Apprentissage statistique : méthodes linéaires pour la régression

C. HELBERT

Plan

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs, prédicteurs)
- Y est la variable à expliquer quantitative.

Méthodes linéaires : on suppose que la relation entre Y et $(X_1, ..., X_p)$ peut s'exprimer comme une combinaison linéaire des $(X_1, ..., X_p)$ (ou de fonctions déterministes de $(X_1, ..., X_p)$).

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs, prédicteurs)
- Y est la variable à expliquer quantitative.

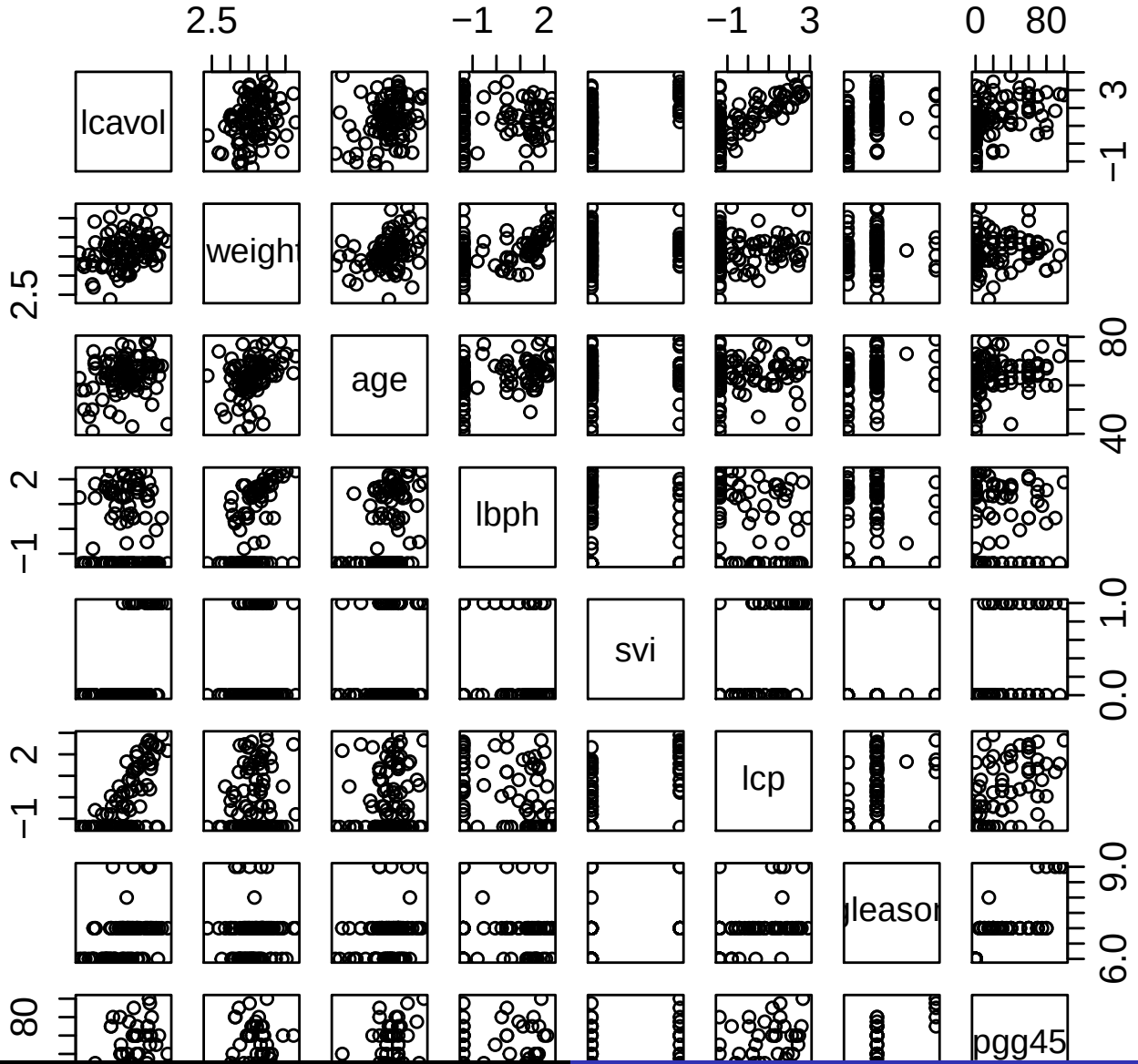
Méthodes linéaires : on suppose que la relation entre Y et $(X_1, ..., X_p)$ peut s'exprimer comme une combinaison linéaire des $(X_1, ..., X_p)$ (ou de fonctions déterministes de $(X_1, ..., X_p)$).

- méthodes simples, les effets des entrées sur la sortie sont interprétables
- robustes dans un contexte avec n petit comparativement à p
- extensions nombreuses en considérant des fonctions déterministes de $(X_1, ..., X_p)$ (**méthodes de régression sur fonctions de base**).

Exemple : cancer de la prostate

- entrées : volume et poids de la tumeur en log (*lcavol* et *lweight*), age (*age*), nombre de tumeurs benignes en log (*lbph)*, invasion des vesicules seminales (*svi*), penetration capsulaire en log (*lcp*), score de Gleason (*gleason* et *pgg45*).
- Y : niveau d'antigène spécifique de la prostate (*lpsa*)

Variables qualitatives, quantitatives et corrélées.
Training : $n_{train} = 67$. Test : $n_{test} = 30$.


lcavol
weight
age
lbph
svi
lcp
lleason
pgg45

Commande Summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.46493    0.08931  27.598  < 2e-16 ***
lcavol       0.67953    0.12663   5.366 1.47e-06 ***
lweight      0.26305    0.09563   2.751  0.00792 **
age         -0.14146    0.10134  -1.396  0.16806
lbph         0.21015    0.10222   2.056  0.04431 *
svi          0.30520    0.12360   2.469  0.01651 *
lcp         -0.28849    0.15453  -1.867  0.06697 .
gleason     -0.02131    0.14525  -0.147  0.88389
pgg45        0.26696    0.15361   1.738  0.08755 .
---
Residual standard error: 0.7123 on 58 degrees of freedom
Multiple R-squared:  0.6944, Adjusted R-squared:  0.6522
F-statistic: 16.47 on 8 and 58 DF,  p-value: 2.042e-12
```

Test des modèles emboités

Influence de *age*, *lcp*, *gleason*, *pgg45* ?

$$F = \frac{\frac{SSEreduit - SSEcomplet}{p-q}}{\frac{SSEcomplet}{n-(p+1)}} = \frac{\frac{32.81-29.43}{8-4}}{\frac{29.43}{67-(8+1)}} = 1.67.$$

Or $Pr(F_{4,58} > 1.67) = 0.17 > 0.05$ donc on conserve H_0 :
"$\beta_{age} = \beta_{lcp} = \beta_{gleason} = \beta_{pgg45} = 0$"

Modèle réduit

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47142    0.08901  27.766  < 2e-16 ***
lcavol       0.59582    0.10910   5.461 8.85e-07 ***
lweight      0.23084    0.09456   2.441   0.0175 *
lbph         0.20313    0.10215   1.988   0.0512 .
svi          0.27814    0.11311   2.459   0.0167 *
---
Residual standard error: 0.7275 on 62 degrees of freedom
Multiple R-squared:  0.6592, Adjusted R-squared:  0.6372
F-statistic: 29.98 on 4 and 62 DF,  p-value: 6.911e-14
```

Pourcentage de variance expliquée sur l'ensemble test :

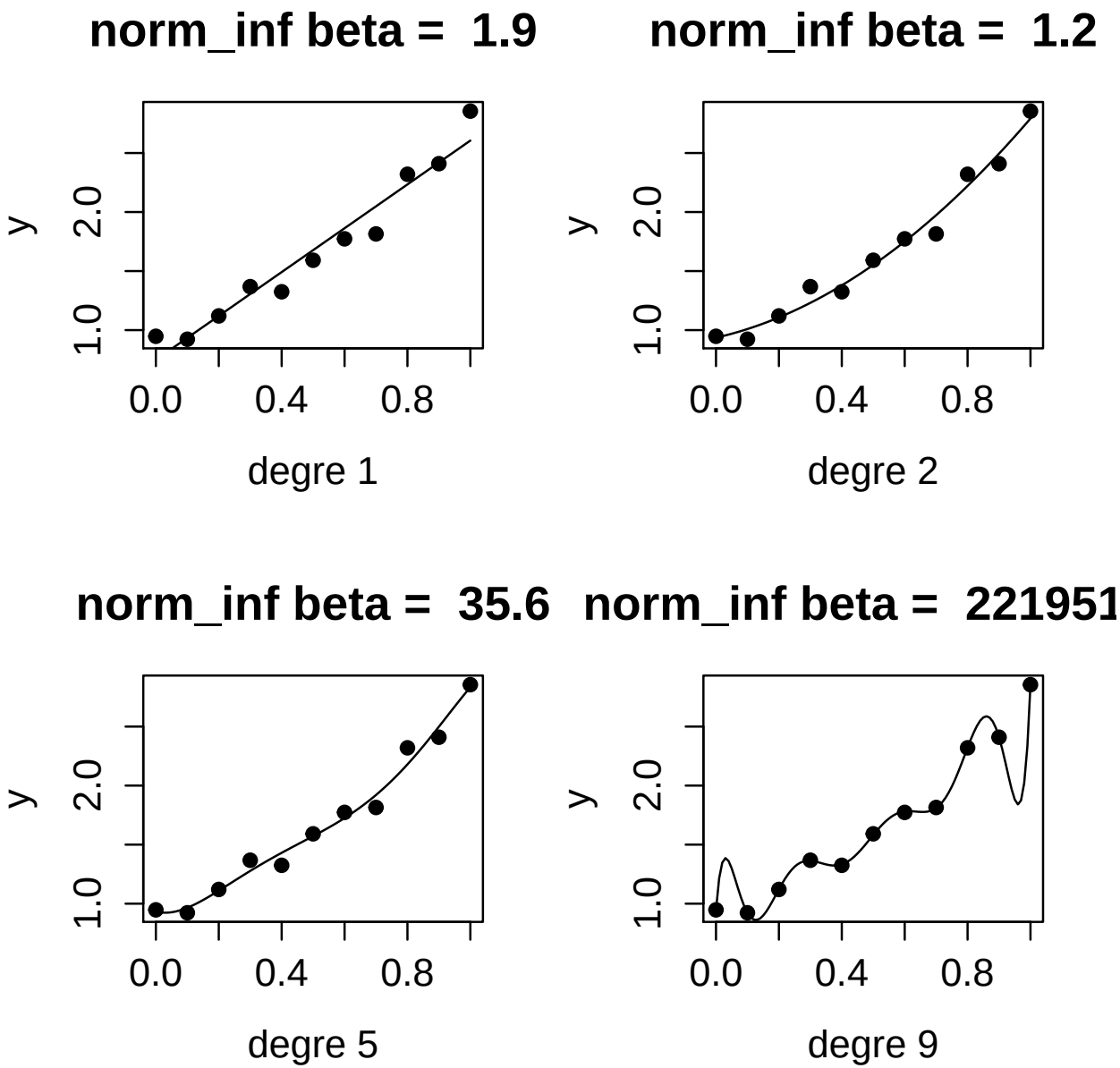
$> R2(prost.test[, 9], predict(mod2, newdata = prost.test))$

[1]0.5652503

Plan

La solution des moindres carrés n'est pas entièrement satisfaisante :

- **Interprétation** : quand p est grand , on veut exhiber les variables les plus influentes (criblage ou **screening**) au dépend des variables les moins influentes.
- **Précision de la prédiction** : quand p est grand ($X_1, ..., X_p$ corrélés) ou degré du polynôme élevé (nombre de fonctions de régression important) la solution des moindres carrés a souvent une forte variance.


norm_inf beta = 1.9
norm_inf beta = 1.2
norm_inf beta = 35.6
norm_inf beta = 221951
y
1.0
2.0
0.0
0.4
0.8
degre 1
degre 2
degre 5
degre 9

Les critères habituels

Quand p n'est pas trop grand, plusieurs critères et méthodes existent pour faire de la sélection de modèle :

- AIC, BIC, Cp de mallows, F-Statistics, Validation croisée ....
- parcours de tous les sous-ensembles (tous les modèles réduits) de façon exhaustive ou uniquement via des méthodes backward ou forward.

Les avancées récentes

Quand p est grand, plusieurs solutions existent pour stabiliser et exploiter le critère des moindres carrés :

- la contraction (**shrinking**)
- le seuillage (**thresholding**)
- méthodes à base de combinaisons linéaires

Ridge Regression

Si on reprend l'exemple 1D ci-dessus, on voit que la norme de β explose avec la dimension, en contexte corrélé.

Première idée : imposer une pénalité sur la norme 2 du vecteur de paramètres.

$$\beta^{ridge} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\},$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

$\Leftrightarrow$

$$\beta^{ridge} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Ridge Regression

Interprétation :

- Il existe une relation directe entre λ et t, plus λ augmente plus t diminue
- Attention : la pénalisation ne s'exerce pas sur β_0. Quand les variables $X_1, ..., X_p$ sont centrées ($\tilde{x}_{ij} = x_{ij} - \bar{x_{.j}}$), β_0 est estimé par $\frac{1}{n}\sum_{i=1}^{n} y_i$
- Dans la suite, on considère le problème de l'estimation de $\beta_1, ..., \beta_p$ quand la matrice $\mathbf{X}$ contient p colonnes centrées. On considère aussi habituellement que les variables sont préalablement réduites pour attribuer la même pénalisation à chaque prédicteur.

Ridge Regression

Le problème peut s'écrire sous la forme matricielle suivante :

$$\beta^{ridge} = \underset{\beta}{argmin} \quad RSS(\lambda)$$

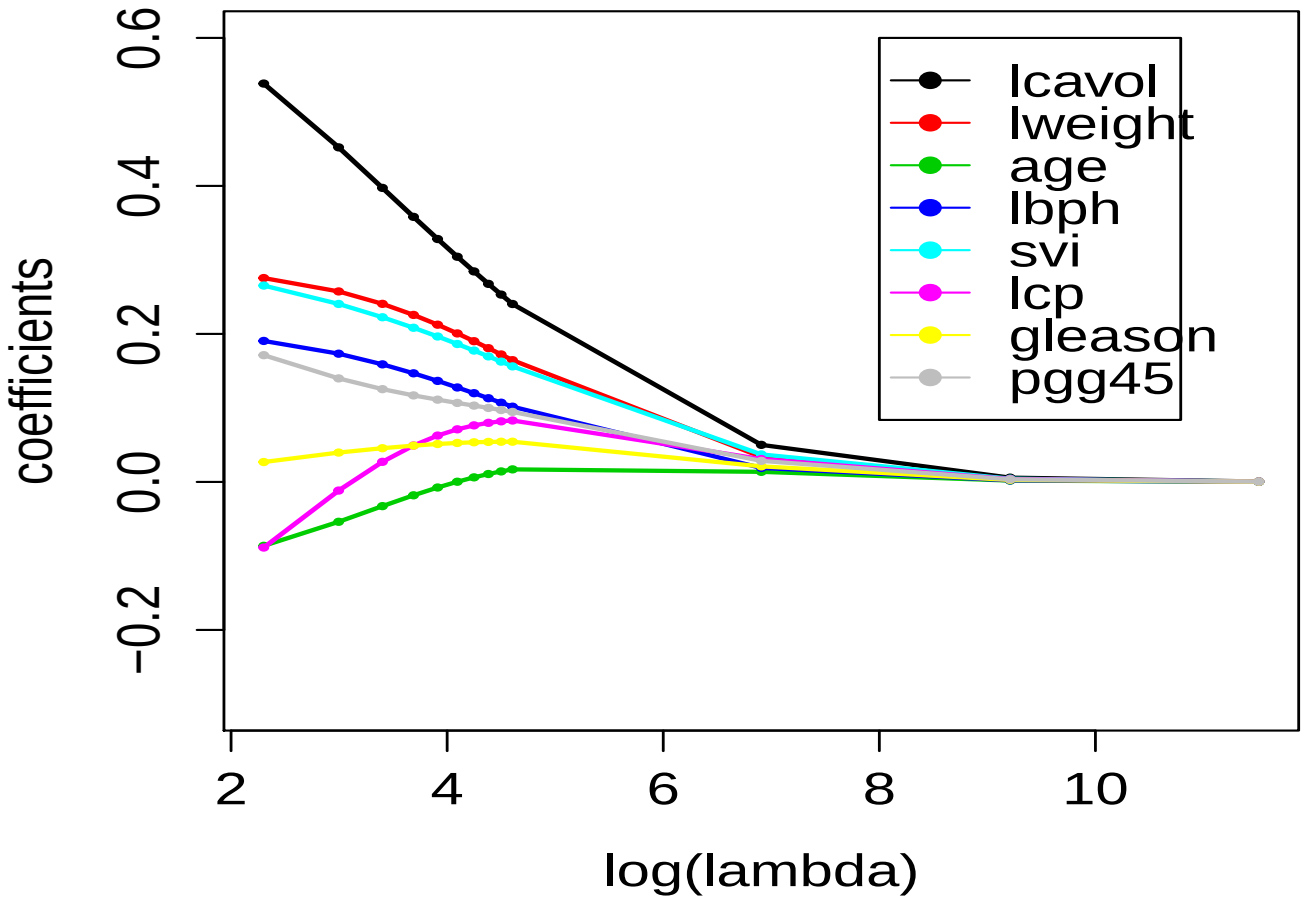
$$\text{où } RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$$

La solution est alors :

$$\beta^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Remarque : la solution dépend de la valeur de λ. L'ajout du terme supplémentaire sur la diagonale rend le problème inversible même quand $\mathbf{X}^T\mathbf{X}$ n'est pas de plein rang.

Ridge Regression sur prostate



Le Lasso

Deuxième idée : imposer une pénalité sur la norme 1 du vecteur de paramètres.

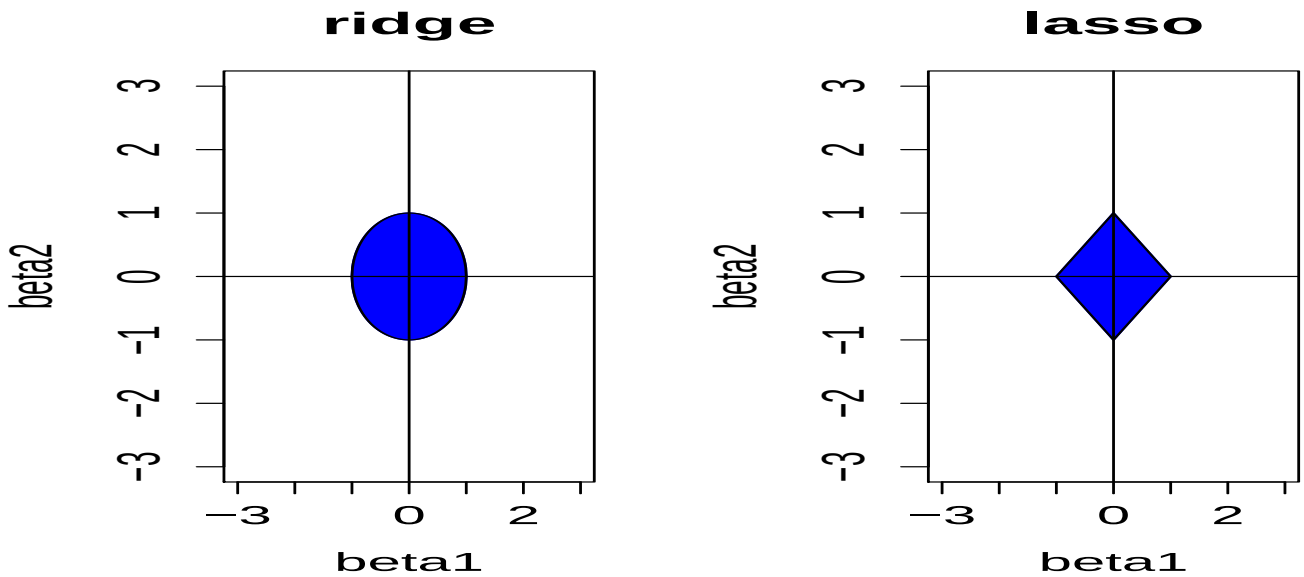
$$\beta^{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\},$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

$\Leftrightarrow$

$$\beta^{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Norme 1 vs. Norme 2



Le Lasso

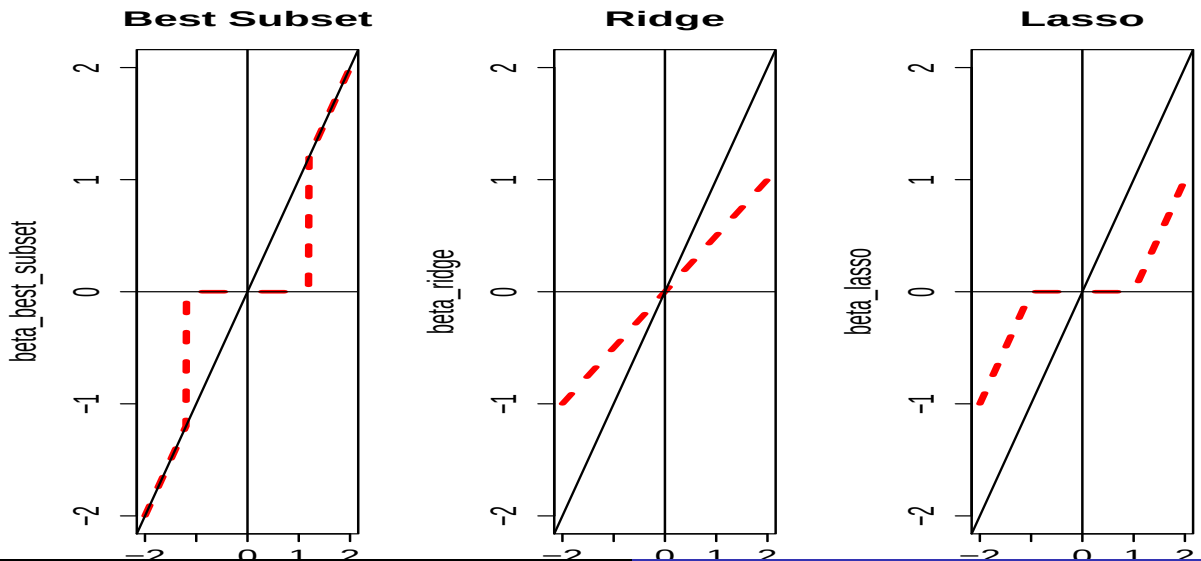
Conséquence :

- la solution n'est plus linéaire en **y**
- il n'y a plus d'expression analytique pour la solution
- La nature de la contrainte entraine que certains coefficients seront exactement 0 ! Le lasso est donc une méthode "continue" de sélection de modèle

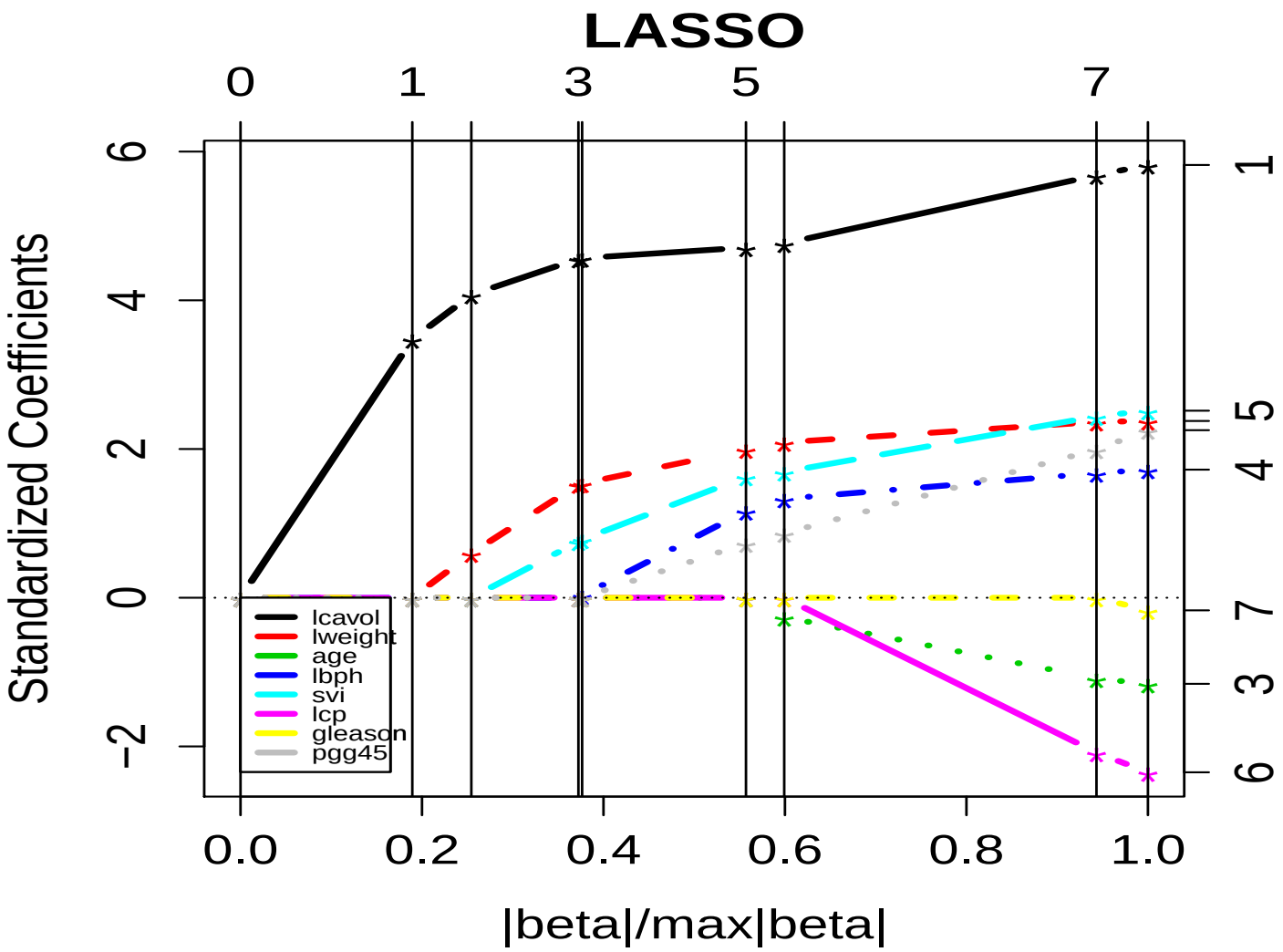
Lasso vs. Ridge

Dans le cas orthogonal :

Estimateur	Formule
Best subset (size M)	$\hat{\beta}_j . \mathbf{1}(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_M\rvert)$
Ridge	$\frac{\hat{\beta}_j}{(1+\lambda)}$
Lasso	$sign(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



Le Lasso sur prostate



Extensions

- Pénalité L^q : pénalité intermédiaire entre Ridge et Lasso quand $q \in]1, 2[$. Si $q > 1$ la méthode ne joue pas le rôle de sélection

$$\tilde{\beta} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- Elasticnet

$$\tilde{\beta} = \underset{\beta}{argmin} \left\{ ... + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right) \right\}$$

algorithme LAR "Least Angle Regression" , Package *lars*

- Stratégie très proche de la stratégie forward en régression
- Implémentation efficace d'une sélection très proche du lasso

Contexte :

- le nombre de variables d'entrée p est grand
- variables fortement corrélées entre elles (redondance d'information). C'est par exemple le cas d'une entrée fonctionnelle discrétisée.

Principe de la méthode :

- Trouver un petit nombre de combinaisons linéaires des variables d'entrée, $Z_1, .., Z_M$ et qui remplacent les p entrées.

Remarque : dans la suite les variables sont préalablement centrées.

PCR : Principal Component Regression

- Dans cette approche, les variables $Z_1, .., Z_p$ sont les composantes principales de la matrice $\mathbf{X}$ telles que $\mathbf{z}_m = \mathbf{X}v_m$
- On sélectionne uniquement les M composantes les plus corrélées à Y et on régresse $\mathbf{y}$ sur $\mathbf{z}_1, ..., \mathbf{z}_M$:

$$\hat{\mathbf{y}}_{(M)}^{PCR} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m \text{ où } \hat{\theta}_m = \frac{< \mathbf{z}_m, y >}{< \mathbf{z}_m, \mathbf{z}_m >}$$

On a alors :

$$\hat{\beta}_{(M)}^{PCR} = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

Il n'y a donc pas systématiquement de sélection de variables parmi les variables d'entrées.

PLS : Partial Least Squares

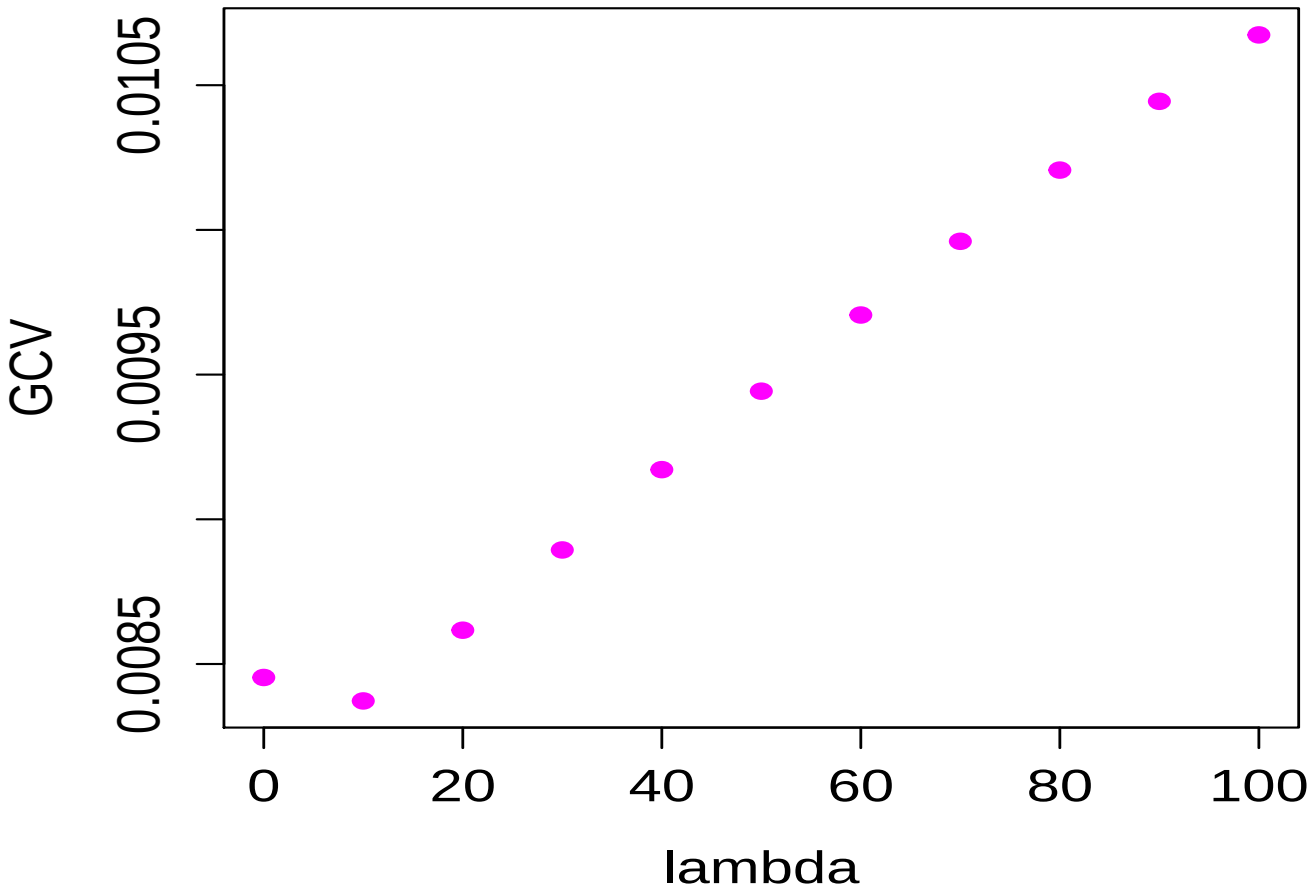
- Dans cette approche, on utilise la variable Y pour construire les combinaisons linéaires des variables d'entrée
- On construit les combinaisons linéaires de sorte à ce qu'elles soient orthogonales et les plus corrélées à Y

PLS : Partial Least Squares

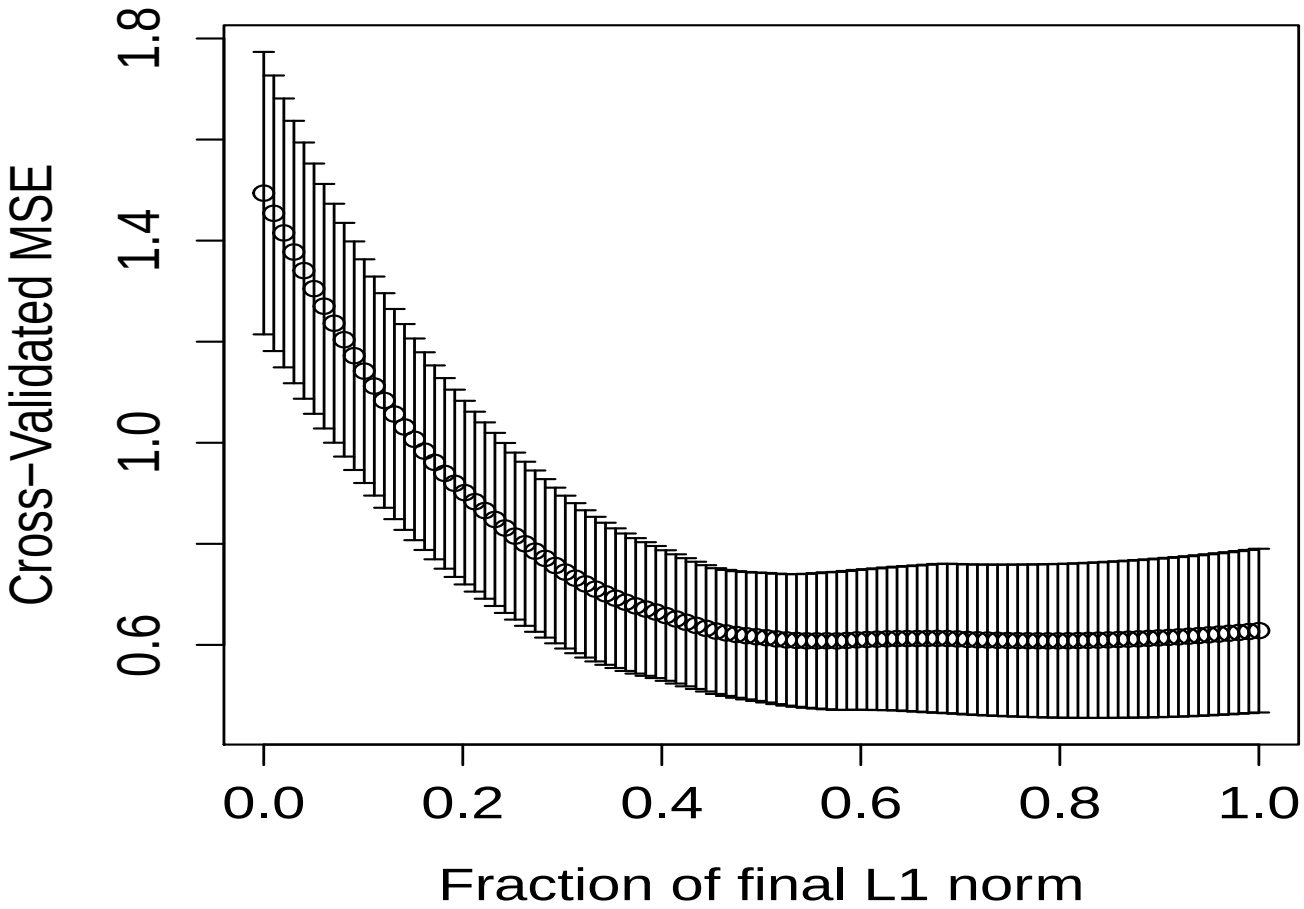
1. Centrer réduire les variables $\mathbf{x}_j$. Mettre $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$ et $\mathbf{x}_j^{(0)} = \mathbf{x_j}, j = 1, ...p$
2. Pour $m = 1, 2..., p$
   a. $\mathbf{z}_m = \sum_1^p \hat{\varphi}_{mj}\mathbf{x}_j^{(m-1)}$ où $\hat{\varphi}_{mj} = < \mathbf{x}_j^{(m-1)}, \mathbf{y} >$
   b. $\hat{\theta}_m = \frac{<\mathbf{z}_m, y>}{<\mathbf{z}_m, \mathbf{z}_m>}$
   c. $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$
   d. Orthogonalisation de $\mathbf{x}_j^{(m-1)}$ par rapport à $\mathbf{z}_m$: $\mathbf{x}_j^m = \mathbf{x}_j^{(m-1)} - \frac{<\mathbf{z}_m, \mathbf{x}_j^{(m-1)}>}{<\mathbf{z}_m, \mathbf{z}_m>}\mathbf{z}_m$
3. La linéarité de la construction en $\mathbf{x}_j$ assure que $\hat{\mathbf{y}}_m = \mathbf{X}\hat{\beta}^{PLS}(m)$.

Plan

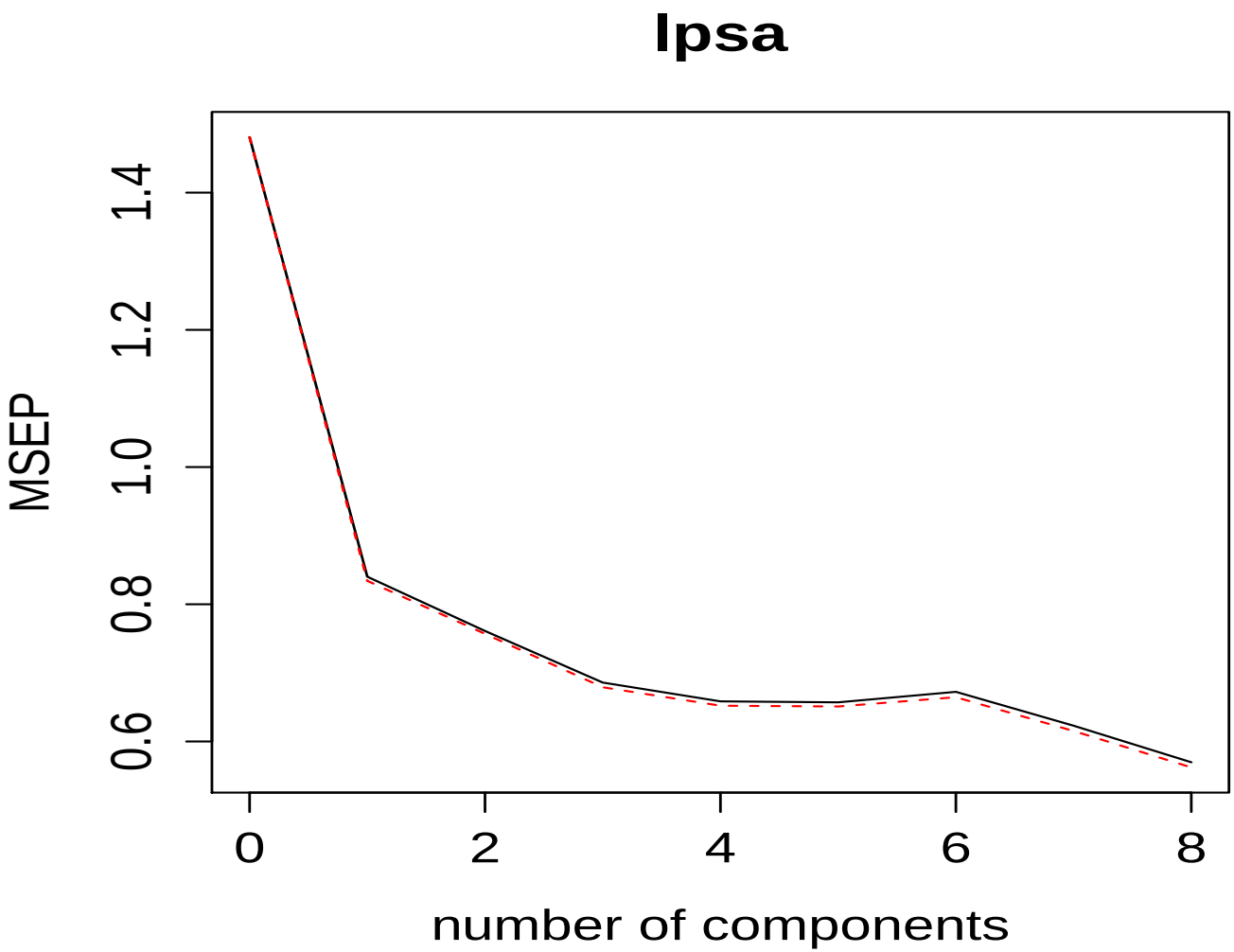
Cross Validation : RIDGE



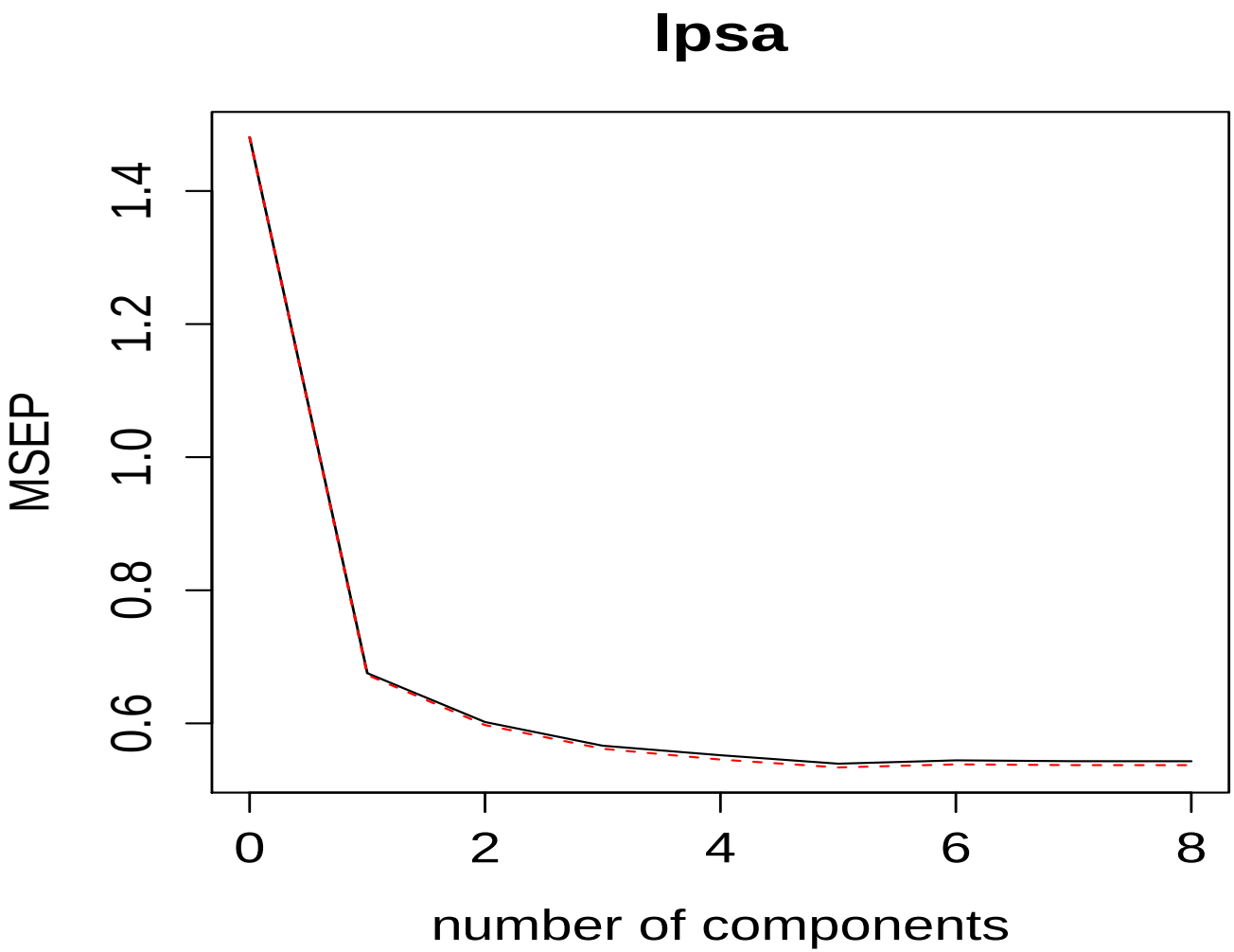
Cross Validation : LASSO



Cross Validation : PCR



Cross Validation : PLS



Comparaison des méthodes

	Ridge $\lambda = 10$	Lasso $s = 0.54$	PCR $nb = 4$	PLS $nb = 3$
lcavol	0.538	0.548	0.273	0.576
lweight	0.276	0.216	0.344	0.279
age	-0.086		-0.076	-0.179
lbph	0.191	0.129	0.194	0.201
svi	0.265	0.188	0.237	0.299
lcp	-0.089		0.236	-0.036
gleason	0.027		0.017	0.006
pgg45	0.171	0.079	0.112	0.117
MSE (test)	0.486	0.456	0.520	0.427

Plan

De multiples extensions existent :

- "The grouped lasso" : quand un prédicteur est codé sur plusieurs composantes : plusieurs gènes correspondant au même phénotype, variables qualitatives à plus de 2 modalités, variables fonctionnelles décomposées sur des bases de fonctions, etc...

$$\underset{\beta}{argmin} \left\{ \left|\left| y - \beta_0 \mathbf{1} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \beta_\ell \right|\right|_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\, ||\beta_\ell||_2 \right\}$$

- Le Lasso permet une sélection des variables influentes, mais il introduit un biais sur l'estimation des coefficients $=>$ Une possibilité est de procéder à deux lasso successifs : le λ sera élevé au premier tour puis nettement moins élevé (biais moins grand) au deuxième tour.