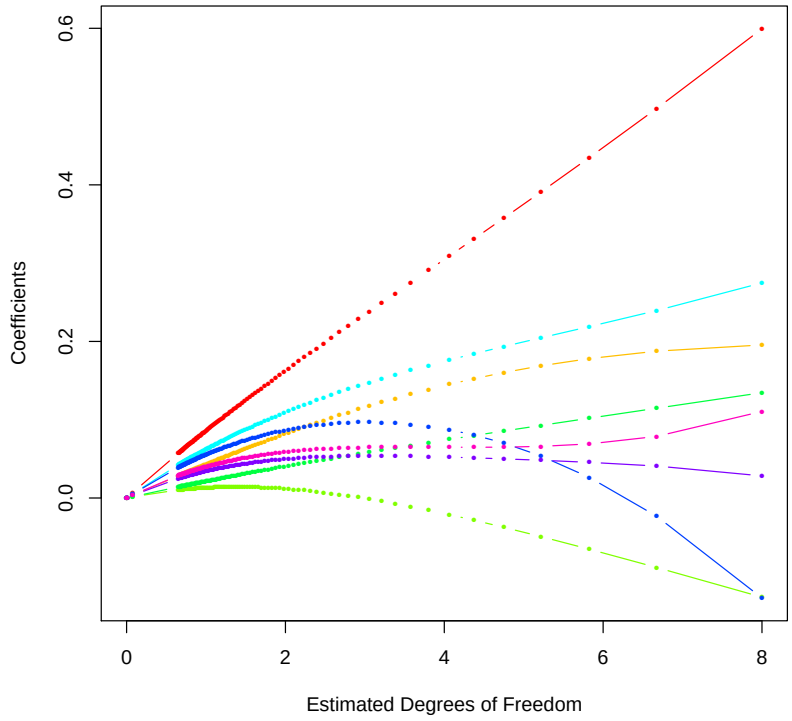
Méthodes de Régression Avancées



Yannig Goude

EDF R&D - Université Paris-Saclay

Ridge Regression

Comme nous l'avons vu précédemment, l'estimateur des **M**oindres **C**arrés **O**rdinaires (MCO) est la solution de:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i \beta)^2 = \min_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2$$

Et s'écrit:

$$\widehat{\beta} = (X'X)^{-1} X'Y$$

sous l'hypothèse que X est de plein rang

Rq: on considèrera ici que les variables sont centrées pour plus de commodité.

En pratique, cet estimateur ne fonctionne pas:

- si les $x_{.,j}$ sont corréléees entre elles, X n'est pas de plein rang
- si $p >> n$

dans ces cas $X'X$ doit être régularisée pour pouvoir être inversée et on ajoute une pénalisation $\rightarrow$ ridge, lasso...

Régression Ridge:

On résout le pb:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda ||\beta||^2$$

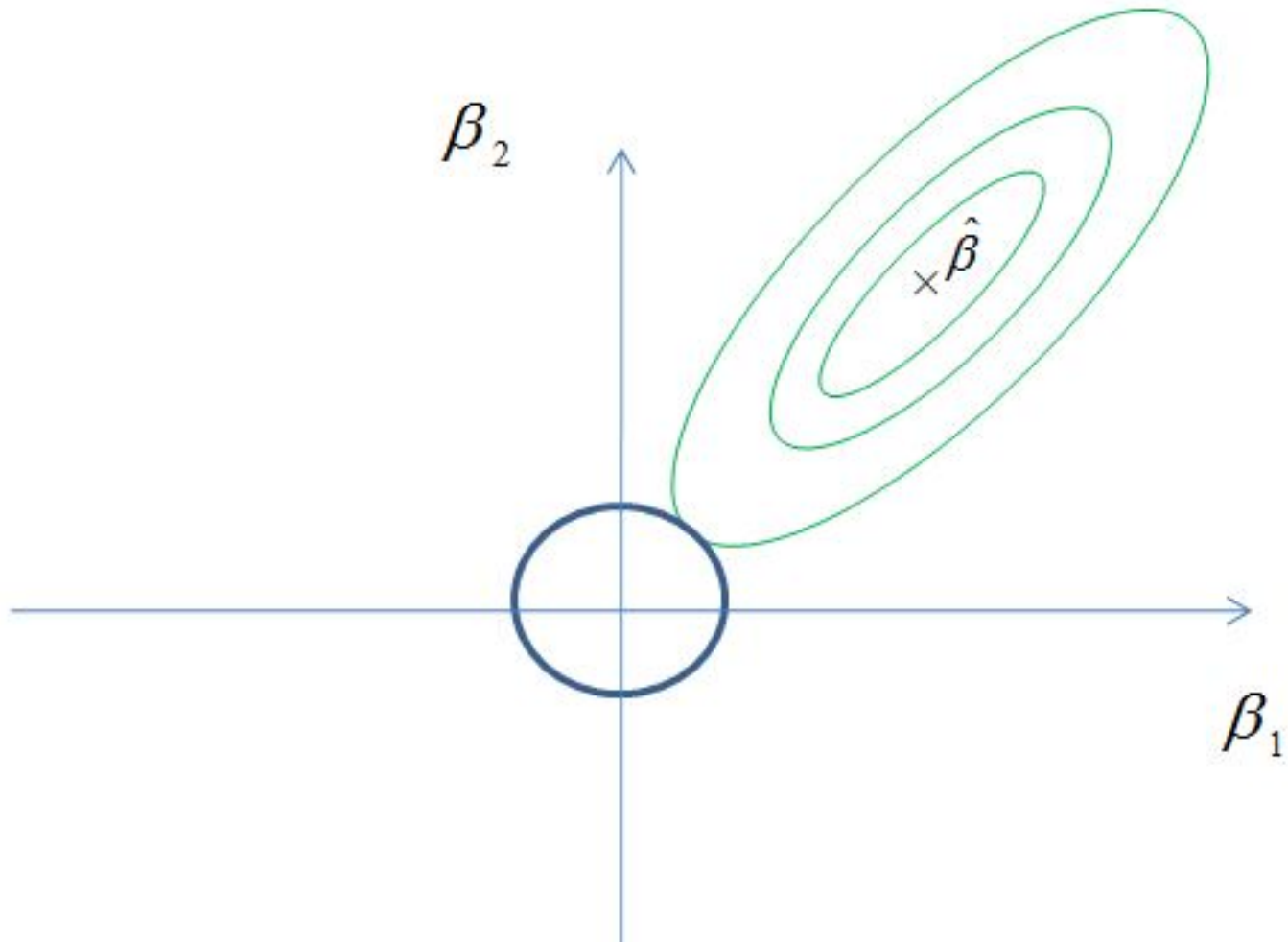
équivalent au pb suivant -dual, Lagrange-:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

$$\text{s. c. } ||\beta||^2 \leqslant t$$

Rq:

- bijection entre t et λ
- les solution du pb ne sont pas invariante par changement d'échelle, usuellement on standardise les variables avant
- les variables sont centrées ($\sim$ on ne pénalise pas la constante)



l'estimateur des coefficients de la régression ridge est donné par:

$$\widehat{\beta}_{ridge} = (X'X + \lambda I)^{-1} X'Y$$

C'est un estimateur **biaisé** de $\widehat{\beta}$!

- $$\mathbb{E}(\widehat{\beta}_{ridge}) = (X'X + \lambda I)^{-1} X'X \mathbb{E}(\widehat{\beta}) = \beta - \lambda(X'X + \lambda I)^{-1}\beta$$
- $$Var(\widehat{\beta}_{ridge}) = \sigma^2 (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1}$$

Preuve:

- $g(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta'\beta$
- $\frac{\mathrm{d}g}{\mathrm{d}\beta} = 2'(X\beta - Y) + 2\lambda\beta$
- s'annule en $(X'X + \lambda)\beta = X'Y$
- soit en $\beta = (X'X + \lambda)^{-1}X'Y$

Preuve:

- $g(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta'\beta$
- $\frac{\mathrm{d}g}{\mathrm{d}\beta} = 2'(X\beta - Y) + 2\lambda\beta$
- s'annule en $(X'X + \lambda)\beta = X'Y$
- soit en $\beta = (X'X + \lambda)^{-1}X'Y$

Preuve:

- $g(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta'\beta$
- $\frac{\mathrm{d}g}{\mathrm{d}\beta} = 2'(X\beta - Y) + 2\lambda\beta$
- s'annule en $(X'X + \lambda)\beta = X'Y$
- soit en $\beta = (X'X + \lambda)^{-1}X'Y$

Preuve:

- $g(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta'\beta$
- $\frac{\mathrm{d}g}{\mathrm{d}\beta} = 2'(X\beta - Y) + 2\lambda\beta$
- s'annule en $(X'X + \lambda)\beta = X'Y$
- soit en $\beta = (X'X + \lambda)^{-1}X'Y$

- $\widehat{\beta}_{ridge}$ est biaisé, son biais vaut $-\lambda(X'X + \lambda I)^{-1}\beta$
- l'inversion de $X'X$ est remplacée par celle de $(X'X + \lambda I)$, estimation plus robuste si les variables explicatives sont corrélées empiriquement

Intuition:

- pénaliser permet de réduire l'espace S: réduire la variance (grande dans les cas évoqué) mais augmente le biais
- le compromis biais-variance revient à trouver la bonne pénalité λ

Choix de λ: VC, GCV, pénalisation (AIC, BIC, Cp), rééchantillonage, bootstrap, analyse du λ-path...

Degrés de liberté estimés

On généralise la notion de degrés de liberté du linéaire

- sans pénalisation: $\widehat{\beta}_0 = (X^T X)^{-1} X^T y$
- avec pénalisation: $\widehat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$
- $\widehat{\beta}_\lambda = F_\lambda \widehat{\beta}_0$

Où

$$F_\lambda = (X^T X + \sum \lambda I)^{-1} X^T X$$

$tr(F_\lambda)$: degrés de liberté estimés

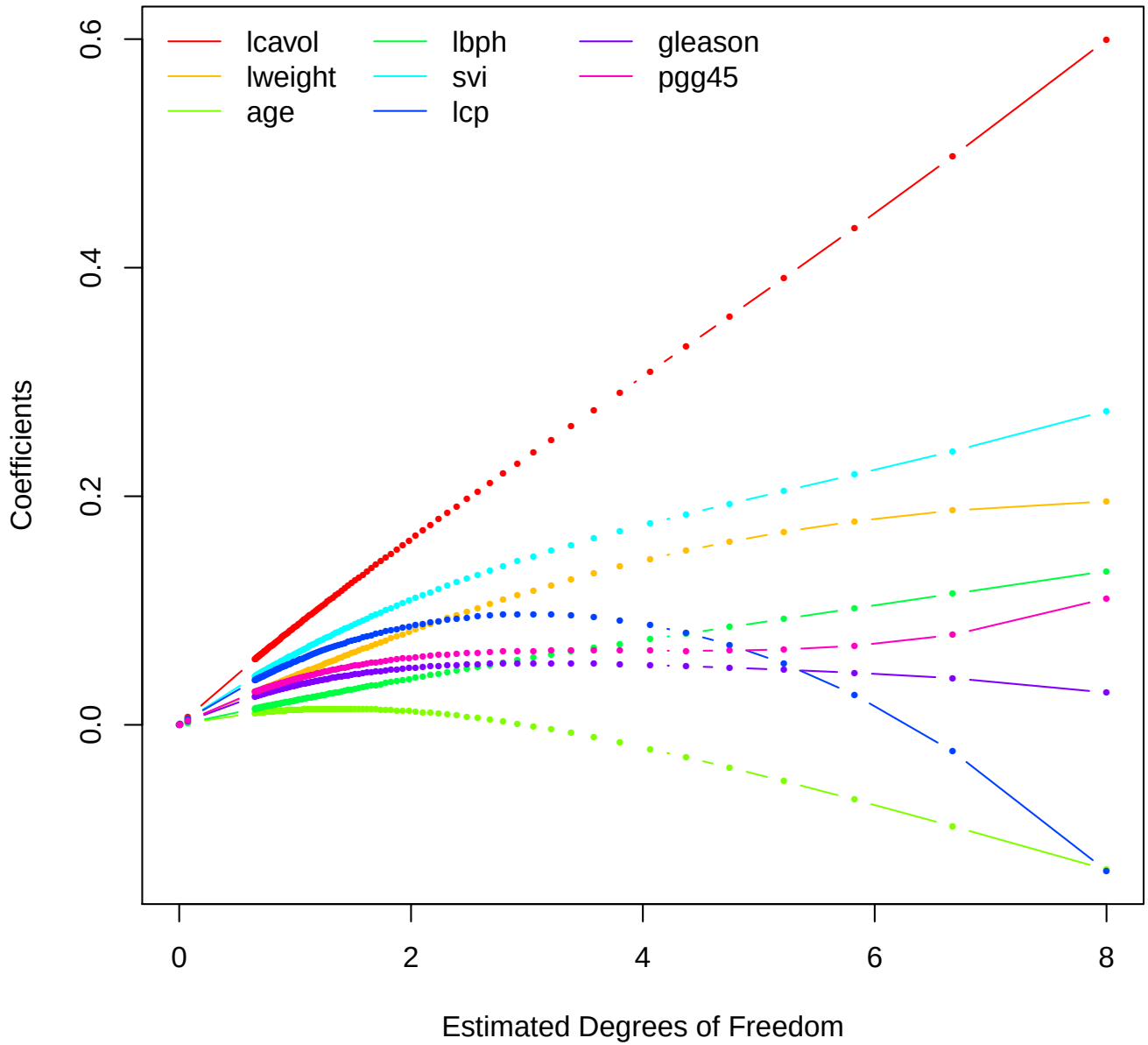
on a également par simple calcul matriciel: $tr(F_\lambda) = tr(H_\lambda)$, ou $H_\lambda = X(X'X + \lambda I)^{-1} X'$

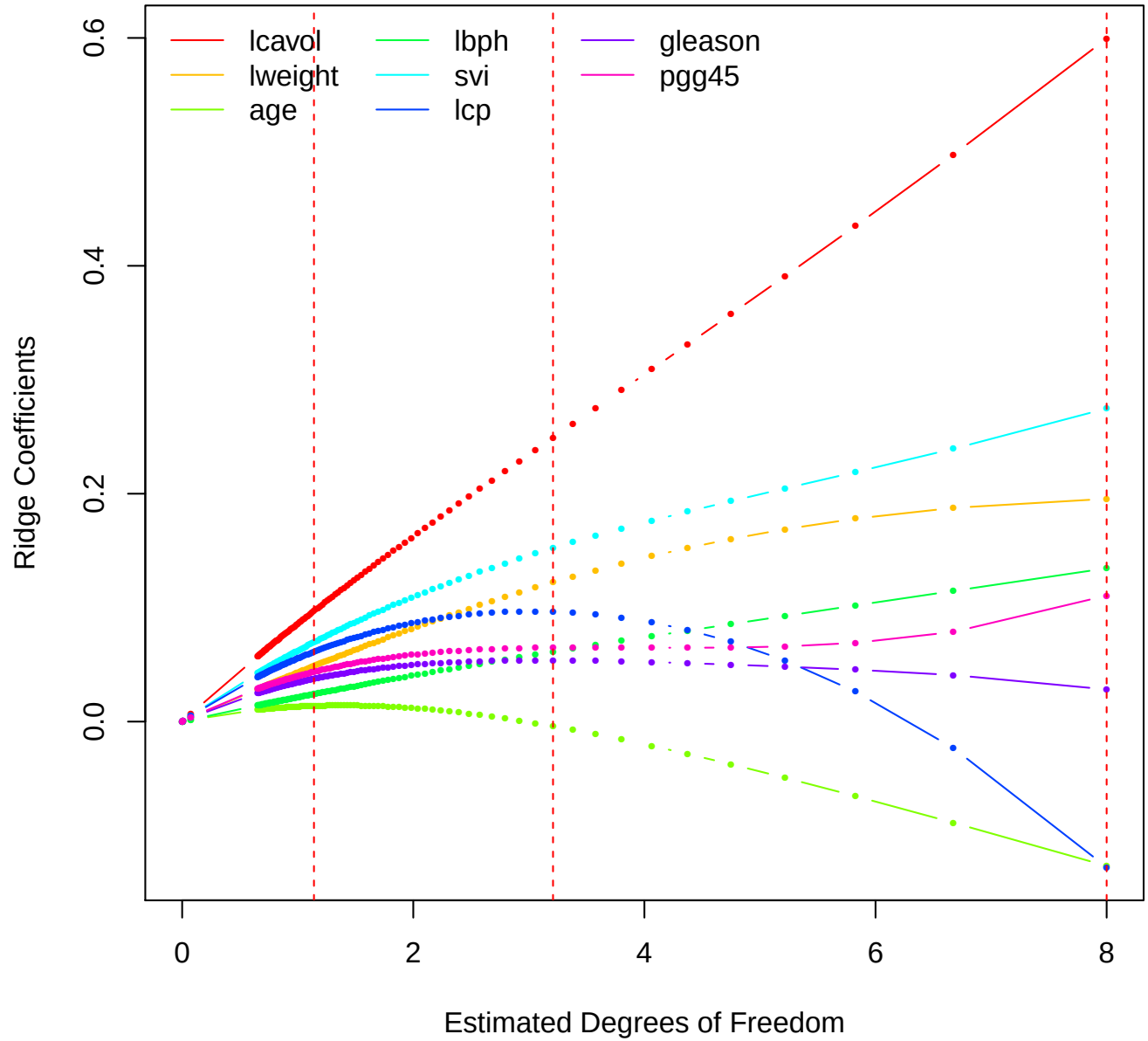
Exemple:Prostate Data -voir Hastie, Tibshirani, Friedman 2009-

réponse: lpsa log(prostate specific antigen)

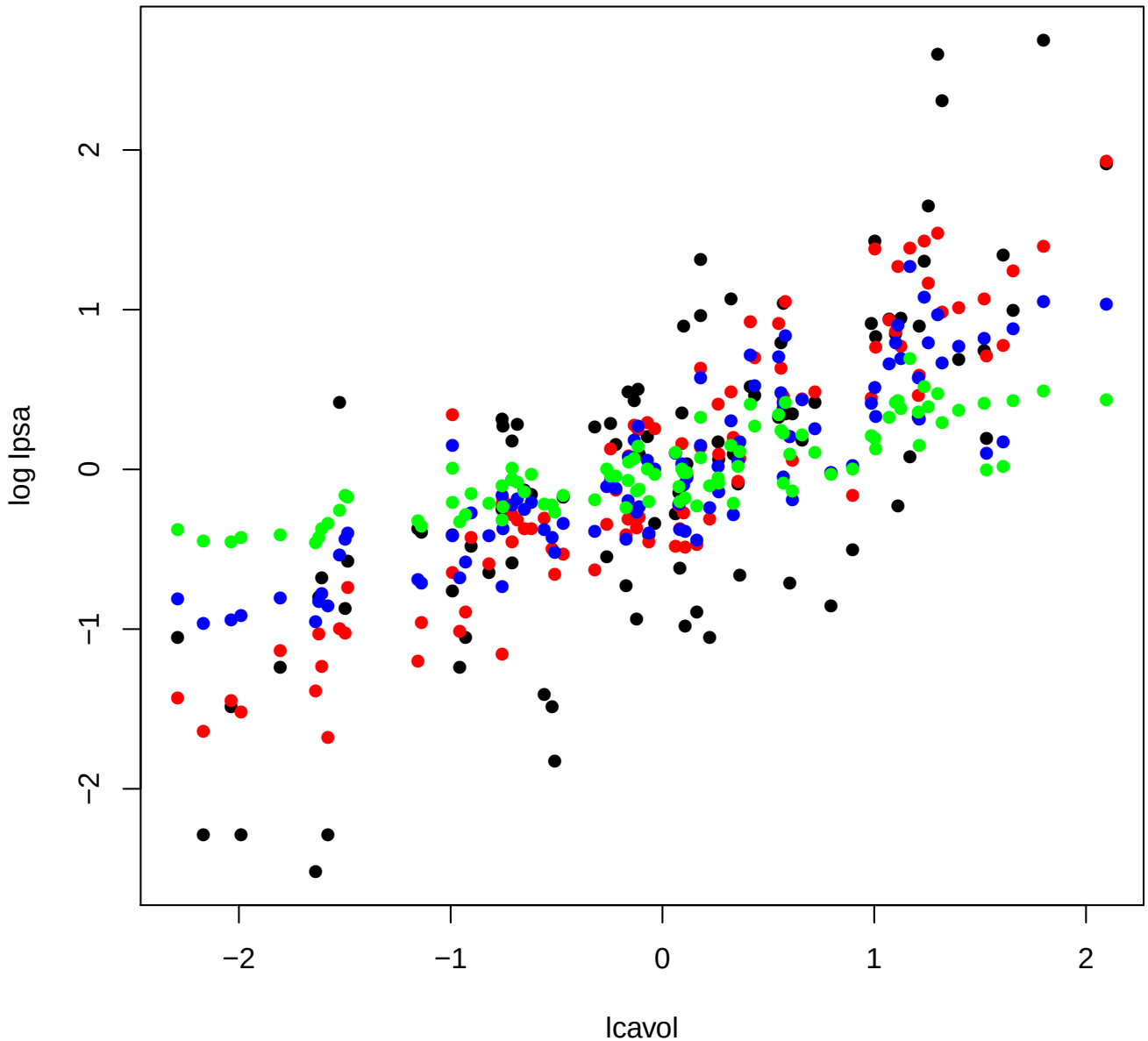
var. expl.: lcavol log(cancer volume); lweight log(prostate weight); age; lbph log(benign prostatic hyperplasia amount); svi seminal vesicle invasion; lcp log(capsular penetration);gleason Gleason score; pgg45 percentage Gleason scores 4 or 5

Exemple de λ-path obtenu sur ces données




lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Ridge Coefficients
0.0
0.2
0.4
0.6
0
2
4
6
8
Estimated Degrees of Freedom

Exemples:



Lasso Regression

Régression Lasso -least absolute shrinkage and selection operator-:

On résout le pb:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda |\beta|$$

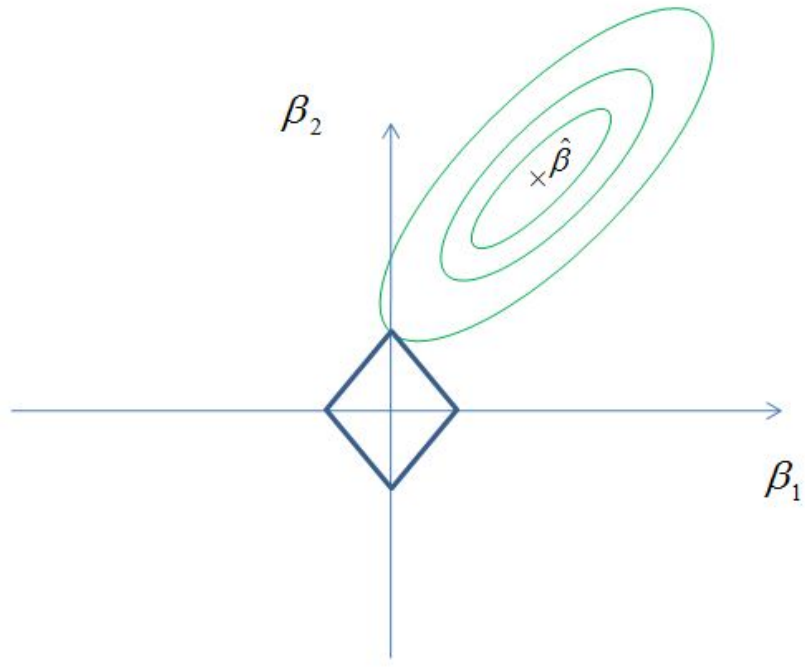
Avec $|\beta| = \sum_{j=1}^{p} |\beta_j|$

équivalent au pb suivant:

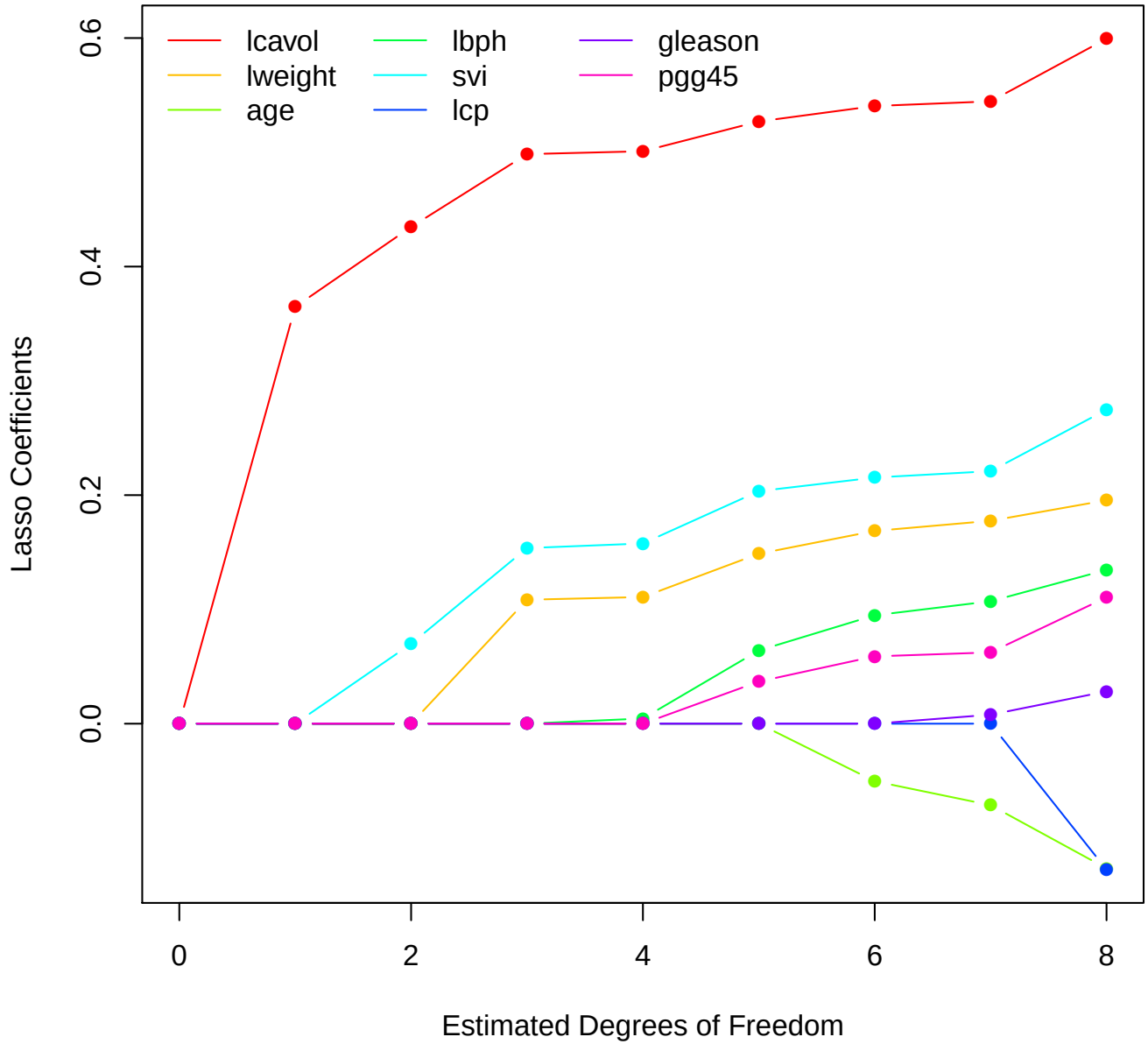
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i \beta)^2$$

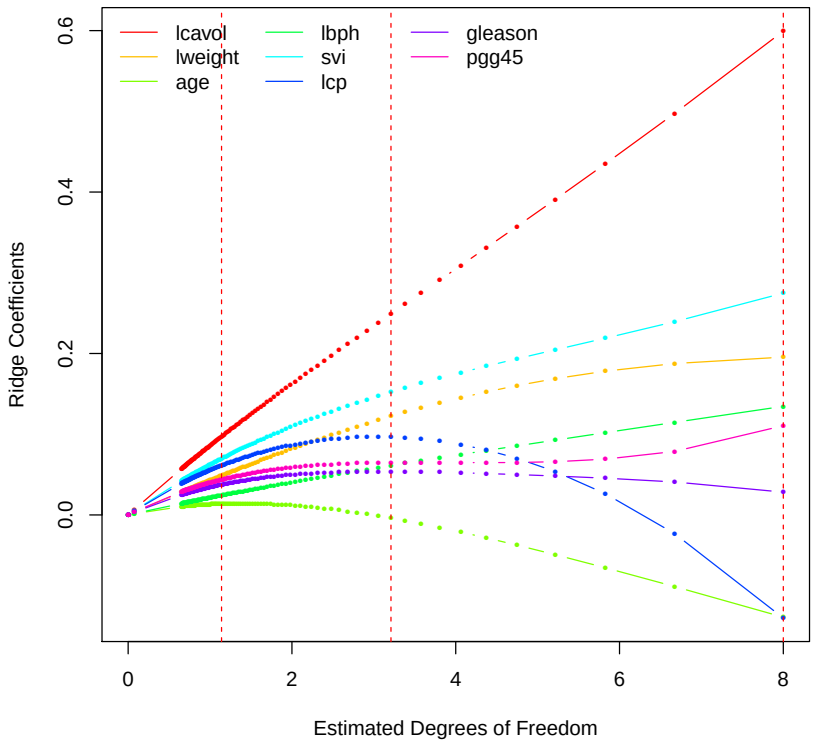
$$\text{s. c. } |\beta|^2 \leqslant t$$

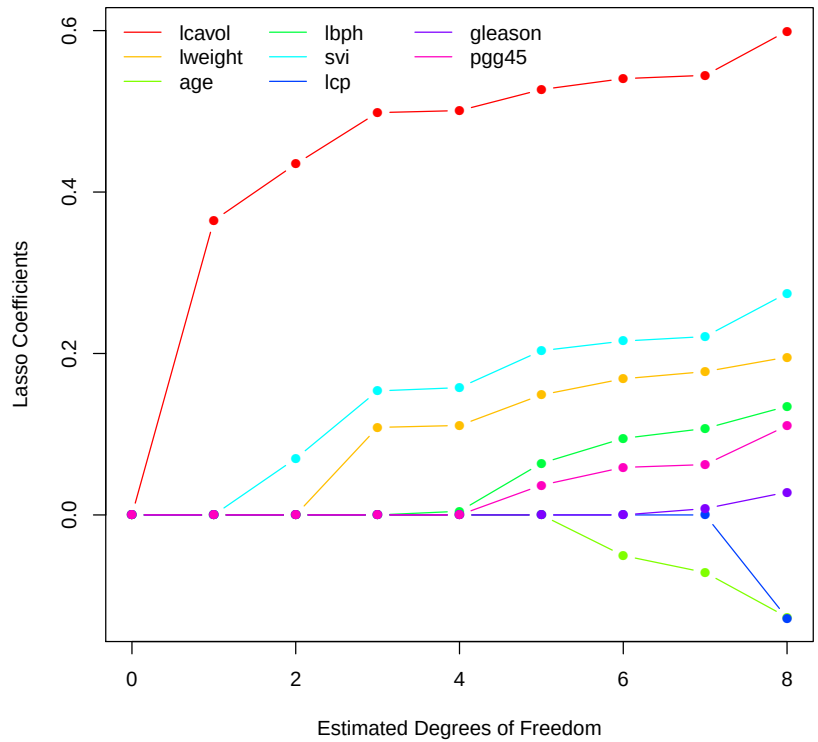
Problème similaire à ridge mais la pénalité L_2 de ridge est ici remplacée par une pénalité en norme L_1: la solution de ce problème n'est plus linéaire en y



La pénalité L_1 a comme propriété de "tronquer" les coefficients faibles, donc de les mettre à 0. Cela permet une sorte de choix de modèle.


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Lasso Coefficients
0.0
0.2
0.4
0.6
0
2
4
6
8
Estimated Degrees of Freedom


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Ridge Coefficients
0.0
0.2
0.4
0.6
0
2
4
6
8
Estimated Degrees of Freedom


lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Lasso Coefficients
0.0
0.2
0.4
0.6
0
2
4
6
8
Estimated Degrees of Freedom

Algorithme LARS: Least Angle Regression

l'algorithme LARS a été introduit par Efron et Al. (2004). Il permet d'obtenir rapidement le λ-path.

L'intuition derrière l'algorithme est la même que pour la régression stepwise: à chaque étape on identifie la "meilleure" variable à inclure dans le modèle -active set-.

Il s'inspire de la "stagewise regression". On suppose ici que les variables y et x sont centrées réduites, on note $\widehat{\mu} = X\widehat{\beta}$.

1. initialisation à $\widehat{\mu} = 0$
2. calcul des corrélations $\widehat{c} = X'(y - \widehat{\mu})$ puis de $\widehat{j} = \operatorname{argmax} |\widehat{c}_j|$
3. mise à jour $\widehat{\mu} \leftarrow \widehat{\mu} + \varepsilon \operatorname{sign} \widehat{c}_j$
4. retour à l'étape 1

avec ε une "petite" constante à choisir.

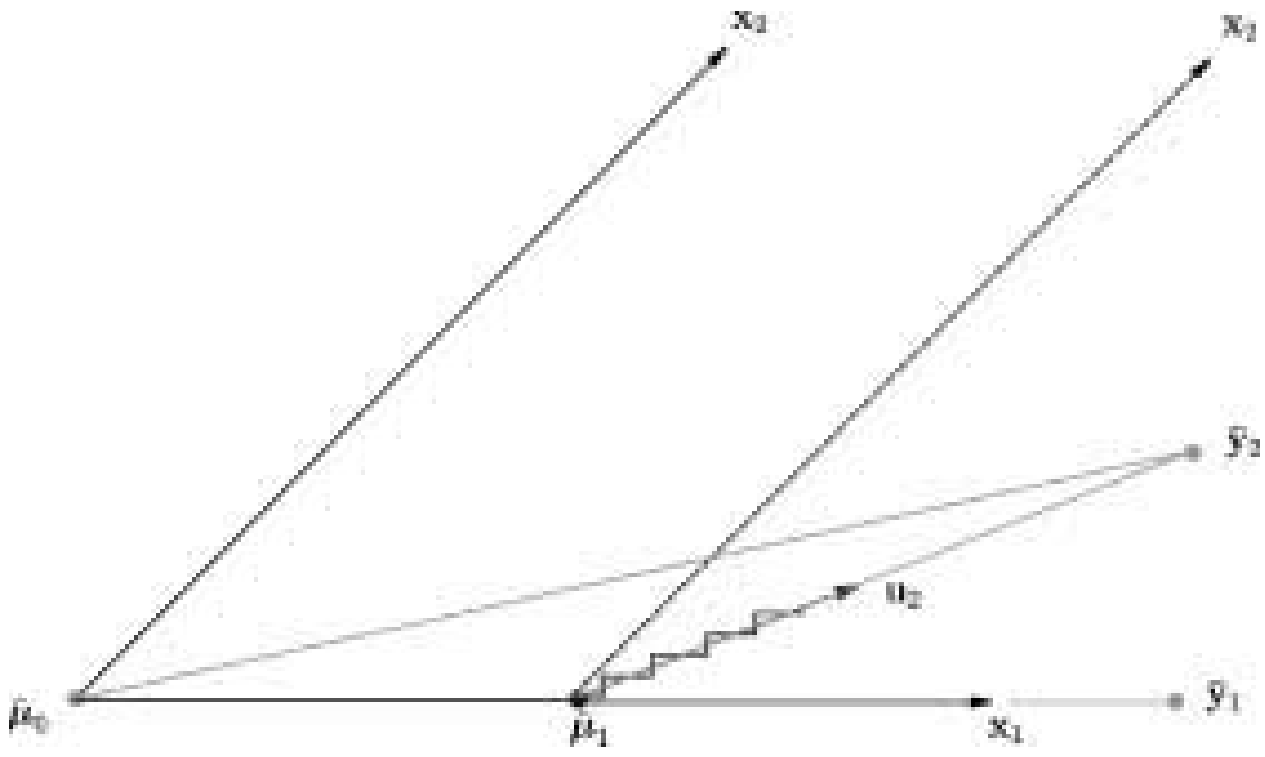
Algorithme LARS: Least Angle Regression

1. initialisation à $\widehat{\mu} = 0$, $\widehat{\beta}_0 = 0$
2. $r = y - \widehat{\mu}$, x_j la variable la plus corrélée à r
3. faire varier β_j de 0 dans la direction $c_j = x_j' r$ jusqu'à ce qu'une variable x_k soit plus corrélée à r que x_j, $|c_k| \geqslant |c_j|$
4. faire varier β_j et β_k (de 0) dans la direction $\delta = (X'X)^{-1}X'r$ jusqu'à ce qu'une variable x_l soit plus corrélée à r

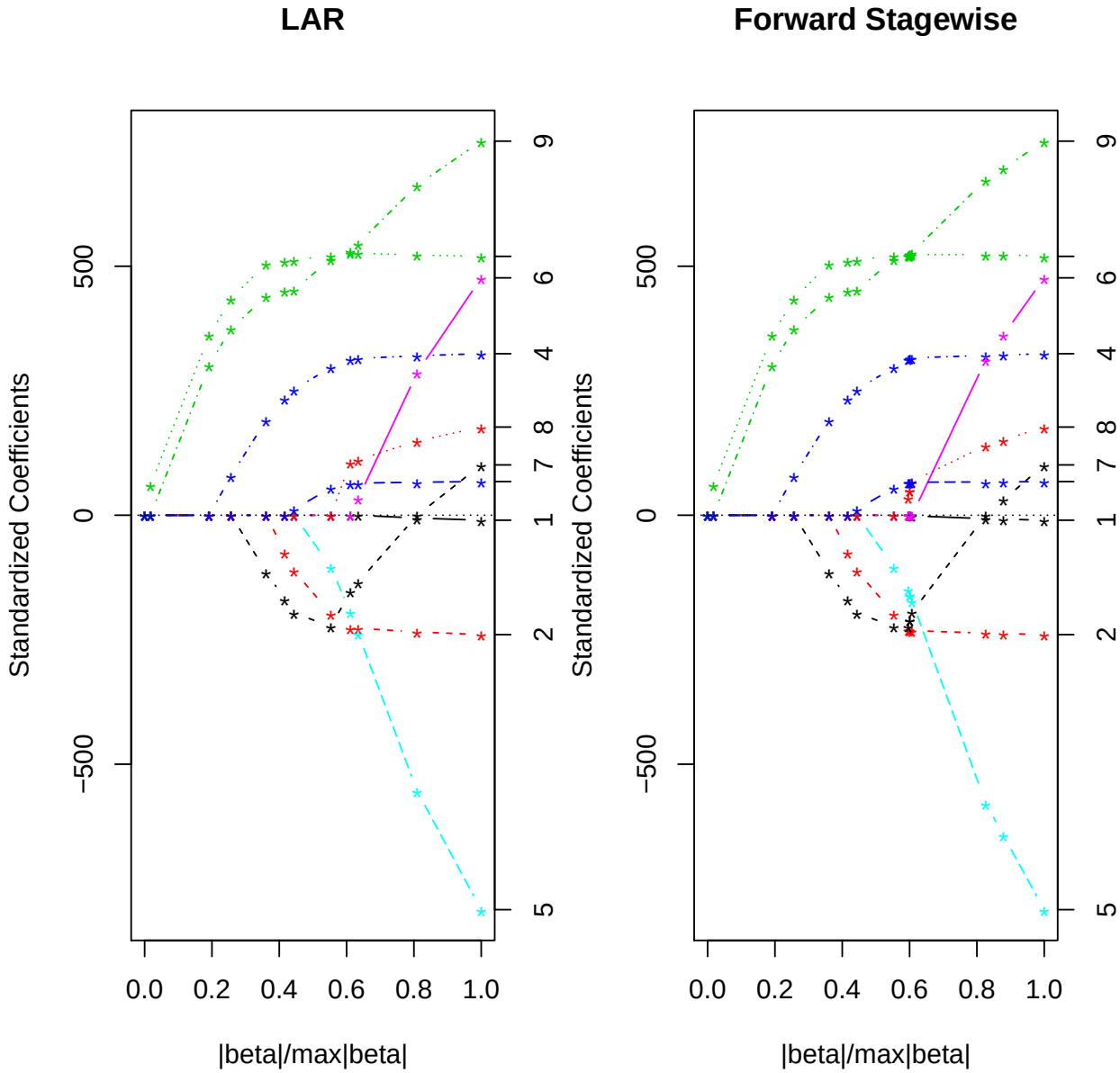
Si $\mathcal{A}_k$ l'ensemble des variables inclues dans le modèle à l'étape k de l'algorithme, $\widehat{\beta}_{\mathcal{A}_k}$ le vecteur de coefficients associé.

$r_k = y - X_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ le résidu à cette étape et $\delta_k = (X'_{\mathcal{A}_k}X_{\mathcal{A}_k})^{-1}X'_{\mathcal{A}_k}r_k$. Les coefficients évoluent ainsi: $\beta_{\mathcal{A}_k} = \beta_{\mathcal{A}_k} + \gamma\delta_k$.

- l'évolution des coefficients est donc linéaire par morceau, à chaque morceau correspond l'ajout d'une nouvelle variable
- γ, le "pas" de l'algorithme à chaque étape est calculable à l'avance (dépend de la covariance des variables): gain d'efficacité par rapport à la reg. stagewise



Efron,et al. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.


LAR
Forward Stagewise
Standardized Coefficients
500
0
−500
9
6
4
8
7
1
2
5
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|

Régression sur composantes principales

Régression sur composantes principales: PCR

Le principe de cette méthode est de:

- calculer les vecteurs propres de la matrice $X'X/n$
- classer ces vecteurs par leurs valeurs propres
- sélectionner le nombre de vecteurs propres (le nombre de "variables" dans la régression) à inclure dans le modèle

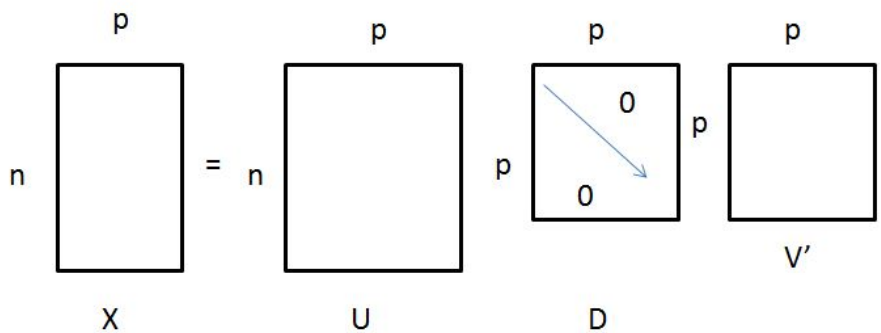
SVD

On transforme ici la matrice des var. explicatives X (n lignes, p colonnes) en effectuant une décomposition en valeurs singulières (SVD, pour Singular Value Decomposition).

$$X = UDV'$$

avec

- U matrice orthogonale $n \times p$
- V matrice orthogonale $p \times p$
- D matrice diagonale dont les éléments d_i sont $d_1 \geqslant d_2 \geqslant ... \geqslant d_p \geqslant 0$
- si au moins une valeur $d_j = 0$, alors X est singulière



Alors la régression linéaire de Y sur X s'écrit:

$$\widehat{y} = X\widehat{\beta} = X(X'X)^{-1}X'Y = UU'Y$$

U étant une base orthogonale de l'espace engendré par X, $U'Y$ les coefficients de la projection de Y sur cette base

Les composantes principales de X sont les vecteurs propres de la matrice de covariance: $X'X/n$ (rappelons que l'on a centré les variables).

Celle-ci s'exprime, après décomposition SVD et à un facteur $1/n$ près:

$$X'X = (VDU')(UDV') = VD^2V'$$

Les vecteurs propres v_j sont également appelés direction de karhunen-loeve de X.

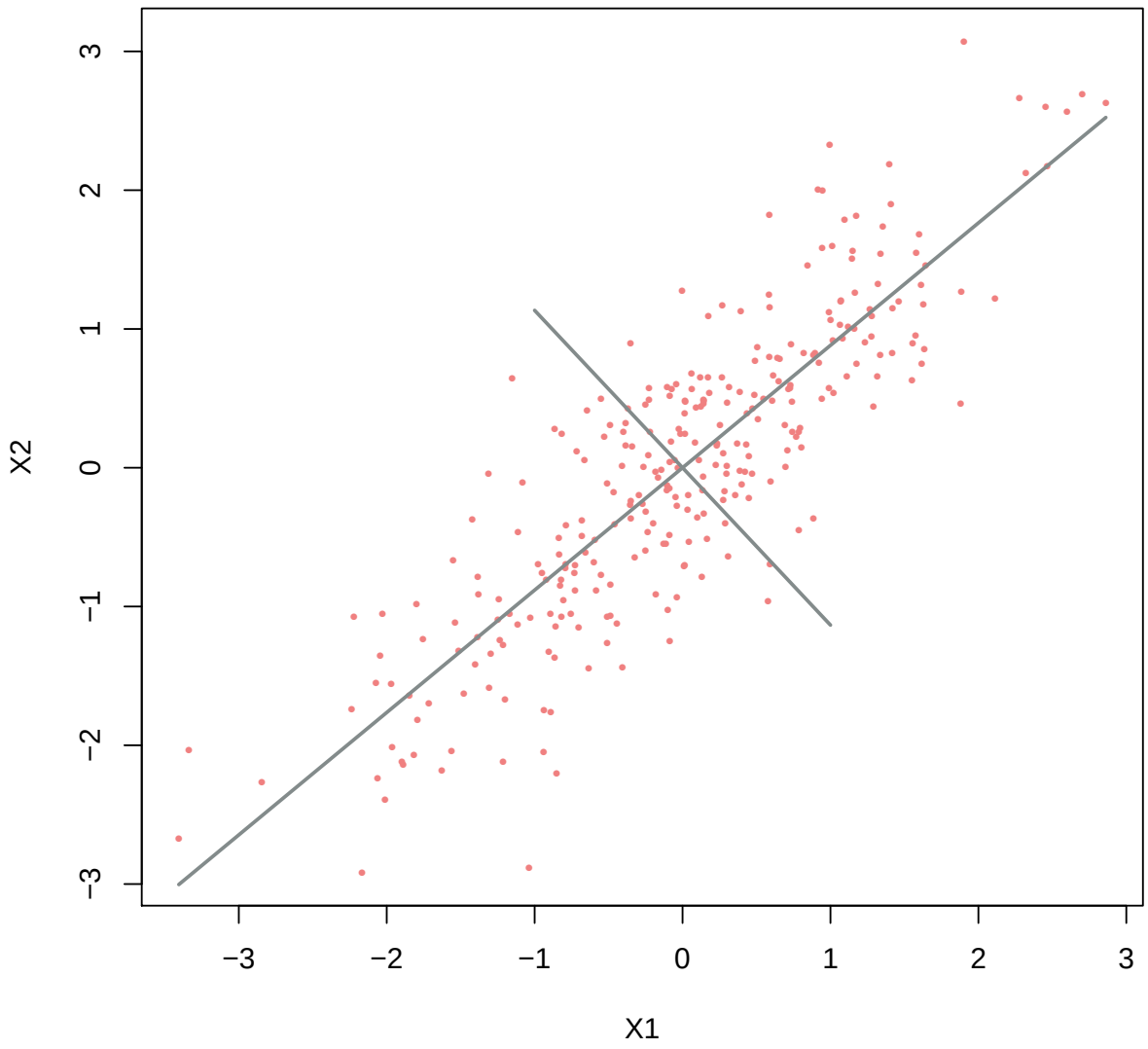
Avantage: simplification, accélération des calculs, possibilité de conserver seulement les axes "importants" de la matrice de covariance

La projection de X sur la j^e composante principale v_j vaut:

$$z_j = Xv_j = UDV'v_j = u_j d_j$$

on a aisément: $\mathrm{Var}(z_1) \geqslant \mathrm{Var}(z_2) \geqslant ... \geqslant \mathrm{Var}(z_p)$

La première composante d'un jeux de données X est la direction qui maximise la variance des données projetées.


X2
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
X1

Pour la régression sur composante principale, on procède ainsi:

- calcul de la SVD de X
- les composantes principales sont obtenus par: $z_j = Xv_j = u_j d_j$
- les coefficients de projection sur ces composantes sont donnés par: $\beta = D^{-1} U' Y$
- on sélectionne les "premières composantes": graphe des valeurs propres d_j puis critère du coude par exemple, sélection de modèles emboités

Lien avec la régression ridge

On a vu que pour la régression ridge:

$$X\widehat{\beta}_{ridge} = X(X'X + \lambda I)^{-1}X'y$$

en remplaçant par la SVD de X on a:

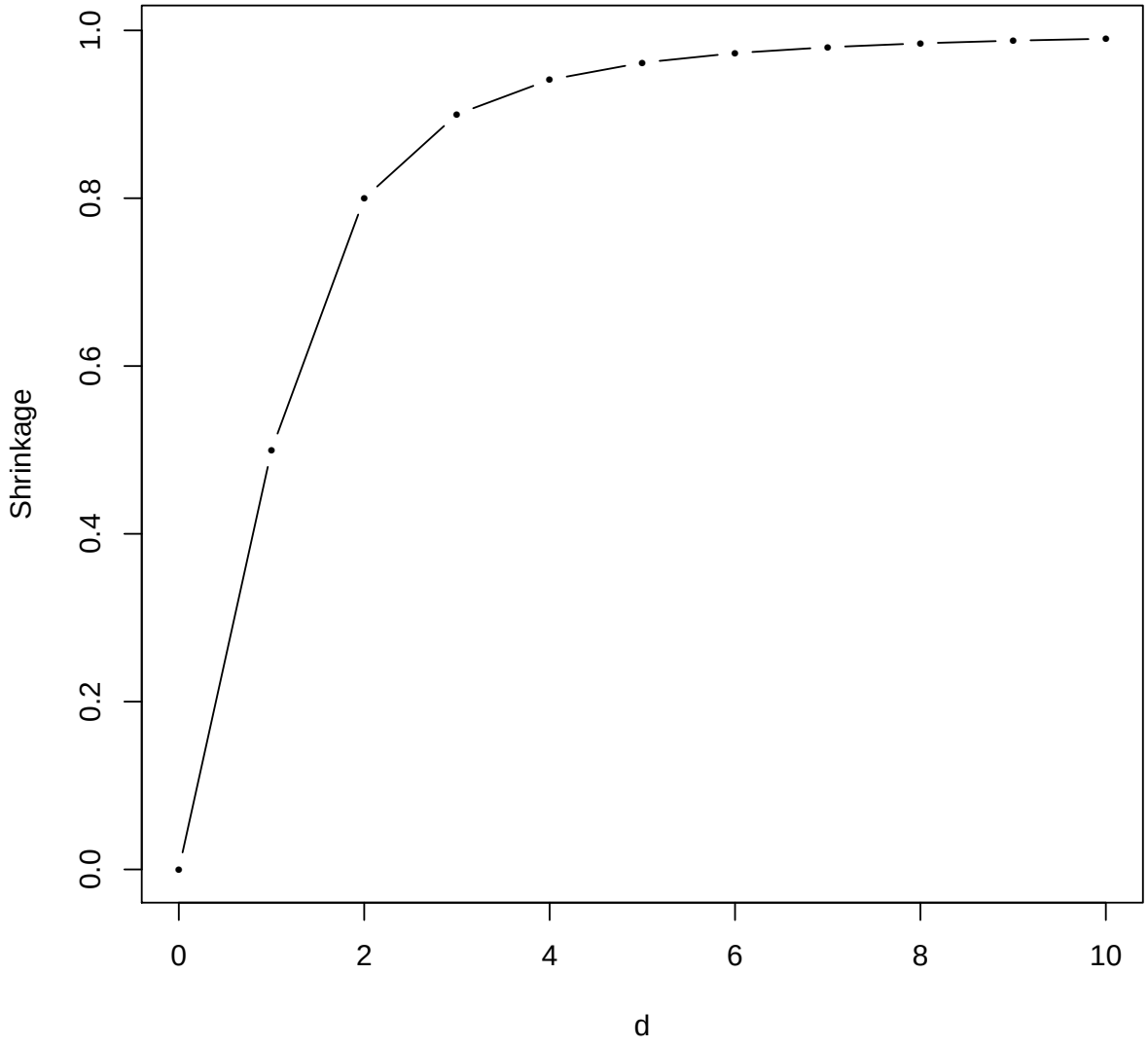
$$X\widehat{\beta}_{ridge} = UDV'(VD^2V' + \lambda I)^{-1}VDU'y$$

$$= UD(D^2 + \lambda I)^{-1}DU'y$$

Soit:

$$X\widehat{\beta}_{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j'y$$

Comme $\lambda \geqslant 0$, $\frac{d_j^2}{d_j^2+\lambda} \leqslant 1$: les coefficients de la régression linéaire classique sur les PC sont simplement réduits de ce facteur. Les coefficients associés aux d_j^2 les plus faibles sont les plus réduits (directions de faible variance dans le plan X).


Shrinkage
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
d

Choix du paramètre de pénalisation

Choix du paramètre de pénalisation

On procède, de même que pour les procédures de choix de modèle vu en régression linéaire simple:

- validation croisée
- critère du coude
- critère de "pénalisation" dépendant -souvent proportionnel- de la dimension du modèle (le nombre de degrés de liberté): ici on a besoin d'un estimateur de ce degré de liberté qui n'est pas simplement le nombre de paramètre comme ds la régression

Validation croisée

De même que pour la régression on a une expression du critère de VC:

$$\widehat{f}^{-i} = \sum_{j \neq i} \frac{H_{i,j}(\lambda)}{1 - H_{i,i}} y_j$$

et

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \widehat{f}_\lambda)^2}{(1 - H_{i,i})^2}$$

D'ou en utilisant l'estimateur du degré de liberté introduit pour la ridge regression:

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \widehat{f}_\lambda)^2}{(1 - tr(H)/n)^2}$$

le critère de VC est donc une moyenne de l'erreur d'estimation pondérée par "l'importance" de chaque observation. Le GCV est une erreur de VC dans laquelle chaque observation à le même "poids".

Cp de Mallows

On considère le modèle $y = f + \varepsilon$, $\widehat{f} = Hy$

$||f - \widehat{f}||^2 = ||f - Hy||^2 = ||y - Hy - \varepsilon||^2$

$= ||y - Hy||^2 + ||\varepsilon||^2 - 2\varepsilon'(y - Hy)$
$= ||y - Hy||^2 + ||\varepsilon||^2 - 2\varepsilon'(f + \varepsilon) + 2\varepsilon'(Hf + H\varepsilon)$

D'ou:

$E||f - \widehat{f}||^2 = E||y - Hy||^2 + n\sigma^2 - 2n\sigma^2 + 2E(\varepsilon' H\varepsilon)$

$$E||f - \widehat{f}||^2 = E||y - Hy||^2 - n\sigma^2 + 2tr(H)\sigma^2$$

L'heuristique de mallows vue en régression linéaire se généralise ici et on peut choisir le modèle qui minimise:

$$||y - Hy||^2 - n\sigma^2 + 2tr(H)\sigma^2$$

ie qui minimise $C_p(\lambda) = ||y - Hy||^2/n + 2tr(H)\sigma^2/n$ Si on ne connait pas σ^2, on l'estime par $\widehat{\sigma} = ||Y - H_{\lambda^*} Y||^2/(n - tr(H^*_\lambda))$ avec λ^* relativement "petit".

Comparaison du GCV et du Cp:

- $C_p = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f})^2 + \frac{2\sigma^2 tr(H_\lambda)}{n}$
- $GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f})^2/(1 - tr(H)/n)^2$

En utilisant l'approximation: $1/(1 - tr(H)/n)^2 \sim 1 + 2tr(H)/n$ on obtient:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f})^2 + 2\frac{tr(H)}{n^2}\sum_{i=1}^{n}(y_i - \widehat{f})^2$$

On remarque que le GCV est proche du critère de Mallows pour lequel on prendrait comme estimateur de la variance $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \widehat{f})^2/n$

Degrés de liberté

Degrés de liberté

Une définition générale de la notion de degrés liberté est donnée par:

$$df(\widehat{y}) = \frac{1}{\Sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{y}_i)$$

Cette définition est valable pour toute méthode d'estimation-prévision. Plus $\widehat{y}$ se rapproche des données -plus on "apprend" les données- plus $df(\widehat{y})$ est grand.

Degrés de liberté

- dans le cas de la régression linéaire, on a aisément: $df(\widehat{y}) = k$ le nombre de variable dans le modèle
- ridge: on retrouve $df(\widehat{y}) = tr(H_\lambda)$ *à faire*
- dans le cas de la régresion "best subset": pas calculable aisément mais intuitivement on sent que $df(\widehat{y}) > k$ même si on a sélectionné que k variable dans le modèle
- lasso: pour l'algorithme LAR, à la k^e étape de l'algorithme $df(\widehat{y}) = k$

Degrés de liberté

$$df(\widehat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{y}_i)$$

$$df(\widehat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \sum_{j=1}^{n} H_{i,j} y_j)$$

$$df(\widehat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} H_{i,i} \, \mathrm{Var}(y_i)$$