



INTRODUCTION AU MACHINE LEARNING

ISFA, Master 2 Actuariat (Sept.-Déc. 2020)

Xavier Milhaud

xavier.milhaud@univ-lyon1.fr

www.xaviermilhaud.fr

PLAN DU COURS

- 1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension
- 2 Actuariat - données et assurance
- 3 Philosophie de l'apprentissage statistique
- 4 Première brique en Machine Learning : arbres de décision
- 5 Bagging + randomization de CART : forêts aléatoires
- 6 Agrégation de modèles par boosting
- 7 Réseau de neurones et Deep Learning
- 8 Extensions

BIBLIOGRAPHIE - EXEMPLES

De livres :

- An introduction to Statistical Learning, (with Applications in R), ;James, Witten, Hastie, Tibshirani
- The Elements of Statistical Learning : Data Mining, Inference and Prediction ; Hastie, Tibshirani, Friedman
- Classification & Regression Trees ; Breiman, Friedman, Olshen, Stone
- Artificial Intelligence : A Modern Approach ; Russell and Norvig
- Speech and Language Processing ; Jurafsky and Martin
- Pattern Recognition and Machine Learning ; Bishop C.

D'articles :

- Model selection for CART regression trees, Gey, Nedelec
- Consistency of random survival forests, Ishwaran, Kogalur
- Quantile regression forests, Meinshausen
- Tree-based methods : a useful tool for life insurance, Olbricht

De mémoires :

- Modélisation des arbitrages dynamiques par approche machine learning, Berrada (2017)
- Construction d'un modèle prédictif des comportements d'arbitrage euro-UC : prise en compte de facteurs psychologiques, Douillard
- Support Vector Machines : Machine Learning, the SVM algorithm and applications in Health Insurance Pricing, Francisco Miguelez (2017)
- Machine learning algorithms (regression trees and random forests) to monitor the performance of claim partners, Barbry (2016)
- Open Data et RC corporelle automobile, Dubert et Hilpert (2018)
- Estimation de la probabilité de maintien d'activité de courtiers, à l'aide de méthodes d'apprentissage automatique, Lesage (2017)

De tutoriels et MOOC :

- Machine Learning :
<https://github.com/ujjwalkarn/Machine-Learning-Tutorials>
- Machine Learning (Stanford Univ, via Coursera) :
<https://www.classcentral.com/course/coursera-machine-learning-835>
- Data Science (Harvard Univ) : <http://cs109.github.io/2014/>

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
 - Estimation et grande dimension
 - Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
 - Notions de biais et variance d'un estimateur
 - Réduction de dimension : pénalisation ex-post
 - Contraste pénalisé - pénalisation ex-ante
 - Pénalisation L^2 : régression Ridge
 - Hypothèse de parcimonie et pénalisation L^1 : LASSO
 - La pénalisation Elastic-net

CONTEXTE CLASSIQUE D'ETUDE DES RISQUES

L'analyse d'engagements d'un assureur nécessite de comprendre l'impact de caractéristiques **X** sur le risque **Y**.

Les bases de données des assureurs comportent généralement

- les **caractéristiques** de l'assuré,
- les **options** du contrat,
- les conditions de **marché**.

Informations **X** *jouent un rôle crucial* dans les prév. de sinistralité **Y**
⇒ **méthodes doivent tenir compte de ces caractéristiques**
(historiquement modélisation paramétrique par régression).

GENERALITES

Pourquoi modéliser ?

⇒ A partir d'une série d'observations, phénomène trop complexe pour une description analytique par un modèle déterministe...

Objectif en statistique : modélisation, parfois décomposable, pour

- ➊ **explorer** : décrire variables, leurs liaisons, positionner obs. ;
- ➋ **expliquer** : tester l'influence d'une variable ds un modèle supposé connu ;
- ➌ **prévoir** et sélectionner : un meilleur ensemble de prédicteurs.

Historiquement, modèles paramétriques avec var. expl. + bruit ⇒ inférer les paramètres depuis les observ. en contrôlant au mieux les propriétés (comportement) de la partie aléatoire.

MOTIVATION DU COURS

Observons n réalisations de $(\mathbf{X}, Y) \in \mathbb{R}^{p+1}$.

D'habitude, on considère que

- le rapport des dimensions (n, p) est raisonnable,
- les hyp. du modèle sont vérifiées (échantillon/résidus supposés suivre des lois sous la forme d'une famille connue),

Alors les techniques statistiques tirées du modèle linéaire général sont optimales (max. de vraisemblance)... Avec des échantillons de taille restreinte \Rightarrow difficile de faire beaucoup mieux.

Mais dès que hyp. distributionnelles ne sont pas vérifiées / relations entre les variables ou la variable à modéliser ne sont pas linéaires, ou encore dès que le volume des données est important, d'autre méthodes viennent concurrencer la stat. classique...

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- **Estimation et grande dimension**
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

PARAMETRIQUE VS NON-PARAMETRIQUE

Cadre : on veut estimer 1 fct. m , par ex. $m(x) = E[Y | X = x]$, ou $m(x) = P(Y = 1 | X = x)$.

- **Estimation paramétrique** : on cherche m parmi une famille indexée par un param. de dim. finie \rightarrow **ex** : rég. lin., $m(x) = a + bx$. Un candidat s'identifie à 2 paramètres (a, b) .
- **Estimation non paramétrique** : pas d'hypothèse (ou bc $-$), cherche $m(x)$ parmi ttes les fonct. possibles (**dim. infinie**) \Rightarrow décompositions dans des bases fonctionnelles (ex GAM) :

$$y = m(x) = \sum_{k=0}^{\infty} w_k g_k(x) \quad \text{et donc} \quad \hat{m}(x) = \sum_{k=0}^{h^*} \hat{w}_k g_k(x)$$

LA DIMENSION, FACTEUR LIMITANT

Paramètres importants du problème : ses dimensions... **Notons** :

- n nombre d'observations ou taille de l'échantillon,
- p nombre de variables observées sur cet échantillon.

→ n grand : pas de pb a priori, bien au contraire (théo asymptot.) !
→ p grand pose problème (fléau de la dimension) !

L'estimateur du max. de vrais. conserve sa prop. de normalité asymptotique si $p^2/n \rightarrow 0$ lorsque $p, n \rightarrow \infty$ (Portnoy, 1988).

⇒ Données “massives” : $p > \sqrt{n}$.

Concept de sparsité \simeq dimension effective \Rightarrow compter le nb de var. expl. réel du pb, à défaut de compter le nb total de var. expl. !

MALEDICTION DE LA DIMENSION

On fait allusion au **nombre de variables p** . Ce fléau est décrit dans le livre de Bellman R.E. (1957), *Dynamic Programming* (Pinceton University Press).

Explication intuitive : le volume de la sphère unité, en dim. p , tend vers 0 lorsque $p \rightarrow \infty \Rightarrow$ l'espace est "sparse" (clairsemé), ce qui signifie que la proba. de trouver un point proche d'un autre point devient de + en + faible.

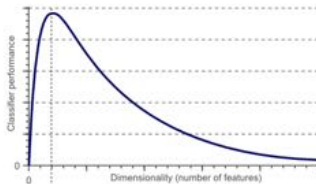
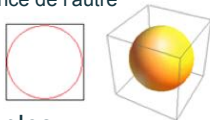
\Rightarrow **Autrement dit** : le volume qu'il faut considérer pour avoir une proportion donnée d'observations **augmente avec p** ...

\Rightarrow Nécessite une sélection des variables ! Sinon, si p grand, il faut énormément d'observations...

FLEAU DE LA DIMENSION, SUITE

“Curse of dimensionality”

- Beaucoup d'algos de Machine Learning utilisent des distances
 - SVM, k-nn, Recommandation par filtrage Collaboratif
- En grandes dimensions toutes les distances sont les mêmes
 - 2 points sont proches si l'un est à une certaine (petite) distance de l'autre
 - En 2 D, le cercle couvre 78% du carré
 - 52% en 3 D, ...0.24% en 10-D
- Quand la dimension augmente, les classifieurs overfittent → il faut augmenter le nombre d'exemples d'apprentissage



FLÉAU DE LA DIMENSION - MONDE PARAMETRIQUE

En statistique “classique” : pas de problème apparent...

→ Ex. modèle linéaire (notons σ^2 la variance des résidus) :

$$m_0(x) = E_0[Y | X = x] = \beta_0^T x.$$

→ On a l'estimateur MCO : $\hat{\beta}_0 = (X^T X)^{-1} X^T Y$.

X : matrice schéma (design matrix), Y : réponses observées.

Propriétés : $\hat{\beta}_0 \approx \mathcal{N}(\beta_0, \sigma^2(X^T X)^{-1})$.

En quittant le modèle lin. gaussien, on n'a pas de forme explicite pour l'estimateur EMV mais il conserve les mêmes propriétés asymptotiques, se calcule numériquement (moyennant qlq hyp.).

POURQUOI LA CV N'EST PAS SI ÉVIDENTE ?

Derrière le résultat de convergence en loi de β intervient implicitement l'hypothèse que $p \ll n$.

- **1^{ère} façon de voir le pb** : mauvais conditionnement de $(X^T X)$ (matrice de variance-covariance empirique des régresseurs).
Rappel : conditionnement de $A = \kappa(A) = \|A^{-1}\| \|A\|$.

Rq : si $n \gg p$, diminue les chances que la corrélation entre les variables explicatives soit très grande...

→ En pratique, X est généralement de plein rang colonne $p \Rightarrow X^T X$ est symétrique définie positive, donc inversible \Rightarrow le problème MCO possède donc bien une solution $\hat{\beta}$ unique !

→ Toutefois, les facteurs de risque peuvent présenter une forte corrélation entre eux (ex : facteurs liés aux taux d'intérêt à différentes maturités)...

⇒ colinéarité entre colonne de $X \Rightarrow$ mauvais conditionnement de $X^T X \Rightarrow$ coef. de rég. très élevés en valeur absolue sur certains facteurs de risque et des signes peu intuitifs...

⇒ Grande sensibilité de la solution $\hat{\beta}$ à de faibles variations de X ou Y ... \Rightarrow pas souhaitable (variance β) !

→ Illustration grâce à la décomposition en valeurs singulières (SVD) : notons r le rang de X , alors on peut écrire

$$X = U \Sigma V^T, \quad \text{avec}$$

$U \in \mathbb{R}^{n \times n}$ une matrice orthogonale : $U^T U = U U^T = I_n$,

$V \in \mathbb{R}^{p \times p}$ une matrice orthogonale : $V^T V = V V^T = I_p$,

$\Sigma \in \mathbb{R}^{n \times p}$ une matrice de la forme

$$\Sigma = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{où} \quad \Sigma_r = \begin{pmatrix} \sigma_1 & & \\ & \dots & \\ & & \sigma_r \end{pmatrix}$$

Eléments de la diagonale $\sigma_1 > \dots > \sigma_r > 0$: valeurs singulières de $X \Rightarrow$ leur carré correspond aux valeurs propres de $X^T X$.

La solution MCO, lorsqu'elle est unique (donc $r = p$), peut s'écrire

$$\hat{\beta} = V \Sigma_p^{-1} U^T Y.$$

$$\text{Or} \quad \Sigma_p^{-1} = \begin{pmatrix} 1/\sigma_1 & & \\ & \dots & \\ & & 1/\sigma_p \end{pmatrix}.$$

→ Si colonnes de X correspondent à des réalisations de v.a. fortement corrélées entre elles \Rightarrow valeurs singulières σ_k très proches de 0...

\Rightarrow certains éléments de la diag. auront valeurs très élevées (car $1/\sigma_k$)

\Rightarrow composantes extrêmes dans le vecteur solution $\hat{\beta}$...

→ Avec la norme 2 (MCO) et $X^T X$ diagonalisable, alors $\kappa(X^T X) = \sigma_1/\sigma_p \Rightarrow$ potentiellement très grand !

→ Soit $\hat{\beta}$ (resp. $\beta + \Delta\beta$) solution de $X\beta = Y$ (resp. ΔY),

$$\frac{\|\Delta\hat{\beta}\|}{\|\beta\|} \leq \kappa(X^T X) \frac{\|\Delta Y\|}{\|Y\|}$$

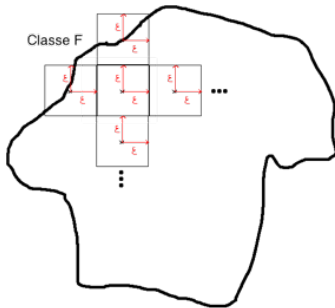
→ **Mauvais conditionn. \Rightarrow erreur relative de solution + gde !**

- 2^{ème} façon : via la complexité des fonctions de régression.

Q : comment mesurer la richesse d'un ensemble de fonctions ?

→ Soit \mathcal{F} une classe de fonctions, munie d'une norme $\|\cdot\|$.

→ Introduisons $N(\epsilon, \mathcal{F})$, le nombre d'ensembles de taille ϵ pour recouvrir \mathcal{F} : + la classe est complexe, + N sera grand !



Classe de Vapnik-Chervonenkis : il s'agit d'une classe \mathcal{F} telle que

$$N(\epsilon, \mathcal{F}) \leq C \epsilon^v$$

où v est appelé **indice VC**.

- Supposons qu'on cherche $\hat{m} \in \mathcal{F}$ avec $m \in \mathcal{F}$ (bonne spécification), alors **vitesse de CV de l'ordre de** $O\left(\sqrt{\frac{v}{n}}\right)$
- Typiquement, si on cherche m dans un espace paramétrique de dimension p , on a **$v \approx p$** .

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- **Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons**
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

THE P-VALUE PROBLEM

“A key issue with applying small-sample statistical inference to large samples is that even minuscule effects can become statistically significant. The increased power leads to a dangerous pitfall as well as to a huge opportunity. The issue is one that statisticians have long been aware of : the p -value problem. Chatfield (1995, p. 70) comments, question is not whether differences are significant (they nearly always are in large samples), but whether they are interesting. Forget statistical significance, what is the practical significance of the results?”

Mingfeng Lin, Henry Lucas, Jr. et Galit Shmueli , 2010 galitshmueli.com

Source : blog d'Arthur Charpentier.

Idée : bonne puissance de test implique qu'un gd échantillon (n grand) fait systématiquement conclure à un effet significatif d'un facteur de risque, quand bien même cet effet serait négligeable...

RAPPEL SUR LA PUISSANCE D'UN TEST

On peut résumer le rôle des probabilités de bonne et mauvaise décision dans le tableau suivant (β est la **puissance** du test) :

Vérité Décision	H_0	H_1
H_0	$1 - \alpha$	$1 - \beta$
H_1	α	β

Risque / Erreur 1^{ère} espèce : décider H_1 vraie alors que H_0 vraie (proba. α).

Erreur seconde espèce : décider H_0 vraie alors que H_1 vraie (proba. erreur de seconde espèce : $1 - \beta$).

La puissance β dépend

- 1 du **nombre d'observations** (d'individus),
- 2 du risque α : en general quand $\alpha \nearrow$, la puissance $\beta \nearrow$ aussi : on ne gagne pas partout !
- 3 et de l'ampleur de l'effet (différence entre les 2 groupes pour un essai clinique par ex.) relativement aux autres grandeurs.

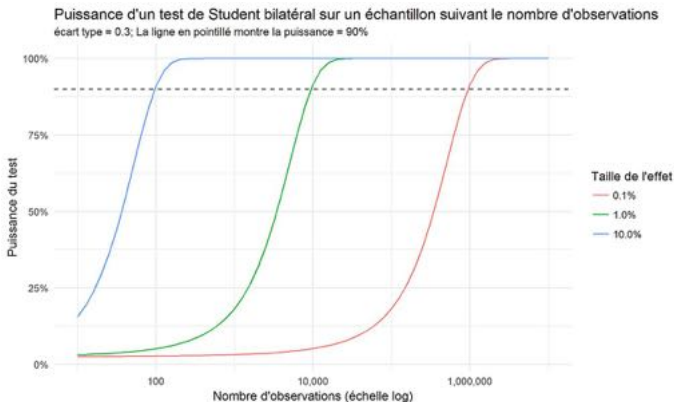
Remarque 1 : puissance statistique β permet de calculer le nb d'observations nécessaire dans une étude (on fixe β désirée, le risque de 1^{ère} espèce et les paramètres associés aux groupes).

Remarque 2 : calcul de la puissance peut s'appliquer à grand nombre de tests statistiques (comparaison de moyennes, comparaison de proportions, modèle logistique, modèle de régression, ...), lorsque l'hyp. alternative est assez restrictive.

ILLUSTRATION AVEC UN TEST DE STUDENT

Peut servir comme test sur les coefficients d'une rég. linéaire.

Même avec un effet faible (1%), on dispose souvent en assurance de + de 10 000 observ., donc d'une bonne puissance...



AUTRE FORMULATION DU MEME PB : FALSE DISCOVERY RATIO (FDR)

Le test de significativité,

$$H_0 : \hat{\beta}_k = 0 \quad \text{VS} \quad H_1 : \hat{\beta}_k \neq 0$$

est basé sur le **test de Student**, issu de la statistique $t_k = \frac{\hat{\beta}_k}{se_{\hat{\beta}_k}}$.

Cette statistique suit une loi de Student, T , à ν degrés de liberté (où $\nu = d + 1$, avec d le nombre de paramètres) : $T \sim t_\nu$.

La p-valeur du test correspond à $P(|T| > |t_k|)$.

En grande dimension, l'intérêt est **limité car le FDR est grand...**

Exemple : avec un niveau de significativité de 5%, 5% des variables sont faussement significatives !

Application : supposons que nous disposons de 100 variables explicatives, avec seulement 5 d'entre elles réellement significatives...

→ Normalement, ces 5 variables passeront le test de Student.

→ Mais 5 autres le passeront aussi (test faussement positif) \Rightarrow 10 variables sont donc détectées significatives !

\Rightarrow Le FDR est de 50% !

Pour corriger cet effet, on peut consulter [BH95]...

AUTRE EXEMPLE ET CONCLUSION

Un coefficient de corrélation égal à 0,002 est significativement différent de 0 si $n = 10^6$, mais il est totalement inutile...

“A researcher might choose to retain a causal covariate which has a strong theoretical justification even if is statistically insignificant”

“Statistical significance plays a minor or no role in assessing predictive performance. In fact, it is sometimes the case that removing inputs with small coefficients, even if they are statistically significant, results in improved prediction accuracy” (Shmueli, 2010)

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- **Notions de biais et variance d'un estimateur**
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

ERREUR D'UNE MODELISATION

On peut décomposer l'erreur dans la modélisation de $m(x)$:

Erreur de spécification + Erreur d'estimation du modèle.

→ **Erreur spécification** : vient d'hyp. sur la classe d'estimateurs de la fct m . Inmesurable par déf. puisque m inconnue.

→ **Erreur d'estimation** du modèle (si le modèle est "vrai", cad bien spécifié). Erreur d'autant + importante que la technique est compliquée et/ou nécessite beaucoup de données.

Rq : un modèle non paramétrique a une erreur de spécification $\simeq 0$, au prix d'une éventuelle inflation de l'erreur d'estimation.

DECOMPOSITION DE L'ERREUR D'ESTIMATION

Soit un estimateur $\hat{\theta}$ (var. aléatoire) de θ .

On a coutume de considérer comme mesure d'erreur d'estimation le **risque quadratique d'un estimateur** (MSE : erreur quadratique moyenne ; ou MSEP : MSE sur de nouvelles données n'ayant pas servi à construire l'estimateur), par

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

Cette erreur se décompose en 2 termes, biais et variance :

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)]^2 + Var(\hat{\theta}),$$

soit approximativement son **biais au carré plus sa variance**.

Globalement, + un modèle est complexe, + son biais diminuera et + sa variance grandira.

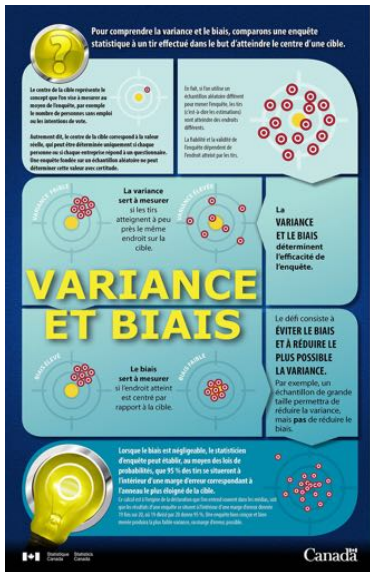
⇒ **Il faut optimiser le dosage entre biais et variance !**

⇒ Cela revient à **contrôler la complexité** du modèle !

Ex : contrôler le nb de variables (explicatives) dans le cadre paramétrique ⇒ a conduit à la déf. de critères de sélection tels que le Cp de Mallows, Akaïke (AIC), Schwartz (BIC), ...

Rq : **hormis la classe, choix du bon modèle dans une classe est primordial**. Pb d'optimisation doivent donc prendre en compte la complexité de la classe dans laquelle la solution est recherchée.

DILEMME BIAIS - VARIANCE



LIEN ENTRE CES NOTIONS

Quelque soit la méthode, tous les auteurs soulignent l'importance de construire des **modèles parcimonieux** (dimension raisonnable).

En effet + un modèle est complexe, + il est flexible \Rightarrow faible erreur d'ajustement (bon "fit") \Rightarrow synonyme d'un biais faible...

Par contre ce modèle peut s'avérer **défaillant pour généraliser**, s'appliquer à des données nouvelles (synonyme de gde variance).

\Rightarrow **Combinaison de modèles** (bagging, boosting) contourne ce pb au prix d'une \nearrow du volume de calculs et de l'interprétabilité.

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

SÉLECTION DE MODÈLE PÉNALISATION A POSTERIORI

Comment savoir où s'arrêter (i.e. à quelle dimension d) ?=

- **En régression linéaire** : un critère de qualité de modèle est le $R^2 = 1 - SCR/SCT$, issu de la décomp. de la variance :

$$\frac{1}{n} \sum_i (y_i - \bar{y})^2 = \frac{1}{n} \sum_i (y_i - \hat{y})^2 + \frac{1}{n} \sum_i (\hat{y}_i - \bar{y})^2,$$

→ terme gauche : $SCT = \sum$ carrés totaux (variance totale),

→ 1^{er} terme droite : $SCR = \sum$ carrés résid. (var. résiduelle),

→ dernier terme \simeq variance expliquée par le modèle.

⇒ Si R^2 proche de 1, le modèle est “bon”.

- **Problème** : R^2 est une **fonction croissante de d** (SCR \searrow lorsqu'on prend un modèle de dimension + grande) \Rightarrow ce critère ne suffit pas !
- **Pénalisation** : compenser la baisse naturelle de SCR par une pénalité qui défavorise les modèles de grande dimension ; i.e. on cherche à minimiser

$$SCR(d) + pen(d)$$

où $pen(d)$ est une fonction croissante de d .

- On considère le R^2 ajusté : $R^2_{adj} = R^2 - (1 - R^2) \frac{d-1}{n-d}$.
- **Voc.** : on appelle degré de liberté (d.d.l., ou d.f. en anglais pour “degree of freedom”) du modèle la quantité $(n - d)$.

CRITÈRES D'INFORMATION - PENALISATION EX-POST

Pénalisations classiques en régression linéaire / GLM :

- critère d'**Akaike** : $AIC = deviance + 2d$.
- critère d'**Akaike gde dimension** : $AIC = deviance + 2d \frac{n}{n-d-1}$.
- critère **bayésien** : $BIC = deviance + d \log n$.

→ BIC pénalise + dès que $n \geq 8$) : sélection modèles + petite dim.

→ AIC car fondé sur des hypothèses “– discutables” que BIC...

→ Critères utilisables pour comparer 2 modèles **emboîtés**.

⇒ Pb : **impossible considérer ttes collections** modèles emboîtés
(2^d) ⇒ trouver une stratégie simplificatrice !

IDÉE DU CRITÈRE AIC

Approximation de la divergence de Kullback-Leibler entre la vraie distribution f et le meilleur choix dans une famille paramétrée.

AIC **asymptotiquement optimal** lorsque l'on veut sélectionner notre modèle sur un critère d'EQM, en faisant l'hypothèse (raisonnable) que notre **modèle n'est pas bien spécifié** (données générées par un autre modèle), voir Yang (2005).

La vitesse de CV de l'AIC est optimale dans un certain sens, ce qui en fait le critère de sélection le plus utilisé en pratique !

Rq : la vraisemblance d'un modèle pas toujours calculable...

IDÉE DU CRITÈRE BIC

Bayesian Information Criterion : choix bayésien parmi des modèles paramétriques a priori équiprobables.

Le BIC est un **critère explicatif**, au contraire de l'AIC qui est un critère prédictif. Il est donc illogique de les utiliser simultanément.

- Si $n \rightarrow \infty$, alors la proba que BIC sélectionne le vrai modèle tend vers 1 (pas pour l'AIC).
- Si n fini, BIC a tendance à choisir un modèle trop simple à cause de sa pénalisation + forte...

Remarque : plus loin dans la théorie \Rightarrow cf thèse..

STRATEGIES DE SELECTION

En pratique, on ne peut pas parcourir l'ensemble des 2^p modèles possibles \Rightarrow sélection selon 3 stratégies (modèles emboîtés) :

- **sélection ascendante** : on part du modèle nul, puis on ajoute 1 à 1 les var. explicatives. La variable ajoutée à chq étape conduit à la + forte \searrow d'AIC, et arrêt lorsque l'AIC ne \searrow plus.
- **sélection descendante** : on part du modèle complexe et on supprime les variables avec un ordre défini par la même stratégie que précédemment.
- **mélange des 2** : sélection ascendante avec possibilité (à chq étape) de suppr. 1 var. déjà ajoutée en cours de process).

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- **Contraste pénalisé - pénalisation ex-ante**
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

CONTRASTE PENALISE - PENALISATION EX-ANTE

Dans cette approche, le raisonnement est **très différent** !

Au lieu d'inférer des estimateurs par les données observées, puis de pénaliser le contraste du modèle par la dimension du modèle....

... on infère des estimateurs par les données observées **en tenant compte de la pénalisation (une norme) lors de la procédure d'optimisation** ! Ce sont donc des estimateurs sous contrainte.

Rq : le **Machine Learning utilise une pénalisation ex-ante**.

RAPPELS SUR LES NORMES

En dimension finie (espace vectoriel), on rappelle quelques normes classiques d'un vecteur $\beta = (\beta_0, \beta_1, \dots, \beta_d)$:

- la norme 1 : $\|\beta\|_1 = |\beta_0| + |\beta_1| + \dots + |\beta_d|$
- la norme 2 : $\|\beta\|_2 = \sqrt{|\beta_0|^2 + |\beta_1|^2 + \dots + |\beta_d|^2}$
- la norme infinie : $\|\beta\|_\infty = \max(|\beta_0|, \dots, |\beta_d|)$.

On utilisera ces normes pour construire des **shrinkage estimators** : plutôt que de maximiser un contraste, on maximisera un contraste pénalisé (ou contraint). Voir le mémoire de Marc Delord pour alimenter cette partie.

ILLUSTRATION : CHOIX DE LA NORME L^p

Il y a plusieurs type de régularisations (pénalisations) :



$$p = \infty$$



$$p = 2$$



$$p = 1$$



$$0 < p < 1$$



$$p = 0$$

Comme vu précédemment avec les critères AIC et BIC, on retrouve ce concept de pénalisation en sélection de modèle avec de nouveaux critères : LASSO, ...

ILLUSTRATION DES NORMES EN DIMENSION 2

On considère le vecteur $\beta = (\beta_0, \beta_1)$ sur le carré unitaire. La valeur de l'aire globale des différentes normes dépend de la valeur du paramètre de régularisation λ (cf LASSO, RIDGE, ...).

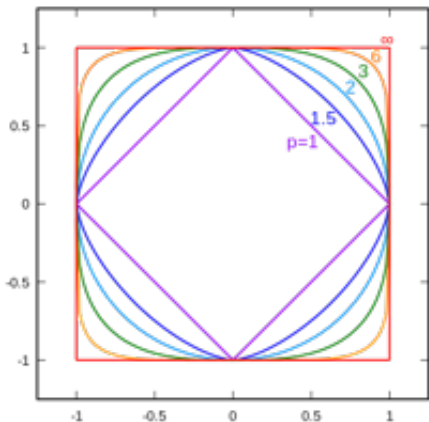
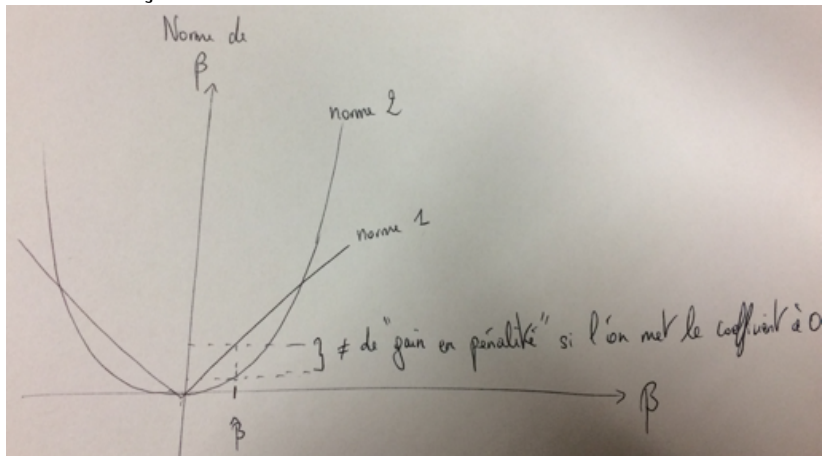


ILLUSTRATION DU GAIN EN PENALITÉ

Gain en forçant des coefficients à 0 entre les normes 1 et 2 :



⇒ LASSO forcera bien + les coef. à valoir 0 que le Ridge !

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- **Pénalisation L^2 : régression Ridge**
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

RÉGULARISATION DE TIKHONOV

Une façon de voir la régression Ridge : corriger le mauvais conditionnement de la matrice des régresseurs $X^T X$.

Estimateur Ridge : dans un modèle de régression linéaire, l'estimateur Ridge $\widehat{\beta}_R$ de β_0 est la solution du programme d'optimisation

$$\widehat{\beta}_R = \arg \min_{\beta=(\beta_1, \dots, \beta_d)} \left(\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda \sum_{j=1}^d \beta_j^2 \right)$$

Il faut bien comprendre que la pénalisation intervient ici lors de la **minimisation**, et non pas a posteriori (après avoir estimé les paramètres) comme pour AIC, ... λ : **paramètre de tuning** !

Solution explicite !

$$\widehat{\beta_R} = (X^T X + \lambda I_d)^{-1} X^T Y,$$

où I_d est l'identité en dimension d .

- **Avantages** : solution explicite, existe toujours !
- Choix λ : validations croisées en général.
- Rajouter λI_d à $(X^T X)$ permet d'améliorer le conditionnement de la matrice (on rajoute $+\lambda$ à toutes les valeurs propres).
- Estimateur biaisé mais régularisation \Rightarrow diminution variance
- **Point de vue bayésien** : Ridge = on a l'a priori que $\|\beta_0\|_2$ (norme 2 du vecteur β théorique inconnu) pas trop grande.
- **La rég. ridge ne résout pas le pb de la gde dimension** : elle ↘ de façon simultanée les coefficients de paramètres corrélés.

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- **Hypothèse de parcimonie et pénalisation L^1 : LASSO**
- La pénalisation Elastic-net

HYPOTHÈSE DE “SPARSITÉ”

Constat pragmatique : si bc de variables explicatives sont significatives, alors on ne parviendra pas à bien estimer les coefficients associés, car leur nb est proche de n .

- **Hypothèse de parcimonie** : on suppose que le nb de coef. $\beta_{0,j}$ du vecteur théorique inconnu β_0 qui sont non nuls est égal à k , avec

$$k \ll n.$$

- Si on connaissait quels sont les coefficients non nuls, on reviendrait à un modèle en petite dimension.

Remarque : rien ne permet de tester cette hypothèse !

LA RÉGRESSION LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) : méthode Ridge où l'on remplace la pénalité L^2 par une pénalité L^1 .

Estimateur LASSO : ds un modèle de rég. linéaire, l'estimateur Lasso $\widehat{\beta}_L$ de β_0 est la solution du programme d'optimisation

$$\widehat{\beta}_L = (\widehat{\beta}_{L,j})_{j=1,\dots,p} = \arg \min_{\beta} \left(\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \right).$$

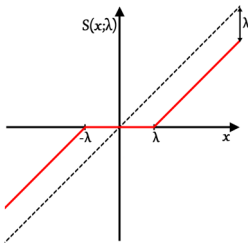
→ **Pas de solution explicite**. En particulier, fonction objectif pas différentiable (L^1 non différentiable en 0 \Rightarrow algo spécifique (LARS). Pas de résultat si $p > n$.

LE LASSO, UN EFFET DE SEUILLAGE

Si X est telle que $X^T X = Id_p$, le LASSO a une **solution explicite**.
L'estimateur correspond alors à un seuillage de la solution des moindres carrés. Notons $\hat{\beta}^{LS}$ la solution MCO, on a

$$\hat{\beta}_j^{LASSO} = \text{sign}(\hat{\beta}_j^{LS}) \times \max(0, |\hat{\beta}_j^{LS}| - \lambda).$$

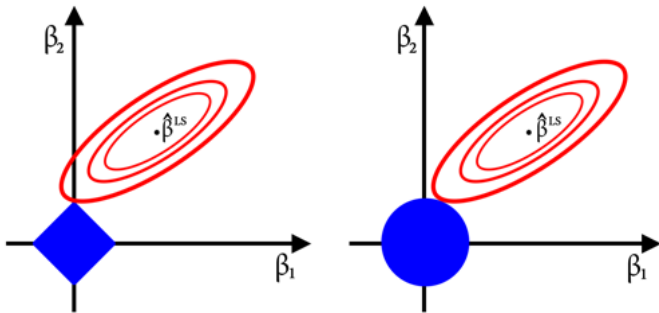
Fonction de seuillage ($\hat{\beta}$ en fonction de β) :



REMARQUES

- λ choisi par cross-validation (ou via un échantillon de test).
- En pratique : fixer seuil min. pour chq β_k en-dessous duquel il sera considéré nul $\Rightarrow ||\beta||$ petite, fait ressortir coef. importants.
- **Fournit un résultat “sparse”** (i.e. avec beaucoup de coefficients nuls) **si l'hypothèse de parcimonie est vérifiée.**
- Théoriquement, proba. que $\widehat{\beta}_{L,j} = 0$ lorsque $\beta_{0,j} \neq 0$ est faible. Sauf parfois : LASSO adaptatif résoud partiellement pb en introduisant des poids sur les coef dans la pénalisation !
- Plutôt que de réduire les coef. corrélés (LASSO assez indifférent aux corrélations), tendance à juste en retenir un.
- Estimation des coef. dépendante des ordre de grandeur : centrer et réduire toutes les covariables (idem Ridge).

LASSO/RIDGE : OPTIMISATIONS SOUS CONTRAINTE



Cas régression linéaire (ellipse : densité gaussienne du MLE) :
En bleu, zones contrainte (pénalité lasso à gauche, ridge à droite).
En rouge : contours de la fonction d'erreur des MCO.

1 Introduction au problème statistique : fléau de la dimension et convergence des estimateurs, réduction de dimension

- Motivation statistique des modèles d'apprentissage
- Estimation et grande dimension
- Extrapolation de la statistique classique (échantillons raisonnables) aux grands échantillons
- Notions de biais et variance d'un estimateur
- Réduction de dimension : pénalisation ex-post
- Contraste pénalisé - pénalisation ex-ante
- Pénalisation L^2 : régression Ridge
- Hypothèse de parcimonie et pénalisation L^1 : LASSO
- La pénalisation Elastic-net

ELASTIC-NET : RIDGE + LASSO

Estimateur Elastic-net : dans un modèle de régression linéaire, l'estimateur Elastic-Net $\widehat{\beta}_{EN}$ de β_0 est la solution du programme d'optimisation

$$\widehat{\beta}_{EN} = \arg \min_{\beta} \left(\sum_{i=1}^n (Y_i - \beta^T X_i)^2 + \lambda_1 \sum_{j=1}^d |\beta_j| + \lambda_2 \sum_{j=1}^d \beta_j^2 \right)$$

où les β_j sont les composantes de β .

- La pénalité L^1 conduit à un **modèle sparse**.
- La pénalité L^2 **enlève la limitation sur le nb des variables**.
- Nécessite le choix de deux paramètres.

APPLICATION POUR MIEUX INTERPRÉTER

On procède par [simulation](#).

Pour comparer les \neq possibilités de pénalisation, on considère :

- un jeu d'entraînement :
 - 100 individus,
 - un modèle linéaire, donc une erreur gaussienne,
 - et 20 variables explicatives candidates (pas d'info sur la simulation de ces covariables).
- un jeu test : 50 individus, même modèle que pour le jeu d'entraînement.

Rq : le TP vous aidera à mieux interpréter ces pénalisations...

Variable	Modèle	Régression	Stepwise	Ridge	LASSO	Elastic Net
Intercepte	2.4	-1.34	-2.58	-3.88	-2.40	-2.38
X2		-0.71		-0.42		
X3		0.68		0.06		
X4		-0.35		-0.60		
X5	-5	-17.13	-17	-15.18	-15.63	-15.58
X6		1.66	1.83	0.63	0.35	0.33
X7	5	14.42	14.43	12.26	12.81	12.75
X8		0.17		-0.54		
X9		-1.26	-1.21	-1.14	-0.24	-0.24
X10		1.43	1.48	0.89		
X11	-3	-46.22	-46.21	-42.46	-44.72	-44.62
X12		1.30	1.30	1.18		
X13		-0.06		0.08		
X14	3	46.71	46.47	42.86	44.52	44.42
X15		-0.45		-0.68		
X16		0.02		0.84		
X17		3.48	3.04	2.96	0.71	0.71
X18	-3	-6.38	-6.21	-6.66	-4.13	-4.16
X19		-1.68		-0.81		
X20		-1.24		0.14		
X21		0.15		0.15		
Erreur Test		-0.86	-0.76	-0.66	-1.14	-1.14
Écart-type		18.59	18.29	18.87	17.23	17.24

Rq : “Modèle” = vrai modèle ; “Stepwise” par AIC.

RÉGRESSIONS NON LINÉAIRES

Les pénalisations Ridge, Lasso, Elastic-Net s'adaptent parfois à des modèles de régression non linéaires (ex : GLM).

Par ex., la rég. logistique ($Y = \{0, 1\}$), et $\log\left(\frac{E[Y|X]}{1-E[Y|X]}\right) = \beta_0^T X$.

- Estim. du max. de vraisembl. : $\text{logit}^{-1}(u) = e^u / (1 + e^u)$,

$$\hat{\beta} = \arg \max_{\beta} \sum_{i=1}^n Y_i \log(\text{logit}^{-1}(\beta^T X_i)) + (1 - Y_i) \log(1 - \text{logit}^{-1}(\beta^T X_i)).$$

- Estimateur pénalisé correspondant :

$$\hat{\beta} = \arg \max_{\beta} \sum_{i=1}^n Y_i \log(\text{logit}^{-1}(\beta^T X_i)) + (1 - Y_i) \log(1 - \text{logit}^{-1}(\beta^T X_i)) + \text{pen}(\beta)$$

ZONES D'OMBRE

Il y a malheureusement encore et toujours des zones d'ombre sur l'utilisation de ces techniques...

- **Comment valider l'hypothèse de parcimonie ?**
- Convergence des estimateurs lorsque l'hypothèse de parcimonie est violée ?
- Comme toujours dans ce type de techniques, les méthodes pour choisir les constantes λ de pénalisation sont plus ou moins discutées...

Remarque : en R, on utilise la librairie `glmnet` pour ces méthodes.

CONCLUSIONS SUR CES MÉTHODES PARCIMONIEUSES

- Des hypothèses lourdes... mais indispensables pour pouvoir travailler.
- Adapté à de nombreux problèmes (cf Travaux Pratiques).
- Algorithmes d'optimisation rapides.
- Autres types de pénalisations existent avec propriétés voisines (exemple : pénalisation SCAD).
- Le choix de la forme de pénalisation la plus adaptée est aussi sujette à question.

EXTENSIONS EN FONCTION DES COVARIABLES

Adapter la pénalisation en fct du type des covariables...

Matching regularization to type of risk factor

- Ordinal risk factors: fused lasso

$$\sum_j w_j |\beta_{j+1} - \beta_j|.$$

- Nominal risk factors: generalized fused lasso

$$\sum_{i>k} w_{i,k} |\beta_i - \beta_k|.$$

- Spatial risk factor: graph guided fused lasso

$$\sum_{(i,k) \in \mathcal{G}} w_{i,k} |\beta_i - \beta_k|.$$