Apprentissage Statistique en Grande Dimension

F. Panloup
LAREMA-Université d'Angers

—

Régression linéaire en grande dimension

—

October 12, 2018

Régression Linéaire Classique : Rappels

Modèle

On considère une variable à expliquer Y (supposée réelle pour l'instant) et p covariables (ou variables explicatives) $X_1, \dots, X_p$. L'objectif de la *régression linéaire* est de modéliser Y comme une combinaison linéaire bruitée de $X_1, \dots, X_p$ de la forme :

$$Y = \theta_0 + \sum_{i=1}^{p} \theta_i X_i + \varepsilon$$

où $\theta = (\theta_0, \theta_1, \dots, \theta_p)^T$ appartient à $\mathbb{R}^{p+1}$ et ε est une variable aléatoire centrée (bruit) indépendante de la variable aléatoire X. Quitte à considérer le vecteur $(1, X_1, \dots, X_p)$, on remarque que l'on peut mettre ce modèle sous la forme plus synthétique :

$$Y = X\theta + \varepsilon.$$

Pour un couple (X, Y) donné, le (un des) but(s) est alors de déterminer :

$$\theta^\star = \mathrm{Argmin}_{\theta \in \mathbb{R}^{p+1}} \mathbb{E}[(Y - X\theta)^2].$$

Remarque : Dans la suite, on note pour simplifier $\theta = (\theta_1, \dots, \theta_p)^T$ et $X = (X_1, \dots, X_p)$ (quitte à supposer que la première composante de X est égale à 1).

Régression et Apprentissage

Si l'on fait l'hypothèse que le "vrai" modèle est de la forme $Y = X\theta^\star + \varepsilon$ (ce que l'on fera dans la suite), alors le prédicteur de Bayes (pour la fonction de perte $\ell(y, y') = (y - y')^2$ est donné par

$$f_{\mathrm{Bayes}}(x) = \mathbb{E}[Y|X = x] = x\theta^\star.$$

Dans ce cas, si $\mathrm{Var}(\varepsilon) = \sigma^2$, alors

$$\inf_{f:\mathbb{R}^p \to \mathbb{R}} \mathcal{R}_f = \mathbb{E}[(Y - X\theta^\star)^2] = \mathbb{E}[\varepsilon^2] = \sigma^2.$$

Soit $\mathcal{D}_n = \{(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})\}$ un échantillon d'apprentissage (en particulier $i.i.d$). On suppose dans la suite que pour $k \in \{1, \ldots, n\}$,

$$Y^{(k)} = X^{(k)}\theta^\star + \varepsilon_k.$$

où $\theta^\star$ est une quantité "à apprendre". Sauf mention contraire, on fera aussi l'hypothèse que $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$.

Estimateur de $\theta^\star$

Pour estimer $\theta^\star$, on utilise l'estimateur

$$\hat{\theta} = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y^{(i)} - X^{(i)}\theta)^2 = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\theta\|^2$$

où $\|\,,\,\|$ désigne la norme euclidienne,

$$\mathbf{Y} = \begin{pmatrix} Y^{(1)} \\ \dots \\ Y^{(n)}) \end{pmatrix} \text{ et } \mathbf{X} = \begin{pmatrix} X_1^{(1)} & \dots & X_p^{(1)} \\ & \dots & \\ X_1^{(n)} & \dots & X_p^{(n)} \end{pmatrix}$$

Attention aux notations : X et Y désignent des variables aléatoire. $\mathbf{X}$ et $\mathbf{Y}$ désignent la matrice et la réponse associées à l'échantillon. Dans la suite, ε désigne le vecteur $(\varepsilon_1, \dots, \varepsilon_n)$.

Loi du vecteur $\mathbf{Y}$ et premières propriétés

Avec ces notations, on a donc l'égalité dans $\mathbb{R}^n$

$$\mathbf{Y} = \mathbf{X}\theta^\star + \varepsilon.$$

En particulier, comme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, on en déduit que conditionnellement à $\mathbf{X}$ (*i.e.* connaissant $\mathbf{X}$)

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\theta^\star, \sigma^2 I_n).$$

Dans la suite, on raisonnera toujours "conditionnellement à $\mathbf{X}$ " de sorte que $\mathbf{X}$ comme une matrice déterministe.

- Dans le modèle linéaire classique, on fait maintenant l'hypothèse que $(\mathbf{X}^T\mathbf{X})$ est inversible, ce qui revient à dire que $\mathrm{rg}(\mathbf{X}) = p$ et donc que $p \leq n$.
- Dans ce cas, si l'on note X_i la i-ème colonne de $\mathbf{X}$, $V_{\mathbf{X}} = \mathrm{Vect}(X_1, \ldots, X_p)$, alors la projection orthogonale sur $V_{\mathbf{X}}$ est donnée par : $\forall U \in \mathbb{R}^n$,

$$\mathrm{Proj}_{V_{\mathbf{X}}}(U) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T U.$$

Calcul de $\hat{\theta}$

Proposition

Si $\mathbf{X}^T\mathbf{X}$ *est inversible, on a :*

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

De plus,

$$\hat{\theta} \sim \mathcal{N}(\theta^\star, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

Pour n observations, le prédicteur est alors défini pour tout $x \in \mathbb{R}^p$ par :

$$\hat{f}_n(x) = x^T\hat{\theta} = \langle x, \hat{\theta} \rangle.$$

Proposition

Si $\mathbf{X}^T\mathbf{X}$ *est inversible, alors (théorème de Cochan),*

$$x^T\hat{\theta} \sim \mathcal{N}(x^T\theta^\star, \sigma^2 x^T(\mathbf{X}^T\mathbf{X})^{-1}x).$$

Propriétés de $\hat{Y}$

On note $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$. $\hat{Y}$ est donc le vecteur des prédictions relatives à l'échantillon d'apprentissage. D'après ce qui précède,

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathrm{Proj}_{V_\mathbf{x}}(Y).$$

On a :

Proposition

$$\|\hat{\mathbf{Y}} - \mathbf{X}\theta^\star\|^2 = \|\mathbf{X}(\hat{\theta} - \theta^\star)\|^2 \sim \sigma^2\chi^2_p$$

Par conséquent,

$$\mathbb{E}\Big[\frac{\|\hat{\mathbf{Y}} - \mathbf{X}\theta^\star\|^2}{n}\Big] = \sigma^2\frac{p}{n}.$$

Preuve : On a vu que

$$\hat{\theta} - \theta^\star \sim \mathcal{N}(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

de sorte que

$$\mathbf{X}(\hat{\theta} - \theta^\star) \sim \mathcal{N}(0, \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathcal{N}(0, \sigma^2 P_{V_\mathbf{x}}).$$

Le théorème de Cochran nous permet alors d'en déduire le résultat (On peut par exemple le vérifier facilement dans le cas où $\mathbf{X}^T\mathbf{X}$ est la matrice I_p).

Commentaires

Lorsque p augmente, on constate que l'erreur d'apprentissage ou plutôt l'erreur d'estimation du "signal non bruité" (sur l'échantillon d'apprentissage) se comporte mal avec p. Par ailleurs, ce qu'on a fait jusqu'ici ne prend pas en compte les phénomènes d'overfitting qui peuvent apparaître lorsque le modèle est trop flexible (puisqu'on n'a pas encore considéré d'échantillon test). Se posent alors plusieurs questions :

- Dans un cadre classique, comment estimer la qualité du modèle et si besoin comment peut-on sélectionner des variables ?
- Lorsque $p > n$ (et donc en particulier lorsque $p >> n$), ce qui précède n'a pas de sens car $\mathbf{X}^T\mathbf{X}$ ne peut être inversible car de rang inférieur ou égal à $\min(n, p)$.

Qualité du modèle linéaire classique

Dans le modèle linéaire classique, une quantité classique est le R^2 basée sur la décomposition suivante de la somme des carrés (SC) :

$$SCT = SCM + SCR$$

où

$$SCT = \sum_{i=1}^{n} (\mathbf{Y}_k - \bar{\mathbf{Y}}_n)^2, \quad SCM = \sum_{i=1}^{n} (\hat{\mathbf{Y}}_k - \bar{\mathbf{Y}}_n)^2, \quad SCR = \sum_{i=1}^{n} (\mathbf{Y}_k - \hat{\mathbf{Y}}_k)^2.$$

Remarque : Cette décomposition de la variance empirique s'écrit sous forme condensée de la manière suivante :

$$\|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2.$$

On remarque en particulier que le "double-produit" a disparu. Ceci est dû à la propriété suivante :

Coefficient de détermination

Proposition

$\bar{\mathbf{Y}} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mathbf{Y}}_k$. *En particulier,* $\bar{\mathbf{Y}} \in V_{\mathbf{X}}$ *et* $\hat{\mathbf{Y}} - \bar{\mathbf{Y}} \in V_{\mathbf{X}}$. *Par conséquent,*

$$\langle \hat{\mathbf{Y}} - \bar{\mathbf{Y}}, \mathbf{Y} - \hat{\mathbf{Y}} \rangle = 0.$$

Pour prouver ce résultat, il faut se rappeler que $\mathbf{1}$ appartient à $V_{\mathbf{X}}$... Le coefficient de détermination est alors défini par :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SST} \quad \text{(en anglais SCR=RSS).}$$

C'est en fait le rapport de carrés de deux longueurs : si l'on "oublie" la moyenne $\bar{\mathbf{Y}}$, c'est le cosinus entre $\mathbf{Y}$ et sa projection $\hat{\mathbf{Y}}$.
Remarque: De cette interprétation géométrique, on comprend facilement que $0 \leq R^2 \leq 1$ et que l'ajout de variables explicatives augmente le R^2. Plus R^2 est proche de 1 plus l'ajustement sur les données observées est bon.

Coefficient de détermination ajusté

Proposition

Soient $\mathcal{M}_1$ et $\mathcal{M}_2$ deux modèles de régression linéaire de la forme :

$$\mathcal{M}_1 : Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$

$$\mathcal{M}_2 : Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \beta_{p+1} X_{p+1} + \varepsilon.$$

Leurs coefficients de détermination satisfont $R_1^2 \leq R_2^2$.

Cette propriété signifie que le R^2 ne détecte pas l'overfitting. On lui préfère souvent le R^2 ajusté défini par :

$$R^2_{adj} = 1 - \frac{\frac{SCR}{n-p-1}}{\frac{SST}{n-1}}.$$

Coefficient de détermination ajusté(suite)

Le R^2 ajusté est en fait un coefficient de détermination normalisé en fonction des paramètres p et n. Pour rappel, si $\mathrm{Vect}(1, \mathbf{X}_1, \ldots, \mathbf{X}_p) = p+1$, alors, on obtient via le théorème de Cochran ($Y - \hat{Y} = P_{V_{\mathbf{x}}^\perp} Y = P_{V_{\mathbf{x}}^\perp} \varepsilon$),

$$\frac{SCR}{\sigma^2} \sim \chi^2(n-p-1).$$

Sous l'hypothèse $(H_0) := \theta = 0$, on a également que

$$\frac{SCM}{\sigma^2} \sim \chi^2(p) \quad \text{et} \quad \frac{SCT}{\sigma^2} \sim \chi^2(n-1)$$

tandis qu'en utilisant que $Y - \hat{Y} = P_{V_{\mathbf{x}}^\perp} Y = P_{V_{\mathbf{x}}^\perp} \varepsilon$, on déduit du théorème de Cochran que

$$\frac{SCR}{\sigma^2} \sim \chi^2(n-p-1).$$

A ce stade, on aborde la "sélection de modèle" au sens où l'on prend en compte la flexibilité dans la comparaison des modèles.

Autres mesures de la qualité du modèle

Le R^2 ajusté prend donc plus en compte la fiabilité du modèle et pose la question : vaut-il mieux un modèle complet moins fiable statistiquement qu'un modèle réduit biaisé mais d'estimation plus précise (question du compromis Biais-Variance) ? Il existe d'autres moyens de comparer les modèles (pénalisant la montée en dimension). On en présente ici succinctement quelques-uns (dans le cadre du modèle linéaire gaussien uniquement) :

- Le C_p de Mallows défini par

$$\frac{1}{n}(SCR + 2p\hat{\sigma}^2) \quad \text{que l'on cherche à minimiser.}$$

- Le critère AIC (Akaike Information Criterion), proportionnel au C_p dans le cas du modèle linéaire (uniquement)

$$AIC = \frac{1}{n\hat{\sigma}^2}(SCR + 2p\hat{\sigma}^2)$$

- Le critère BIC (Bayesian) :

$$BIC = \frac{1}{n}(SCR + \log(n)\hat{\sigma}^2)$$

  Là encore, on cherche à minimiser ce critère.

Et la validation ?? Et dans la pratique ??

- Validation : on a pu constater que dans ce chapitre les aspects "erreur test" et "erreur de validation". Les outils présentés plus hauts peuvent être vus comme des alternatives à la validation. Ces outils ont en particulier été développés pour pallier la difficulté computationnelle générée par la validation croisée. Néanmoins, aujourd'hui, ce problème est moins prégnant avec l'accélération des calculateurs.
- Pratique : qu'on utilise validation croisée, R^2_{adj}, C_p, AIC ou BIC, une question se pose, comment comparer l'ensemble des modèles que l'on peut fabriquer avec p variables ? En effet, il en a 2^p !!
- Réponse : Algorithmes de sélections de variables "Pas à Pas", "Mixtes", "Globaux" selon les situations.

Algorithmes de sélection de variables

Présentons quelques méthodes (qui possèdent toutes leurs variantes) :

- Forward : A chaque pas, une variable est ajoutée au modèle. C'est celle dont le R^2 est maximal (ou de manière équivalente celle dont le SCR est minimal). On obtient ainsi p modèles contenant 1, 2, 3, ..., p variables. On choisit ensuite celui qui minimise un critère pénalisé ($R^2_{adj}, \ldots$).
- Backward : A chaque pas, une variable est retirée au modèle : celle qui augmente le moins le SCR (ou qui diminue le plus moins le R^2.
- On a aussi des approches hybrides (qui mélangent ces deux approches).

LASSO/RIDGE/ELASTIC NET

La pénalisation par la norme de θ

Les méthodes présentées ci-dessus souffrent de deux problèmes :

- Lorsque p est grand, leur mise en oeuvre est coûteuse et l'exploration des modèles reste nécessairement très partielle.
- Lorsque $p > n$, elles ne sont a priori pas applicables puisque l'estimateur $\hat{\theta}$ n'est pas bien défini. En effet, comme $rg(X) < p$, $\mathrm{Ker}X$ est non réduit à 0. Ainsi,

  $$\mathrm{Argmin}_\theta \|\mathbf{Y} - \mathbf{X}\theta\|^2$$

  n'est pas unique.
- Une théorie assez récente permet de pallier le second problème en introduisant une pénalisation dans l'estimateur. Il permet aussi de donner une alternative à la sélection de modèles dans le cadre classique.

Principe général de la régression pénalisée

- Pour un $\lambda > 0$, on cherche à minimiser la fonction

$$\Phi_{\mathcal{P}}(\theta) = \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{n} + \lambda \mathcal{P}(\theta).$$

  $\mathcal{P}$ désigne la *pénalité*.

- Quelques exemples de pénalités :
  - $\mathcal{P}(\theta) = \|\theta\|_0 := \mathrm{Card}\{i, \theta_i \neq 0\}$.
  - $\mathcal{P}(\theta) = \|\theta\|_1 = \sum_{i=1}^{p} |\theta_i|$ (LASSO)
  - $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2} = \frac{1}{2}\sum_{i=1}^{p} |\theta_i|^2$ (Ridge).
  - $\mathcal{P}(\theta) = (1-\alpha)\frac{\|\theta\|_2^2}{2} + \alpha\|\theta\|_1$, $\alpha \in [0,1]$ (Elastic Net).

Rôle de la pénalisation

- On favorise les solutions pour lesquelles $\mathcal{P}(\theta)$ est petit.
- Ainsi, par exemple, en norme $\|\,,\,\|_0$, on pénalise le nombre de variables. On favorise les solutions qui ont peu de variables "allumées".
- **Remarque** : les C_p, *AIC* et *BIC* ressemblent à des pénalisations par cette norme sauf que la pénalisation est "uniforme".
- Interprétation Lagrangienne : la fonction $\Phi_{\mathcal{P}}(\theta)$ est à comprendre comme un *Lagrangien*, *i.e.* une fonction fabriquée pour *régulariser* le problème d'optimisation sous contrainte

$$\min_{\mathcal{P}(\theta)\leq s(\lambda)} \frac{\|\mathbf{Y}-\mathbf{X}\theta\|^2}{n}.$$

(voir plus loin pour quelques rappels)

Représentation graphique

Représentons ici le problème sous contrainte (en dimension 2) :

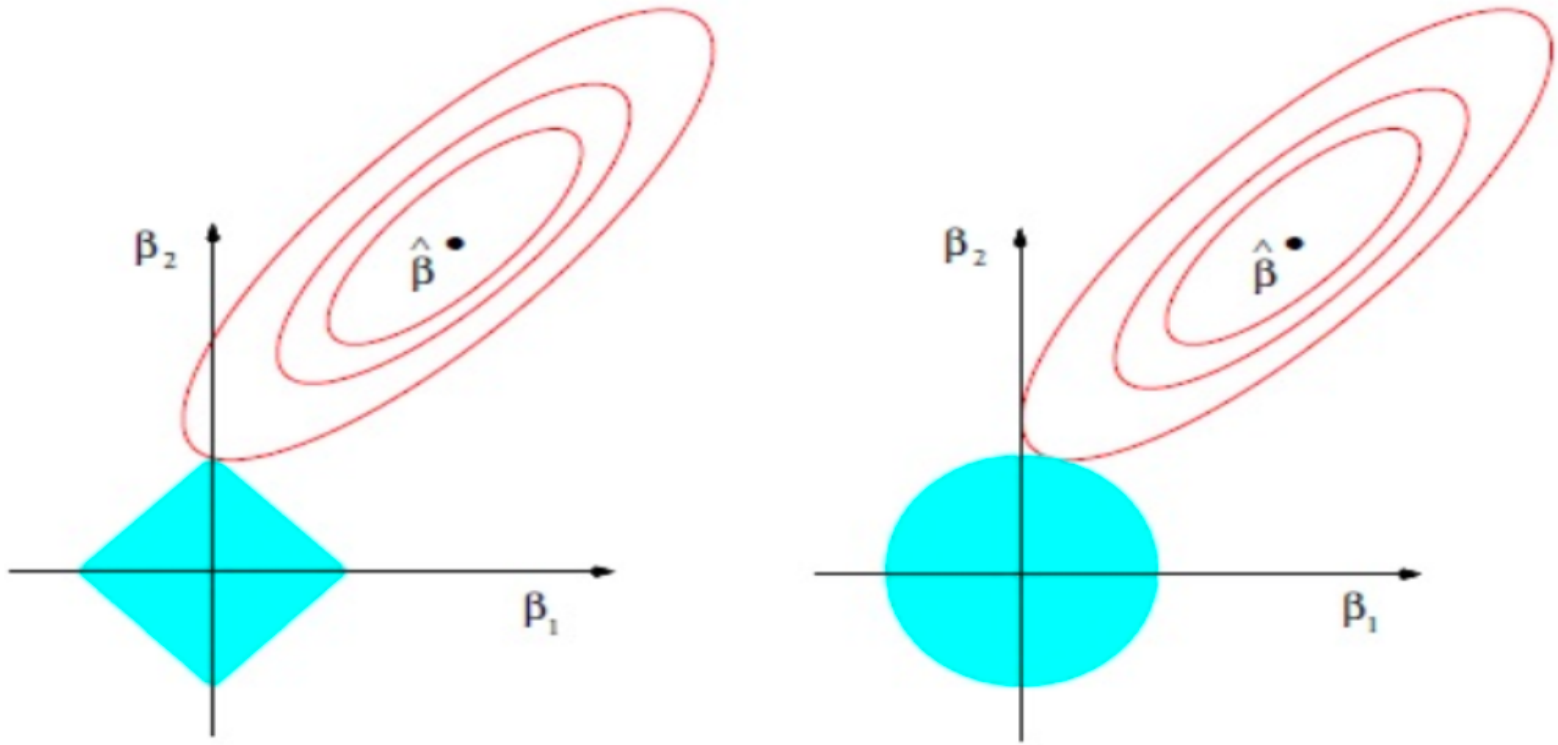

Figure 1: Gauche : en norme 1, Droite, en norme 2 ($\hat{\beta} = \mathrm{Argmin}_{\beta \in \mathbb{R}^2} \|Y - X\beta\|^2$, en rouge, les lignes de niveau de $\|Y - X\beta\|^2$ autour du minimum)

Rappels succincts sur le Lagrangien

On considère le problème suivant : minimiser la fonction J sur un sous-ensemble G défini par $G = \{\theta \in \mathbb{R}^p, f(\theta) = 0\}$ (f supposée régulière ici).

- Exemple : $J(\theta) = \|Y - X\theta\|^2$ et $f(\theta) = \|\theta\|_2^2 - R$.
- Dans ce cas, le théorème des multiplicateurs de Lagrange donne la condition nécessaire suivante :

Théorème

Si $\theta^\star$ est un point de G où J atteint son minimum, alors il existe $\lambda^\star$ tel que

$$\nabla J(\theta^\star) + \lambda^\star \nabla f(\theta^\star) = 0.$$

Ainsi, si l'on note L la fonction définie par

$$L(\lambda, \theta) = J(\theta) + \lambda f(\theta)$$

alors, $(\lambda^\star, \theta^\star)$ est solution de

$$\partial_\theta L(\lambda^\star, \theta^\star) = \partial_\lambda L(\lambda^\star, \theta^\star) = 0.$$

Rappels succincts sur le Lagrangien (suite)

Admettons maintenant que J et f sont convexes. Dans ce cas,

- $\mathcal{C} = \{\theta, f(\theta) \leq 0\}$ est convexe.
- Si $\min_{\theta \in \mathbb{R}^p} J(\theta)$ est en dehors de $\mathcal{C}$, on peut alors vérifier que $\min_{\theta \in \mathcal{C}} J(\theta)$ est nécessairement sur le bord du convexe.
- Les points critiques de $\theta \mapsto L(\lambda, \theta)$ sont des minimums.
- **Pour aller plus loin** :Pour rendre la condition nécessaire du théorème précédent suffisante, on s'appuie souvent sur les conditions "KKT".
- **Important :** Considérer la fonction $\theta \mapsto L(\lambda, \theta)$ permet de considérer un problème d'optimisation non contraint. Lorsque la fonction L est convexe, alors, on peut mettre en oeuvre des algorithmes d'optimisation (type descente de gradient) pour déterminer le minimum de L
- **En pratique** : on cherche le minimum $\theta^\star_\lambda$ de $\theta \mapsto L(\lambda, \theta)$ (à λ fixé) puis on fait varier λ en étudiant l'évolution de $J(\theta^\star)$.
- **Cas limites** :Lorsque $\lambda = 0$, on regarde le problème non pénalisé, lorsque $\lambda \to +\infty$, on cherche à déterminer la meilleure approximation sur une boule de rayon tendant vers 0.

A propos de la norme $\|\,,\,\|_0$

- Si l'on veut pénaliser le nombre de variables pour réduire la variance du modèle, le choix naturel est la norme $\|\,,\,\|_0$.
- **Parcimonie** : on parle dans ce cas d'hypothèse de parcimonie ("sparsity") : on fait l'hypothèse que peu de variables sont réellement explicatives. Si tel est le cas, *i.e.* que le vrai modèle est de la forme

$$Y = X\theta^\star + \varepsilon$$

avec $\|\theta^\star\|_0 = \{j, \theta^\star_j \neq 0\} = s_0$ petit devant p (et si possible devant n également).

- Notons $\mathbf{X}_{s_0}$ la sous-matrice de $\mathbf{X}$ de taille $n \times s_0$ constituée des colonnes "allumées" de $\mathbf{X}$. Si l'on suppose que $\mathbf{X}_{s_0}^T \mathbf{X}_{s_0}$ est inversible, alors, une fois "déterminé ce support", l'application du modèle linéaire standard permet d'obtenir une erreur en $\frac{s_0}{n}$.
- **Problème** : La norme $\|\,,\,\|_0$ n'est pas une fonction convexe !! Chercher le minimum de la fonction $\Phi_{\|\,\|_0}$ n'est pas un problème soluble numériquement car il faut explorer toutes les configurations de manière exhaustive (pour un support de taille s_0, on a s_0^p possibilités).

De la norme $\|\,,\,\|_0$ à la norme $\|\,,\,\|_1$

- **Idée** : Remplacer la norme $\|\,,\,\|_0$ par la norme convexe "minimale" : la norme $\|\,,\,\|_1$ (pour comprendre la notion de minimalité de convexité, penser à l'application $x \mapsto |x|$ en dimension 1).
- **LASSO** : Acronyme de Least Absolute Shrinkage and Selection Operator (introduit par Tibshirani en 1996).
- **Objectifs du LASSO** : Obtenir avec une norme moins "sélective" que la norme $\|\,,\,\|_0$ des résultats qui sont sensiblement similaires à ceux que l'on obtiendrait avec la norme "idéale. Notons

$$\hat{\theta}_\lambda = \mathrm{Argmin}_{\theta \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{n} + \lambda \|\theta\|_1.$$

  - Estimation : $\hat{\theta}_\lambda$ proche de $\theta^\star$?
  - Sélection : si J_0 est le "vrai" support, obtenir avec grande probabilité $J(\theta^\star_\lambda)$ proche de J_0 ?
  - Prédiction : $\frac{1}{n}\|\mathbf{X}(\hat{\theta}_\lambda - \theta^\star)\|_2^2 = O(\frac{s_0}{n})$ avec grande probabilité ?

Lasso vs Ridge et Elastic Net

Ridge et Elastic Net sont aussi basés sur des normes convexes (même strictement convexes) (Ridge a même une solution explicite). Quel est leur rôle statistique ?

- LASSO : tente de mimer la norme $\|,\|_0$ en tentant d'extraire des variables d'intérêt et en annulant les autres.
- Ridge est une méthode plus "douce" qui réduit ("shrinks") la taille des variables, notamment celles qui sont corrélées.
- Elastic Net tente donc de mélanger ces deux approches.
- Dans la suite, on parlera principalement du LASSO et du Ridge.

Evolution avec λ

- La fonction $\lambda \to \Phi_{\mathcal{P}}$ est croissante avec λ.

- Comment choisir λ ? Il s'agit d'un compromis biais-variance. Plus λ est petit, plus la variance est élevée. Plus λ est grand, moins le modèle est flexible et donc plus le biais est élevé.
- On choisit donc λ en minimisant (l'estimation de) l'erreur de prédiction.
- Ce choix est fait par validation (simple ou croisée) (voir figure slide suivant).
- On trace aussi généralement les coefficients sélectionnés en fonction de λ (voir figure slide suivant).

Evolution avec λ (suite)

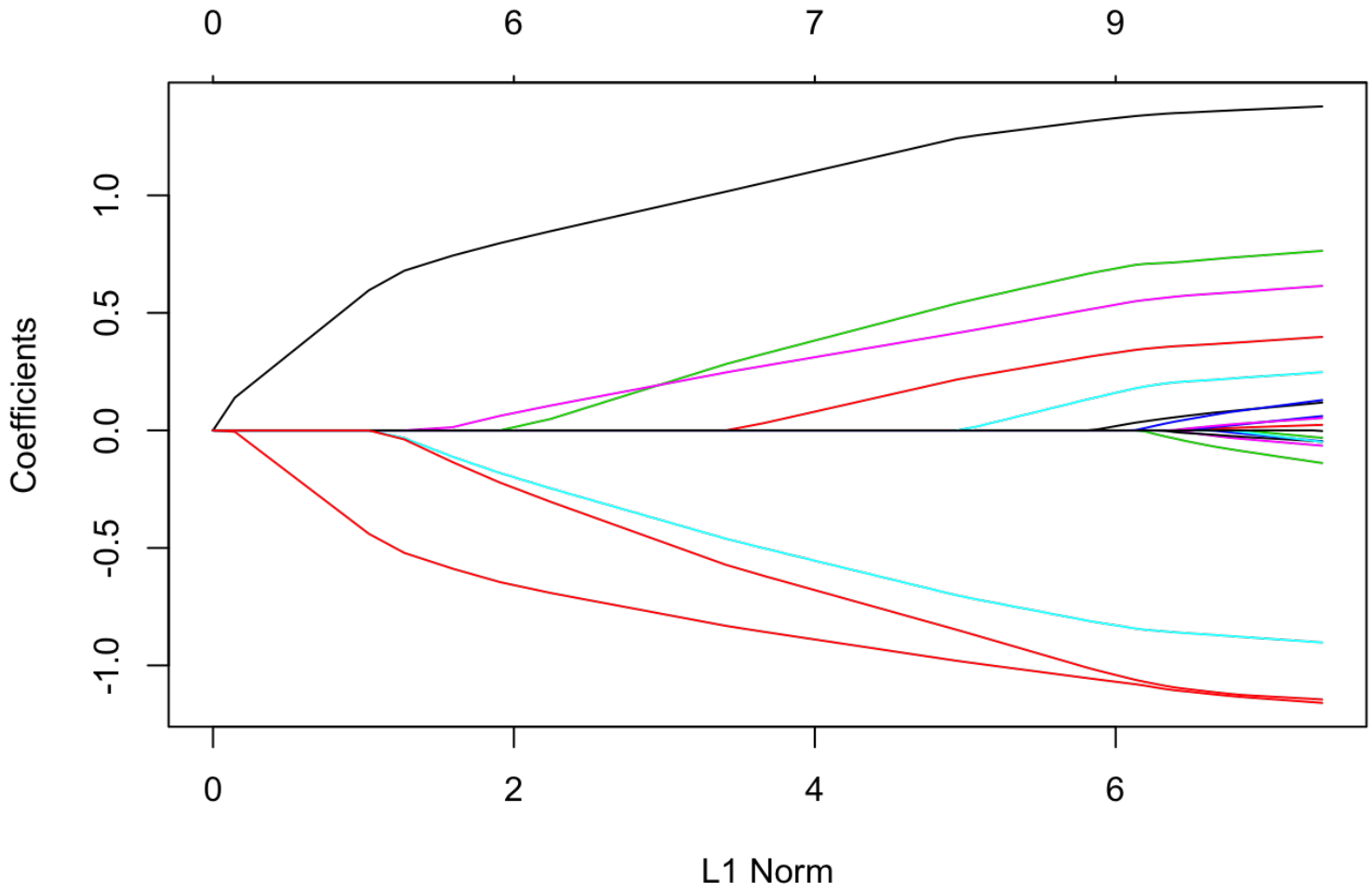

Figure 2: Apparition des coefficients dans le LASSO en fonction de $\|\hat{\theta}_\lambda\|_1$ (remarque : à l'origine $\lambda = +\infty$)

Evolution avec λ (suite)

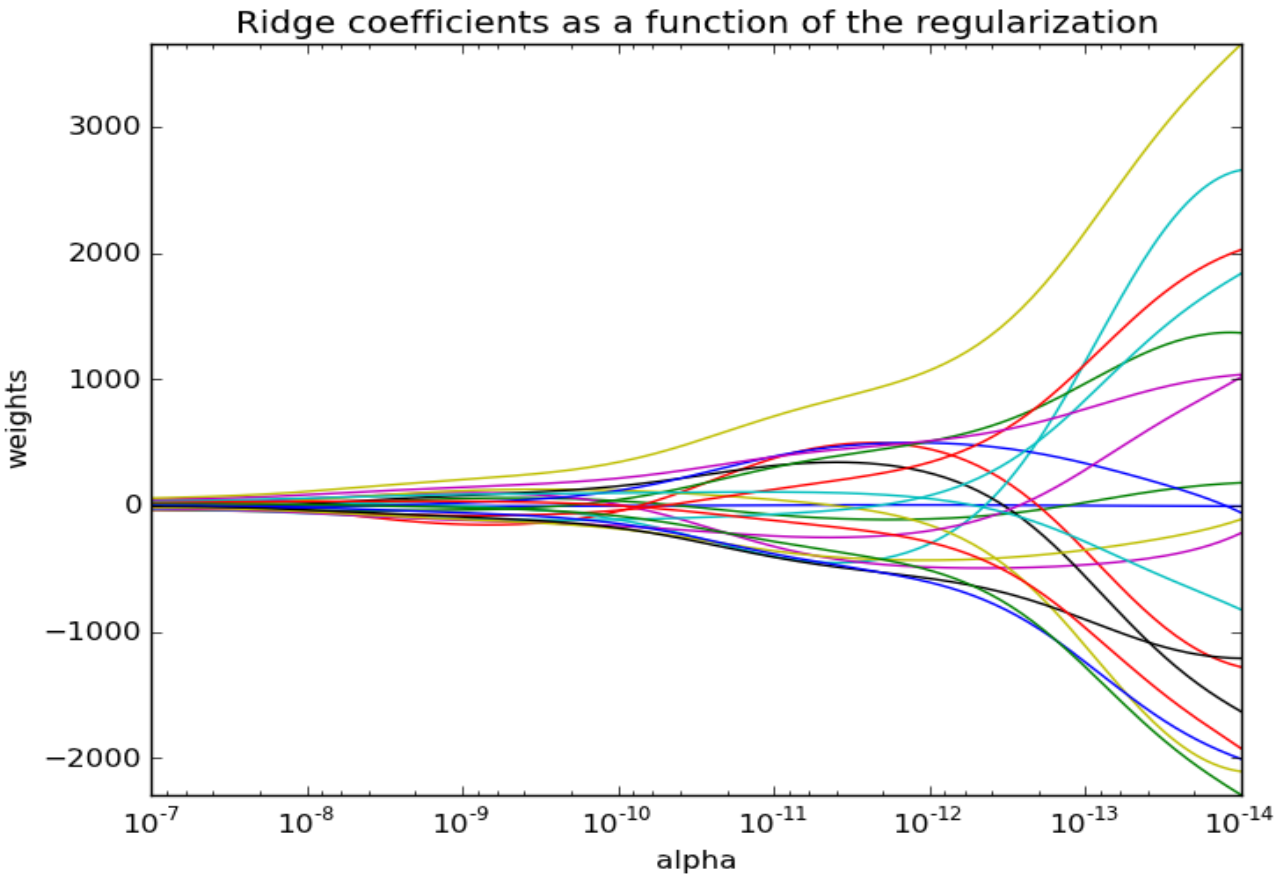

Figure 3: Apparition des coefficients dans le RIDGE en fonction de λ (plus "smooth")/"Chemin de régularisation"

Evolution avec λ (suite)

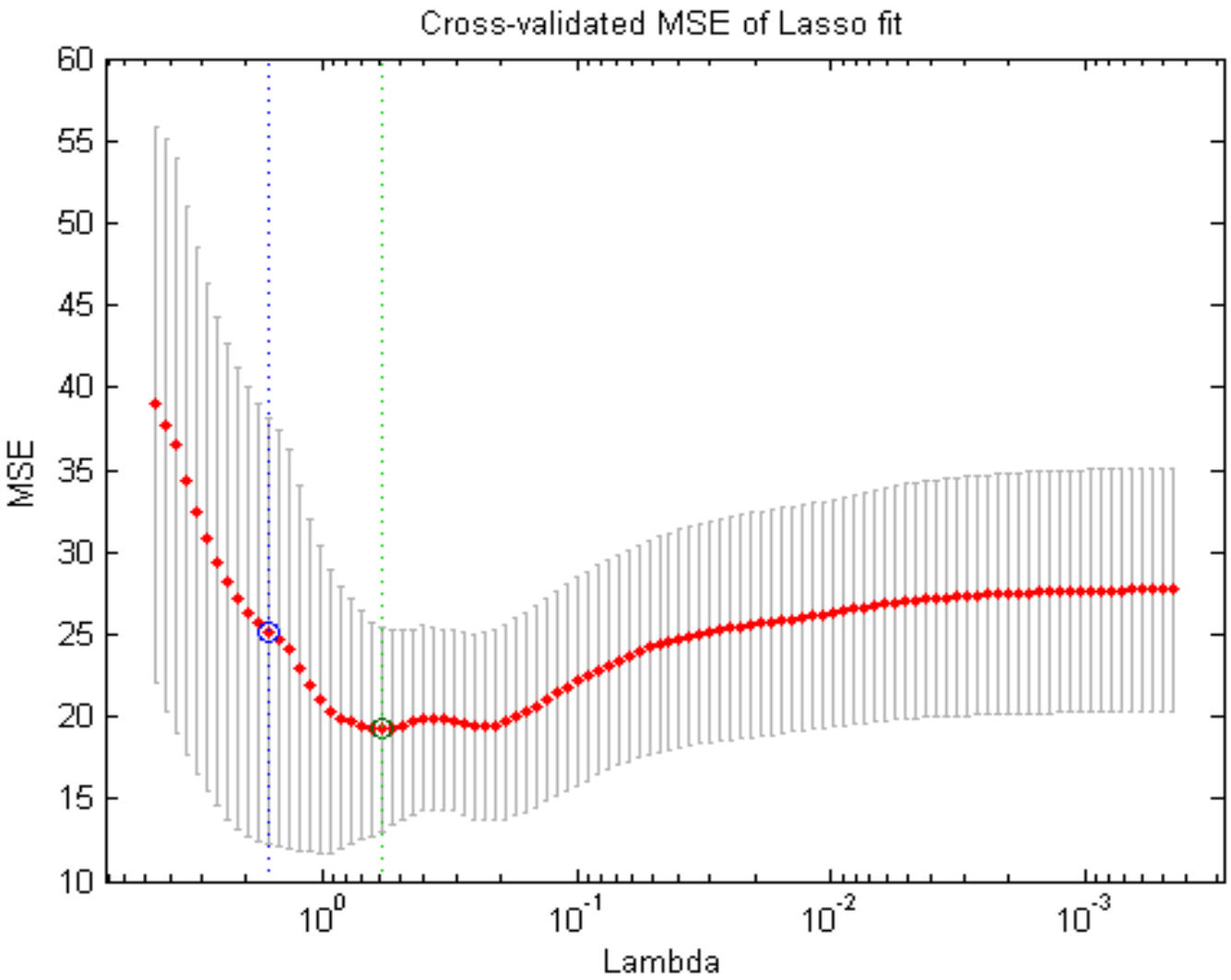

Figure 4: Estimation de la MSE en fonction de λ

Un résultat simple pour le Ridge

- si $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2}$, alors la fonction $\Phi_{\mathcal{P}}$ définie par

$$\Phi_{\mathcal{P}}(\theta) = \frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{n} + \lambda \mathcal{P}(\theta)$$

est strictement convexe de classe $\mathcal{C}^2$ et coercive (tend vers $+\infty$ lorsque $\|\theta\|_2 \to +\infty$). On a alors le résultat suivant :

Proposition

Si $\mathcal{P}(\theta) = \frac{\|\theta\|_2^2}{2}$, alors $\Phi_{\mathcal{P}}$ admet un unique minimum en

$$\hat{\theta}_\lambda = (X^T X + \lambda n I_p)^{-1} X^T Y.$$

Remarque : 1 -Inverse bien définie car $X^T X + \lambda n I_p$ est une matrice définie positive.
2 - On voit l'effet de la pénalisation qui apparaît au "dénominateur" et réduit la taille des coefficients. On voit également que l'on ne "tue" pas les coefficients.

"Calcul" de $\hat{\theta}_\lambda$ pour le LASSO

Le problème est plus difficile car $\|\theta\|_1$ est non strictement convexe et non dérivable sur les axes.

- **Un cas particulier** : $\mathbf{X}^T\mathbf{X} = I_p$ (ce qui implique que $p \leq n$. Dans ce cas, la première forme quadratique est définie positive donc $\Phi_{\mathcal{P}}$ est strictement convexe et coercive. Elle admet donc un unique minimum. On a de plus

Proposition (Seuillage doux (Soft-Thresholding))

Si $\mathbf{X}^T\mathbf{X} = I_p$, alors si l'on note $\hat{\theta} = X^T Y$ la solution du modèle linéaire non pénalisé, on a :

$$\hat{\theta}_\lambda(j) = \mathrm{sgn}(\hat{\theta}(j))(|\hat{\theta}(j)| - \lambda)_+.$$

Remarque : On voit donc sur cet exemple que le LASSO sélectionne les coordonnées de la solution classique qui sont supérieures à λ mais les "shrinke" (les réduit) aussi.

"Calcul" de $\hat{\theta}_\lambda$ pour le LASSO (suite)

Dans le cas général, il n'y a pas nécessairement unicité à la solution du LASSO. De plus, "la" solution n'est pas explicite. On doit faire recours à un algorithme d'optimisation pour déterminer $\hat{\theta}_\lambda$. On peut néanmoins affirmer que :

Proposition

$\theta \in \mathbb{R}^p$ est solution du LASSO de paramètre λ si et seulement si les conditions de "stationnarité" suivantes sont satisfaites : pour tout $j \in \{1, \dots, p\}$

$$\begin{cases} \mathbf{X}_j^T(Y - \mathbf{X}\theta) = \lambda \mathrm{sgn}(\theta_j) & \text{si } \theta_j \neq 0 \\ |\mathbf{X}_j^T(Y - \mathbf{X}\theta)| \leq \lambda & \text{sinon.} \end{cases}$$

En pratique

2 types d'algorithmes pour calculer $\hat{\theta}_{LASSO}$ (voir TD pour plus de détails)

- **L'algorithme LARS** (Least Angle Regression). Calcul par palier des variables allumées (cf graphe d'apparition des coefficients). On fait décroître λ jusqu'à l'apparition d'une première variable (la plus corrélée avec y). Ensuite, on cherche le seuil pour lequel la deuxième variable apparaît (toujours par un principe de corrélation ou d'angle)
- **La descente coordonnées par coordonnées.** Il s'agit d'un algorithme d'optimisation où l'on minimise coordonnées par coordonnées. Etape 1 : on fige $\theta_2, \ldots, \theta_p$ et on regarde la fonction

$$\theta_1 \mapsto \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{ik}\theta_k - x_{i1}\theta_1)^2 + \lambda \sum_{k\neq 1}\theta_k + \lambda|\theta_1|.$$

On en tire un minimum (explicite dans le cas du LASSO) $\hat{\theta}_1$ puis on réitère la minimisation sur la deuxième coordonnée en conservant la valeur de $\hat{\theta}_1 \ldots$. Par construction, il s'agit d'une suite décroissante. Sous conditions standard, l'algorithme converge vers un minimiseur (même principe que EM).

Prédiction/Estimation : Résultats

La condition RE

Definition

Soit $\alpha > 0$ et J_0 le support de $\beta^\star$. On note

$$C_\alpha(J_0) = \{\Delta \in \mathbb{R}^p, \|\Delta^c_{J_0}\|_1 \leq \alpha \|\Delta_{J_0}\|_1\}.$$

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

$$\|\mathbf{X}\Delta\|_2 \geq \kappa \|\Delta\|_2 \quad \forall \Delta \in C_\alpha(J_0).$$

Estimation

Théorème

Supposons que la condition RE vérifiée avec $\alpha = 3$ *et* $\kappa > 0$. *(i) (Résultat algébrique) Alors, si* $\lambda > \|\frac{X^T\varepsilon}{n}\|_\infty$,

$$\|\hat{\theta} - \theta^\star\|_2 \leq \frac{3\lambda\sqrt{s_0}}{\kappa}.$$

(ii) (Résultat probabiliste) Supposons que les ϵ_i *sont i.i.d. de loi* $\mathcal{N}(0, \sigma^2)$ *et que les colonnes de* $\mathbf{X}$ *ont été standardisés de sorte pour tout* $j \in \{1, \ldots, p\}$, $n^{-\frac{1}{2}}\|\mathbf{X}_j\| \leq C$. *Alors, si* $\delta > 0$ *et* $\lambda = 2C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}$, *le point* (i) *est vérifié avec probabilité* $1 - 2e^{-\frac{\delta^2}{2}}$. *Ainsi, avec grande probabilité,*

$$\|\hat{\theta} - \theta^\star\|_2^2 \lesssim \frac{s_0 \log p}{n}$$

où $s_0 = \mathrm{Card}\{i, \beta_i^\star \neq 0\}$.

Idée de preuve de (ii)

$\|\frac{X^T\varepsilon}{n}\|_\infty$ correspond au maximum de p variables gaussiennes. Sous les hypothèses sur les colonnes, la variance de chacune de ces variables est bornée par $C\frac{\sigma^2}{n}$. Ainsi, via un résultat similaire à celui obtenu en TD sur le maximum de variables gaussiennes, on a :

$$\mathbb{P}(\|\frac{X^T\varepsilon}{n}\|_\infty \geq C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}) \leq 2e^{-\frac{\delta^2}{2}}.$$

Le résultat suit.

Prédiction

Théorème

Supposons que les ϵ_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ et que les colonnes de $\mathbf{X}$ ont été standardisés de sorte pour tout $j \in \{1, \dots, p\}$, $n^{-\frac{1}{2}} \|\mathbf{X}_j\| \leq C$. Alors, si $\delta > 0$ et $\lambda = 2C\sigma\sqrt{\frac{2\log p + \delta^2}{n}}$, (i)

$$\frac{\|X(\hat{\theta}_\lambda - \theta^\star\|_2^2}{n} \lesssim \|\beta^\star\|_1 \sigma \sqrt{\frac{s_0 \log p}{n}}.$$

(ii) Si la condition RE est satisfaite, alors

$$\frac{\|X(\hat{\theta}_\lambda - \theta^\star\|_2^2}{n} \lesssim \frac{s_0 \log p}{n}.$$

Recouvrement du support

Théorème

Notons $\mathbf{X}_{J_0}$ la sous-matrice obtenue en conservant les colonnes actives de $\theta^\star$. Si $\mathbf{X}_{J_0}^T \mathbf{X}_{J_0}$ est inversible et qu'une condition d'incohérence mutuelle est satisfaite. Alors, si les colonnes sont normalisées comme précédemment, alors le choix $\lambda \approx \sqrt{\frac{\log p}{n}}$ permet d'assurer avec grande probabilité que le support de $\hat{\theta}_\lambda$ est contenu dans celui de $\theta^\star$.