Régression pénalisée : le Lasso

V. Viallon

M2 Maths en Action

Cadre considéré dans ce cours

- On supposera disposer d'un échantillon $(x_1, Y_1), \dots, (x_n, Y_n)$ tel que

$$Y_i = x_i^T \beta^* + \xi_i, \quad i = 1, \dots, n,$$

où les $\xi_i \sim \mathcal{N}(0, \sigma^2)$ sont i.i.d., les $Y_i \in \mathbb{R}$ sont aléatoires mais les $x_i \in \mathbb{R}^p$ sont déterministes, et le paramètre $\beta^* \in \mathbb{R}^p$ est inconnu.

$\Rightarrow$ régression linéaire sur design fixe, avec erreurs gaussiennes (sans intercept).

- Exemple typique :
  - Y_i : niveau d'expression du gène G chez l'individu i
  - $x_i = (x_{i1}, \dots, x_{ip})^T$: SNPs pour l'individu i (à valeurs dans $\{0, 1, 2\}$ généralement).

Ecriture matricielle

- Le modèle peut se réécrire sous forme matricielle

$$\mathbf{Y} = \mathbf{X}\beta^* + \boldsymbol{\xi}$$

où

- $\mathbf{Y} = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ et $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T \in \mathbb{R}^n$
- $\mathbf{X} = (x_1^T, \dots, x_n^T)^T \in \mathbb{R}^{n \times p}$.

Rq: Dans ce cours, on considère que $p = p(n)$ (typiquement, fonction croissante de n).

Cadre "standard"

- $n \gg p$, et $\text{rang}(\mathbf{X}) = p$
- alors l'estimateur des MCO

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

est donné par

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

- Si les ξ sont iid $\mathcal{N}(0, \sigma^2)$, on a

$$\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{\sigma^2} \sim \chi^2_p$$

et donc *[exercice: utilisez l'inégalité de Tchebychev[1]]*

$$(ii) \quad \frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} = \mathcal{O}_{\mathbb{P}}\left(\frac{p}{n}\right)$$

où $X_n = \mathcal{O}_{\mathbb{P}}(a_n) : \forall \epsilon > 0, \exists M \geqslant 0 : \mathbb{P}(|X_n/a_n| > M) \leqslant \epsilon$.

[1] Soit X, une v.a. d'espérance μ et de variance finie σ^2, alors pour tout $\alpha > 0$, $\mathbb{P}(|X - \mu| \geqslant \alpha) \leqslant \sigma^2/a^2$.

Cadre de la grande dimension

- $p \geqslant n$ (voire $p \gg n$)
- alors $\text{rang}(\mathbf{X}) < p$ (et donc $\mathbf{X}^T\mathbf{X}$ n'est pas inversible)
- l'estimateur des MCO n'est plus unique (même formule avec pseudo-inverse de Moore-Penrose)
- et il "overfit" les données
- notamment

$$\frac{\|\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} = \mathcal{O}_{\mathbb{P}}(1).$$

(Exemple avec $n = p$ et $\mathbf{X} = \mathbf{I}_n$.)

"Solutions"

- Hypothèse de parcimonie : $\boldsymbol{\beta}^*$ est "creux", i.e.

$$\boxed{s_0 := \sharp\{j : \beta_j^* \neq 0\} \ll p \quad (\text{et surtout} \ll n).}$$

- Alors, si l'on connaissait l'ensemble $J_0 := \{j : \beta_j^* \neq 0\}$, on aurait une erreur de prédiction : $\mathcal{O}_{\mathbb{P}}\left(\frac{s_0}{n}\right) \to_{\mathbb{P}} 0$.

$\Rightarrow$ Sélection de variables

- meilleure interprétabilité du modèle
- meilleur pouvoir prédictif aussi

- Autre hypothèse possible : peu de coefficients "grands" (plutôt que peu de coefficients non nuls).

Principe général de la régression pénalisée

Pour un $\lambda \geqslant 0$,

$$\phi_{\mathcal{P}}(\lambda) := \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda \mathcal{P}(\boldsymbol{\beta}).$$

- Si $\lambda = 0$: MCO
- Différents choix pour $\mathcal{P}(\boldsymbol{\beta})$:
  - $\|\boldsymbol{\beta}\|_0 = \sharp\{j : \beta_j \neq 0\}$: Théorie +++, Implémentation –
    - AIC : $\lambda = \sigma^2/n$
    - BIC : $\lambda = \sigma^2 \log(n)/(2n)$
    - Pb "combinatoire" : on doit énumérer les 2^p modèles possibles

Principe général de la régression pénalisée

Pour un $\lambda \geqslant 0$,

$$\phi_{\mathcal{P}}(\lambda) := \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda \mathcal{P}(\boldsymbol{\beta}).$$

- Si $\lambda = 0$: MCO
- Différents choix pour $\mathcal{P}(\boldsymbol{\beta})$:
  - $\|\boldsymbol{\beta}\|_0 = \sharp\{j : \beta_j \neq 0\}$: Théorie +++, Implémentation –
  - $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$ (Lasso) : Théorie : ++, Implémentation ++
  - $\|\boldsymbol{\beta}\|_2^2 = \sum_j \beta_j^2$ (Ridge) : Théorie : +, Implémentation ++
  - ...

Least Absolute Shrinkage and Selection Operator

Pour $\lambda \geqslant 0$,

$$\hat{\boldsymbol{\beta}}(\lambda) \in \operatorname{Arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda \|\boldsymbol{\beta}\|_1. \quad (1)$$

- Problème convexe, mais la solution n'est pas nécessairement unique
- Si $\lambda = 0$: MCO
- la solution est typiquement creuse : plus λ est grand, et plus $\hat{\boldsymbol{\beta}}(\lambda)$ est creux (en "gros").

Propriétés de sélection du Lasso: intuition

Le problème d'optimisation

$$\phi(\lambda) := \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda \|\boldsymbol{\beta}\|_1.$$

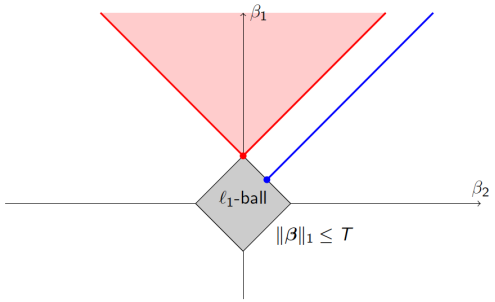
est équivalent, pour une certaine valeur de $T = T(\lambda)$, à

$$\tilde{\phi}(T) := \min_{\|\boldsymbol{\beta}\|_1 \leqslant T} \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n}.$$

Ex: dans le cas où $n = p$ et $\mathbf{X} = \mathbf{I}_n$. On cherche alors à résoudre

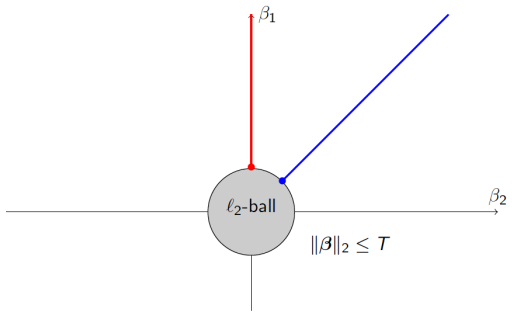
$$\tilde{\phi}(T) := \min_{\|\boldsymbol{\beta}\|_1 \leqslant T} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \beta_i)^2.$$

Cône ℓ_1



Emprunté aux slides de J. Mairal

Cône ℓ_2



Emprunté aux slides de J. Mairal

Propriétés du Lasso : que peut-on espérer ?

On peut espérer qu'avec grande probabilité, sous certaines hypothèses et pour des choix appropriés de λ,

- Estimation : $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^*$
- Sélection : $\widehat{J}(\lambda) \approx J_0$, où $\widehat{J}(\lambda) = \{j : \hat{\boldsymbol{\beta}}_j(\lambda) \neq 0\}$.
- Prédiction : $n^{-1}\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \approx s_0/n$

On précisera plus tard la signification de $\approx$ dans chacun des cas précédents.

Rq1: "Difficulté" pour l'analyse des propriétés des estimateurs Lasso (par rapport aux MCO): pas de forme explicite (on va utiliser des conditions d'optimalité qui caractérisent les solutions du problème (1)).

Rq2: Propriétés non asymptotiques.

Une première condition d'optimalité

Pour simplifier les notations, on suppose que λ est fixé et on pose $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda)$, une solution de (1).

Lemme 2.1

Dénotons le gradient de $(2n)^{-1}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ par $G(\boldsymbol{\beta}) = -\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/n$. Alors une CNS pour que $\hat{\boldsymbol{\beta}}$ soit solution du problème (1) est

$$\begin{aligned} G_k(\hat{\boldsymbol{\beta}}) &= -\lambda \operatorname{sign}(\hat{\boldsymbol{\beta}}_k) \quad si \quad \hat{\boldsymbol{\beta}}_k \neq 0 \\ |G_j(\hat{\boldsymbol{\beta}})| &\leqslant \lambda \quad si \quad \hat{\boldsymbol{\beta}}_j = 0 \end{aligned}$$

Cette caractérisation nous sera utile pour établir les propriétés de sélection du Lasso.

Lasso et soft-thresholding

Elle nous permet également de déduire le résultat suivant.

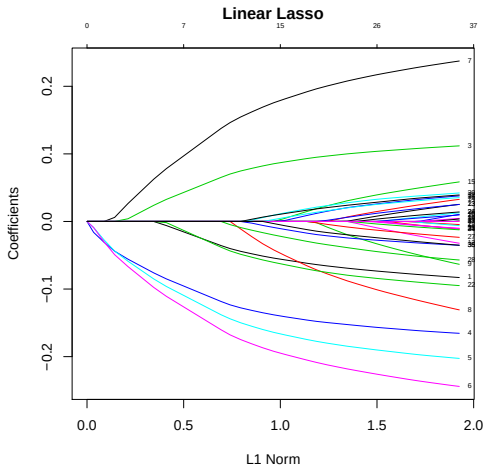
- Si $\mathbf{X}^T\mathbf{X}/n = \mathbf{I}_p$ ($\Rightarrow p \leqslant n$), alors Lasso = soft-thresholding :

$$\hat{\boldsymbol{\beta}}_j(\lambda) = \operatorname{sign}(\tilde{\boldsymbol{\beta}}_j)(|\tilde{\boldsymbol{\beta}}_j| - \lambda)_+.$$

$\Rightarrow$ Rq : Le Lasso sélectionne.. mais shrinke aussi : les estimateurs sont généralement biaisés (cf. regularization path).
Diverses extensions pour débiaiser les estimateurs Lasso

- Adaptive Lasso (Zou): $\|\boldsymbol{\beta}\|_1 \Rightarrow \sum_j |\beta_j|/|\hat{\beta}_j^{\text{init}}|$: on pénalise plus β_j si $|\hat{\beta}_j^{\text{init}}|$ est petit.
- Lasso-OLS Hybrid (= Relaxed Lasso de Meinshausen, avec $\phi = 0$)
- On reviendra sur ces approches plus tard.

Regularization path



Une seconde condition d'optimalité

Lemme 2.2

Une autre CNS pour que $\hat{\boldsymbol{\beta}}$ soit solution de (1) est que pour tout $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{2n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}\|_1$$

En particulier, on a la CN suivante :

$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{2n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}^*\|_1$$

Cette caractérisation nous sera utile pour étudier les propriétés d'estimation et d'erreur de prédiction du Lasso.

"Sparsistency" du Lasso

- Une procédure de sélection de variables est dite consistante en sélection de variables, ou "sparsistent", ssi le support $\widehat{J}_n$ du vecteur estimé est identique au support du vecteur théorique, J_0, avec probabilité tendant vers 1 :

$$\mathbb{P}(\widehat{J}_n = J_0) \to 1 \quad \text{lorsque } n \to \infty.$$

- Intuitivement, 2 types d'hypothèses sont nécessaires pour la sparsistency.
  - conditions d'identifiabilité
  - beta-min conditions

Hypothèses liées à l'identifiabilité

On supposera qu'il existe un paramètre de non-représentabilité $\gamma \in (0, 1]$ et une constante $C_{\min} > 0$ tels que

$$\max_{j \in J_0^c} \|X_j^T \mathbf{X}_{J_0} (\mathbf{X}_{J_0}^T \mathbf{X}_{J_0})^{-1}\|_1 \leqslant (1 - \gamma) \tag{2}$$

$$\Lambda_{\min}\left(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n}\right) \geqslant C_{\min} \tag{3}$$

Remarque : On ne peut pas savoir si ces conditions sont vérifiées a priori puisque J_0 est inconnu.

Interprétations de ces hypothèses

- Hypothèse de valeur propre minimale (3) : identifiabilité du problème restreint à J_0.

- Condition de non-représentabilité (2) :
  - pour tout $j \in J_0^c$, $\|X_j^T \mathbf{X}_{J_0}(\mathbf{X}_{J_0}^T \mathbf{X}_{J_0})^{-1}\|_1$: norme ℓ_1 du paramètre de la régression linéaire de X_j sur $\mathbf{X}_{J_0}$, estimé par MCO.
  - un design idéal est tel que X_j est orthogonal aux colonnes de la matrice $\mathbf{X}_{J_0}$, auquel cas on aurait $\gamma = 1$.
  - En grande dimension, on ne peut pas avoir cette orthogonalité stricte, mais on peut espérer être dans une situation de "quasi-orthogonalité".

Liens avec d'autres hypothèses "classiques"

- Dans ce cours, on se contentera de présenter des résultats obtenus sous la condition de non-représentabilité.
- On peut aussi travailler sous les hypothèses suivantes
  - hypothèse d'incohérence mutuelle : le paramètre d'incohérence de la matrice de design est "petit":

$$\iota^{(1)}(\mathbf{X}) = \max_{j \neq k} \left| \frac{X_j^T X_k}{n} \right| \leqslant \nu.$$

  - RIP (restricted isometry property) : la constante d'isométrie restreinte est "petite"

$$\iota_s^{(2)}(\mathbf{X}) = \inf \left\{ \epsilon : \forall S : |S| \leqslant s, \left\| \frac{\mathbf{X}_S^T \mathbf{X}_S}{n} - \mathbf{I}_{s \times s} \right\|_2 \leqslant \epsilon \right\}.$$

Résultat principal

Soit $\Pi_{\mathbf{X}_{J_0}^\perp} := \mathbf{I}_{n \times n} - \mathbf{X}_{J_0}(\mathbf{X}_{J_0}^T \mathbf{X}_{J_0})^{-1} \mathbf{X}_{J_0}^T$.

Théorème 3.1

Sous les hypothèses (2) et (3) précédentes, l'estimateur Lasso vérifie, pour $\lambda \geqslant (2/\gamma)\|\mathbf{X}_{J_0^c}^T \Pi_{\mathbf{X}_{J_0}^\perp} \xi / n\|_\infty$,

1. Unicité : *Le Lasso (1) a une solution unique* $\hat{\boldsymbol{\beta}}$.
2. Absence de "faux positif" : $\widehat{J} \subseteq J_0$.
3. Borne sur la norme ℓ_∞ : $\|\hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^*_{J_0}\|_\infty \leqslant B(\lambda, \mathbf{X})$ *avec*

$$B(\lambda, \mathbf{X}) = \left\| \left(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \right)^{-1} \mathbf{X}_{J_0}^T \left(\frac{\xi}{n} \right) \right\|_\infty + \lambda \left\| \left(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \right)^{-1} \right\|_\infty$$

4. Absence de "faux négatif" : *le Lasso est sparsistent si* $\min_{k \in J_0} |\beta_k^*| > B(\lambda, \mathbf{X})$.

Corollaire 3.1

On suppose que les $\xi_i \sim \mathcal{N}(0, \sigma^2)$, *et que la matrice de design* $\mathbf{X}$ *est déterministe, vérifie les conditions (2) et (3), et a ses colonnes normalisées, telles que* $n^{-1/2} \max_{j=1,\dots,p} \|X_j\|_2 \leqslant C$, *pour une constante* $C > 0$. *Pour le choix*

$$\lambda = \frac{2C\sigma}{\gamma} \sqrt{\frac{2\log(p - s_0) + \delta^2}{n}},$$

pour une constante $\delta > 0$, *on a le résultat suivant, avec probabilité supérieure à* $1 - 2e^{-\delta^2/2} - 2e^{-\varepsilon^2/2}$: *pour tout* $\varepsilon > 0$, *la solution optimale* $\hat{\boldsymbol{\beta}}$ *est unique, de support* $\widehat{J} \subseteq J_0$ *et telle que*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leqslant \frac{\sigma}{\sqrt{C_{\min}}} \sqrt{\frac{2\log s_0 + \varepsilon^2}{n}} + \frac{\lambda \sqrt{s_0}}{C_{\min}}.$$

Preuve

On doit premièrement montrer que ce choix de λ vérifie, avec grande probabilité, la condition sur le λ du Théorème 3.1. Soit, pour tout $j \in J_0^c$, $V_j = X_j^T \Pi_{\mathbf{X}_{J_0}^\perp} \xi / n$. Ces variables aléatoires sont gaussiennes, centrées, et de variance bornée par

$$\sigma^2 \|\Pi_{\mathbf{X}_{J_0}^\perp} X_j/n\|_2^2 \leqslant \sigma^2 \|X_j/n\|_2^2 \leqslant \sigma^2 C^2/n.$$

On en déduit que[2]

$$\mathbb{P}(\max_{j \in J_0^c} |V_j| \geqslant t) \leqslant 2(p-s_0) e^{-nt^2/(2C^2\sigma^2)}$$

et donc que

$$\mathbb{P}\Big(\Big\| \mathbf{X}_{J_0^c}^T \Pi_{\mathbf{X}_{J_0}^\perp} \frac{\xi}{n} \Big\|_\infty \geqslant C\sigma \sqrt{\frac{2\log(p-s_0)+\delta^2}{n}} \Big) \leqslant 2e^{-\delta^2/2}$$

[2] par l'Union Bound et l'inégalité de concentration pour variable gaussienne : $X \sim \mathcal{N}(\mu, \sigma^2) : \mathbb{P}(|X-\mu| > t) \leqslant 2\exp(-t^2/(2\sigma^2))$

Preuve (suite)

Soit $\widetilde{V}_k = e_k^T \Big(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \Big)^{-1} \frac{\mathbf{X}_{J_0}^T \xi}{n}$, pour tout $k \in J_0$. On montre facilement que les $\widetilde{V}_k$ sont gaussiennes, centrées de variance bornée par

$$\frac{\sigma^2}{n} \Big\| \Big(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \Big)^{-1} \Big\|_2 \leqslant \frac{\sigma^2}{C_{\min} n}.$$

En procédant comme précédemment, il vient donc

$$\mathbb{P}\Big(\max_{k=1,\dots,s_0} |\widetilde{V}_k| \geqslant \frac{\sigma}{\sqrt{C_{\min}}} \Big\{ \sqrt{\frac{2\log s_0 + \varepsilon^2}{n}} \Big\} \Big) \leqslant 2e^{-\varepsilon^2/2}.$$

Comme enfin

$$\Big\| \Big(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \Big)^{-1} \Big\|_\infty \leqslant \sqrt{s_0} \Big\| \Big(\frac{\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}}{n} \Big)^{-1} \Big\|_2 \leqslant \frac{\sqrt{s_0}}{C_{\min}},$$

le résultat du Lemme est donc vérifié avec probabilité supérieure à $1 - 2e^{-\delta^2/2} - 2e^{-\varepsilon^2/2}$.

Corollaire 3.2

On suppose que la matrice de design $\mathbf{X}$ vérifie les hypothèses du Théorème 3.1, que $p = \mathcal{O}(\exp(n^{\delta_3}))$, $s_0 = \mathcal{O}(n^{\delta_1})$, et que $\beta_{\min}^2 > n^{-(1-\delta_2)}$ avec

$$0 < \delta_1 + \delta_3 < \delta_2 < 1.$$

Si $\lambda_n = n^{-(1-\delta_4)/2}$ pour un $\delta_4 \in (\delta_3, \delta_2 - \delta_1)$, alors il existe une constante $c_1 > 0$ telle que $\mathbb{P}(\hat{J} = J_0) \geqslant 1 - \exp(-c_1 n^{\delta_4})$.

- p peut croître exponentiellement avec n
- $s_0/p \approx n^{\delta_1} \exp(-n^{\delta_3})$ décroît exponentiellement avec n.
- Si p (et s_0) fixe, $\lambda = n^{-(1-\delta)/2}$ et $\beta_{\min} \geqslant c_2 n^{-(1-\delta)/2}$ (pour $\delta > 0$) assurent la sparsistency avec probabilité $\geqslant 1 - 2\exp(-c_1 n\delta)$, pour une constante $c_1 > 0$.

Une condition suffisante d'échec pour le Lasso

Théorème 3.2

On suppose que la condition sur la valeur propre minimale (3) est vérifiée et que le vecteur de bruit ξ a une distribution symétrique autour de 0.

1. *Si la condition de non-représentabilité (2) n'est pas vérifiée, en particulier si*

$$\max_{j \in J_0^c} |X_j^T \mathbf{X}_{J_0} (\mathbf{X}_{J_0}^T \mathbf{X}_{J_0})^{-1} \mathrm{sign}(\boldsymbol{\beta}_{J_0}^*)| = 1 + \nu > 1, \qquad (4)$$

alors pour tout $\lambda_n > 0$ et n

$$\mathbb{P}[\mathrm{sign}(\hat{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta}^*)] \leqslant 1/2.$$

Lemme 2.1 "étendu"

Lemme 3.1

1. *Un vecteur* $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ *est optimal ssi* $\exists \hat{\mathbf{z}} \in \partial \|\hat{\boldsymbol{\beta}}\|_1$ *tel que*

$$\frac{\mathbf{X}^T\mathbf{X}}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \frac{\mathbf{X}^T\boldsymbol{\xi}}{n} + \lambda\hat{\mathbf{z}} = 0 \qquad (5)$$

2. *Pour tout* $j \in \widehat{J}^c$, *si* $|\hat{z}_j| < 1$ *alors toute solution optimale* $\bar{\boldsymbol{\beta}}$ *du Lasso est telle que* $\bar{\beta}_j = 0$.

Preuve : D'après la règle de Fermat (cf. cours de N.P.), $\hat{\boldsymbol{\beta}} \in \text{Argmin}_{\boldsymbol{\beta}} (2n)^{-1}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ est équivalent à : $\exists \hat{\mathbf{z}} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ tel que $\frac{1}{n}\mathbf{X}^T(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Y}) + \lambda\hat{\mathbf{z}} = 0$. La partie (1) du Lemme découle ensuite de l'égalité $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$. On en déduit également le résultat du Lemme 2.1 puisque $G_j(\hat{\boldsymbol{\beta}}) = -\lambda\hat{z}_j$.

Preuve (suite)

- Pour le point (2), raisonnons par l'absurde.
- Soit $\acute{\boldsymbol{\beta}}$ une autre solution du problème Lasso (1) et $j \in \widehat{J}^c$ tel que $|\hat{z}_j| < 1$ et $\acute{\beta}_j \neq 0$.
- Puisque le problème Lasso est convexe, l'ensemble de ses solutions est convexe et donc, pour tout $\rho \in [0,1]$,

$$\hat{\boldsymbol{\beta}}_\rho = (1-\rho)\hat{\boldsymbol{\beta}} + \rho\acute{\boldsymbol{\beta}}$$

  est également solution du Lasso.
- Pour tout $\rho \in (0,1]$, on a par ailleurs $\hat{\beta}_{\rho,j} \neq 0$ (par construction), et donc, d'après le résultat du Lemme 2.1, $|G_j(\hat{\boldsymbol{\beta}}_\rho)| = \lambda$.
- En définissant la fonction $f(\rho) = |G_j(\hat{\boldsymbol{\beta}}_\rho)|$, on a donc $f(0) < \lambda$ et $f(\rho) = \lambda$, pour tout $\rho \in (0,1]$.
- Ceci est en contradiction avec la continuité de la fonction f.

Primal-Dual Witness construction

On va chercher à construire une paire $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{z}}) \in \mathbb{R}^p \times \mathbb{R}^p$ de la manière suivante :

1. Soit $\hat{\boldsymbol{\beta}}_{J_0^c} = \mathbf{0}_{p-s_0}$.
2. Soit $(\hat{\boldsymbol{\beta}}_{J_0}, \hat{\mathbf{z}}_{J_0})$, avec $\hat{\boldsymbol{\beta}}_{J_0} \in \mathbb{R}^{s_0}$ une solution du problème Lasso *oraculaire*:

$$\hat{\boldsymbol{\beta}}_{J_0} \in \text{Arg}\min_{\boldsymbol{\beta}_{J_0} \in \mathbb{R}_{s_0}} \left\{ \frac{\|\mathbf{Y} - \mathbf{X}_{J_0}\boldsymbol{\beta}_{J_0}\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}_{J_0}\|_1 \right\}$$

   et $\hat{\mathbf{z}}_{J_0} \in \partial\|\hat{\boldsymbol{\beta}}_{J_0}\|_1$ tel que $G_{J_0}(\hat{\boldsymbol{\beta}}_{J_0}) + \lambda\hat{\mathbf{z}}_{J_0} = 0$.
3. On résout, en $\hat{\mathbf{z}}_{J_0^c} \in \mathbb{R}^{p-s_0}$, l'équation (5), et on vérifie la condition de faisabilité stricte, $|\hat{z}_j| < 1$, pour tout $j \in J_0^c$

PDW et sparsistency du Lasso

- $\neq$ méthode de résolution numérique pour le Lasso !

Lemme 3.2

Si la construction PDW aboutit, alors sous l'hypothèse de valeur propre minimale (3), le vecteur $(\hat{\boldsymbol{\beta}}_{J_0}, \mathbf{0})$ *est l'unique solution du Lasso.*

Preuve : Sous la condition de faisabilité stricte, le Lemme 3.1 assure que toute solution du Lasso $\bar{\boldsymbol{\beta}}$ est telle que $\bar{\beta}_j = 0$ pour tout $j \in J_0^c$. Toute solution est donc de la forme $(\bar{\boldsymbol{\beta}}_{J_0}, \mathbf{0}_{p-s_0})$, et on peut donc obtenir $\bar{\boldsymbol{\beta}}_{J_0}$ en résolvant le Lasso oraculaire. D'autre part, sous l'hypothèse (3), le Lasso oraculaire est strictement convexe, et admet donc une solution unique.

Preuve du Théorème 3.1

- Pour établir les points (1) et (2) du théorème, au vu du Lemme 3.2, il suffit de montrer que la faisabilité stricte est vérifiée dans le PDW.
- En réécrivant (5) par bloc, on a $\hat{\boldsymbol{\beta}}_{J_0}, \hat{\mathbf{z}}_{J_0}$ et $\hat{\mathbf{z}}_{J_0^c}$ qui vérifient:

$$\frac{1}{n}\begin{bmatrix} \mathbf{X}_{J_0}^T\mathbf{X}_{J_0} & \mathbf{X}_{J_0}^T\mathbf{X}_{J_0^c} \\ \mathbf{X}_{J_0^c}^T\mathbf{X}_{J_0} & \mathbf{X}_{J_0^c}^T\mathbf{X}_{J_0^c} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^*_{J_0} \\ \mathbf{0} \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathbf{X}_{J_0}^T\boldsymbol{\xi} \\ \mathbf{X}_{J_0^c}^T\boldsymbol{\xi} \end{bmatrix} + \lambda\begin{bmatrix} \hat{\mathbf{z}}_{J_0} \\ \hat{\mathbf{z}}_{J_0^c} \end{bmatrix} = \mathbf{0}_p.$$

- On a donc

$$\hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^*_{J_0} = \Big(\frac{\mathbf{X}_{J_0}^T\mathbf{X}_{J_0}}{n}\Big)^{-1}\Big[\frac{\mathbf{X}_{J_0}^T\boldsymbol{\xi}}{n} - \lambda\hat{\mathbf{z}}_{J_0}\Big] \qquad (6)$$

et

$$\begin{aligned}\hat{\mathbf{z}}_{J_0^c} &= \mathbf{X}_{J_0^c}^T\mathbf{X}_{J_0}(\mathbf{X}_{J_0}^T\mathbf{X}_{J_0})^{-1}\hat{\mathbf{z}}_{J_0} + \mathbf{X}_{J_0^c}^T\Pi_{\mathbf{X}_{J_0}^\perp}\Big(\frac{\boldsymbol{\xi}}{n\lambda}\Big) \\ &=: \mu + V\end{aligned}$$

Preuve du Théorème 3.1 (suite)

- Sous la condition de non-représentabilité, on a $\|\mu\|_\infty \leqslant (1-\gamma)$.
- La condition sur λ assure pour sa part que $\|V\|_\infty \leqslant \gamma/2$.
- La faisabilité stricte suit facilement en utilisant

$$\|\hat{\mathbf{z}}_{J_0^c}\|_\infty \leqslant \|\mu\|_\infty + \|V\|_\infty \leqslant (1-\gamma/2) < 1.$$

- Pour le point (3) du théorème, il vient de (6) que

$$\|\hat{\boldsymbol{\beta}}_{J_0} - \boldsymbol{\beta}^*_{J_0}\|_\infty \leqslant \Big\|\Big(\frac{\mathbf{X}_{J_0}^T\mathbf{X}_{J_0}}{n}\Big)^{-1}\frac{\mathbf{X}_{J_0}^T\boldsymbol{\xi}}{n}\Big\|_\infty + \lambda\Big\|\Big(\frac{\mathbf{X}_{J_0}^T\mathbf{X}_{J_0}}{n}\Big)^{-1}\Big\|_\infty.$$

- Le point (4) du théorème est directe.

Restricted Eigenvalue condition

Définition 4.1

Soit, pour tout $\alpha > 0$, le "cône" $\mathbb{C}_\alpha(J_0)$ de $\mathbb{R}^p$ défini par

$$\mathbb{C}_\alpha(J_0) = \{\Delta \in \mathbb{R}^p : \|\Delta_{J_0^c}\|_1 \leqslant \alpha\|\Delta_{J_0}\|_1\}$$

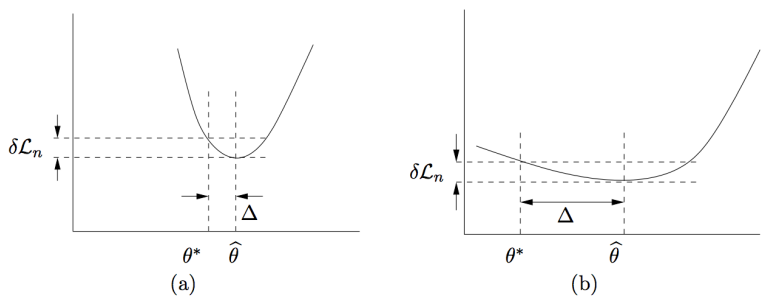
Définition 4.2

La matrice de design $\mathbf{X}$ vérifie la Restricted Eigenvalue condition sur J_0, avec les paramètres (κ, α), avec $\kappa > 0$, si

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geqslant \kappa\|\Delta\|_2^2 \quad \text{pour tout } \Delta \in \mathbb{C}_\alpha(J_0).$$

Intuition pour la RE

- Considérons la version contrainte avec $T = \|\boldsymbol{\beta}^*\|_1$.
- Soit $\mathcal{L}_n(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n)$.
- Si $n \to \infty$, on peut espérer $\mathcal{L}_n(\hat{\boldsymbol{\beta}}) \approx \mathcal{L}_n(\boldsymbol{\beta}^*)$.
- Sous quelles conditions cela implique-t-il que $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^*$?



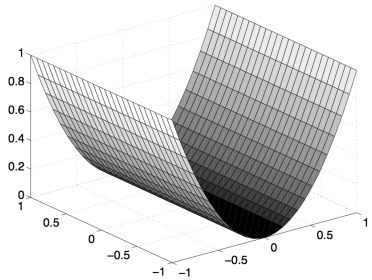
D'après Wainwright (unpublished)

Intuition pour la RE (suite)

- En multivarié, la courbure est liée au Hessien de $\mathcal{L}_n$: $(\mathbf{X}^T\mathbf{X})/n$: si cette matrice est définie positive, i.e.,

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geqslant \kappa\|\Delta\|_2^2 > 0 \quad \text{pour tout } \Delta \in \mathbb{R}^p \setminus \{\mathbf{0}\},$$

  alors $\mathcal{L}_n$ a une courbure élevée dans toutes les directions.
- Impossible en grande dimension ($p \geqslant n$) : il y a forcément au moins $p - n$ directions selon lesquelles $\mathcal{L}_n$ est "plate".



(a)

Résultat principal

Théorème 4.1

On suppose que $\mathbf{X}$ vérifie la condition RE sur J_0 avec les paramètres $(\kappa, 3)$. Alors toute solution du Lasso avec $\lambda \geqslant 2\|\mathbf{X}^T\boldsymbol{\xi}/n\|_\infty$ est telle que $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leqslant 3\lambda\sqrt{s_0}/\kappa$.

Corollaire 4.1

Supposons que les conditions du Théorème 4.1 et les hypothèses de normalité des résidus $\xi_i \sim \mathcal{N}(0, \sigma^2)$ et de standardisation des variables, $n^{-1/2}\max_{j=1,\dots,p}\|X_j\|_2 \leqslant C$ (pour une constante $C > 0$) sont vérifiées. Alors, pour le choix

$$\lambda = 2C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}$$

le résultat du Théorème 4.1 est vérifié avec probabilité supérieure à $1 - 2e^{-\delta^2/2}$.

Preuve du corollaire

Sous les conditions du corollaire, la quantité $\|\mathbf{X}^T\boldsymbol{\xi}/n\|_\infty$ correspond au maximum de la valeur absolue de p variables gaussiennes, centrées et de variance bornée par $C^2\sigma^2/n$.
En procédant comme précédemment, on obtient alors que pour tout $\delta > 0$,

$$\mathbb{P}\left(\|\mathbf{X}^T\boldsymbol{\xi}/n\|_\infty \geqslant C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}\right) \leqslant 2e^{-\delta^2/2}.$$

Le résultat du Théorème 4.1 permet donc de conclure qu'avec une probabilité supérieure à $1 - 2e^{-\delta^2/2}$, on a

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leqslant \frac{6C\sigma}{\kappa}\sqrt{\frac{2s_0\log p + s_0\delta^2}{n}}.$$

Preuve du Théorème 4.1

1. Si $\lambda \geqslant 2\|\mathbf{X}^T\boldsymbol{\xi}/n\|_\infty$, alors $\widehat{\Delta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \in \mathbb{C}_3(J_0)$
   Puisque $\hat{\boldsymbol{\beta}}$ est optimal, on a les propriétés suivantes

$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{2n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}^*\|_1$$
$$\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2}{2n} + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leqslant \frac{\|\boldsymbol{\xi}\|_2^2}{2n} + \lambda\|\boldsymbol{\beta}^*\|_1$$

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} + 2\lambda(\|\boldsymbol{\beta}^*_{J_0} + \widehat{\Delta}_{J_0}\|_1 + \|\widehat{\Delta}_{J_0^c}\|_1) \leqslant 2\frac{\boldsymbol{\xi}^T\mathbf{X}\widehat{\Delta}}{n} + 2\lambda\|\boldsymbol{\beta}^*_{J_0}\|_1$$
$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} + 2\lambda\|\widehat{\Delta}_{J_0^c}\|_1 \leqslant 2\|\widehat{\Delta}\|_1\left\|\frac{\boldsymbol{\xi}^T\mathbf{X}}{n}\right\|_\infty + 2\lambda\|\widehat{\Delta}_{J_0}\|_1$$
$$0 \leqslant \frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leqslant \lambda\{3\|\widehat{\Delta}_{J_0}\|_1 - \|\widehat{\Delta}_{J_0^c}\|_1\}.$$

2. On conclut la preuve du Théorème 4.1 en appliquant la RE pour obtenir $\kappa\|\widehat{\Delta}\|_2^2 \leqslant 3\lambda\|\widehat{\Delta}_{J_0}\|_1 \leqslant 3\lambda\sqrt{s_0}\|\widehat{\Delta}\|_2$.

Résultat principal

Théorème 5.1

Soit $\hat{\boldsymbol{\beta}}$ une solution optimale du problème Lasso (1) avec le choix $\lambda \geqslant 2\|\mathbf{X}^T\boldsymbol{\xi}/n\|_\infty$.

1. *On a toujours la vitesse lente suivante:*
$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leqslant 12\|\boldsymbol{\beta}^*\|_1\lambda.$$
2. *Si le support de $\boldsymbol{\beta}^*$, J_0, est tel que $|J_0| = s_0$ et que la matrice de design $\mathbf{X}$ vérifie la condition RE avec paramètres $(\kappa, 3)$ sur J_0, on a alors la vitesse rapide suivante*
$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leqslant \frac{9}{\kappa}s_0\lambda^2.$$

Corollaire

En procédant comme précédemment, on montre que si les $\xi_j \sim \mathcal{N}(0, \sigma^2)$, et sous l'hypothèse de variables normalisées, $n^{-1/2}\max_j \|X_j\| \leqslant C$, alors le choix $\lambda = 2C\sigma\sqrt{(2\log p + \delta^2)/n}$ est "valide" avec probabilité supérieure à $1 - \exp(-\delta^2/2)$, et alors :

1. la partie (1) du Théorème implique que
$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leqslant 24\|\boldsymbol{\beta}^*\|_1 C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}.$$
   - sous la seule contrainte $\|\boldsymbol{\beta}^*\|_1 \leqslant T$, cette borne ne peut pas être améliorée.
2. sous les hypothèses de sparsité et RE, alors on obtient la borne
$$\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{n} \leqslant \left(\frac{36C^2\sigma^2}{\kappa}\right)\frac{2s_0\log p + s_0\delta^2}{n}.$$

Preuve du point (2) du Théorème 5.1

En procédant comme dans la preuve du Théorème 4.1, il vient

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n} \leqslant 3\lambda\|\widehat{\Delta}_{J_0}\|_1 \leqslant 3\lambda\sqrt{s_0}\|\widehat{\Delta}_{J_0}\|_2.$$

D'autre part, comme on a toujours $\widehat{\Delta} \in \mathbb{C}_3(J_0)$, on peut appliquer une nouvelle fois la condition RE$-(\kappa, 3)$:

$$\|\widehat{\Delta}\|_2^2 \leqslant \frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{n\kappa},$$

ce qui, combiné à l'inégalité précédente, conduit à

$$\frac{\|\mathbf{X}\widehat{\Delta}\|_2}{\sqrt{n}} \leqslant \frac{3\lambda\sqrt{s_0}}{\sqrt{\kappa}}$$

Preuve du point (1) du Théorème 5.1

1. Montrons que $\|\widehat{\Delta}\|_1 \leqslant 4\|\boldsymbol{\beta}^*\|_1$.

   En procédant comme précédemment on obtient aisément
$$0 \leqslant \frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{2n} \leqslant \frac{\boldsymbol{\xi}^T\mathbf{X}\widehat{\Delta}}{n} + \lambda\{\|\boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}\|_1\}. \qquad (7)$$
   D'après l'inégalité d'Hölder, le choix de λ, puis l'inégalité triangulaire, il vient
$$\left|\frac{\boldsymbol{\xi}^T\mathbf{X}\widehat{\Delta}}{n}\right| \leqslant \left\|\frac{\mathbf{X}^T\boldsymbol{\xi}}{n}\right\|_\infty\|\widehat{\Delta}\|_1 \leqslant \frac{\lambda}{2}(\|\boldsymbol{\beta}^*\|_1 + \|\hat{\boldsymbol{\beta}}\|_1). \qquad (8)$$
   En combinant ces deux inégalités, il vient $\|\hat{\boldsymbol{\beta}}\|_1 \leqslant 3\|\boldsymbol{\beta}^*\|_1$, et donc, via l'inégalité triangulaire,
$$\|\widehat{\Delta}\|_1 \leqslant \|\hat{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\beta}^*\|_1 \leqslant 4\|\boldsymbol{\beta}^*\|_1.$$

Preuve du point (1) du Théorème 5.1 (suite)

2. En combinant les résultats obtenus dans les dérivations des équations (7) et (8), il vient également

$$\begin{aligned}\frac{\|\mathbf{X}\widehat{\Delta}\|_2^2}{2n} &\leqslant \frac{\lambda}{2}\|\widehat{\Delta}\|_1 + \lambda\{\|\boldsymbol{\beta}^*\|_1 - \|\boldsymbol{\beta}^* + \widehat{\Delta}\|_1\} \\ &\leqslant \frac{3\lambda}{2}\|\widehat{\Delta}\|_1 \\ &\leqslant 6\lambda\|\boldsymbol{\beta}^*\|_1\end{aligned}$$

où la 2ème ligne vient de l'inégalité triangulaire

$$\|\boldsymbol{\beta}^* + \widehat{\Delta}\|_1 \geqslant \|\boldsymbol{\beta}^*\|_1 - \|\widehat{\Delta}\|_1,$$

et la 3ème ligne de la borne $\|\widehat{\Delta}\|_1 \leqslant 4\|\boldsymbol{\beta}^*\|_1$.

Sur les conditions RE, MI, etc.

Table : Récapitulatif des liens "Résultats-conditions"

Propriétés	conditions sur $\mathbf{X}$	beta-min
Prédiction Vitesse Lente	"Rien"	Non
Prédiction Vitesse Rapide	RE (pas nécess.)	Non
$J_0 \subseteq \widehat{J}$	RE	Oui
$J_0 = \widehat{J}$	Non-Repres.	Oui

- On peut montrer que RE n'est pas nécessaire pour vitesse rapide en prédiction (mais il faut quand même certaines hypothèses sur la matrice de design, contrairement à d'autres approches telles que ℓ_0)
- Pour sparsistency (et estimation), il faut des conditions, liées à l'identifiabilité.
- Cas de design aléatoire

Qq conditions proposées, et leurs inter-relations....

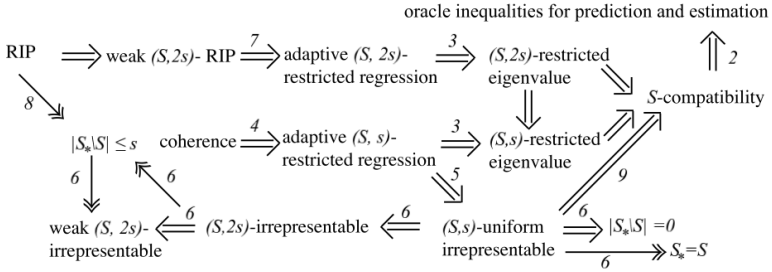


FIG 1. *A double arrow (⇒) indicates a straight implication, whereas the more fancy arrowheads mean that the relation is under side-conditions. The numbers indicate the section where the result is (re)proved.*

D'après van de Geer et Bühlmann, EJS, 2009.

Modèles Linéaires généralisés

- Pour ces modèles, la vraisemblance s'exprime généralement sous la forme $\sum_i \mathcal{L}(Y_i, \mathbf{X}_i^T\boldsymbol{\beta})$
  - Régression logistique : $\mathcal{L}(y, \eta) = y\eta - \log(1 + e^\eta)$
- Le Lasso se généralise alors :

$$\phi(\lambda) := \min_{\boldsymbol{\beta}\in\mathbb{R}^p} -\sum_{i=1}^n \mathcal{L}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1.$$

ou

$$\phi(\lambda) := \max_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \mathcal{L}(Y_i, \mathbf{x}_i^T\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1.$$

En pratique

- On travaille généralement sur variables "standardisées"
- Différentes approches ont été proposées, à partir du Lasso, pour l'améliorer
  - Adaptive Lasso
  - Thresholded Lasso
  - Relaxed Lasso
  - BoLasso
  - ...

Biblio

1. Wainwright. *Sharp thresholds for high-dimensional ...*, IEEE Trans. Inform. Theory, 2009.
2. Wainwright. Livre en cours d'écriture.
3. Lounici. *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*. EJS, 2008.
4. Bickel, Ritov, Tsybakov. *Simultaneous analysis of Lasso and Dantzig Selector*, AoS, 2009.
5. van de Geer et Bühlmann. *Statistics for high-dimensional data*, (Chap. 2, 6, 7), Springer, 2011.
6. van de Geer et Bühlmann. *On the conditions used to prove oracle results for the Lasso*, EJS, 2009.
7. Giraud. *High-dimensional statistics*, 2013. http://www.cmap.polytechnique.fr/~giraud/MSV/LectureNotes.pdf
8. Horn et Johnson. *Matrix Analysis, 2nd Ed.*, Cambridge Univ. Press, 2013.