Université de Nantes
UFR des Sciences et Techniques
Département de Mathématiques

TD Statistique en grande dimension.
M2 Ingénierie Statistique
Frédéric Lavancier

Références

- ESL : "The elements of statistical learning", T. Hastie, R. Tibshirani, J. Friedman.
- "An introduction to statistical learning with applications in R", G. James, D. Witten, T. Hastie, R. Tibshirani.
- "Introduction to high-dimensional statistics", C. Giraud.

Exercice 1. On suppose la relation linéaire $Y = X\beta + \epsilon$ vérifiée, où Y est un vecteur de taille n, X une matrice non aléatoire de taille (n, p) de rang $\min(n, p)$ et ϵ est d'espérance nulle et de variance $\sigma^2 I_n$.

1. Que vaut l'estimateur par MCO $\hat{\beta}$? On prendra soin de distinguer le cas $p \leq n$ et $p \geq n$.
2. Que vaut $\hat{Y}$, le projeté de Y sur l'espace vectoriel engendré par les colonnes de X?

On se place à présent dans le cas $p \leq n$.

3. Montrer que $\hat{\beta}$ est sans biais et calculer sa variance.
4. Montrer que $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est un estimateur sans biais de σ^2.
5. Montrer que si l'on suppose ϵ Gaussien, alors $\hat{\beta}$ correspond à l'estimateur du maximum de vraisemblance de β.
6. Rappeler les définitions du R^2, R^2_a, C_p, AIC et BIC. Que valent ces critères lorsque $p \geq n$?

Exercice 2. On se place dans le même contexte que dans l'exercice précédent et on considère un nouvel individu *new* satisfaisant la même relation linéaire, i.e.

$$y_{new} = x'_{new}\beta + \epsilon_{new}$$

où x_{new} est un vecteur de taille p et ϵ_{new} est une variable aléatoire centrée, non corrélée avec le vecteur ϵ et de même variance σ^2.

On observe Y, X, x_{new} et on souhaite prédire au mieux y_{new}

1. Si on utilise la prédiction naive $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$, quelle est l'erreur quadratique commise, i.e. $E(y_{new} - \bar{y}_n)^2$?

2. Si on utilise la prédiction $\hat{y}_{new} = x'_{new}\hat{\beta}$, où $\hat{\beta}$ est l'estimateur des moindres carrés calculé à partir de Y et X, quelle est l'erreur quadratique commise?

3. On suppose que le conditions sont réunies pour impliquer que $\hat{\beta}$ converge en moyenne quadratique vers β lorsque $n \to \infty$. Comparer les deux erreurs précédentes lorsque $n \to \infty$.

4. On suppose à présent que l'on a utilisé les mauvaises variables explicatives dès le départ. Précisément, on suppose que la vraie relation est $Y = X^*\beta^* + \epsilon$ avec $X^* \neq X$. Il en est de même pour la nouvelle observation : $y_{new} = x^{*'}_{new}\beta^* + \epsilon_{new}$. Ayant régressé Y sur X, on considère toujours la prévision $\hat{y}_{new} = x'_{new}\hat{\beta}$, où $\hat{\beta}$ est comme précédemment. Calculer $E(\hat{\beta})$ et $V(\hat{\beta})$ et montrer que l'erreur quadratique de prévision se dégrade, plus précisément qu'elle vaut celle de la question 2 plus le terme de biais $[E(x'_{new}\hat{\beta}) - x^{*'}_{new}\beta^*]^2$.

Exercice 3. On dispose d'un échantillon i.i.d $(y_1, x_1), \ldots, (y_n, x_n)$ où, pour tout i, $y_i \in \mathbb{R}$ et $x_i \in \mathbb{R}^p$ est un vecteur de taille p. On suppose qu'une régression linéaire a été effectuée à partir de cet échantillon afin d'expliquer les y_i en fonction des x_i, et on note par $\hat{\beta}$ le coefficient de régression estimé et par $\hat{y}_i = x'_i\hat{\beta}$ les valeurs ajustées de y_i par ce modèle. On se place dans la situation où on ne sait pas si ce modèle linéaire est valide.

On considère le risque empirique

$$R = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x'_i\hat{\beta})^2.$$

1. De quel problème de minimisation les $\hat{y}_i$ sont-ils solution?

2. On considère à présent un échantillon test $(\tilde{y}_1, \tilde{x}_1), \ldots, (\tilde{y}_n, \tilde{x}_n)$ indépendant du précédent et identiquement distribué. Le risque de prévision (ou erreur test) empirique est

$$\tilde{R} = \frac{1}{n}\sum_{i=1}^n (\tilde{y}_i - \tilde{x}'_i\hat{\beta})^2.$$

   Montrer qu'en moyenne le risque empirique est toujours inférieur au risque de prévision, i.e. $E(R) \leq E(\tilde{R})$.

3. On suppose de plus que tous les couples (x_i, y_i) suivent la même loi. Montrer que le résultat précédent reste vrai si l'échantillon test contient un nombre $m \neq n$ d'observations.

4. Donner un exemple de couple (n, p) pour lequel on a $R = 0$ quel que soit l'échantillon d'apprentissage et la validité ou non du modèle linéaire estimé. Aura-t-on $\tilde{R} = 0$?

Exercice 4. Soit $Z_1, \ldots, Z_M$ les composantes principales utilisées dans une PCR ou une PLS, construites à partir d'une matrice X dont les variables ont été centrées et réduites. Le modèle PCR ou PLS s'écrit donc

$$Y = \gamma_0 + \sum_{j=1}^{M} \gamma_j Z_j + \epsilon,$$

où $\gamma_0, \gamma_1, \ldots, \gamma_p$ sont les paramètres à estimer. Montrer que les estimateurs par moindres carrés ordinaires valent

$$\hat{\gamma}_0 = \bar{Y}, \qquad \hat{\gamma}_j = \frac{Y' Z_j}{Z_j' Z_j}. \tag{1}$$

Exercice 5. On suppose que les p variables de taille n formant la matrice X sont centrées et réduites. On rappelle qu'étant donné un vecteur Y de $\mathbb{R}^n$, les composantes de la régression PLS de Y sur X sont définies pour $j = 1, \ldots, p$ par $Z_j = X\alpha_j$ où

$$\alpha_j = \underset{\alpha \in \mathbb{R}^p}{argmax}\ \mathrm{Cov}(X\alpha, Y) \tag{2}$$

sous les contraintes $\|\alpha\| = 1$ et $\alpha' X' X \alpha_l = 0$ pour tout $l = 1, \ldots, j-1$.

1. Montrer que la première composante est $Z_1 = XX'Y/\|X'Y\|$.
2. Montrer que Z_2 appartient nécessairement à l'espace vectoriel engendré par $P_{[Z_1]^\perp} X$ où $P_{[Z_1]^\perp}$ désigne la matrice de projection sur l'orthogonal de $[Z_1]$. Que vaut cette matrice?
3. En déduire l'expression de Z_2.
4. L'algorithme général pour construire les axes de la PLS est le suivant : partant de $X_{(1)} = X$ et de $j = 1$, tant que $j \leq p$ et $\alpha_j \neq 0$
   (a) $\alpha_j = X'_{(j)} Y / \|X'_{(j)} Y\|$.
   (b) $Z_j = X_{(j)} \alpha_j$
   (c) $X_{(j+1)} \leftarrow X_{(j)} - Z_j Z_j' X_{(j)} / (Z_j' Z_j)$ et retour à 1 avec $j \leftarrow j + 1$.

   Si $\alpha_j = 0$ pour un certain j, $Z_k = 0$ pour tout $k \geq j$.

   Justifier que cet algorithme construit bien les composantes de la PLS.

Exercice 6. On suppose que les variables composant la matrice X sont centrées, réduites et orthogonales entre elles.

1. Que vaut dans ce cas $X'X$?
2. Montrer que les composantes de la PLS sont toutes nulles sauf la première.

3. Que vaut le coefficient de régression de X_j issu de la régression PLS dans ce cas?

4. Déterminer les composantes principales de l'ACP construites à partir des variables explicatives.

5. Que donnerait une régression PCR sur M composantes ($M \leq p$)?

Exercice 7. Soit un modèle de régression linéaire multiple

$$y = X\beta + \epsilon,$$

où $\beta \in \mathbb{R}^p$, X est une matrice de taille $n \times p$ et où les variables ϵ_i, $i = 1, \ldots, p$, sont centrées, homoscédastiques et non-corrélées.
Pour $\lambda \geq 0$, on considère l'estimateur $\hat{\beta}_R = (X'X + \lambda I_p)^{-1} X'Y$ où I_p désigne la matrice identité de taille p.

1. Comment s'appelle l'estimateur $\hat{\beta}_R$? A quel problème de minimisation sous contrainte est-il la solution? Dans quel cas est-il égal à l'estimateur des MCO?

   On suppose dans la suite que les variables explicatives ont été centrées, réduites et qu'elles sont non-corrélées entre elles.

2. Exprimer $\hat{\beta}_R$ en fonction de l'estimateur par MCO $\hat{\beta}$.

3. En déduire l'espérance et la matrice de variance de $\hat{\beta}_R$.

4. Soit $\hat{\beta}_{R,i}$, $\hat{\beta}_i$ et β_i la i-ème composante des vecteurs $\hat{\beta}_R$, $\hat{\beta}$ et β respectivement ($i = 1, \ldots, p$). En supposant que n est suffisamment grand pour garantir $n\beta_i^2 > \sigma^2$, montrer que $EQM(\hat{\beta}_{R,i}) \leq EQM(\hat{\beta}_i)$ si et seulement si $\lambda \leq 2\sigma^2/(\beta_i^2 - \sigma^2/n)$. Dans tous les cas (quel que soit n), montrer que la condition $\lambda \leq 2\sigma^2/\beta_i^2$ est suffisante pour garantir $EQM(\hat{\beta}_{R,i}) \leq EQM(\hat{\beta}_i)$.

5. Montrer de façon plus générale qu'en supposant n suffisamment grand, la condition $\lambda \leq 2\sigma^2/(\|\beta\|^2 - \sigma^2/n)$ implique que $\hat{\beta}_R$ est meilleur que $\hat{\beta}$ au sens du coût quadratique.

6. Quelle condition suffisante, montrée en cours, implique que $\hat{\beta}_R$ est meilleur que $\hat{\beta}$ au sens du coût quadratique dans le cas général où les variables X_j ne sont plus supposées non-corrélées entre elles? Vérifier que la condition précédente est plus faible.

7. Les conditions établies précédemment sont-elles utilisables en pratique? Quelle procédure peut-on mettre en place en pratique pour choisir λ?

Exercice 8. On souhaite étudier quelques propriétés de l'estimateur Lasso de la régression de Y sur X. On rappelle les notions d'optimisation suivantes.

Soit f une fonction convexe de $\mathbb{R}^d$ dans $\mathbb{R}$. Un sous-gradient $s \in \mathbb{R}^d$ de f au point $x \in \mathbb{R}^d$ vérifie

$$f(x+h) \geq f(x) + s'h, \quad \forall h \in \mathbb{R}^d.$$

Si la fonction f est différentiable en x, alors le sous-gradient de f en x est unique et correspond à son gradient en x. Dans le cas contraire, on note $\partial f(x)$ l'ensemble (éventuellement vide) des sous-gradients possibles de f en x.

Selon la règle de Fermat (ou condition d'optimalité du premier ordre), le point x^* est le minimum d'une fonction convexe f ssi $0 \in \partial f(x^*)$.

1. Montrer que si $f(x) = |x|$, $x \in \mathbb{R}$, alors

$$\partial f(x) = \begin{cases} sign(x) & \text{si } x \neq 0 \\ [-1,1] & \text{si } x = 0 \end{cases},$$

où $sign(x) = 1$ si $x > 0$ et $sign(x) = -1$ si $x < 0$.

On rappelle que l'estimateur Lasso β^* associé au paramètre $\lambda > 0$ minimise

$$L(\beta) = \frac{1}{2}\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

2. En utilisant la règle de Fermat, montrer que β^* est solution du problème précédent ssi, pour tout $j = 1, \ldots, p$,

$$\begin{cases} 0 = (X'X\beta^*)_j - X_j'Y + \lambda sign(\beta_j^*) & \text{si } \beta_j^* \neq 0, \\ 0 \in \left[(X'X\beta^*)_j - X_j'Y - \lambda;\, (X'X\beta^*)_j - X_j'Y + \lambda\right] & \text{si } \beta_j^* = 0. \end{cases}$$

On se place dans la suite dans la situation où les variables composant la matrice X sont centrées, réduites et non corrélées.

3. Que vaut alors $X'X$? Que vaut la j-ème composante $\hat{\beta}_j$ de l'estimateur par MCO du modèle de régression.

4. Déduire des deux questions précédentes que dans ce contexte $\beta_j^* = 0$ ssi $|\hat{\beta}_j| \leq \lambda/n$.

5. Montrer enfin que $\beta_j^* = \hat{\beta}_j - \lambda/n$ lorsque $\hat{\beta}_j > \lambda/n$ et $\beta_j^* = \hat{\beta}_j + \lambda/n$ lorsque $\hat{\beta}_j < -\lambda/n$.

6. En guise de comparaison, rappeler l'expression de l'estimateur Ridge dans ce contexte (voir exercice 7). Les régressions Ridge et Lasso peuvent être vues comme des techniques de régularisation de l'estimateur par MCO $\hat{\beta}$, dans le sens où elles réduisent la valeur des coefficients de $\hat{\beta}$ (shrinkage) et/ou les seuillent (thresholding ou selection). Illustrer ces propriétés à l'aide des résultats de cet exercice.

Exercice 9. *Effet de la multicolinéarité*

On considère p variables $X_1, \dots, X_p$ non colinéaires et une variable à expliquer Y. Chaque vecteur est de taille $n > p$.

1. On suppose que la régression par MCO de Y sur les variables $X_1, \dots, X_p$ a donné l'estimateur $\hat{\beta}$. On ajoute à présent une variable X_p^* copie exacte de X_p, i.e. $X_p^* = X_p$. Que vaut l'espace vectoriel engendré par $X_1, \dots, X_p, X_p^*$? En déduire le projeté de Y sur cet espace et décrire l'ensemble des estimateurs par MCO de Y sur $X_1, \dots, X_p, X_p^*$.

2. On suppose que la régression Lasso de Y sur les variables $X_1, \dots, X_p$, associée au paramètre de régularisation $\lambda > 0$, a donné l'estimation $\hat{\beta}(\lambda)$. Montrer que l'ensemble des solutions de la régression Lasso de Y sur $X_1, \dots, X_p, X_p^*$, associée au même paramètre de régularisation $\lambda > 0$, s'écrit $(\hat{\beta}_1(\lambda), \dots, \hat{\beta}_{p-1}(\lambda), \hat{\gamma}_p(\lambda), \hat{\gamma}_{p^*}(\lambda))$ où $\hat{\gamma}_p(\lambda) + \hat{\gamma}_{p^*}(\lambda) = \hat{\beta}_p(\lambda)$. On pourra utiliser une récurrence sur les étapes de l'algorithme LARS-Lasso.

3. On suppose que la régression Ridge de Y sur les variables $X_1, \dots, X_p$, associée à un paramètre de régularisation $\lambda > 0$, a donné l'estimateur $\hat{\beta}(\lambda)$. On suppose de plus que X_p est orthogonale aux autres variables explicatives. Pour le même λ, que vaut l'estimateur Ridge de Y sur $X_1, \dots, X_p, X_p^*$?

4. Commenter sur la capacité de chaque méthode à gérer les problèmes de multicolinéarité.

Exercice 10. On suppose que l'on effectue la régression d'une variable Y centrée sur deux variables X_1 et X_2 qui sont chacune centrée et de norme 1. On suppose que la corrélation empirique entre X_1 et X_2 vaut ρ et que l'estimateur des MCO de cette régression est

$$\hat{\beta}^{MCO} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

1. Montrer que la matrice $X = (X_1\ X_2)$ vérifie $X'X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ et que $X'Y = \begin{pmatrix} 4+2\rho \\ 4\rho+2 \end{pmatrix}$.

Dans ce contexte, T. Hastie, R. Tibshirani et J. Friedman présentent dans la section 3.6 de leur ouvrage ESL la représentation reproduite dans la figure 1. Elle décrit, pour les deux situations où $\rho = 0.5$ et $\rho = -0.6$ (la figure présentée dans l'ouvrage est indiquée pour $\rho = -0.5$ et non $\rho = -0.6$, mais il y a manifestement une erreur), l'évolution des estimations de β_1 et β_2 pour les méthodes suivantes :

- Best Subset : estimation basée sur le meilleur modèle à k variables (au sens où la SCR est minimale). Cette méthode fournit donc 3 points (reliés entre eux dans la figure) : le point 0 associé à $k = 0$, l'estimateur associé à $k = 1$, et celui associé à $k = 2$ qui correspond forcément à $\hat{\beta}^{MCO}$.

- PCR : estimation basée sur la PCR à k composantes. Il y a 3 points (pour $k = 0$, $k = 1$ et $k = 2$).

- PLS : estimation basée sur la PLS à k composantes (3 points).
- Ridge : estimation basée sur l'estimateur ridge lorsque λ varie de $+\infty$ à 0. L'ensemble des estimateurs est une courbe variant de 0 (le cas $\lambda = +\infty$) à $\hat{\beta}^{MCO}$ (le cas $\lambda = 0$).
- Lasso : estimation basée sur l'estimateur Lasso lorsque λ varie de $+\infty$ à 0. L'ensemble des estimateurs est une courbe linéaire par morceaux variant de 0 à $\hat{\beta}^{MCO}$.

2. Retrouver les coordonnées de tous les points et l'équation de la courbe Ridge dans les deux cas. On pourra s'aider pour PLS des exercices 5 et 4, et s'appuyer pour Lasso sur l'algorithme LARS (voir cours). Vérifier vos résultats en reproduisant le même type de graphique que dans la figure 1 avec R.

Les auteurs commentent cette figure de la façon suivante (voir Section 3.6 de ESL) :

"In the top panel, starting at the origin, ridge regression shrinks the coefficients together until it finally converges to least squares. PLS and PCR show similar behavior to ridge, although are discrete and more extreme. Best subset overshoots the solution and then backtracks. The behavior of the lasso is intermediate to the other methods. When the correlation is negative (lower panel), again PLS and PCR roughly track the ridge path, while all of the methods are more similar to one another."

Et ils concluent ainsi sur le comportement général de ces méthodes en pratique :

"To summarize, PLS, PCR and ridge regression tend to behave similarly. Ridge regression may be preferred because it shrinks smoothly, rather than in discrete steps. Lasso falls somewhere between ridge regression and best subset regression, and enjoys some of the properties of each."

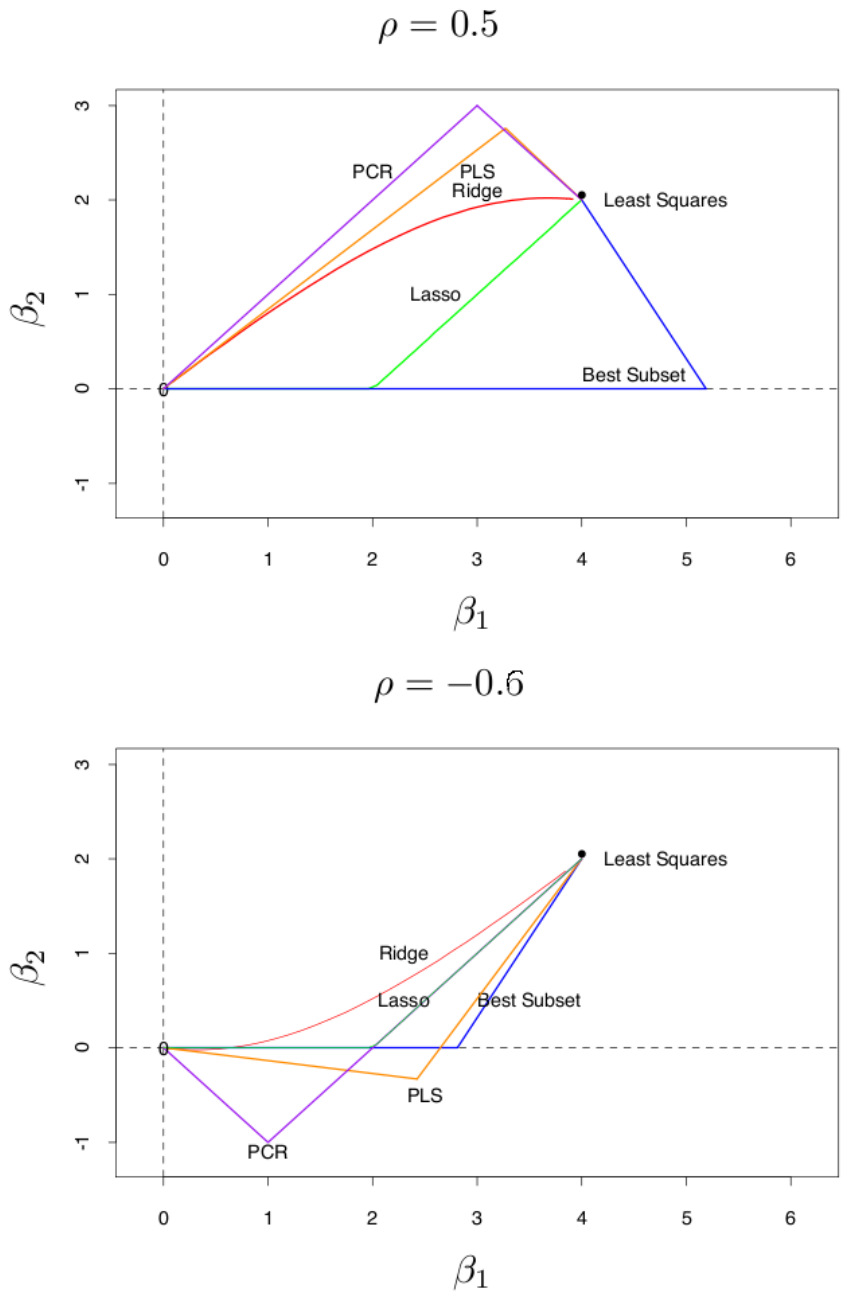

Figure 1: Evolution des estimations de β_1 et β_2 selon différentes méthodes lorsqu'on augmente le nombre de composantes (pour Best Subset, PCR et PLS) ou que le paramètre de régularisation décroit de $+\infty$ à 0 (pour Ridge et Lasso). Les détails sont donnés dans l'exercice 10.