Problèmes sparses en statistique

GDR MASCOT NUM - IHP, Paris

Jérémie Bigot
Institut de Mathématiques de Toulouse, UPS

Avril 2008

Modèle linéaire

Observations : $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ telles que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \text{ avec } \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n),$$

$\mathbf{X} = [X_1, \ldots, X_p]$ matrice $n \times p$ connue, et $\boldsymbol{\beta}^* \in \mathbb{R}^p$ vecteur de paramètres à estimer.

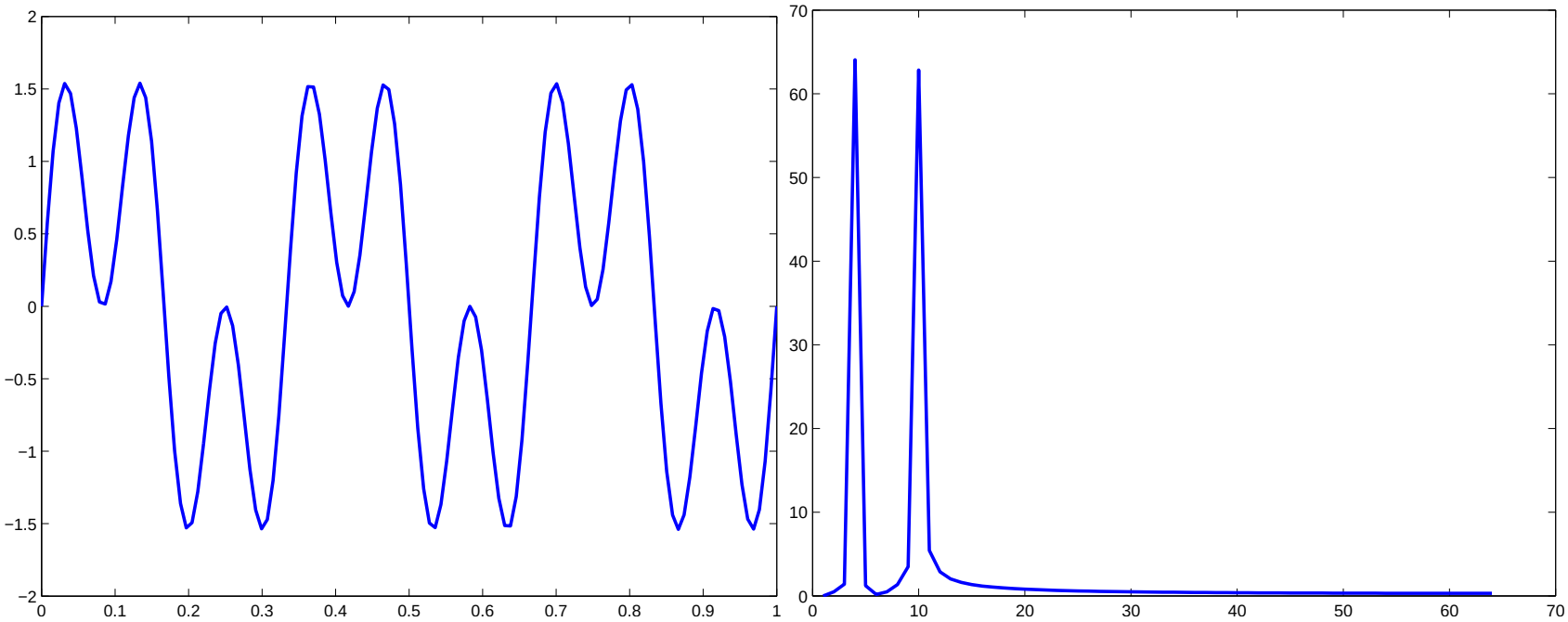
Exemple : régression nonparamétrique et décomposition dans une base de Fourier

$$\begin{aligned} Y_j &= f(x_j) + \epsilon_j,\ j = 1, \ldots, n, \quad x_j = \frac{\ell_j}{p},\ \text{où } \ell_j \in \{1, \ldots, p\} \\ &= \sum_{k=1}^{p} \beta_k^* e^{-i2\pi(k-1)x_j} + \epsilon_j = \sum_{k=1}^{p} \beta_k^* X_k^j + \epsilon_j, \end{aligned}$$

avec $X_k^j = e^{-i2\pi(k-1)x_j}$.

Cas orthogonal : $n = p$ et $\mathbf{X}$ matrice orthogonale

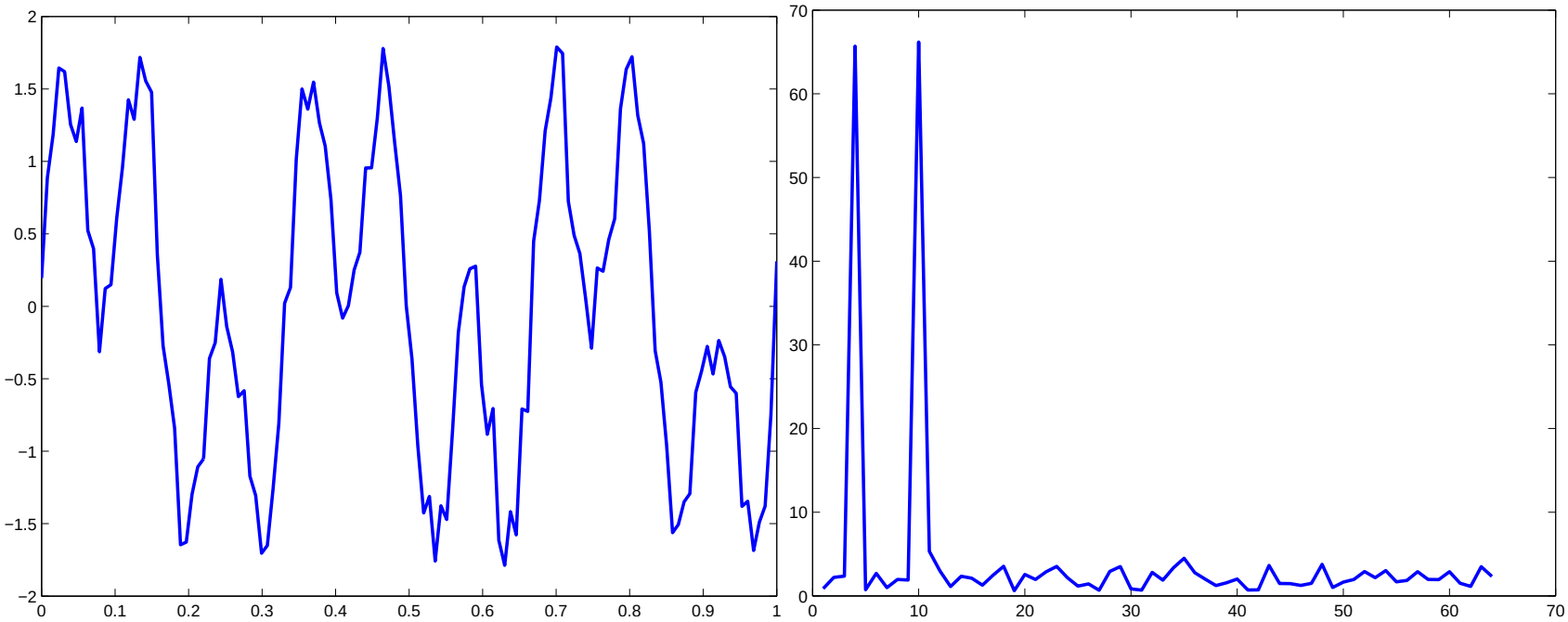
Décomposition en Fourier (sans bruit)



$$n = p = 128$$

Remarque : $\#\{k\,;\beta_k^* \neq 0\}$ est petit !
Le vecteur β^* est dit **sparse** (creux) dans ce cas.

Décomposition en Fourier (avec bruit)



$$n = p = 128$$

Remarque : $\#\{k\,;\beta_k^* \neq 0\}$ est petit !
Le vecteur β^* est dit **sparse** (creux) dans ce cas.

Choix d'un modèle

Soit $m = (i_1, \ldots, i_{|m|})$ un sous-ensemble d'indices dans $1, \ldots, p$.

Modèle : supposer que $\beta_k = 0$ pour tout $k \notin m$.

Observations :

$$\mathbf{Y} = \mathbf{X}^m \boldsymbol{\beta}^m + \boldsymbol{\epsilon},$$

avec $\mathbf{X} = [X_{i_1}, \ldots, i_{|m|}]$ et $\boldsymbol{\beta}^m \in \mathbb{R}^{|m|}$.

Estimation par moindres carrés :

$$\|\mathbf{Y} - \mathbf{X}^m \hat{\boldsymbol{\beta}}^m\|^2 = \min_{\boldsymbol{\beta} \in \mathbb{R}^{|m|}} \|\mathbf{Y} - \mathbf{X}^m \boldsymbol{\beta}\|^2.$$

Problème : comment choisir le meilleur modèle m ?

Risque de l'estimateur : $R(m) = \mathbb{E}\|\mathbf{X}^m \hat{\boldsymbol{\beta}}^m - \mathbf{X}\boldsymbol{\beta}^*\|^2$

$$m_{ideal} = \arg \min_{m \subset \{1,\ldots,p\}} R(m) \quad \text{non calculable en pratique !}$$

Choix d'un modèle

Idée naive : $\hat{R}(m) = \|\mathbf{Y} - \mathbf{X}^m \hat{\boldsymbol{\beta}}^m\|^2$

$$\hat{m}_{naif} = \arg \min_{m \subset \{1,\dots,p\}} \hat{R}(m)$$

mais $\hat{R}(m)$ n'est pas un bon estimateur de $R(m)$ car

$$\mathbb{E}\hat{R}(m) \approx R(m) - 2\sigma^2 |m|$$

Minimisation d'un critère pénalisé :

$$\begin{aligned} \hat{m} &= \arg \min_{m \subset \{1,\dots,p\}} \|\mathbf{Y} - \mathbf{X}^m \hat{\boldsymbol{\beta}}^m\|^2 + 2\sigma^2 |m| \\ \hat{\boldsymbol{\beta}}_{\hat{m}} &= \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sigma^2 \|\boldsymbol{\beta}\|_0, \end{aligned}$$

où $\|\boldsymbol{\beta}\|_0 = \sum_{k=1}^p \mathbb{1}_{\{\beta_k \neq 0\}}$.

Minimisation de la norme ℓ_1

Problème : minimisation sur l'ensemble des 2^p modèles possibles : impossible en pratique car problème NP-difficile si p est très grand (ex : $n = 100, p = 1000$) !

Minimisation d'un critère pénalisé par norme ℓ_1 (critère LASSO) :

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda \|\boldsymbol{\beta}\|_1,$$

où $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$ et $\lambda \geq 0$ est un paramètre de régularisation.

Intérêt de la norme ℓ_1 :

- conduit à des choix de modèles sparses i.e. de nombreuses composantes de $\hat{\boldsymbol{\beta}}$ sont nulles
- problème de minimisation convexe qui se réduit à de la programmation linéaire : mise en oeuvre rapide !

Norme ℓ_1 et seuillage

Cas orthogonal : si $n = p$ et $\mathbf{X}$ orthognale alors

Minimisation d'un critère pénalisé par norme ℓ_1 :

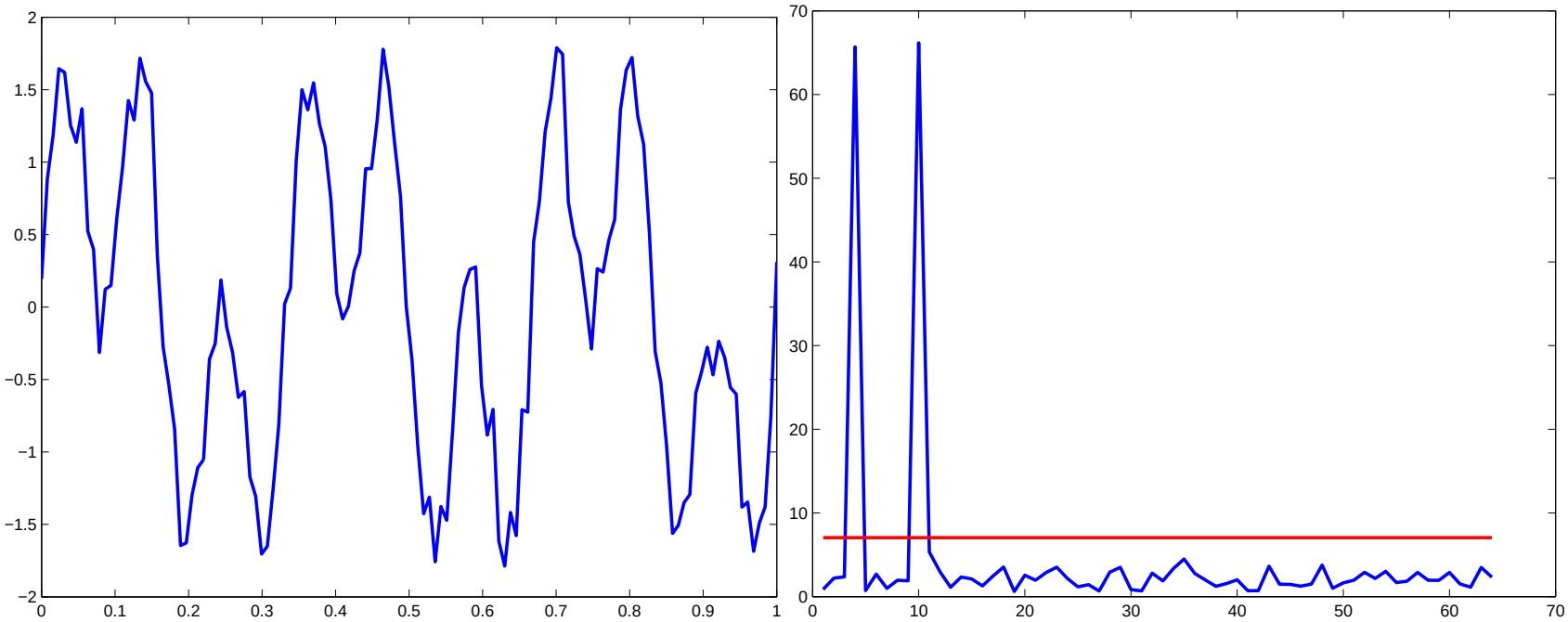
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1,$$

est équivalent à faire du seuillage doux pour $k = 1, \ldots, p$

$$\hat{\boldsymbol{\beta}}_k = \begin{cases} y_k - \lambda & si \quad y_k > \lambda \\ 0 & si \quad |y_k| \leq \lambda \\ y_k + \lambda & si \quad y_k < -\lambda \end{cases}$$

où $(y_1, \ldots, y_p)^t = \mathbf{X}^t\mathbf{Y}$.

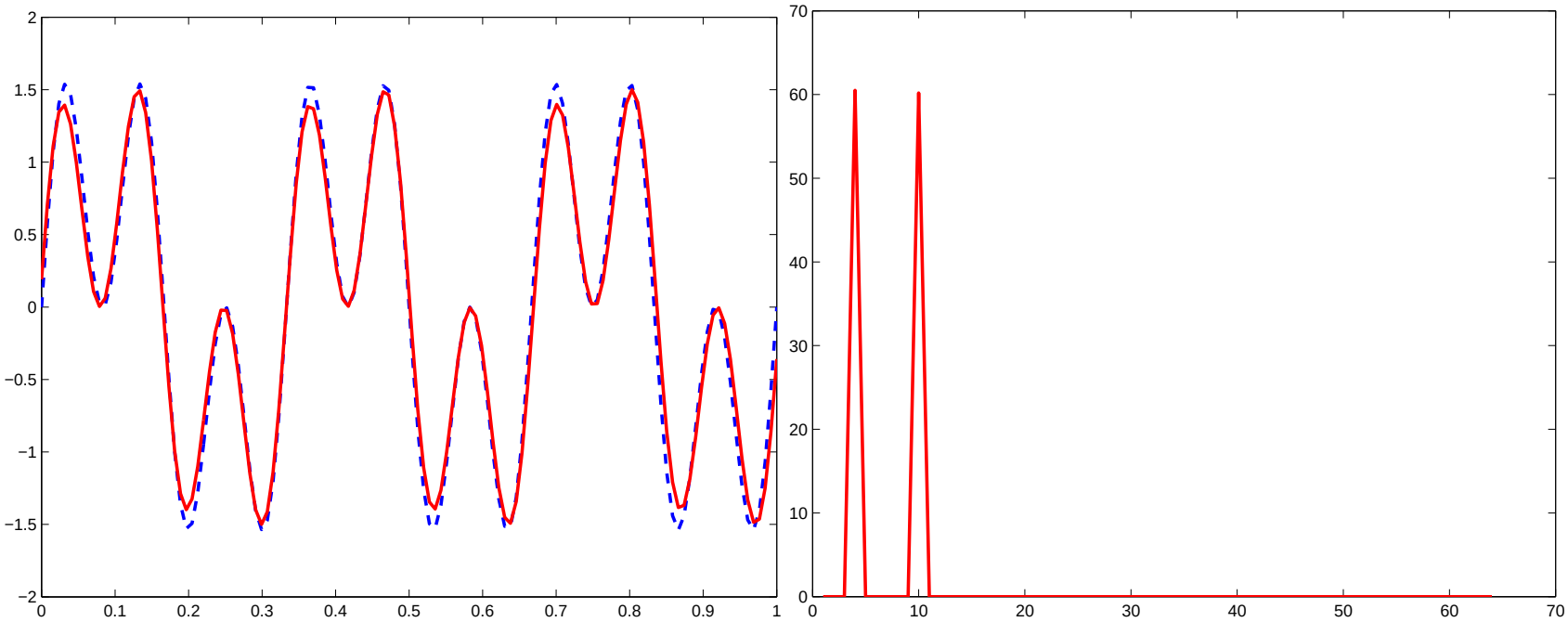
Débruitage par seuillage des coefficients de Fourier



$$n = p = 128$$

Seuillage doux des coefficients : $\lambda = \sigma\sqrt{2\log(n)}$.

Débruitage par seuillage des coefficients de Fourier



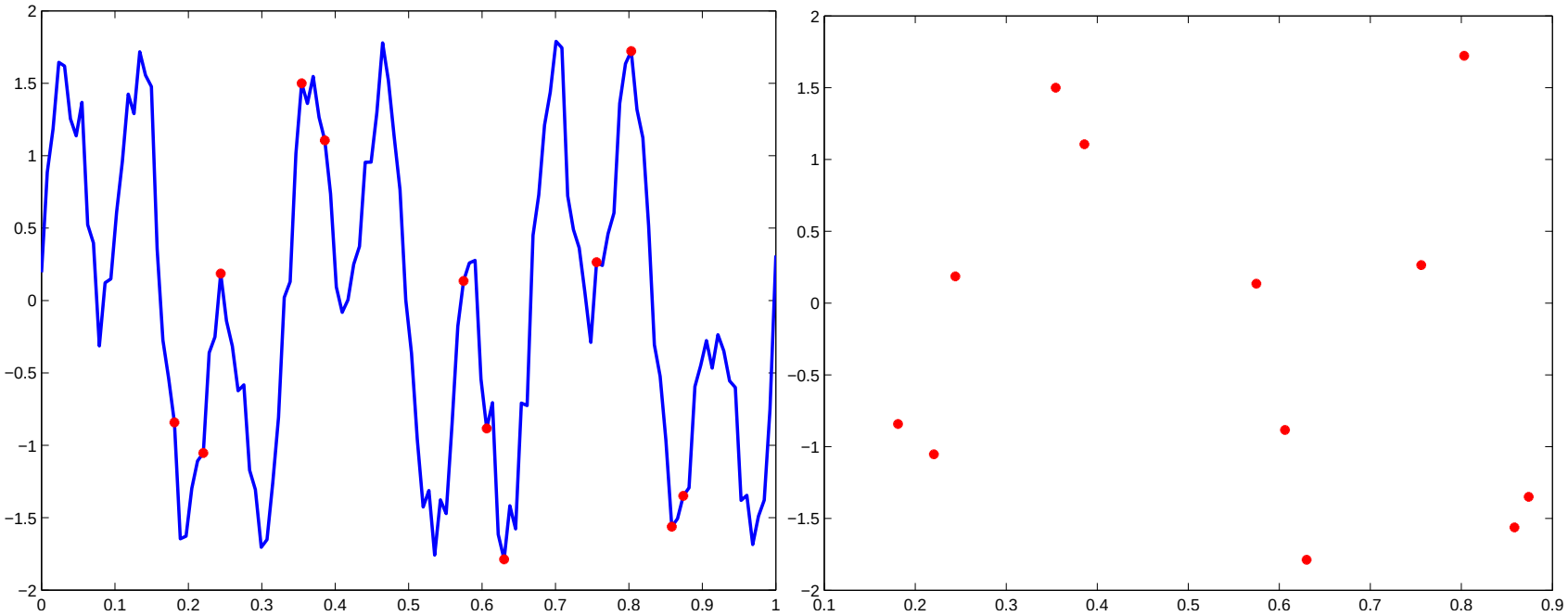
$$n = p = 128$$

Seuillage des coefficients : $\lambda = \sigma\sqrt{2\log(n)}$.

Compressive sensing (Candès, Donoho, Tao, Romberg...)

Question : peut-on reconstruire un signal à partir de l'observation de quelques valeurs ?

Observations : $Y_j = f(x_j) + \epsilon_j,\ j = 1, \ldots, n$, où n est petit et les x_j sont des points tirés au hasard parmi $\{\frac{k}{p}, k = 1, \ldots, p\}$



$p = 128, n = 12$

Compressive sensing (Candès, Donoho, Tao, Romberg...)

Observations : $\mathbf{Y} = \mathbf{X}\beta^* + \epsilon$ avec $\mathbf{X} = [X_1, \ldots, X_p]$ sous-matrice de Fourier de taille $n \times p$ composée de n lignes tirées au hasard parmi p possibles.

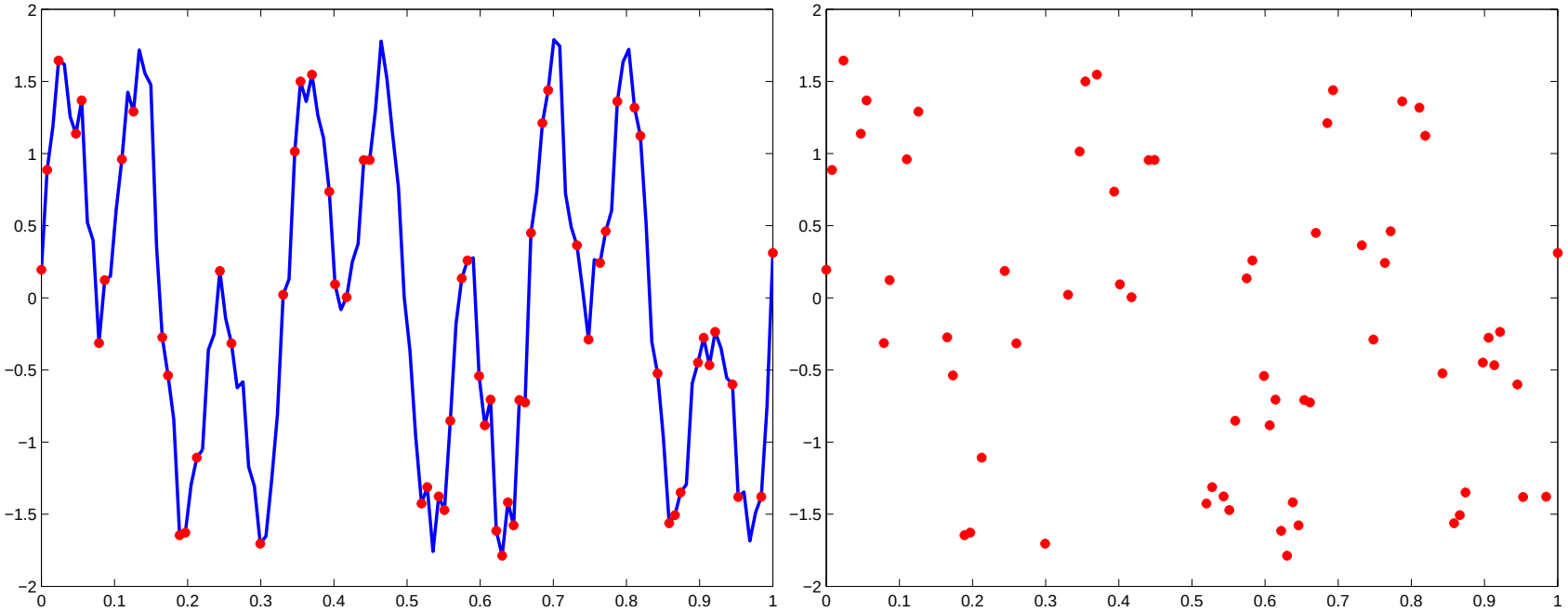
Problème de statistique en grande dimension : cas $n << p$

LASSO : $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + 2\lambda\|\beta\|_1$

Dantzig Selector :

$$\left\{ \begin{array}{lll} \hat{\beta}_{DS} & = & \arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \\ & & \text{tel que} \|\mathbf{X}^t(\mathbf{Y} - \mathbf{X}\beta)\|_\infty \leq \lambda \end{array} \right.$$

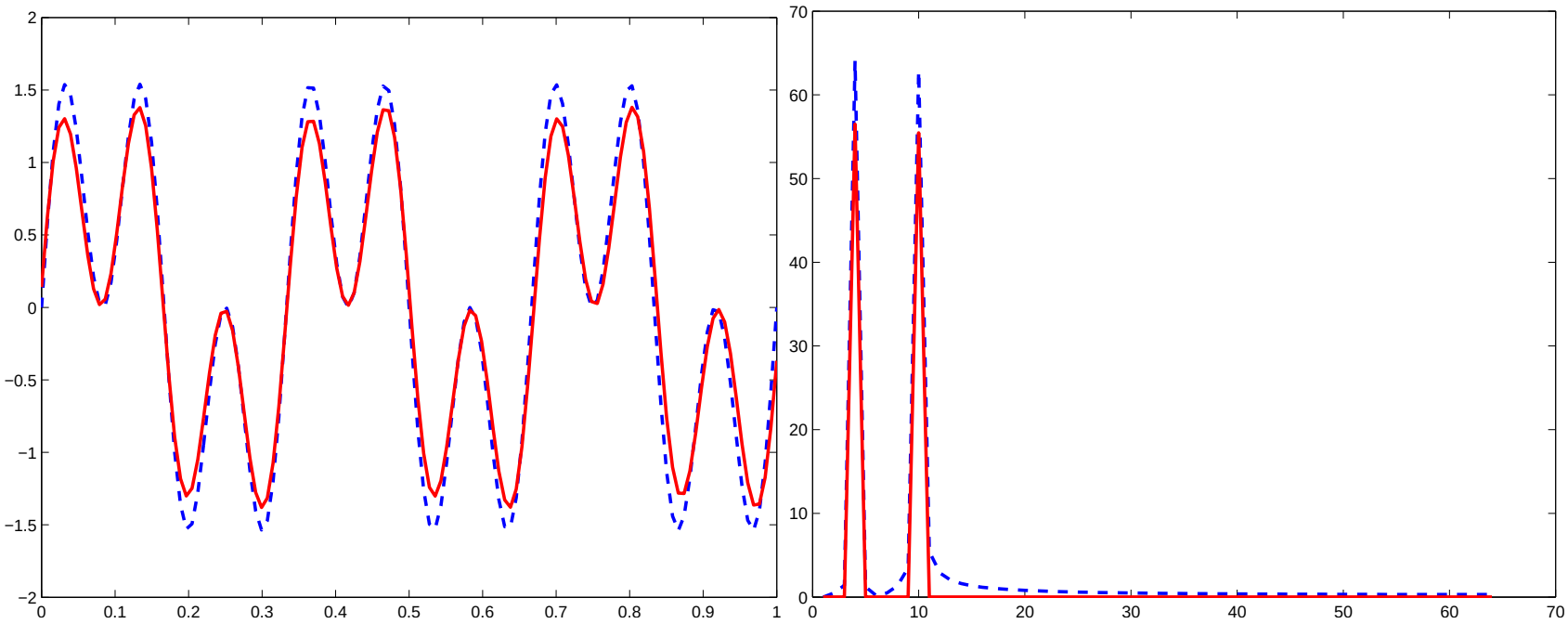
Dantzig selector



$$p = 128, n = p/2$$

Observations partielles - Design aléatoire.

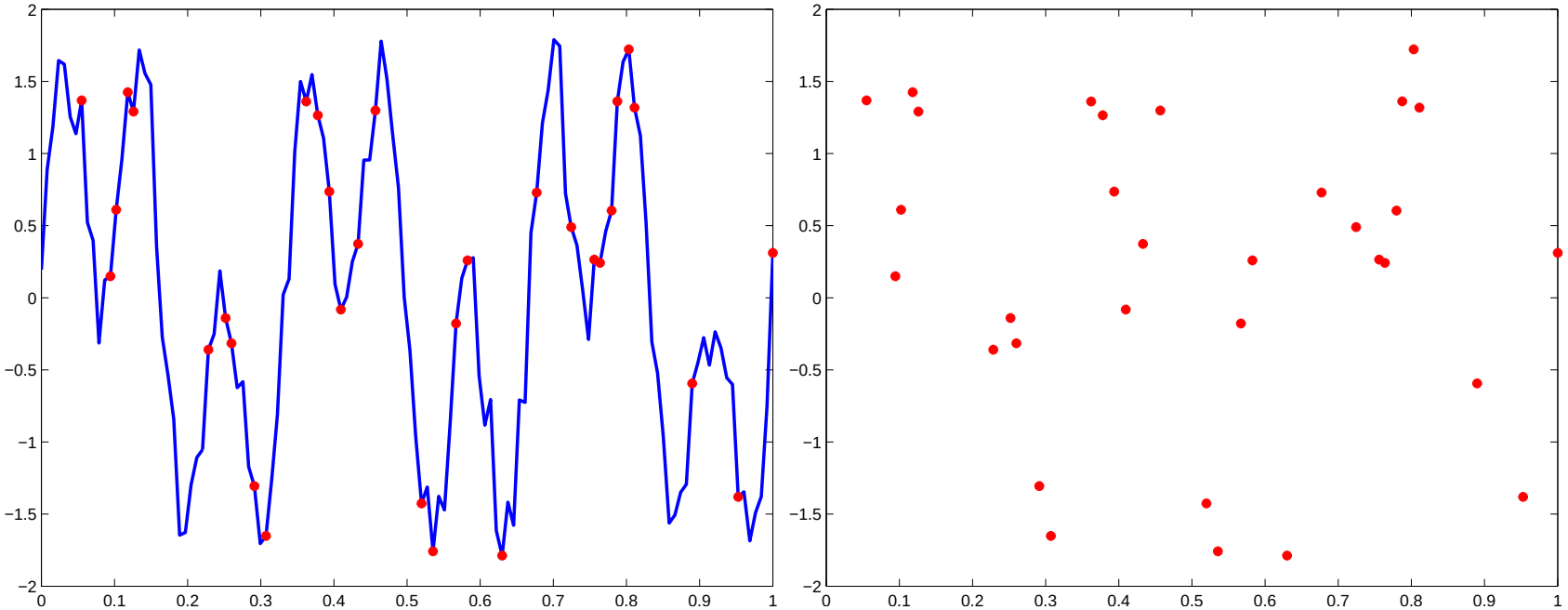
Dantzig selector



$$p = 128, n = p/2$$

Minimisation de la norme ℓ_1 sous contraintes.

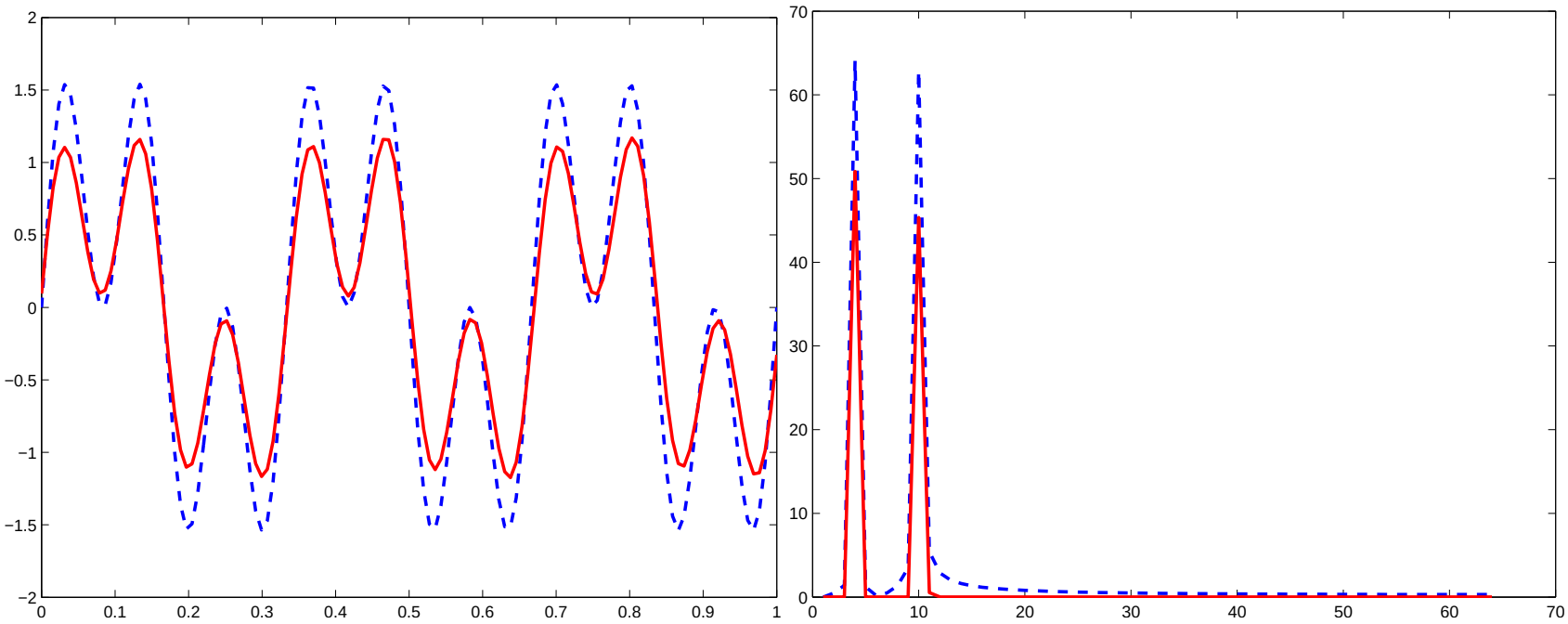
Dantzig selector



$$p = 128, n = p/4$$

Observations partielles - Design aléatoire.

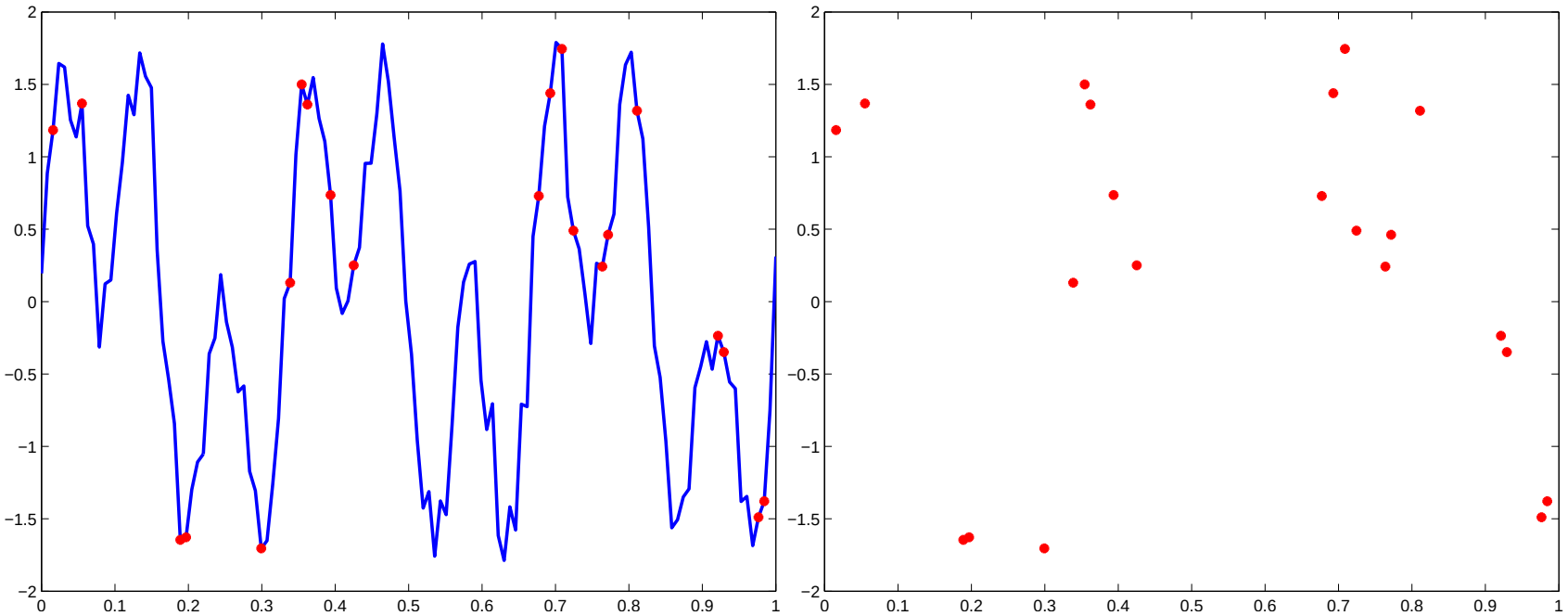
Dantzig selector



$$p = 128, n = p/4$$

Minimisation de la norme ℓ_1 sous contraintes.

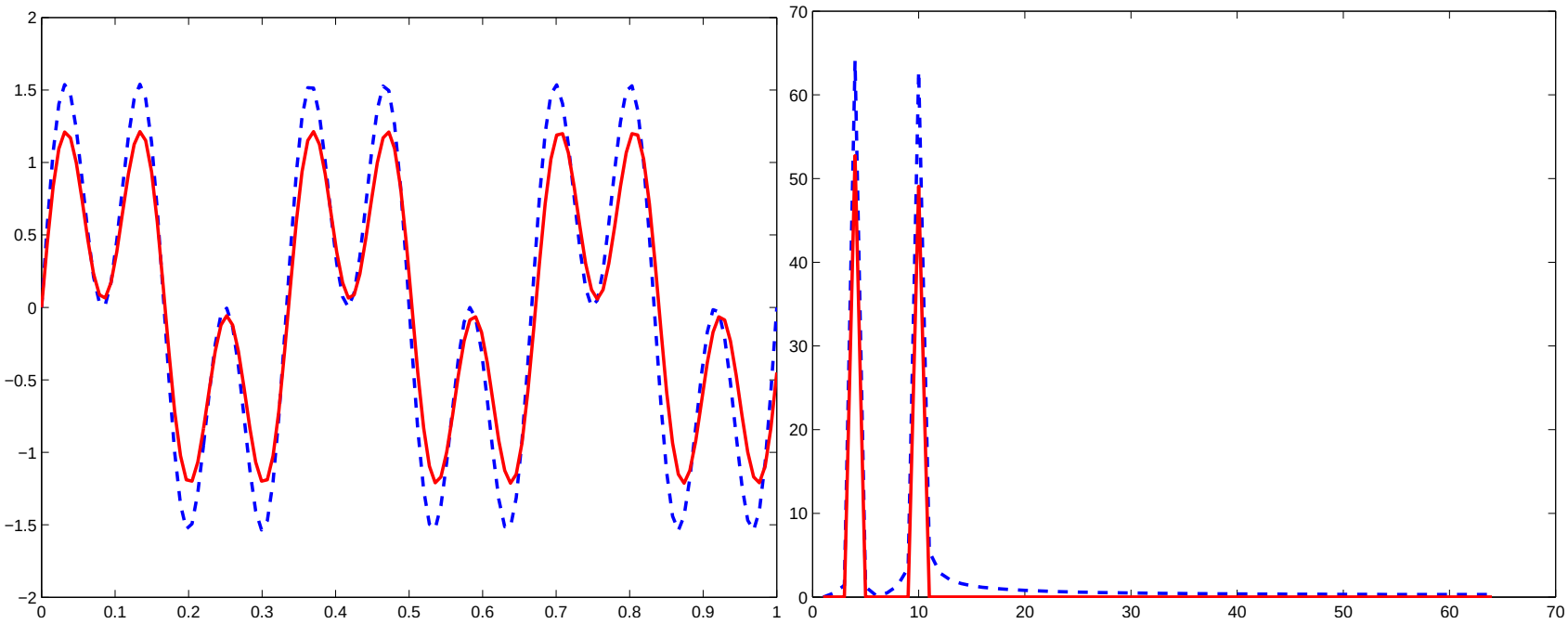
Dantzig selector



$$p = 128, n = p/6$$

Observations partielles - Design aléatoire.

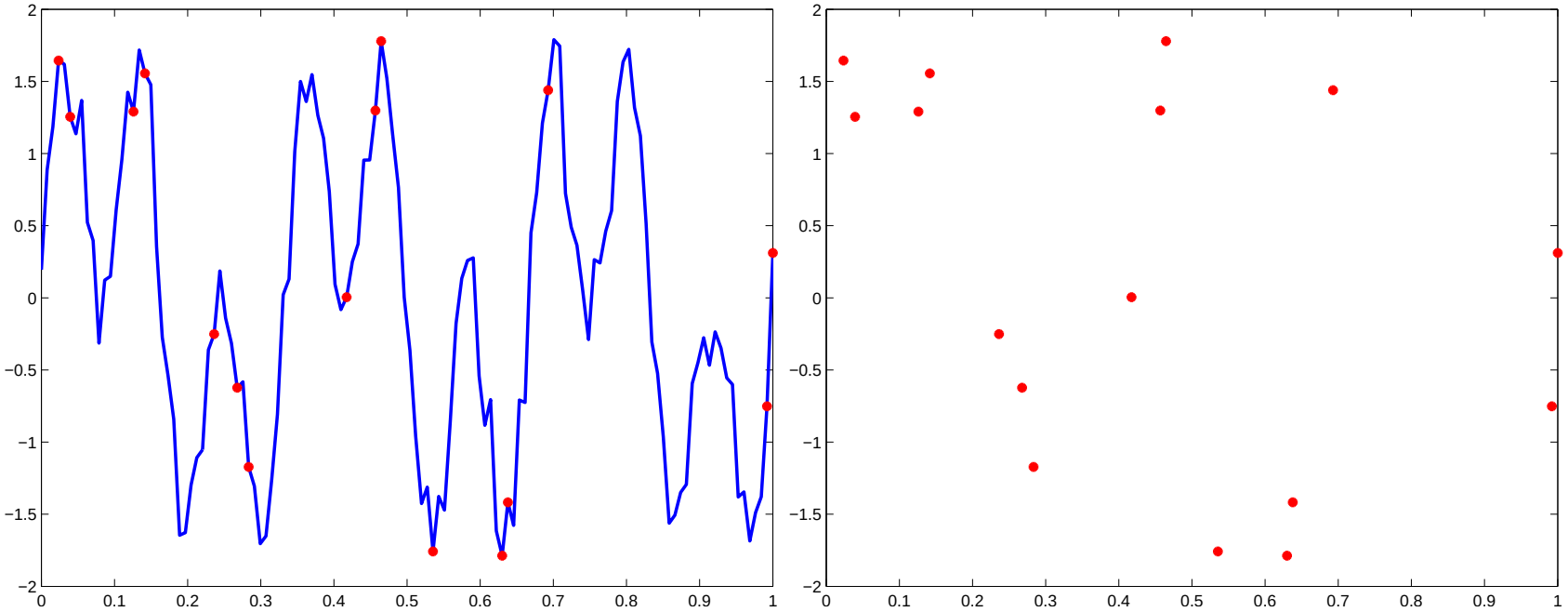
Dantzig selector



$$p = 128, n = p/6$$

Minimisation de la norme ℓ_1 sous contraintes.

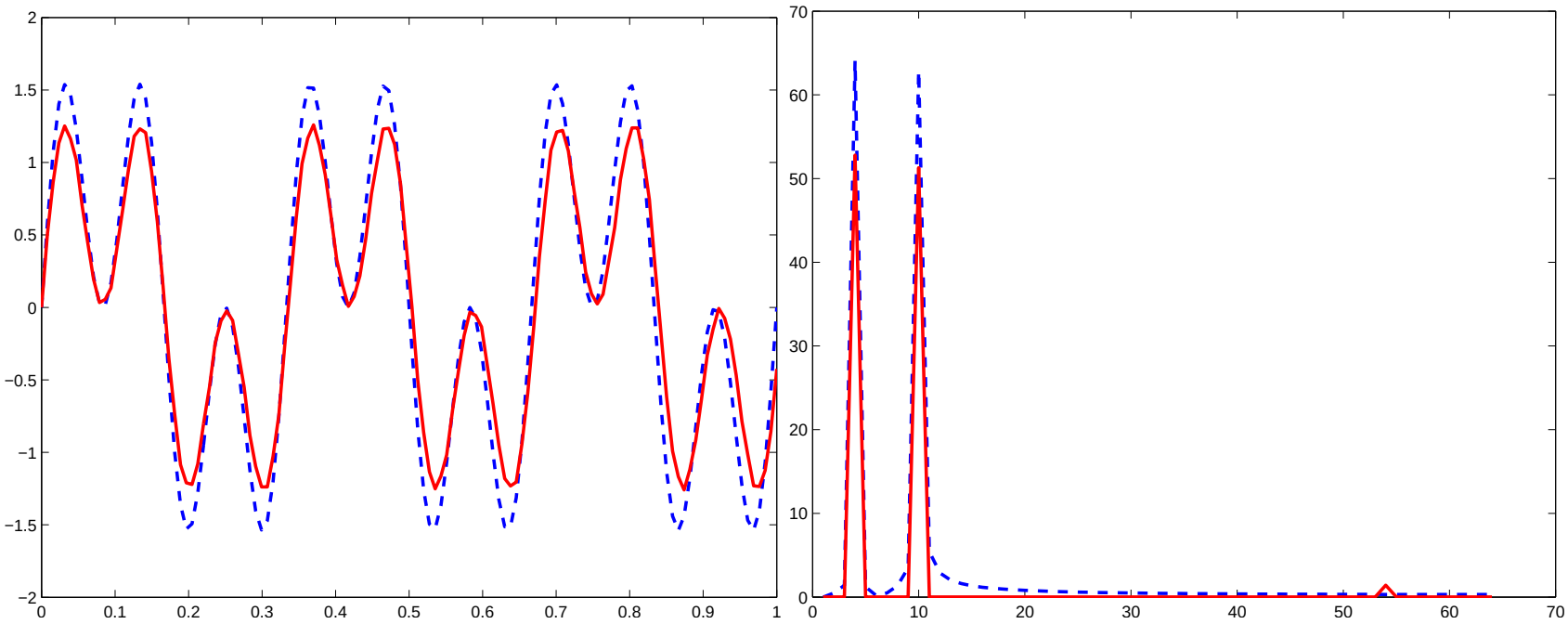
Dantzig selector



$$p = 128, n = p/8$$

Observations partielles - Design aléatoire.

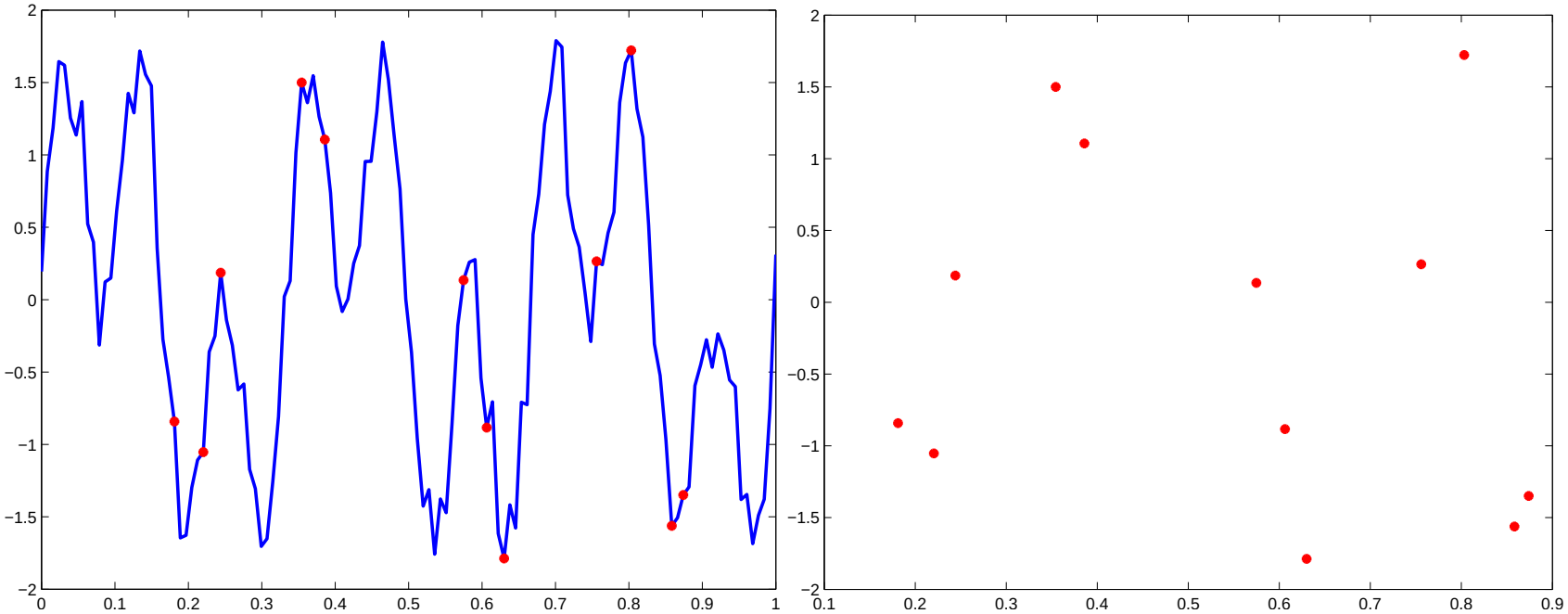
Dantzig selector



$$p = 128, n = p/8$$

Minimisation de la norme ℓ_1 sous contraintes.

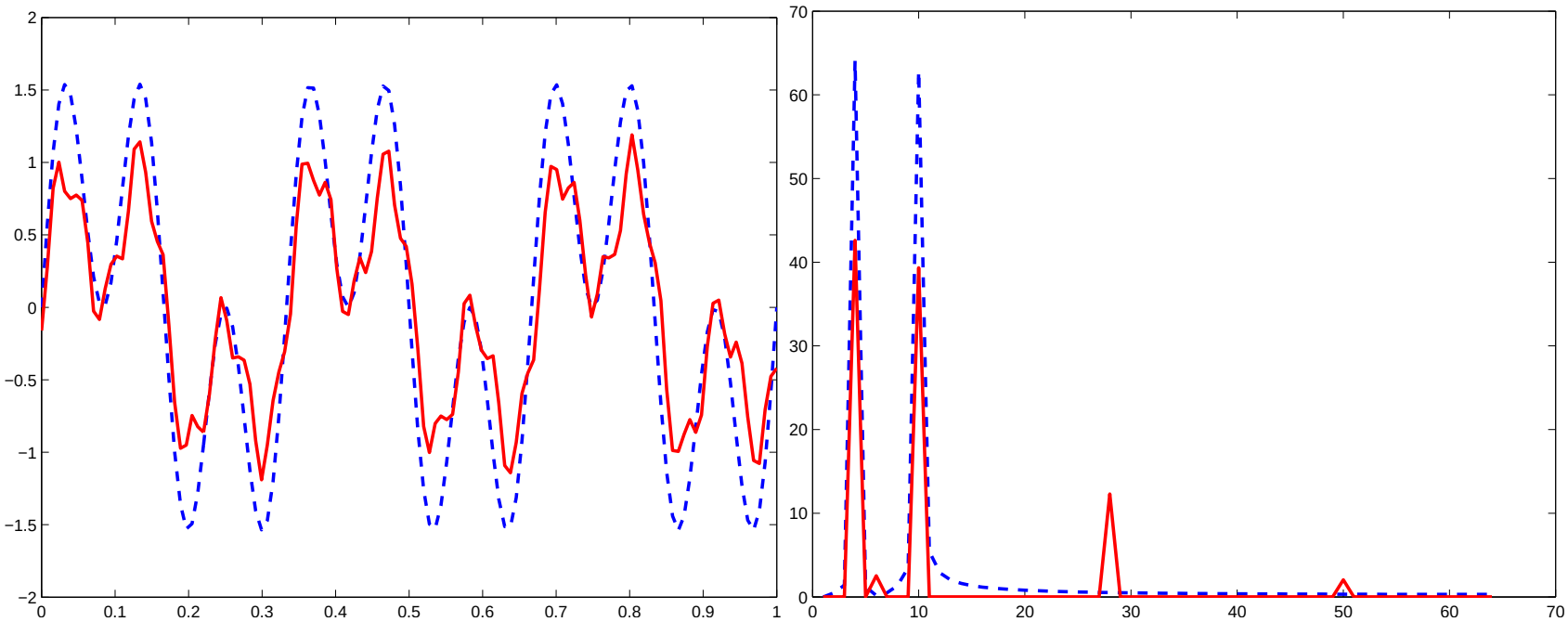
Dantzig selector



$$p = 128, n = p/10$$

Observations partielles - Design aléatoire.

Dantzig selector



$$p = 128, n = p/10$$

Minimisation de la norme ℓ_1 sous contraintes.

Conditions de consistence des estimateurs

Question : quelles sont les conditions sur n, p, la matrice $\mathbf{X}$ et le vecteur β qui garantissent une bonne estimation ?

Que veut-on estimer ? On peut s'intéresser à

$$\begin{aligned} \|\hat{\beta} - \beta^*\|_p^p \leq \quad ? \quad (p = 1, 2) \\ support(\hat{\beta}) = support(\beta^*) \quad ? \\ \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^*\|^2 \leq \quad ? \end{aligned}$$

(résultats en espérance ou en probabilité)

D'où vient l'aléatoire ?

- absence de bruit additif (vecteur ϵ) et matrice $\mathbf{X}$ aléatoire
- présence de bruit additif + β est un vecteur composé de S composantes non-nulles dont les positions sont tirées au hasard de sorte que $\mathbb{E}(S) \leq C(n, p)$, et $\mathbf{X}$ matrice déterministe

Un peu d'heuristique

Conditions sur $\mathbf{X}$

- les sous-matrices de $\mathbf{X}$ de dimension $n \times S$ doivent être des quasi-isométries avec $S \leq C(n,p)$
- le maximum des produit scalaires $|\langle X_k, X_{k'} \rangle|$ doit être plus petit que $C \log(p)^{-1}$

Conditions sur β

- β doit être suffisamment sparse $\#\{k\,;\beta_k^* \neq 0\} \leq Cn$ avec $C < 1$
- les β_k non-nuls doivent être d'amplitude suffisamment élevée