$

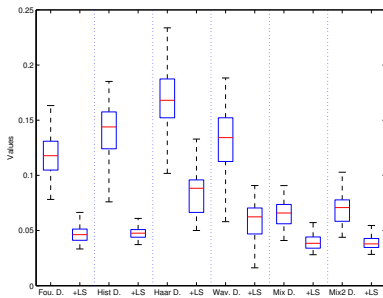
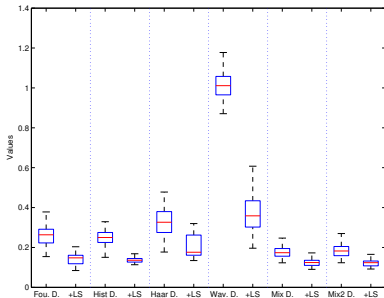
$n = 2000$

Dantzig / Dantzig+LS f_1/f_2

f_1



f_2

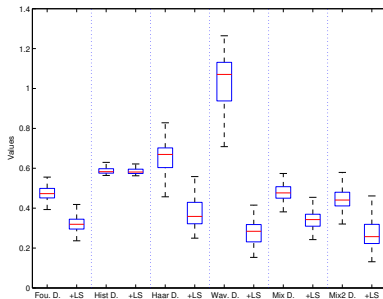


$n = 500$

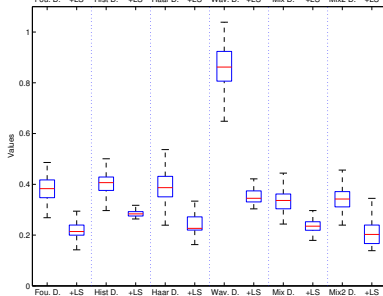
$n = 2000$

Dantzig / Dantzig+LS f_3/f_4

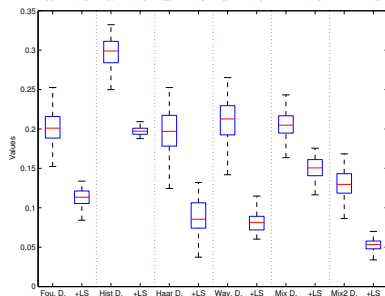
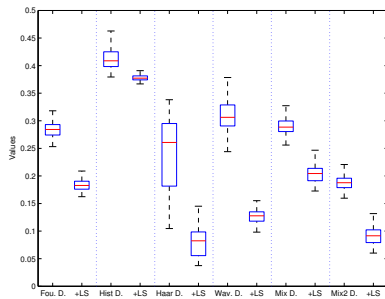
f_3



f_4



$n = 500$



$n = 2000$

Hypothèses sur le dictionnaire

Hypothèses de structure sur le dictionnaire

- Indispensable pour obtenir des résultats.
- Variante autour des bases orthonormées...

Hypothèse locale

- Hypothèse « minimale » pour obtenir un résultat.
- Soit J_0 un support, l'hypothèse locale $HL(J_0, \kappa_{J_0}, \mu_{J_0})$ est :

$$\|f_\lambda\|_2 \geq \frac{\kappa_{J_0} \|\lambda_{J_0}\|_{\ell_1} - \mu_{J_0} \left(\|\lambda_{J_0^c}\|_{\ell_1} - \|\lambda_{J_0}\|_{\ell_1} \right)_+}{\sqrt{|J_0|}}$$

est vérifiée pour tout λ avec $\kappa_{J_0} > 0$ et $\mu_{J_0} \geq 0$.

- Pour un dictionnaire $\mathcal{D}$ orthonormé, $\kappa_{J_0} = 1$ and $\mu_{J_0} = 0$ convient :

$$\|f_\lambda\|_2 = \|\lambda\|_{\ell_2} \geq \|\lambda_{J_0}\|_{\ell_2} \geq \frac{\|\lambda_{J_0}\|_{\ell_1}}{\sqrt{|J_0|}}$$

- RIP, RE et leurs variantes $\implies HL(J_0, \kappa_{J_0}, \mu_{J_0})$ dès que $|J_0| < c \frac{n}{\log n}$ avec κ_{J_0} et μ_{J_0} ne dépendant que de la taille du support.

Théorème Dantzig

Théorème

- Sous Ω_{γ_1} , événement de proba $\geq 1 - 2 \left(\frac{1}{p}\right)^{\gamma_1 - 1}$, si l'hypothèse locale $HL(J_0, \kappa_{J_0}, \mu_{J_0})$ est satisfaite alors pour tout $\lambda \in \mathbb{R}^p$

$$\begin{aligned} \|\hat{f}^{D, \gamma_0} - f_0\|_2^2 &\leq \|f_\lambda - f_0\|_2^2 + 2 \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{J_0^c}\|_{\ell_1}^2}{|J_0|} \\ &\quad + \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D, \gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+^2}{|J_0|} \\ &\quad + 4 \left(1 + \frac{1}{\kappa_{J_0}^2}\right) |J_0| (\|\eta^{\gamma_0}\|_{\ell_\infty} + \|\eta^{\gamma_1}\|_{\ell_\infty})^2. \end{aligned}$$

Interprétation

- Biais (Approx. par f_λ + Approx. en dehors de J_0) / Variance
- “Faisabilité” de la contrainte de Dantzig

Variance

Terme de variance

- Dans le théorème : $4 \left(1 + \frac{1}{\kappa_{J_0}^2} \right) |J_0| (\|\eta^{\gamma_0}\|_{\ell_\infty} + \|\eta^{\gamma_1}\|_{\ell_\infty})^2$

- Sous Ω_{γ_1} ,

$$\begin{aligned} (\|\eta^{\gamma_0}\|_{\ell_\infty} + \|\eta^{\gamma_1}\|_{\ell_\infty})^2 &\leq 8(\gamma_0 + \gamma_1) \log p \frac{\sup_k \sigma_k^2}{n} \\ &\quad + 64 \log^2 p (\gamma_0 \gamma_1 + \frac{49}{36} \gamma_0^2) \frac{\sup_k \|\phi_k\|_\infty^2}{n^2}. \end{aligned}$$

- Simplification dans le théorème !

Théorème Dantzig simplifié

Théorème

- Sous Ω_{γ_1} , événement de proba $\geq 1 - 2 \left(\frac{1}{p}\right)^{\gamma_1-1}$, si l'hypothèse locale $HL(J_0, \kappa_{J_0}, \mu_{J_0})$ est satisfaite alors pour tout $\lambda \in \mathbb{R}^p$

$$\begin{aligned} \|\hat{f}^{D, \gamma_0} - f_0\|_2^2 &\lesssim \|f_\lambda - f_0\|_2^2 + 2 \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{J_0^c}\|_{\ell_1}^2}{|J_0|} \\ &+ \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D, \gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+^2}{|J_0|} \\ &+ 32(\gamma_0 + \gamma_1) \left(1 + \frac{1}{\kappa_{J_0}^2}\right) |J_0| \log p \frac{\sup_k \sigma_k^2}{n}. \end{aligned}$$

Interprétation

- Biais (Approx. par f_λ + Approx. en dehors de J_0) / Variance
- “Faisabilité” de la contrainte de Dantzig

Contrainte de Dantzig

"Faisabilité" de la contrainte de Dantzig

- Dans le théorème : $\left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+^2}{|J_0|}$
- Contrôle difficile : pas de lien a priori entre λ et $\hat{\lambda}^{D,\gamma_0}$...
- Disparition si $\|\lambda\|_{\ell_1} \geq \|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1}$.
- Disparition si λ satisfait la contrainte de Dantzig.

Corollaire

- Si $\gamma_1 \leq \gamma_0$, sous Ω_{γ_1} , événement de proba $\geq 1 - 2\left(\frac{1}{\rho}\right)^{\gamma_1-1}$, si l'hypothèse locale $HL(J_0, \kappa_{J_0}, \mu_{J_0})$ est satisfaite et si $f_{\lambda_0} = P_{\mathcal{D}} f_0$ alors

$$\begin{aligned} \|\hat{f}^{D,\gamma_0} - f_0\|_2^2 &\lesssim \|P_{\mathcal{D}} f_0 - f_0\|_2^2 + 2\left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{0J_0^c}\|_{\ell_1}^2}{|J_0|} \\ &+ 32(\gamma_0 + \gamma_1) \left(1 + \frac{1}{\kappa_{J_0}^2}\right) |J_0| \log p \frac{\sup_k \sigma_k^2}{n}. \end{aligned}$$

Théorème Dantzig/Lasso

Dantzig/Lasso

- Proximité des estimateurs $\implies$ proximité des solutions ?
- Proximité déjà prouvée dans un sens car $\hat{f}^{L,\gamma_0}$ est faisable...
- Contrôle plus fin possible.

Théorème (sans le terme de Dantzig)

- Sous Ω_{γ_1} , événement de proba $\geq 1 - 2 \left(\frac{1}{p}\right)^{\gamma_1-1}$, si l'hypothèse locale $HL(J_0, \kappa_{J_0}, \mu_{J_0})$ est satisfaite alors

$$\begin{aligned} & \left| \|\hat{f}^{D,\gamma_0} - f_0\|_2^2 - \|\hat{f}^{L,\gamma_0} - f_0\|_2^2 \right| \\ & \leq 2 \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}} \right)^2 \frac{\|\hat{\lambda}_{J_0^c}^{L,\gamma_0}\|_{\ell_1}^2}{|J_0|} \\ & \quad + 32(\gamma_0 + \gamma_1) \left(1 + \frac{1}{\kappa_{J_0}^2} \right) |J_0| \log p \frac{\sup_k \sigma_k^2}{n}. \end{aligned}$$

Pénalité minimale

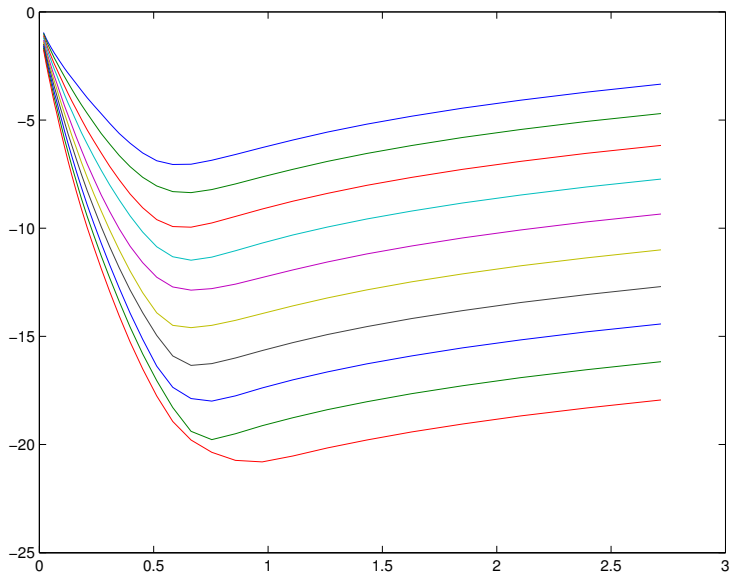
Choix de γ_0

- Comment choisir γ_0 ?
- Paramètre de pénalisation :
 - γ_0 grand : surlissage (grand biais),
 - γ_0 petit : instabilité (grande variance).
- Inégalité oracle pour tout γ_0 :
 - γ_0 grand : surpénalisation de la variance dans l'inégalité oracle,
 - γ_0 petit : présence du terme Dantzig en $(\|\lambda^{S, \gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+$ difficile à contrôler.

Pénalité minimale

- Bon calibrage : $\gamma_0 = 1$!
- Résultat théorique :
 - Si $\gamma_0 > 1$: avec proba $\geq 1 - 2 \left(\frac{1}{p}\right)^{\gamma_0 - 1}$, inégalité oracle sans le terme Dantzig pour tout λ tel que $f_\lambda = P_{\mathcal{D}} f_0$.
 - Si $\gamma_0 < 1$: théorème montrant que l'estimateur est mauvais (sur un cas simple d'estimation de la densité uniforme sur $[0, 1]$ dans la base de Haar).

Pénalité minimale



Conclusion

Estimation de densité par Dantzig adaptatif

- Nouvelle méthode d'estimation de densité par pénalisation ℓ_1 .
- Utilisation d'inégalités de concentration fine pour obtenir une pénalisation adaptative.
- Inégalité oracle en probabilité sur l'estimateur.
- Lien avec le Lasso adaptatif.
- Calibration de la pénalité : pénalité minimale.
- Hypothèses assez faibles sur la structure du dictionnaire.

Perspectives

- Performance de la méthode en deux étapes.
- Passage à des résultats en espérance.
- Relaxation des hypothèses sur la structure par passage à des résultats en proba sur les signaux (modélisation probabiliste des signaux).