Density estimation by ℓ_1 minimal penalization

Erwan Le Pennec
LPMA, Univ. Paris Diderot / SELECT, INRIA Saclay - Univ. Paris Sud
joint work with Karine Bertin (Univ. de Valparaiso, Chili)
and Vincent Rivoirard (Univ. Paris Sud / ENS Paris)

December 2009

Outline

1. Framework and estimators
   - Density estimation, Dantzig and Lasso estimators
   - Concentration inequalities and adaptive estimators
2. Numerical experiments
3. Theoretical results
   - Assumptions on the dictionary
   - Oracle inequalities
   - Minimal penalty

Outline

1. Framework and estimators
   - Density estimation, Dantzig and Lasso estimators
   - Concentration inequalities and adaptive estimators
2. Numerical experiments
3. Theoretical results
   - Assumptions on the dictionary
   - Oracle inequalities
   - Minimal penalty

Plan

1. Framework and estimators
   - Density estimation, Dantzig and Lasso estimators
   - Concentration inequalities and adaptive estimators
2. Numerical experiments
3. Theoretical results
   - Assumptions on the dictionary
   - Oracle inequalities
   - Minimal penalty

Framework

Density estimation

- $X_1, \ldots, X_n$ i.i.d of law $f_0(x)dx$.
- Assumption: $f_0 \in L^2$.
- No assumption $f_0 \in L^\infty$ or $\|f_0\|_\infty$ known.

Dictionary approach

- Arbitrary dictionary $\mathcal{D} = \{\phi_k\}_{1 \leq k \leq p}$.
- Assumptions: $\phi_k \in L^2 \cap L^\infty$ and $\|\phi_k\|_\infty$ known.
- Notation: $f_\lambda = \sum_{k=1}^{p} \lambda_k \phi_k$.

Goals

- Estimation of f_0 by a linear combination f_λ of the dictionary $\mathcal{D}$.
- Risk measured by the L^2 norm.
- $p \gg n$ case possible...

Dantzig estimator

Scalar products

- Scalar products and means: $\beta_k = \langle f_0, \phi_k \rangle = \mathbb{E}(\phi_k(X))$.
- Empirical scalar products: $\hat{\beta}_k = \frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i) \simeq \beta_k$.
- G: Gram matrix of the dictionary $\mathcal{D}$ ($G_{k,k'} = \langle \phi_k, \phi_{k'} \rangle$).
- For $f_\lambda = \sum_{k=1}^{p} \lambda_k \phi_k$, $\langle f_\lambda, \phi_k \rangle = (G\lambda)_k$

Constraints and regularization

- Search for λ such that $G\lambda \simeq \hat{\beta} \simeq \beta$.
- Dantzig constraint: $\forall k \in \{1, \ldots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k$.
- $+$ Regularization by the ℓ_1 norm of λ.
- Dantzig estimator:

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

Lasso estimator / SPADES

Quadratic norm

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

Contrast and penalization

- Minimization of the empirical contrast penalized by $2\sum_k \eta_k |\lambda_k|$.
- Lasso estimator:

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

Link Dantzig/Lasso

- First order optimality condition of Lasso
  $=$ Dantzig constraint, $\forall k \in \{1, \ldots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k$
- Strong link between the two estimators...

Sparse bibliography

ℓ_1 penalization in statistics

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

Sparsity and underdetermined system

- Donoho (2004) Sparsity and Gaussian matrix.
- Candes, Tao and Romberg (2004) Sparsity and Fourier matrix.

Sparsity and ℓ_1 penalization in statistics

- Candes and Tao (2007) Dantzig.
- Huge literature: Bickel, Tsybakov, Ritov, Bunea, Wegkamp, Plan, Temlyakov, van de Geer, Lounici, Hebiri...

Concentration inequalities

Choice of η_k

- How to choose η_k such that $\mathbb{P}\left(|\beta_k - \hat{\beta}_k| > \eta_k\right) \leq 2e^{-u^2/2}$?
- Concentration of an empirical process around its mean.

Hoeffding

- $\eta_k = u\dfrac{\|\phi_k\|_\infty}{\sqrt{n}} \implies \mathbb{P}\left(|\beta_k - \hat{\beta}_k| > \eta_k\right) \leq 2e^{-u^2/2}$.
- Often used but crude upper bound that does not use f_0 and strong limitation on the norm of ϕ_k.

Bernstein

- Use the variance: $\sigma_k^2 = \mathrm{Var}\left(\phi_k(X)\right)$.
- $\eta_k = u\dfrac{\sigma_k}{\sqrt{n}} + \dfrac{u^2}{3}\dfrac{\|\phi_k\|_\infty}{n} \implies \mathbb{P}\left(|\beta_k - \hat{\beta}_k| > \eta_k\right) \leq 2e^{-u^2/2}$.
- Sharp use of f_0 and much weaker limitation on the norm of ϕ_k.

Concentration inequalities

Bernstein (continuation)

- Issue: As f_0 is unknown, the variance $\sigma_k^2 = \mathrm{Var}\left(\phi_k(X)\right)$ is unknown...
- Upper bound possible but loss of sharpness:
  - $\sigma_k \leq \|\phi_k\|_\infty$ ($\simeq$ Hoeffding)
  - $\sigma_k \leq \|f\|_\infty^{1/2} \|\phi_k\|_2$ (Issue as $\|f\|_\infty$ is not known)

Adaptation/Auto renormalization

- Empirical variance: $\hat{\sigma}_k^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} \left(\phi_k(X_i) - \hat{\beta}_k\right)^2$
- $\eta_k = u \dfrac{\hat{\sigma}_k}{\sqrt{n}} + 7 \dfrac{u^2}{3} \dfrac{\|\phi_k\|_\infty}{n} \implies \mathbb{P}\left(|\beta_k - \hat{\beta}_k| > \eta_k\right) \leq 2e^{-u^2/2}.$
- Adaptation to the unknown f_0.
- Moreover, $\mathbb{P}\left(|\hat{\sigma}_k - \sigma_k| > 2u \frac{\|\phi_k\|_\infty}{n}\right) \leq 2e^{-u^2/2}.$

Control on the coefficients with high probability

Control on all the coefficients

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

The event Ω_{γ_1}

- For all $\gamma_1 > 0$, event Ω_{γ_1} of probability $\geq 1 - 2\left(\frac{1}{p}\right)^{\gamma_1 - 1}$:

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

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

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

Estimators

Dantzig constraint

- For $\gamma_0 > 0$, $\forall k \in \{1, \dots, p\}$
$$|(G\lambda)_k - \hat{\beta}_k| \leq \eta_k^{\gamma_0} = \sqrt{2\gamma_0 \log p} \frac{\hat{\sigma}_k}{\sqrt{n}} + \frac{14}{3}\gamma_0 \log p \frac{\|\phi_k\|_\infty}{n}.$$

Adaptive penalized estimators

- Dantzig:
$\hat{\lambda}^{D,\gamma_0} = \operatorname{argmin}\|\lambda\|_{\ell_1}$ under $\forall k \in \{1, \dots, p\}, |(G\lambda)_k - \hat{\beta}_k| \leq \eta_k^{\gamma_0}$.
- Lasso: $\hat{\lambda}^{L,\gamma_0} = \operatorname{argmin} \frac{1}{2}\left(-2\lambda^* \hat{\beta} + \lambda^* G\lambda\right) + \sum_{k=1}^{p} \eta_k^{\gamma_0} |\lambda_k|$.

Variations

- Non adaptive estimators by replacing $\eta_k^{\gamma_0}$ by $\tilde{\eta}_k^{\gamma_0}$:

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

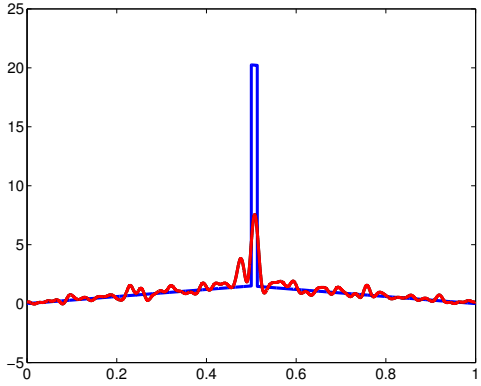
- Improvement possible with a further least square type step on the support of $\hat{\lambda}$ (à la Gauss-Dantzig).

Plan

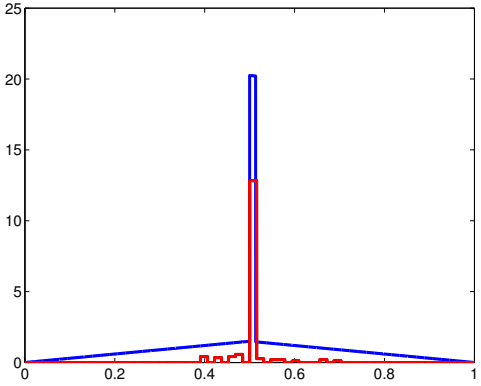
1. Framework and estimators
   - Density estimation, Dantzig and Lasso estimators
   - Concentration inequalities and adaptive estimators
2. Numerical experiments
3. Theoretical results
   - Assumptions on the dictionary
   - Oracle inequalities
   - Minimal penalty

Various dictionaries

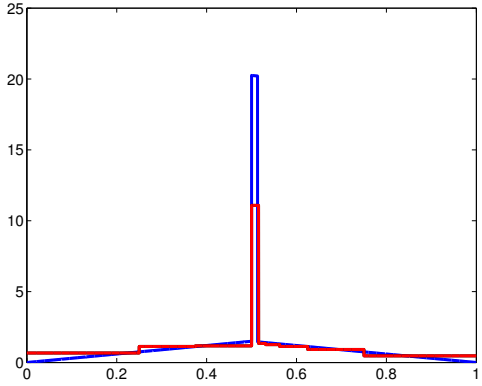
Fourier

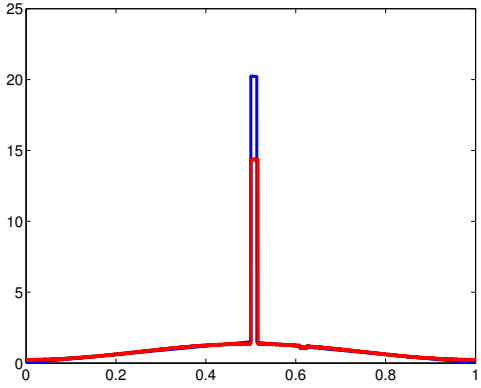


Boxes



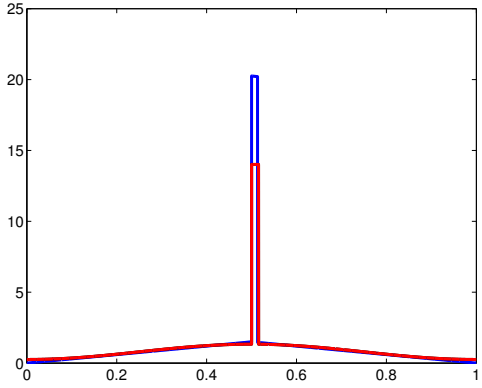
Haar





Mix

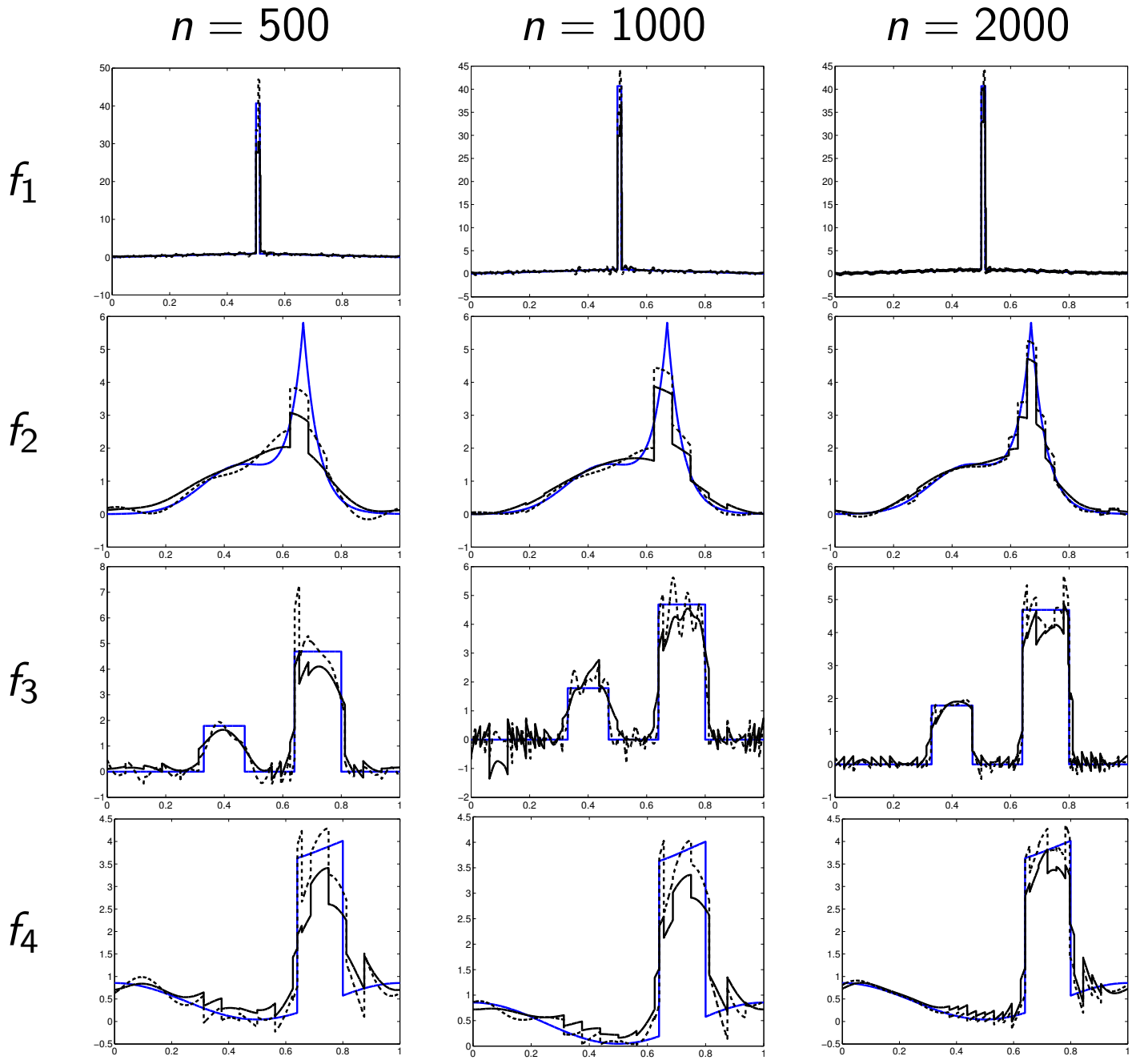
Fourier + Boxes



Mix2

Fourier + Boxes + Haar

Various functions and estimators



Numerous results

Algorithm

- Homotopy path type algorithm (LARS).
- Dantzig constraint with γ_0=1.

Dictionary

- Works with "mixed" dictionaries.
- Best results obtained with "mixed" dictionaries.

Dantzig/Lasso

- Very similar performance...

Adaptive/Non adaptive

- Adaptation of the constraint yields a significant improvement.

Further Least Square step variant

- Significant improvement (reduces the bias of Lasso type method).

Plan

1. Framework and estimators
   - Density estimation, Dantzig and Lasso estimators
   - Concentration inequalities and adaptive estimators
2. Numerical experiments
3. Theoretical results
   - Assumptions on the dictionary
   - Oracle inequalities
   - Minimal penalty

Assumptions on the dictionary

Structure assumptions on the dictionary

- Required to obtain results.
- Variations around "orthonormal" bases...

Local assumption

- "Minimal" assumption to obtain a result.
- Let J_0 a support, the local assumption $LA(J_0, \kappa_{J_0}, \mu_{J_0})$ is:

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

  is satisfied for all λ with $\kappa_{J_0} > 0$ and $\mu_{J_0} \geq 0$.
- For an orthonormal dictionary $\mathcal{D}$, $\kappa_{J_0} = 1$ and $\mu_{J_0} = 0$ is sufficient:

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

- RIP, RE and their variations $\Longrightarrow$ $LA(J_0, \kappa_{J_0}, \mu_{J_0})$ with κ_{J_0} and μ_{J_0} depending only on the size of the support.

Dantzig Theorem

Theorem

- Under Ω_{γ_1}, event of probability $\geq 1 - 2\left(\frac{1}{p}\right)^{\gamma_1 - 1}$, if the local assumption $LA(J_0, \kappa_{J_0}, \mu_{J_0})$ is satisfied then for all $\lambda \in \mathbb{R}^p$

$$\|\hat{f}^{D,\gamma_0} - f_0\|_2^2 \leq \|f_\lambda - f_0\|_2^2 + 2\left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{|J_0|}$$

$$+ \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+^2}{|J_0|}$$

$$+ 4\left(1 + \frac{1}{\kappa_{J_0}^2}\right) |J_0| (\|\eta^{\gamma_0}\|_{\ell_\infty} + \|\eta^{\gamma_1}\|_{\ell_\infty})^2 .$$

Interpretation

- Bias (Approx. by f_λ + Approx. outside J_0) / Variance
- "Feasibility" of the Dantzig constraint

Variance

Variance term

- In the Theorem: $4\left(1+\frac{1}{\kappa_{J_0}^2}\right)|J_0|\,(\|\eta^{\gamma_0}\|_{\ell_\infty}+\|\eta^{\gamma_1}\|_{\ell_\infty})^2$
- Under Ω_{γ_1},

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

- Simplified version of the theorem!

Simplified Dantzig Theorem

Theorem

- Under Ω_{γ_1}, event of probability $\geq 1 - 2\left(\frac{1}{p}\right)^{\gamma_1 - 1}$, if the local assumption $LA(J_0, \kappa_{J_0}, \mu_{J_0})$ is satisfied then for all $\lambda \in \mathbb{R}^p$

$$\|\hat{f}^{D,\gamma_0} - f_0\|_2^2 \lesssim \|f_\lambda - f_0\|_2^2 + 2\left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{J_0^C}\|_{\ell_1}^2}{|J_0|}$$

$$+ \left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+^2}{|J_0|}$$

$$+ 32(\gamma_0 + \gamma_1)\left(1 + \frac{1}{\kappa_{J_0}^2}\right)|J_0| \log p \frac{\sup_k \sigma_k^2}{n}.$$

Interpretation

- Bias (Approx. by f_λ + Approx. outside J_0) / Variance
- "Feasibility" of the Dantzig constraint

Dantzig constraint

"Feasibility" of the Dantzig constraint

- In the theorem: $\left(1+\frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{(\|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1}-\|\lambda\|_{\ell_1})_+^2}{|J_0|}$
- Difficult to control...
- Disappears if $\|\lambda\|_{\ell_1} \geq \|\hat{\lambda}^{D,\gamma_0}\|_{\ell_1}$.
- Disappears if λ satisfies the Dantzig constraint.

Corollary

- If $\gamma_1 \leq \gamma_0$, under Ω_{γ_1}, event of probability $\geq 1-2\left(\frac{1}{p}\right)^{\gamma_1-1}$, if the local assumption $LA(J_0,\kappa_{J_0},\mu_{J_0})$ is satisfied and if $f_{\lambda_0}=P_{\mathcal{D}}f_0$ then

$$\|\hat{f}^{D,\gamma_0}-f_0\|_2^2 \lesssim \|P_{\mathcal{D}}f_0-f_0\|_2^2 + 2\left(1+\frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\lambda_{0J_0^c}\|_{\ell_1}^2}{|J_0|}$$

$$+ 32(\gamma_0+\gamma_1)\left(1+\frac{1}{\kappa_{J_0}^2}\right)|J_0|\log p \frac{\sup_k \sigma_k^2}{n}.$$

Dantzig/Lasso Theorem

Dantzig/Lasso

- Estimators close $\Longrightarrow$ Estimates close?
- One sided proximity already obtained as $\hat{f}^{L,\gamma_0}$ is feasible...
- Better control possible.

Theorem

- Under Ω_{γ_1}, event of probability $\geq 1 - 2\left(\frac{1}{p}\right)^{\gamma_1 - 1}$, if the local assumption $LA(J_0, \kappa_{J_0}, \mu_{J_0})$ is satisfied then

$$\left| \|\hat{f}^{D,\gamma_0} - f_0\|_2^2 - \|\hat{f}^{L,\gamma_0} - f_0\|_2^2 \right|$$

$$\leq 2\left(1 + \frac{2\mu_{J_0}}{\kappa_{J_0}}\right)^2 \frac{\|\widehat{\lambda}^{L,\gamma_0}_{J_0^C}\|_{\ell_1}^2}{|J_0|}$$

$$+ 32(\gamma_0 + \gamma_1)\left(1 + \frac{1}{\kappa_{J_0}^2}\right)|J_0| \log p \frac{\sup_k \sigma_k^2}{n}.$$

Minimal penalty

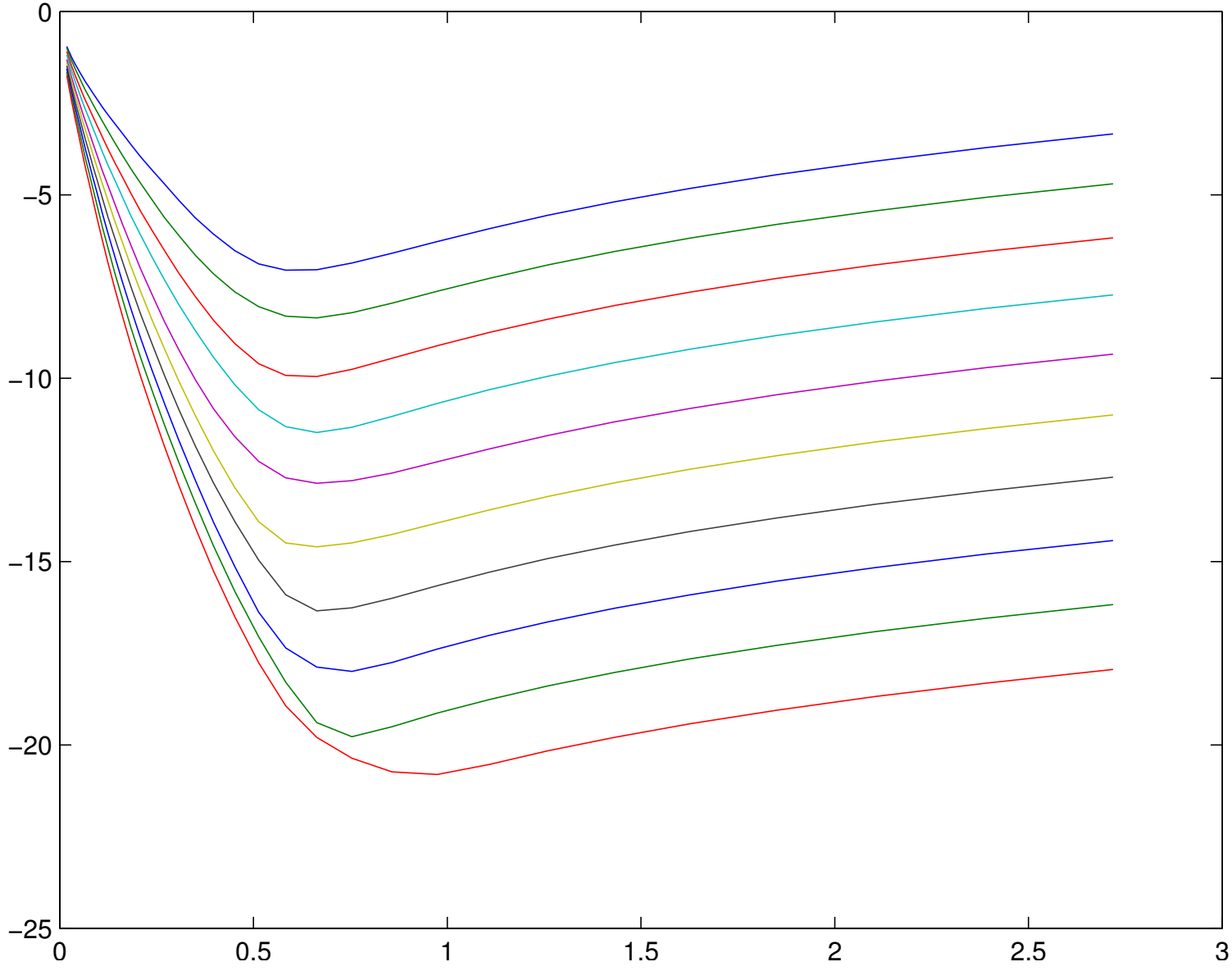
Choice of γ_0

- How to choose γ_0?
- Regularization parameter heuristic:
  - γ_0 large: oversmoothing (large bias),
  - γ_0 small: overfitting (large variance).
- Oracle inequality valid for all γ_0:
  - γ_0 large: overpenalization of the variance in the oracle inequality,
  - γ_0 small: Dantzig term, $(\|\widehat{\lambda}^{D,\gamma_0}\|_{\ell_1} - \|\lambda\|_{\ell_1})_+$, hard to control.

Minimal penalty

- Good calibration: $\gamma_0 = 1$!
- Theoretical result:
  - If $\gamma_0 > 1$: with probability $\geq 1 - 2\left(\frac{1}{p}\right)^{\gamma_0}$, oracle inequality without the Dantzig term for λ such that $f_\lambda = P_{\mathcal{D}} f_0$.
  - For $\gamma_0 < 1$: Theorem showing that the estimator behaves badly (on a simple uniform density on $[0,1]$ estimation with a Haar basis)

Minimal penalty



Conclusion

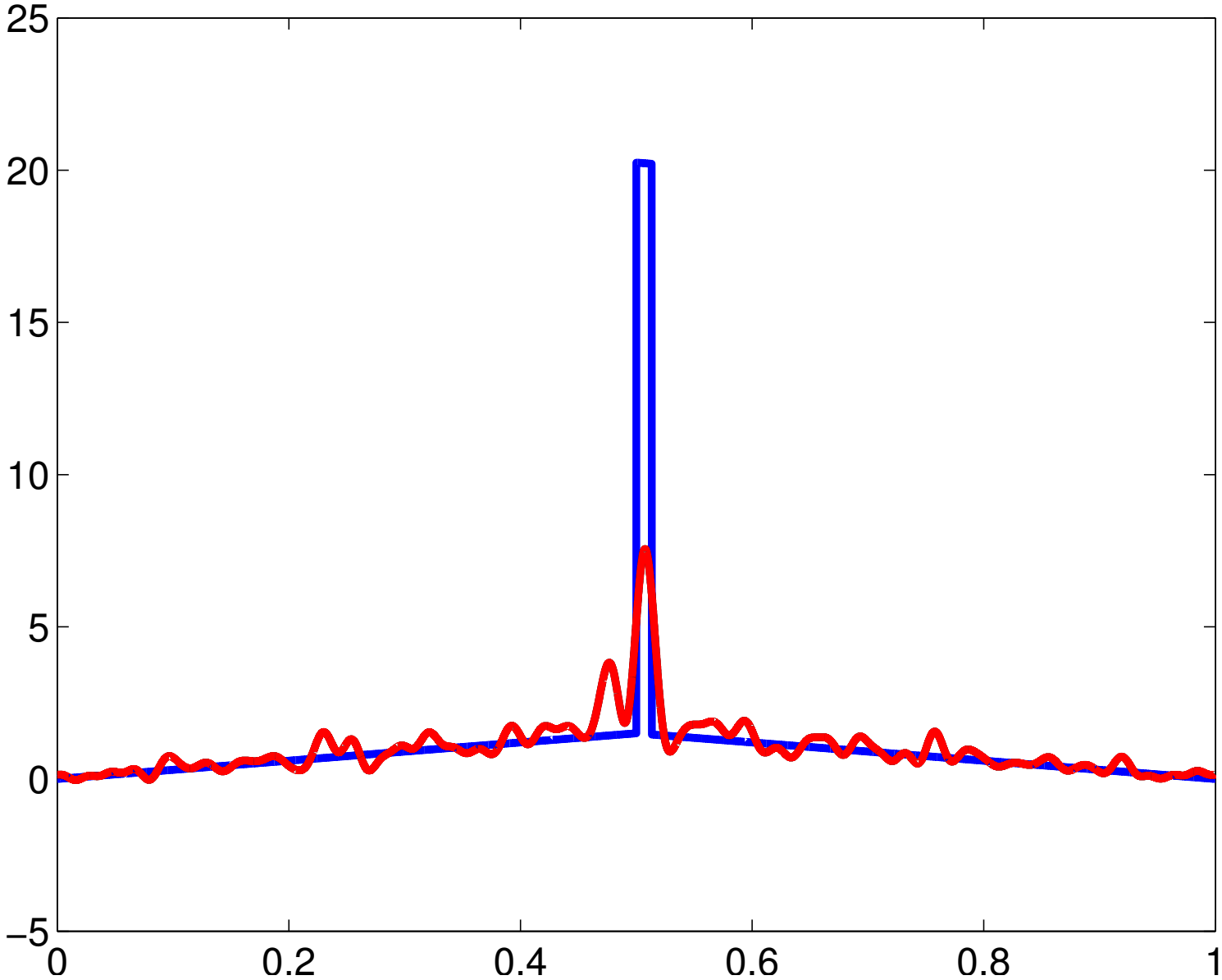
Adaptive Dantzig density estimation

- New density estimation method by ℓ_1 penalization.
- Use of sharp concentration inequalities to obtain an adaptive penalization.
- Oracle inequality in probability for the estimator.
- Link with an adaptive Lasso.
- Penalty calibration: minimal penalty.
- Weak assumptions on the dictionary structure.

Prospects

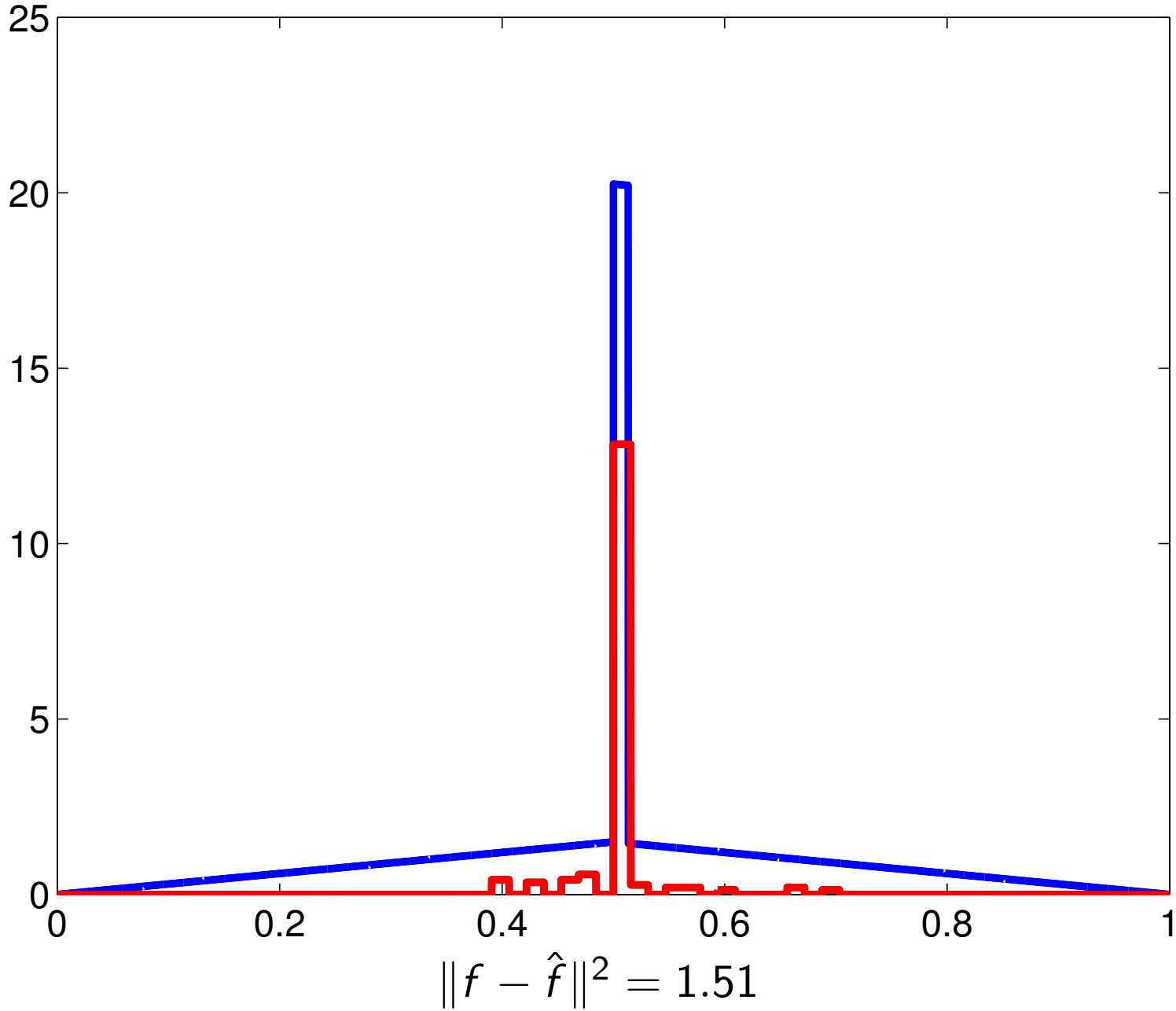
- Expectation result.
- Analysis of the two steps method.
- Further weakening of the structure assumptions with a probabilistic model of signals: results with only high probability on the signals.

Fourier



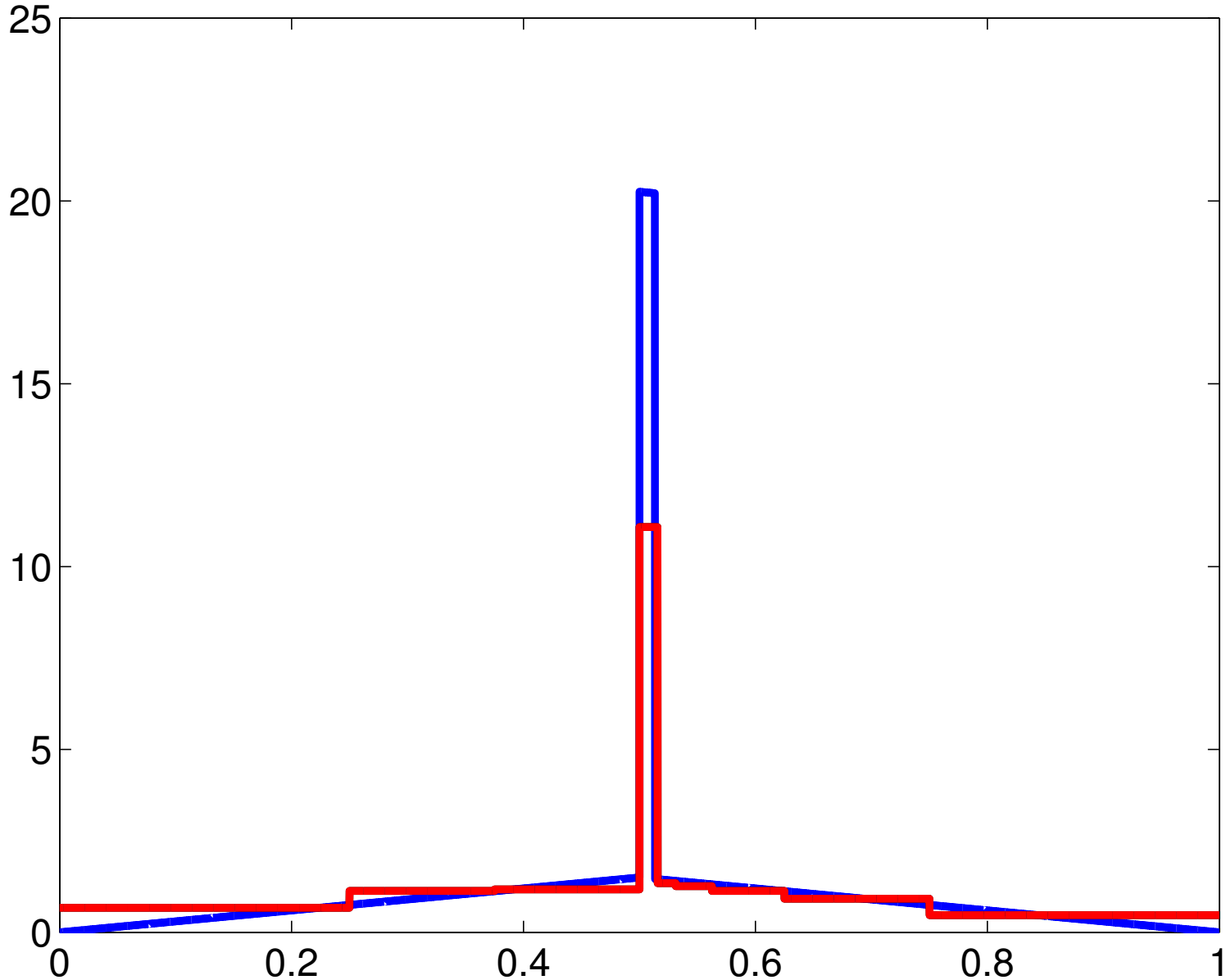
$\|f - \hat{f}\|^2 = 2.16$

Boxes



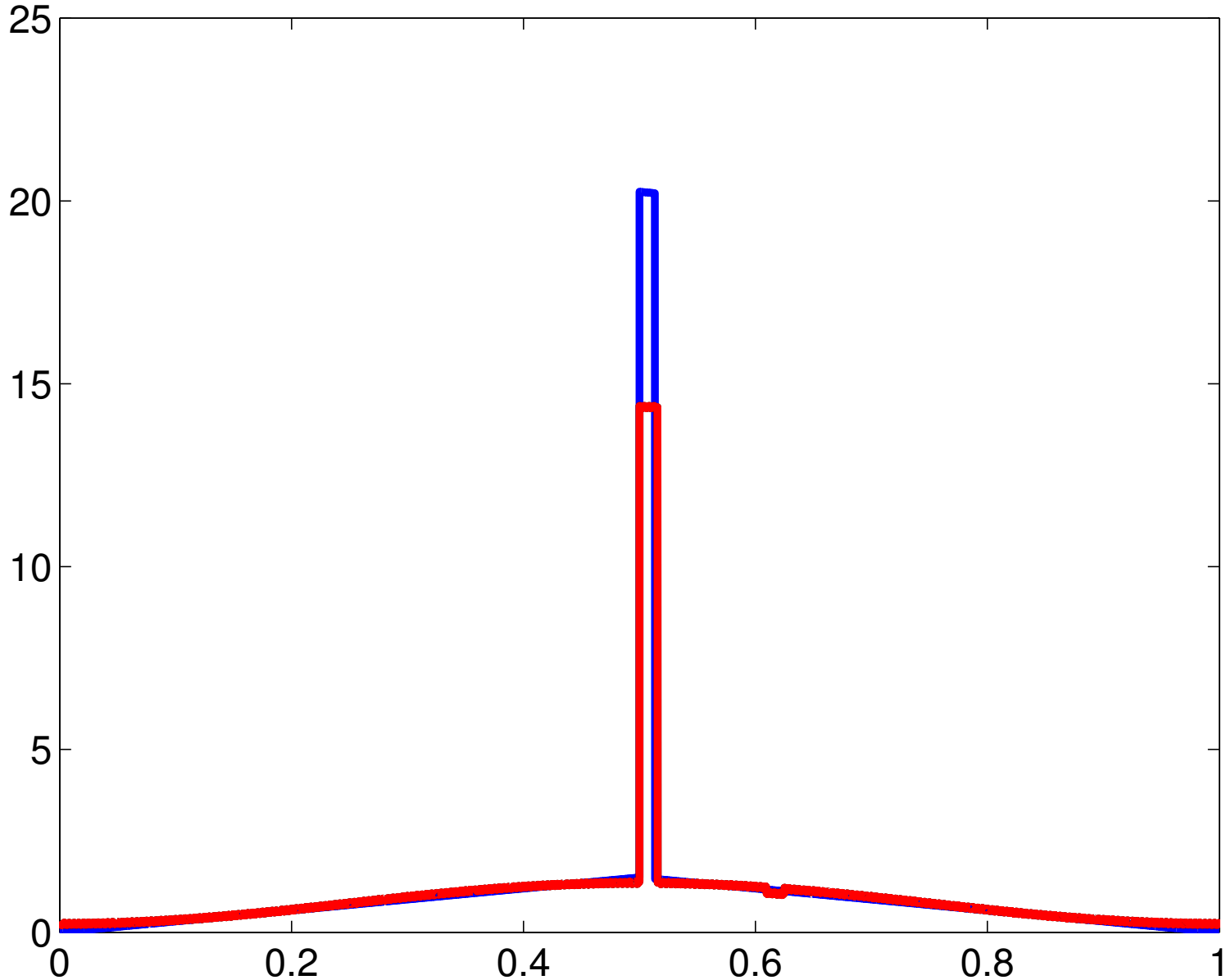
$$\|f - \hat{f}\|^2 = 1.51$$

Haar



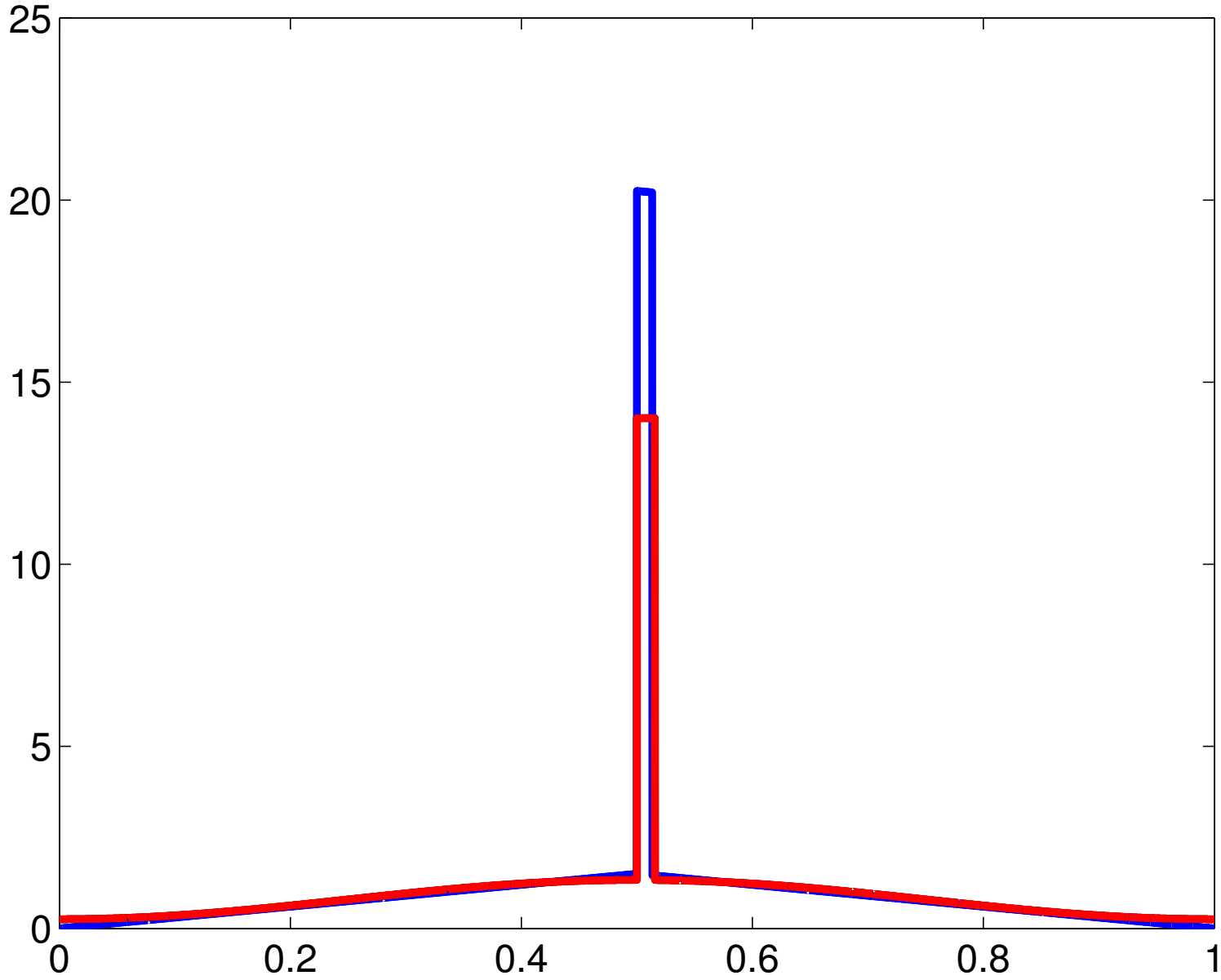
$\|f - \hat{f}\|^2 = 1.09$

Mix = Fourier + Boxes



$\|f - \hat{f}\|^2 = 0.74$

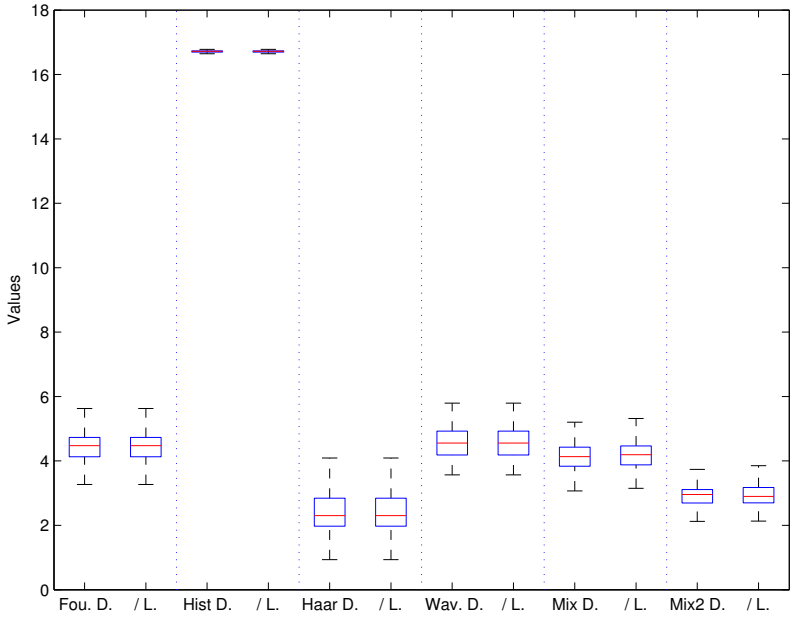
Mix2 = Fourier + Boxes + Haar

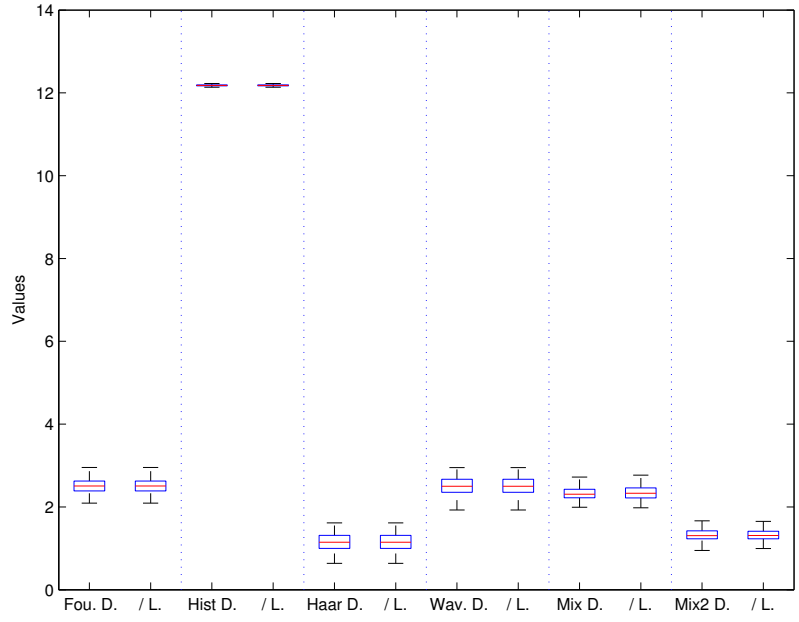


$\|f - \hat{f}\|^2 = 0.75$

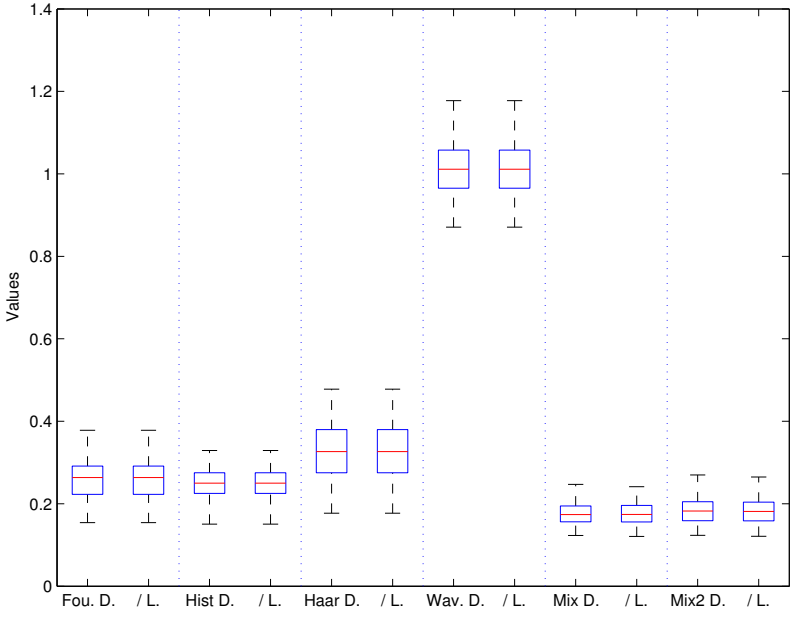
Dantzig / Lasso f_1/f_2

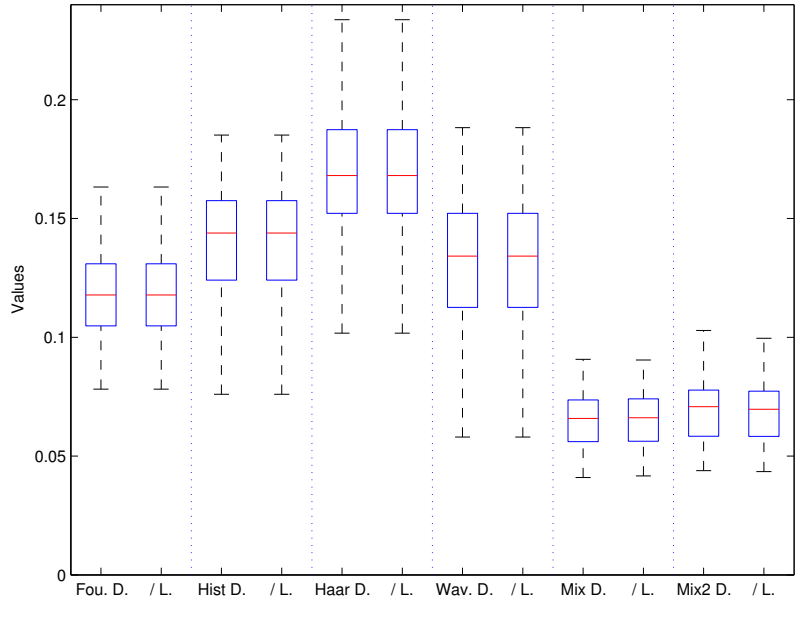
f_1





f_2

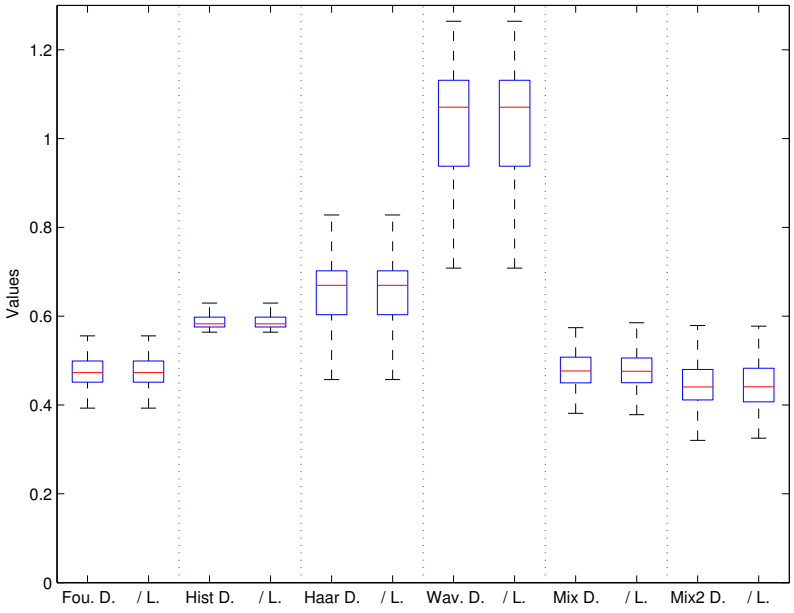


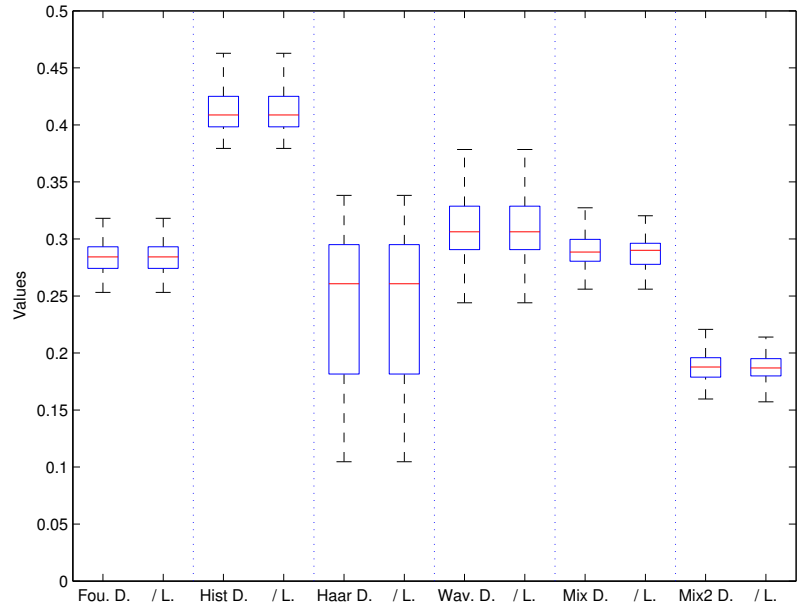


$n = 500$ $n = 2000$

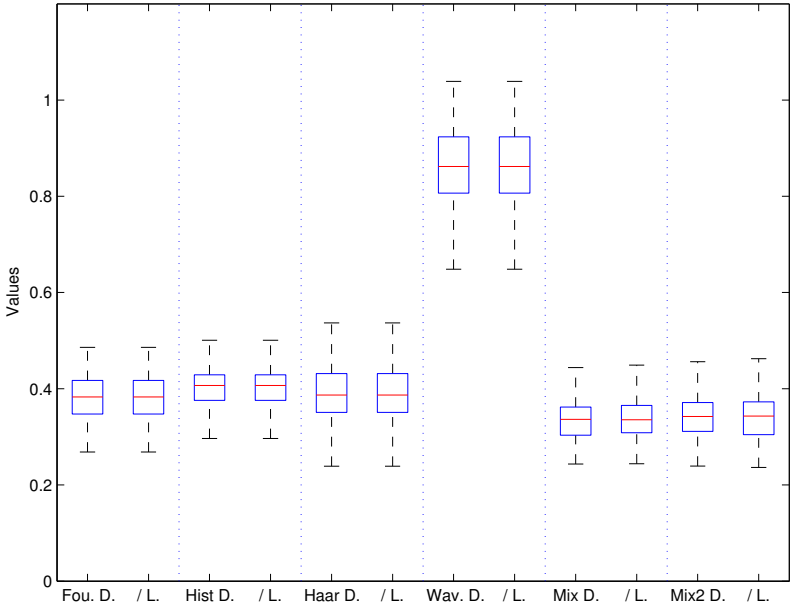
Dantzig / Lasso f_3/f_4

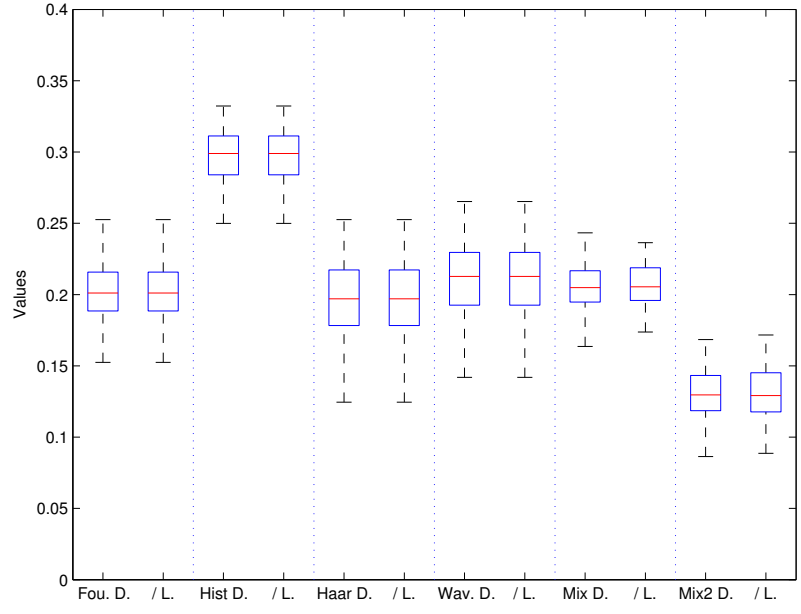
f_3





f_4

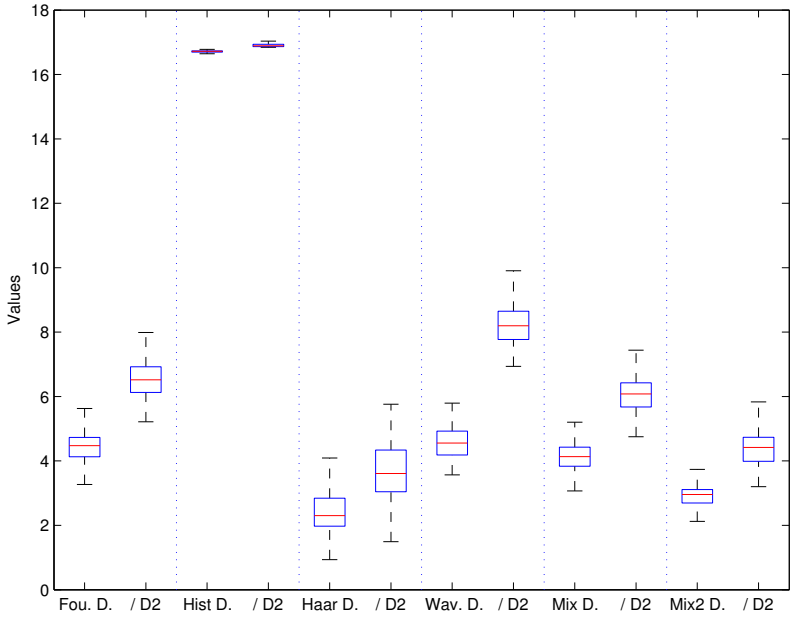


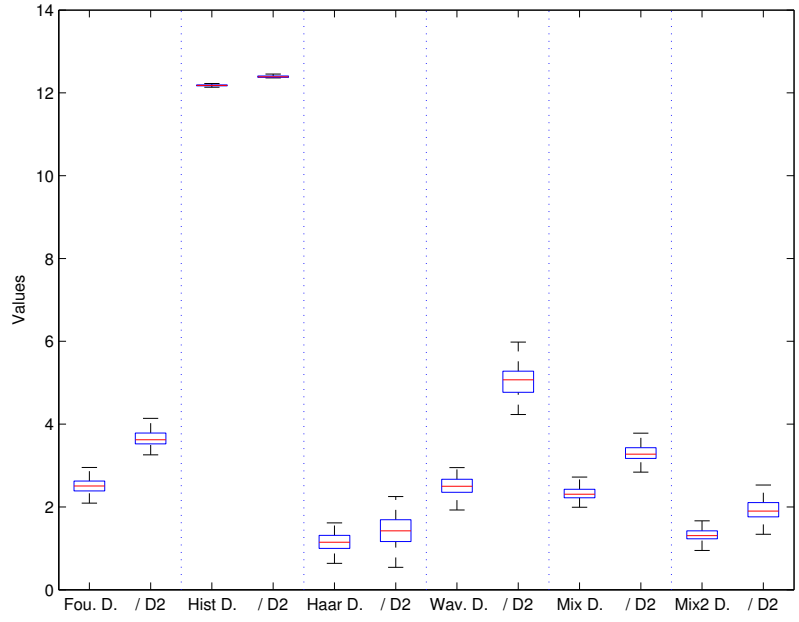


$n = 500$ $n = 2000$

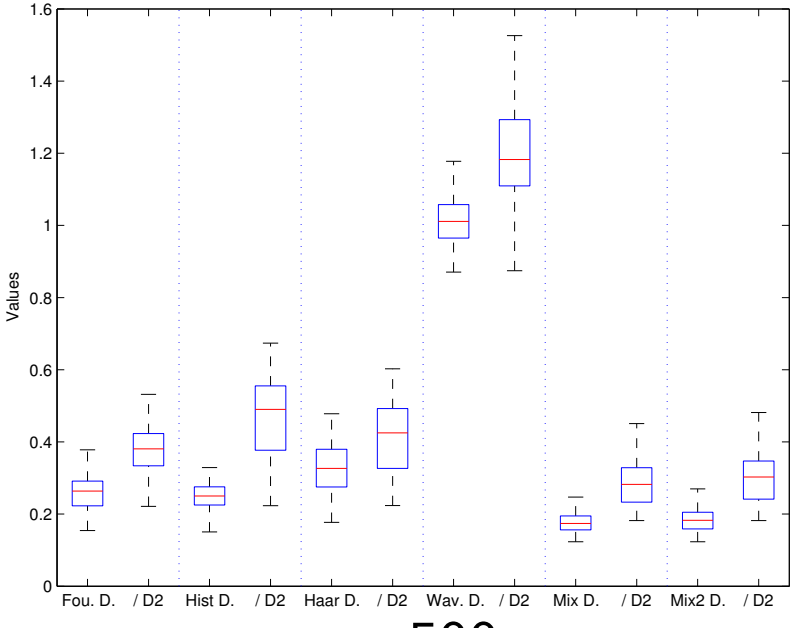
Dantzig / Non adaptive D. f_1/f_2

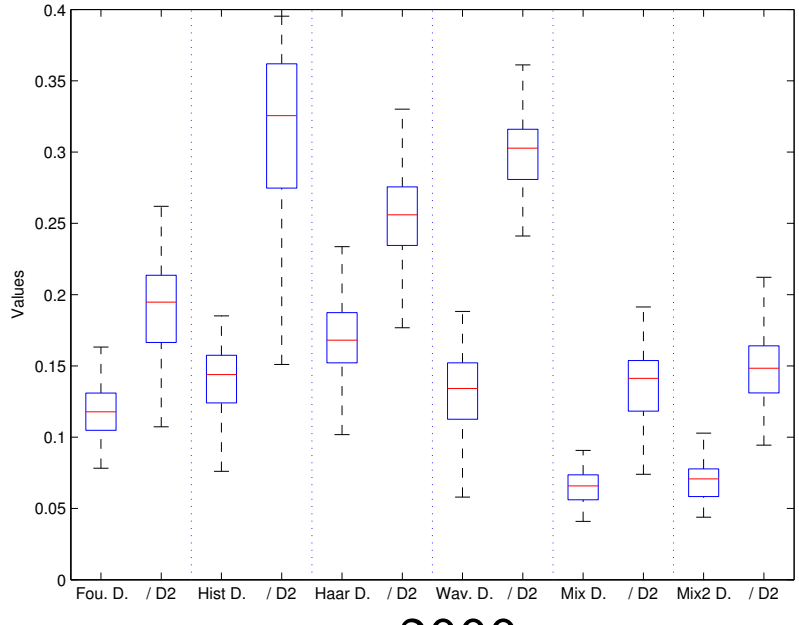
f_1





f_2

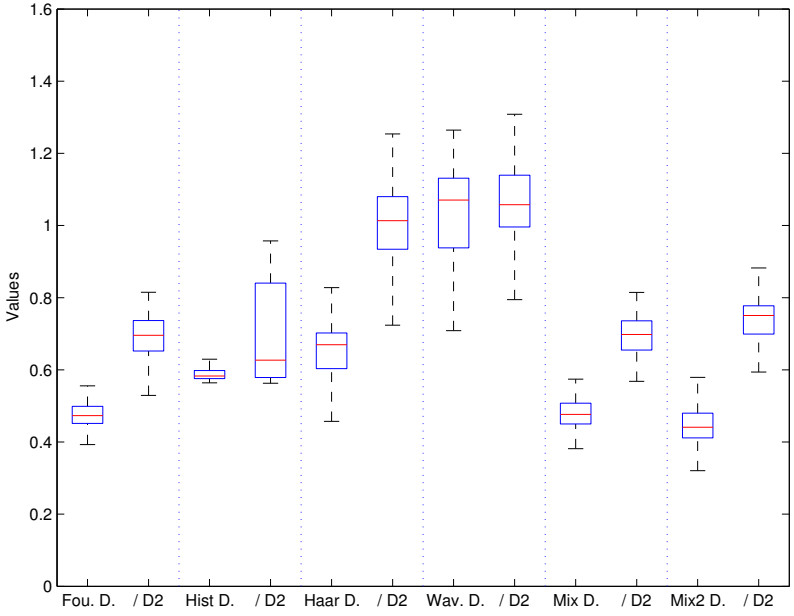


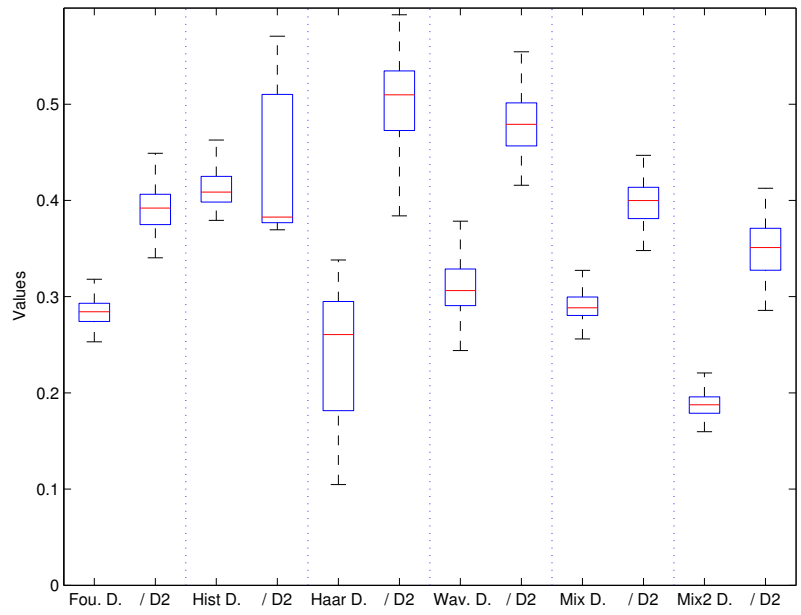


$n = 500$ $n = 2000$

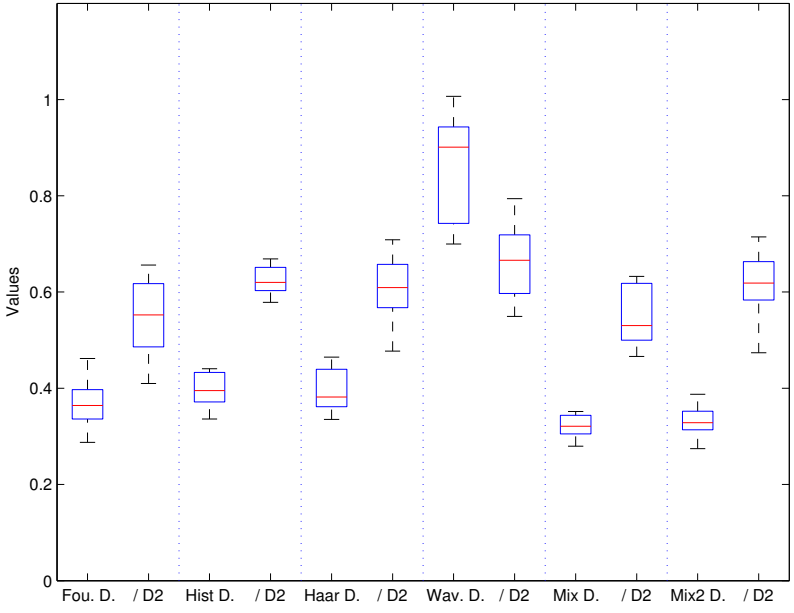
Dantzig / Non adaptive D. f_3/f_4

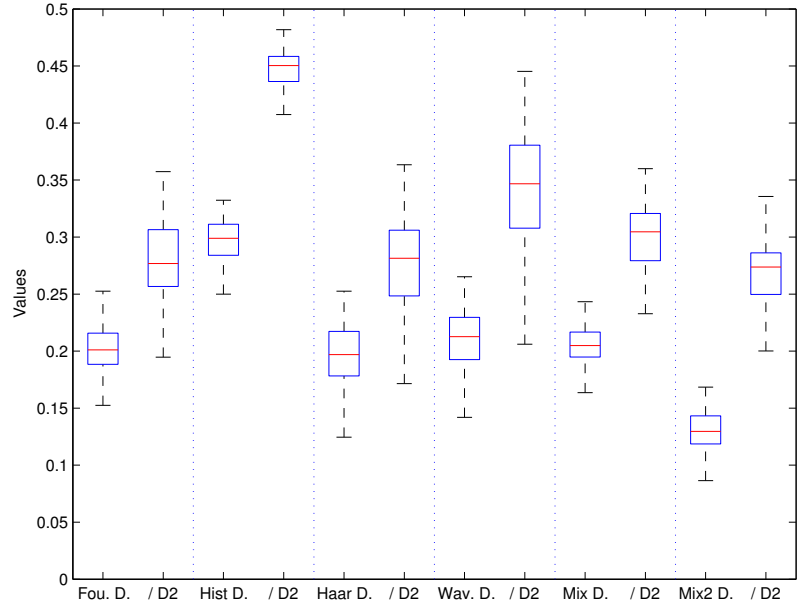
f_3





f_4

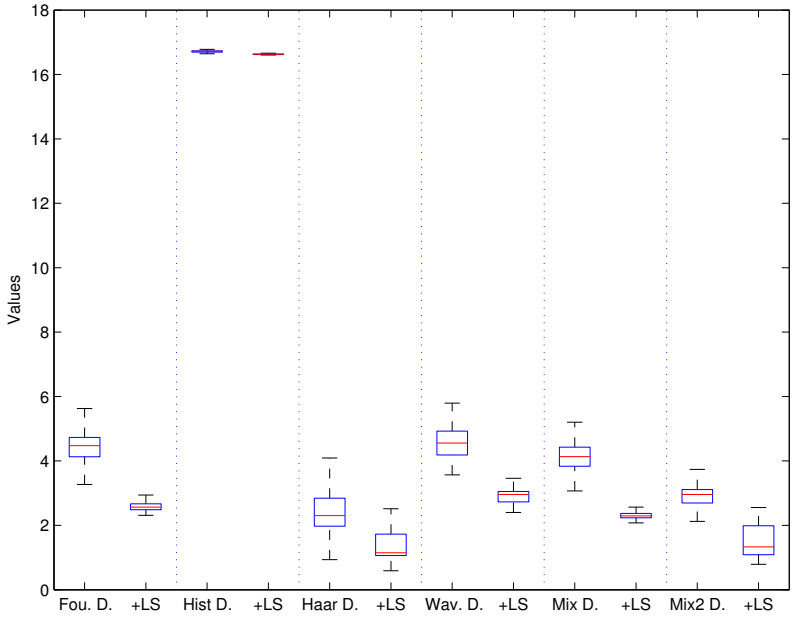


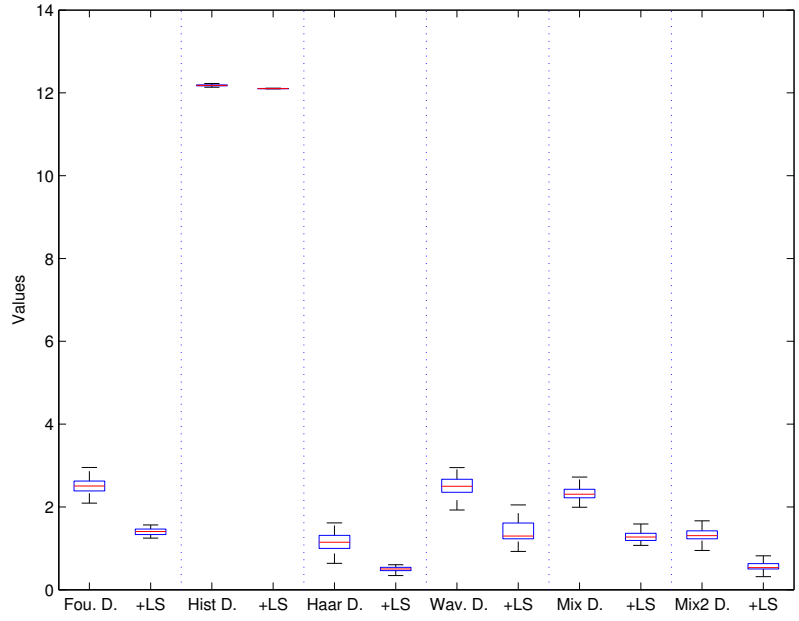


$n = 500$ $n = 2000$

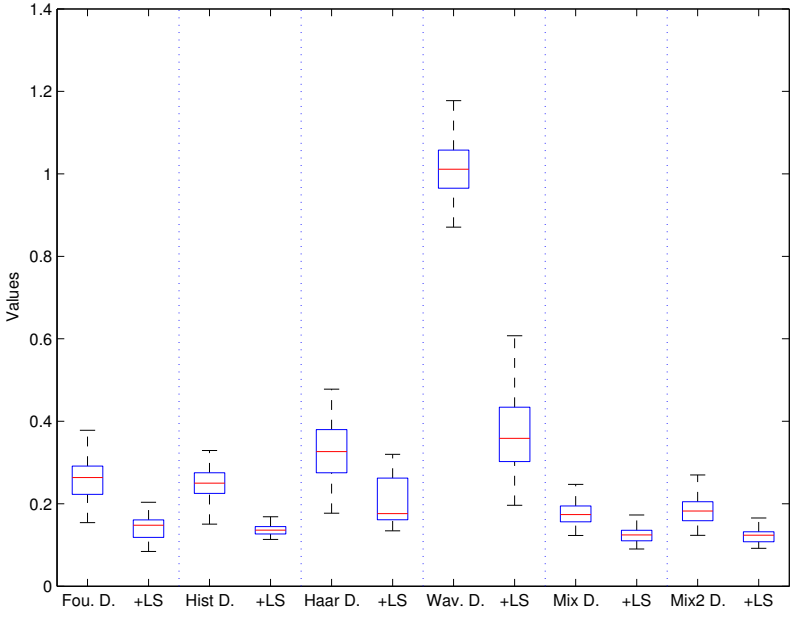
Dantzig / Dantzig+LS f_1/f_2

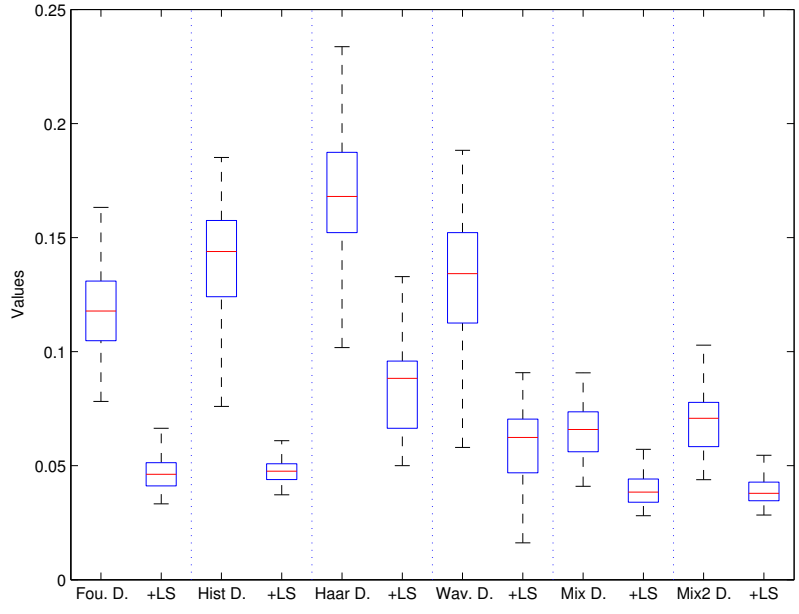
f_1





f_2

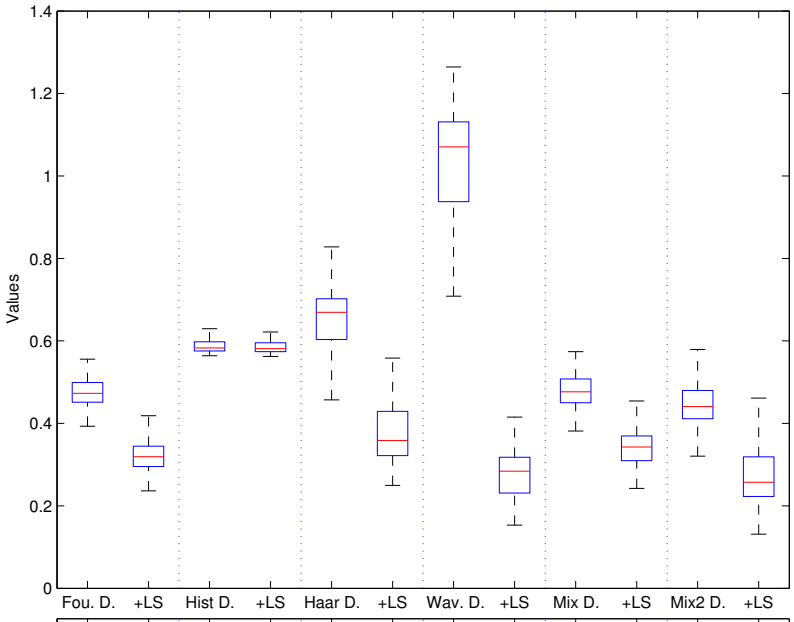


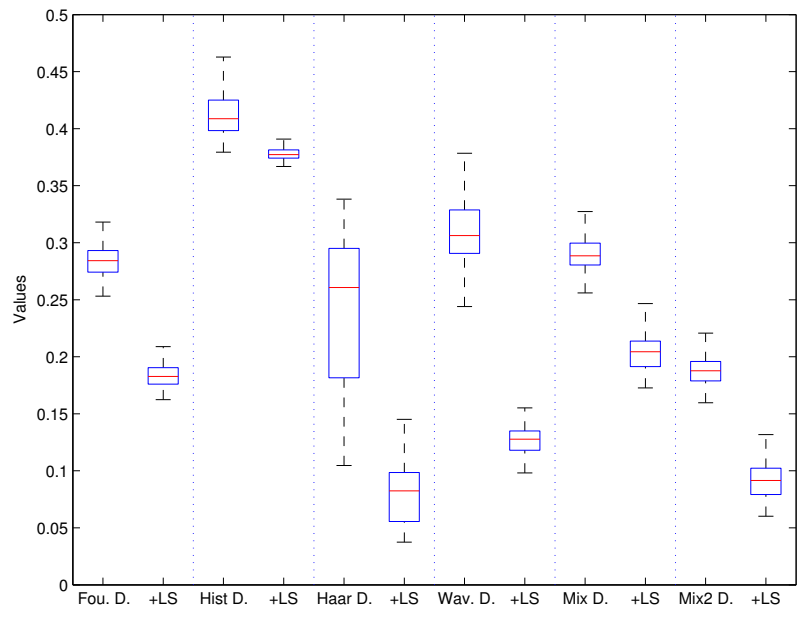


$n = 500$ $n = 2000$

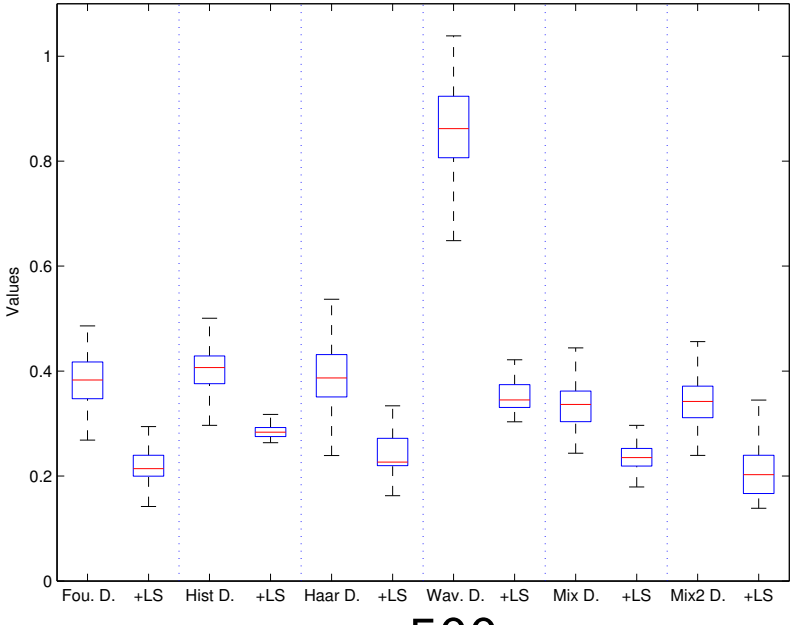
Dantzig / Dantzig+LS f_3/f_4

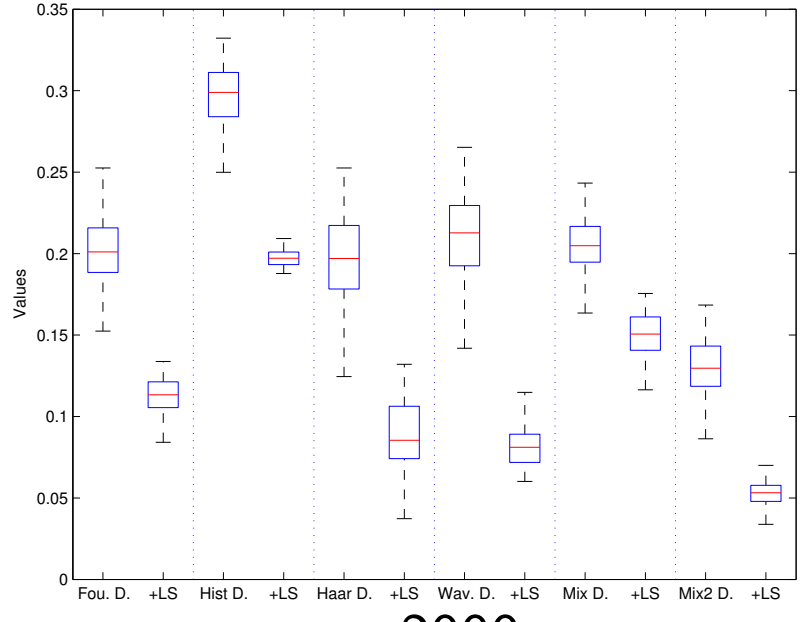
f_3





f_4





$n = 500$ $n = 2000$