Contrôle : Statistique pour les big data

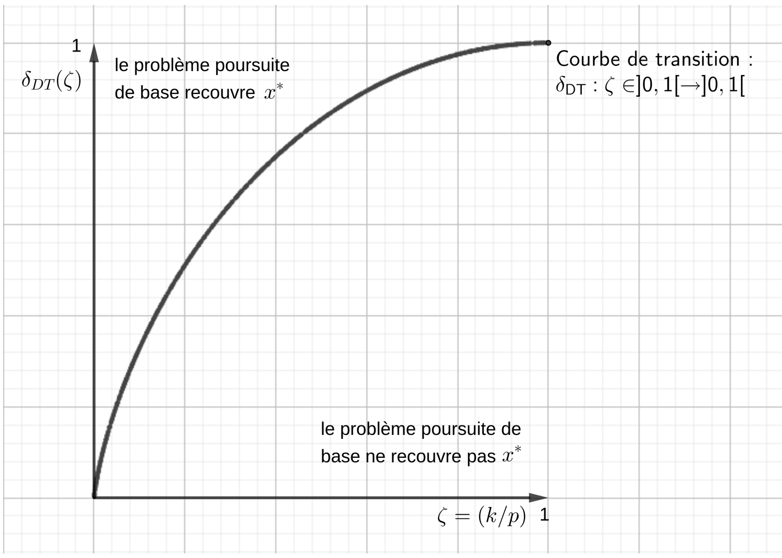
Patrick Tardivel, Université de Bourgogne

Notation

La notation $\|x\|_0 := |\{i \in \{1,,p\}; x_i \neq 0\}|$ représente la “norme” ℓ_0 de $x \in \mathbb{R}^p$; c'est-à-dire le nombre de composantes non-nulles de x.

Exercice 1 (questions de cours)

1. *Soit $A \in \mathbb{R}^{n\times p}$ une matrice telle que $\dim(\ker(A)) \geq 1$. On rappelle que le germe de la matrice A est défini par*
$$\mathrm{germe}(A) := \min\{\|h\|_0 : h \in \ker(A) \text{ et } h \neq 0\}.$$
Soit $b \in \mathrm{im}(A)$ et x^* une solution du système linéaire d'équations $Ax = b$. Montrer que si $\|x^*\|_0 < \mathrm{germe}(A)/2$ alors x^* est l'unique solution la plus parcimonieuse du système $Ax = b$ (i.e. $S_{A,\mathrm{parc}}(b) = \{x^*\}$).
2. *Soit $A \in \mathbb{R}^{n\times p}$ avec $\dim(\ker(A)) \geq 1$, $b \in \mathrm{im}(A)$ et x^* une solution du système linéaire d'équations $Ax = b$. Montrer que si l'inégalité suivante*
$$\forall h \in \ker(A) \setminus \{0\} \quad \sum_{i \in \mathrm{supp}(x^*)} |h_i| < \sum_{i \notin \mathrm{supp}(x^*)} |h_i|$$
est satisfaite alors x^* est l'unique solution poursuit de base du système $Ax = b$.
3. *Soit $Z \in \mathbb{R}^{n\times p}$ une matrice dont les coefficients sont iid $\mathcal{N}(0,1)$ et $x^* \in \mathbb{R}^p$ tel que $\|x^*\|_0 = k$. Pour chacun des cas suivants justifier si la probabilité que x^* soit identifiable relativement à Z et à la norme ℓ_1 est proche de 0 ou proche de 1.*
a) Lorsque $k = 120, n = 500$ et $p = 1000$.
b) Lorsque $k = 500, n = 1000$ et $p = 2000$.
Pour cette question on pourra utiliser la courbe de transition de Donoho-Tanner donnée ci-dessous :



Quelques rappels sur le LASSO

Soit $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ et $\lambda > 0$, l'ensemble $S_{X,\lambda}(y)$ des solutions LASSO est défini par

$$S_{X,\lambda}(y) := \underset{b\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1.$$

On rappelle que $\hat{\beta} \in S_{X,\lambda}(y)$ si seulement si l'une des caractérisations suivantes est satisfaite

$$\begin{cases} \|X^T(y - X\hat{\beta})\|_\infty \leq \lambda \\ \forall i \in \mathrm{supp}(\hat{\beta}) \quad X_i^T(y - X\hat{\beta}) = \lambda \mathrm{sign}(\hat{\beta}_i) \end{cases} \Leftrightarrow \begin{cases} \|X^T(y - X\hat{\beta})\|_\infty \leq \lambda \\ \hat{\beta}^T X^T(y - X\hat{\beta}) = \lambda\|\hat{\beta}\|_1 \end{cases}.$$

On pose $\hat{r} = y - X\hat{\beta}$, on rappelle également que $\hat{r}$ est la solution du problème suivant

$$\hat{r} = \underset{r\in\mathbb{R}^n}{\mathrm{argmin}}\ \|y - r\|_2 \text{ sous la contrainte } \|X^T r\|_\infty \leq \lambda.$$

Exercice 2 *Soit $\lambda = 2$ et $X \in \mathbb{R}^{2\times 3}$ la matrice suivante*

$$X = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 1 & 4 \end{pmatrix}.$$

1. *Tracer le polyèdre nul du LASSO : $\{r \in \mathbb{R}^2 : \|X^T r\|_\infty \leq \lambda\}$.*

2. *En utilisant le polyèdre nul du LASSO, justifier si $S_{X,\lambda}(y) = \{0\}$ pour les vecteurs suivants : a) $y = (1/4, 1/4)^T$ et b) $y = (1, 1/2)^T$.*
3. *En utilisant le polyèdre nul du LASSO, justifier que pour tout $y \in \mathbb{R}^2$ et pour tout $\hat{\beta} \in S_{X,\lambda}(y)$ on a $\hat{\beta}_2 = 0$.*
4. *En utilisant le polyèdre nul du LASSO, déterminer $S_{X,\lambda}((2, 1/4)^T)$.*

Problème : la constante d'isométrie restreinte

Définition 1 *Soit $A \in \mathbb{R}^{n\times p}$ et $s \in \{1,\dots,p\}$. La constante d'isométrie restreinte δ_s est définie par*

$$\delta_s := \inf\left\{l \geq 0 : (1-l)\|x\|^2 \leq \|Ax\|^2 \leq (1+l)\|x\|^2 \quad \forall x \in \mathbb{R}^p, \|x\|_0 \leq s\right\}.$$

La constante d'isométrie restreinte donne une condition sous laquelle une solution est à la fois la plus parcimonieuse et la solution de poursuite de base.

Proposition 1 (condition d'isométrie restreinte [1]) *Soit $A \in \mathbb{R}^{n\times p}$, $b \in \mathrm{im}(A)$, x^* une solution du système linéaire d'équations $Ax = b$ et $s := \|x^*\|_0$. Si $\delta_s < 1/3$ alors x^* est simultanément l'unique solution poursuite de base et l'unique solution la plus parcimonieuse.*

Le but de ce problème est d'étudier un exemple montrant que lorsque $\delta_s \geq 1/3$ alors, en toute généralité, il n'est pas possible de conclure que x^* est simultanément l'unique solution poursuite de base et l'unique solution la plus parcimonieuse. On utilisera la notation $\langle\cdot,\cdot\rangle$ pour le produit scalaire usuel dans $\mathbb{R}^p$.

Soit $v_1,\dots,v_p$ une base orthonormée de $\mathbb{R}^p$ où $v_1 = (1/\sqrt{2s},\dots,1/\sqrt{2s},0\dots,0)$ et $\|v_1\|_0 = 2s$.

1. Soit $x \in \mathbb{R}^p$ tel que $\|x\|_0 \leq s$. En utilisant (astucieusement) l'inégalité de Cauchy-Schwarz montrer l'inégalité suivante
$$|\langle x, v_1\rangle| \leq \frac{1}{\sqrt{2}}\|x\|_2.$$

Soit $f : \mathbb{R}^p \to \mathbb{R}^p$ l'application linéaire telle que

$$f(x) = \sqrt{4/3}\left(\langle x, v_2\rangle v_2 + \ldots + \langle x, v_p\rangle v_p\right).$$

On pose $A \in \mathbb{R}^{p\times p}$ la matrice de f dans la base canonique de $\mathbb{R}^p$ ($i.e$ $f(x) = Ax$).

2. Montrer l'identité suivante
$$\forall x \in \mathbb{R}^p \quad \|Ax\|_2^2 = \frac{4}{3}\left(\|x\|_2^2 - \langle x, v_1\rangle^2\right)$$
3. En déduire que pour la matrice A la constante d'isométrie restreinte vaut $\delta_s = 1/3$, c'est-à-dire établir l'inégalité suivante
$$\forall x \in \mathbb{R}^p, \|x\|_0 \leq s \quad \frac{2}{3}\|x\|_2^2 \leq \|Ax\|_2^2 \leq \frac{4}{3}\|x\|_2^2.$$
4. Soit $\tilde{x} \in \mathbb{R}^p$ et $x^* \in \mathbb{R}^p$ définis par
$$x^* = (\underbrace{1\ \ldots\ 1}_{\text{s termes}}, 0\ \ldots\ 0)^T \text{ et } \tilde{x} = (\underbrace{0,\dots,0}_{\text{s termes}}\ \underbrace{-1,\dots,-1}_{\text{s termes}}, 0,\dots,0)^T$$

   Montrer que $Ax^* = A\tilde{x}$ puis conclure.

Références

[1] T Tony Cai and Anru Zhang. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1) :74–93, 2013.