L'algorithme Ridge-Partial Least Squares

Application à la classification de puces à ADN.

Gersende Fort (CNRS/UJF)

Collaboration : Sophie Lambert-Lacroix (UJF)

Preprint, codes MATLAB et Résultats Numériques, disponibles à l'adresse

http://www-lmc.imag.fr/lmc-sms/Gersende.Fort/

Puces à ADN : Oligonucleotide Microarrays (Affymetrix)

$\hookrightarrow$ Image de l'état de la cellule par l'étude du transcriptome (quantité de gènes transcrits dans la cellule)

► Description : gènes $\sim$ segments d'ADN

Sur un support, brins d'ADN synthétisés *in situ* [sondes].

Dépôt sur la puce d'un mélange de brins d'ADN (issus de la cellule), marqués par fluorescence [cibles].

Hybridation.

Lavage, détection des hybridations par scanner.

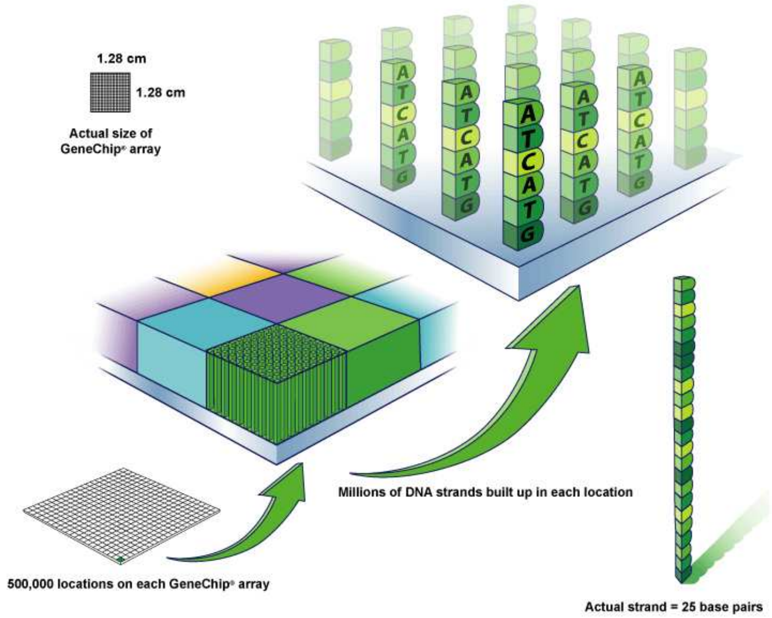
► Usages : gènes $\longmapsto$ protéines $\sim$ agents fonctionnels des organismes

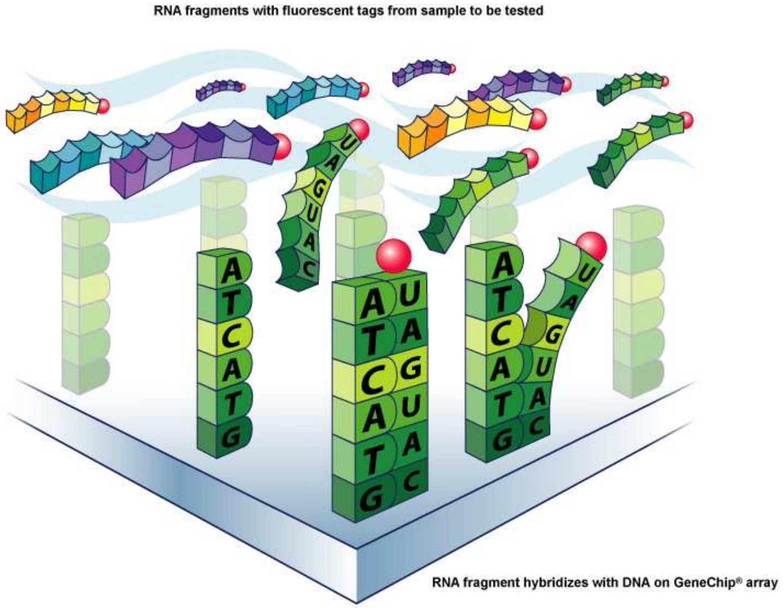
Aide à la compréhension plus fine du génôme et de sa régulation : identification de cibles thérapeutiques pour la recherche pharmaceutique.

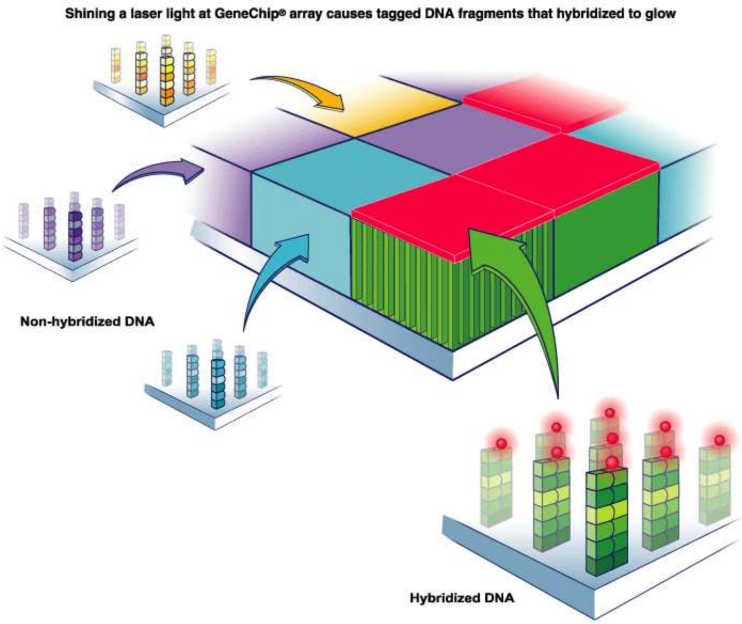
Pharmacogénomique : identifier le gènes impliqués dans l'efficacité d'un produit (ou ses effets indésirables).

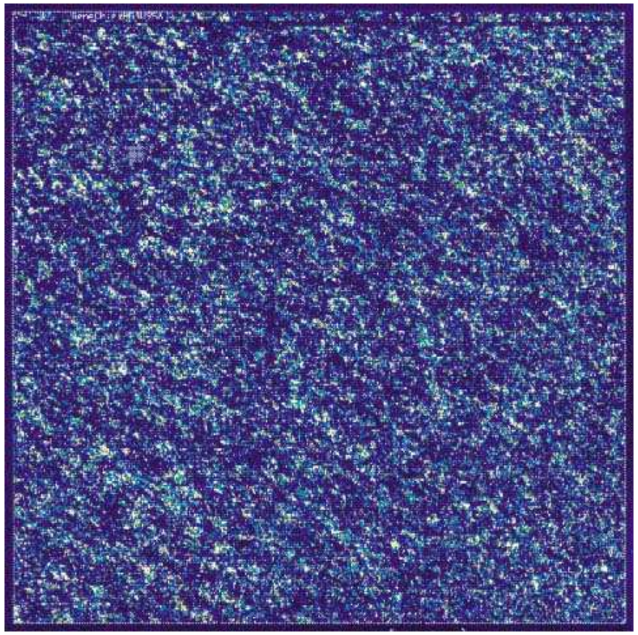
Diagnostic des maladies infectieuses et génétiques :

- ex. : détecter la présence du virus de l'hépatite C + type (I,II,III).
- ex. : déceler les mutations génétiques indiquant que le patient a de fortes probabilités d'être atteint d'une maladie.


1.28 cm
1.28 cm
Actual size of GeneChip® array
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs


RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip® array


Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA



Diagnostic de maladies : Pour chaque individu,

Niveau d'expression de plusieurs milliers de gènes,

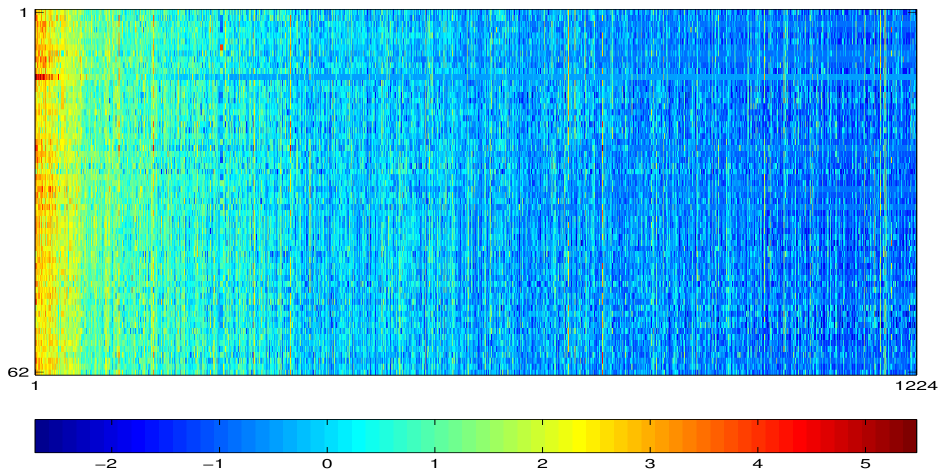
Connaissance de la pathologie dont il est atteint.

- Jeu de données LEUKEMIA [Golub *et al.*(1999)]

  Niveau d'expression de 7129 gènes; 72 individus : 47 ALL (Acute Lymphoblastic) et 25 AML (Acute Myeloid).

- Jeu de données COLON [Alon *et al.*(1999)]

  Niveau d'expression de 2000 gènes; 62 individus : 22 tissus sains et 40 tissus tumoraux.



Colon Data set :

1224 gènes après pré-filtrage

22 tissus sains puis 40 tumoraux

$\hookrightarrow$ Discrimination logistique linéaire.

Discrimination Logistique Linéaire :

$Y \in \{0,1\}$: Label, (réponse) $x \in \mathbb{R}^{p_{\max}}$: Expression des gènes, (covariables).

Pour $\theta \in \mathbb{R}^{p_{\max}+1}$,

$$\mathbb{P}\left(Y=1|x;\theta\right) = \left[1+\exp(-\eta(x;\theta))\right]^{-1},$$

de prédicteur linéaire

$$\eta(x;\theta) = \theta_0 + \theta_1 x_1 + \cdots + \theta_{p_{\max}} x_{p_{\max}}.$$

Pour $\hat{\theta} \in \mathbb{R}^{p_{\max}+1}$, $x \in \mathbb{R}^{p_{\max}}$,

$$\hat{Y} = 1 \quad \text{ssi} \quad \mathbb{P}\left(Y=1|x;\hat{\theta}\right) \geq 0.5.$$

$\hookrightarrow$ Inférence dans les GLM : n individus (qq dizaines) et $p_{\max}$ régresseurs (qq milliers),

Cadre statistique *"grande dimension"* $n << p$.

En conclusion,

- Inférence dans un modèle linéaire généralisé, de fonction de lien canonique :
  par Maximum de Vraisemblance,
  - ▷ Identifiabilité,
  - ▷ Forte colinéarité des régresseurs.

- Analyse qui doit combiner:
  - ▷ une étape de réduction de dimension,
  - ▷ une étape de régularisation.

$\hookrightarrow$ Algorithme Ridge-Partial Least Squares.

A. Partial Least Squares dans les modèles linéaires.
   (i) Description algorithmique,
   (ii) Interprétation de la méthode,
   (iii) Quelques propriétés.

B. Inférence dans les modèles linéaires généralisés (cas Bernoulli).
   (i) Algorithme *Iteratively Reweighted Least Squares* (IRLS).
   (ii) Existence du maximum de vraisemblance.

C. Algorithme Ridge-Partial Least Squares.
   (i) Comment étendre PLS aux GLM,
   (ii) Description algorithmique.

Généralités : Méthode de réduction de dimension et de régression :

- Etant donnés un entier $\kappa > 0$, $\mathrm{Y} \in \mathbb{R}^n$, $\mathrm{X} \in \mathbb{R}^{n\times p}$,

  (i) produire κ PLS-scores, $\mathbf{t}_1, \cdots, \mathbf{t}_\kappa \in \mathrm{Vect}(\mathrm{X})$ tq

  $$\mathbf{t}_j \perp \mathbb{1}_n, \qquad \text{PLS-scores } 2 \text{ à } 2 \text{ orthogonaux}$$

  i.e. $\mathbf{t}_1, \cdots, \mathbf{t}_\kappa \in \mathrm{Vect}(\mathrm{X}^c)$, $\qquad \mathrm{X}^c$: X centrée en colonnes.

  (ii) régresser Y sur $[\mathbb{1}_n \; \mathbf{t}_1 \; \cdots \; \mathbf{t}_\kappa]$,

  $$\hat{\mathrm{Y}} = q_0 \mathbb{1}_n + q_1 \mathbf{t}_1 + \cdots + q_\kappa \mathbf{t}_\kappa, \qquad q_0 = \frac{\langle \mathrm{Y}, \mathbb{1}_n \rangle}{\langle \mathbb{1}_n, \mathbb{1}_n \rangle} = \mathrm{mean}(\mathrm{Y}),$$

  i.e. régresser Y^c sur $[\mathbf{t}_1 \; \cdots \; \mathbf{t}_\kappa]$, $\qquad \mathrm{Y}^c$: Y centré.

- D'où la définition de $\hat{\theta}$ tq

  $$\hat{\mathrm{Y}} = \hat{\theta}_0 \mathbb{1}_n + \mathrm{X}^c\, \hat{\theta}_{1:p} \qquad \text{avec } \hat{\theta} \left| \begin{array}{l} \text{unique si } \mathrm{rang}(\mathrm{X}^c) = p, \\ \text{celui de norme minimale sinon.} \end{array} \right.$$

- En conséquence, dans la suite de la présentation PLS,

  $\mathrm{Y} \in \mathbb{R}^n$ centré $\qquad$ $\mathrm{X} \in \mathbb{R}^{n\times p}$ centrée en col., de rang p $\qquad$ $\hat{\mathrm{Y}} = \mathrm{X}\, \hat{\theta}^{\mathrm{PLS},\kappa}$.

Algorithme

Pour (Y, X) donné, $\mathrm{Y} \in \mathbb{R}^n$ centré, $\mathrm{X} \in \mathbb{R}^{n \times p}$ centrée en col. de rang p,

- Méthode itérative

$$\mathrm{f}_0 \leftarrow \mathrm{Y} \qquad \mathrm{E}_0 \leftarrow \mathrm{X}$$

$$\mathbf{t}_1 \leftarrow \mathrm{E}_0\ \mathrm{E}_0'\mathrm{f}_0$$

$$q_1 \leftarrow \mathrm{f}_0'\mathbf{t}_1/\|\mathbf{t}_1\|^2 \qquad \text{Projection de } \mathrm{f}_0 \text{ sur } \mathbf{t}_1$$

$$\mathrm{f}_1 \leftarrow \mathrm{f}_0 - q_1\mathbf{t}_1 \qquad \text{Projection de } \mathrm{f}_0 \text{ sur } \mathbf{t}_1^\perp$$

$$\mathrm{E}_1 \leftarrow \mathrm{E}_0 - \mathbf{t}_1\ \mathrm{E}_0'\mathbf{t}_1/\|\mathbf{t}_1\|^2 \qquad \text{Projection des col. de } \mathrm{E}_0 \text{ sur } \mathbf{t}_1^\perp$$

Reboucler

- Ainsi,

(i) $\mathrm{Y} = q_1\mathbf{t}_1 + \cdots + q_\kappa\mathbf{t}_\kappa + f_\kappa \qquad f_\kappa \perp \mathrm{Vect}(\mathbb{1}_n,\ \mathbf{t}_1,\ \cdots,\mathbf{t}_\kappa)$.

(ii) Lorsque $\kappa = \kappa_{\max}$, $\hat{\theta}^{\mathrm{PLS},\kappa} = \hat{\theta}^{\mathrm{OLS}}$.

(iii) Géométrie à poids *i.e.* $\langle\cdot;\cdot\rangle_{\mathrm{W}} \Leftrightarrow (\mathrm{Y}, \mathrm{X}) \leftarrow (\sqrt{\mathrm{W}}\mathrm{Y}, \sqrt{\mathrm{W}}\mathrm{X})$ [mod. hétéroscédastique W^{-1}].

Interprétation de la méthode

- Définition des PLS-scores $\mathbf{t}_j \in \mathbb{R}^n$:

$$\mathbf{t}_1 = \mathrm{E}_0 \;\; \mathrm{E}_0' \mathrm{f}_0, \qquad\qquad \mathrm{E}_0 \longleftrightarrow \mathrm{X}\text{centrée}_{\text{ en colonnes}}, \qquad\qquad \mathrm{f}_0 \longleftrightarrow \mathrm{Y}\text{centré};$$

qui est obtenu par

$$\mathbf{t}_1 \propto \mathrm{E}_0 \;\; c, \qquad c \text{ tq.} \begin{cases} \max \left|\langle \mathrm{E}_0 \; c, \mathrm{f}_0 \rangle\right|, \\ \|c\| = 1. \end{cases}$$

Rmqs. : (i) Norme à poids $\|c\|_{\Sigma^2} \longleftrightarrow$ Scaling Σ^{-1} sur X (ii) norme égale à '1'.

- Sur la base des $\overrightarrow{\text{vp}}$ de XX',

$$\mathbf{t}_1 = \sum_{k=1}^{p} d_k^2 \;\; \langle U_{k\cdot}, \mathrm{Y} \rangle \;\; U_{k\cdot} \qquad \text{où } \; \mathrm{X} = UDV'.$$

Propriétés de l'estimateur $\hat{\theta}^{\mathrm{PLS},\kappa}$

$\mathrm{Y} \in \mathbb{R}^n$ centré et $\mathrm{X} \in \mathbb{R}^{n\times p}$ centrée en col. rang plein p.

- Suite de Krylov : $\mathrm{R}_{xx} = \mathrm{X}'\mathrm{X}$ et $\mathrm{R}_{xy} = \mathrm{X}'\mathrm{Y}$,

$$\hat{\theta}^{\mathrm{PLS},\kappa} \in \mathrm{Vect}\left(\mathrm{R}_{xy}, \mathrm{R}_{xx}\mathrm{R}_{xy}, \cdots, \mathrm{R}_{xx}^{\kappa-1}\mathrm{R}_{xy}\right) =: \mathrm{Vect}\left(\mathcal{K}_\kappa\right);$$

$$\hat{\theta}^{\mathrm{PLS},\kappa} \text{ sol. de } \begin{cases} \min_\theta \|\mathrm{Y} - \mathrm{X}\theta\|, \\ \theta \in \mathcal{K}_\kappa. \end{cases} \quad \text{i.e. } \hat{\theta}^{\mathrm{PLS},\kappa} = \mathcal{K}_\kappa \left(\mathcal{K}_\kappa'\mathrm{X}'\mathrm{X}\mathcal{K}_\kappa\right)^{-1}\mathcal{K}_\kappa\mathrm{X}'\mathrm{Y}.$$

- Filtres : $\mathrm{X} = UDV'$

$$\hat{\theta}^{\mathrm{PLS},\kappa} = \sum_{l=1}^{p} \omega_{l,\kappa} \frac{\langle \mathrm{Y}; U_{l\cdot}\rangle}{d_l} V_{l\cdot}$$

ex. PCR : $\omega_{l,\kappa} = 1 \qquad l \leq \kappa$,

Ridge : $\omega_l = d_l^2/(d_l^2 + \lambda)$.

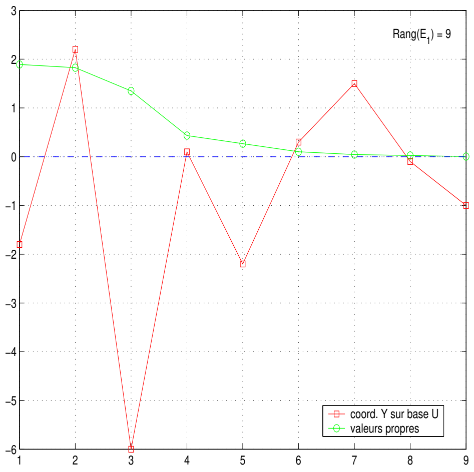
Propriétés de l'estimateur $\hat{\theta}^{\mathrm{PLS},\kappa}$ (suite)

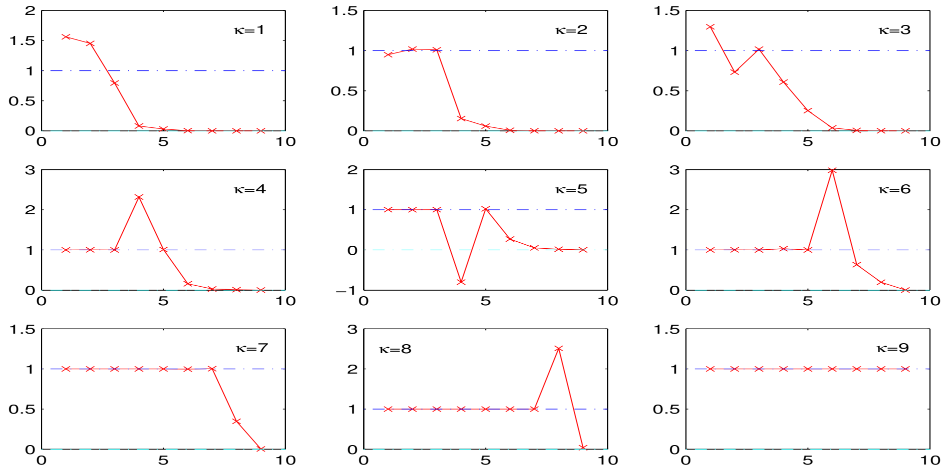
$$\hat{\theta}^{\mathrm{PLS},\kappa} = \sum_{l=1}^{p} \omega_{l,\kappa} \frac{\langle \mathrm{Y}; U_{l}. \rangle}{d_l} V_{l}.$$

- Filtres et Facteurs de Ritz [Lingjærde *et al.*(2000)]

$$\omega_{l,\kappa} = 1 - \prod_{j=1}^{\kappa} \left(1 - \frac{d_l^2}{\nu_{j,\kappa}}\right)$$

$\nu_{j,\kappa}$ vp de $\mathcal{K}'_\kappa \mathrm{X}'\mathrm{X}\mathcal{K}_\kappa$, $\mathcal{K}_\kappa$orthonormée.





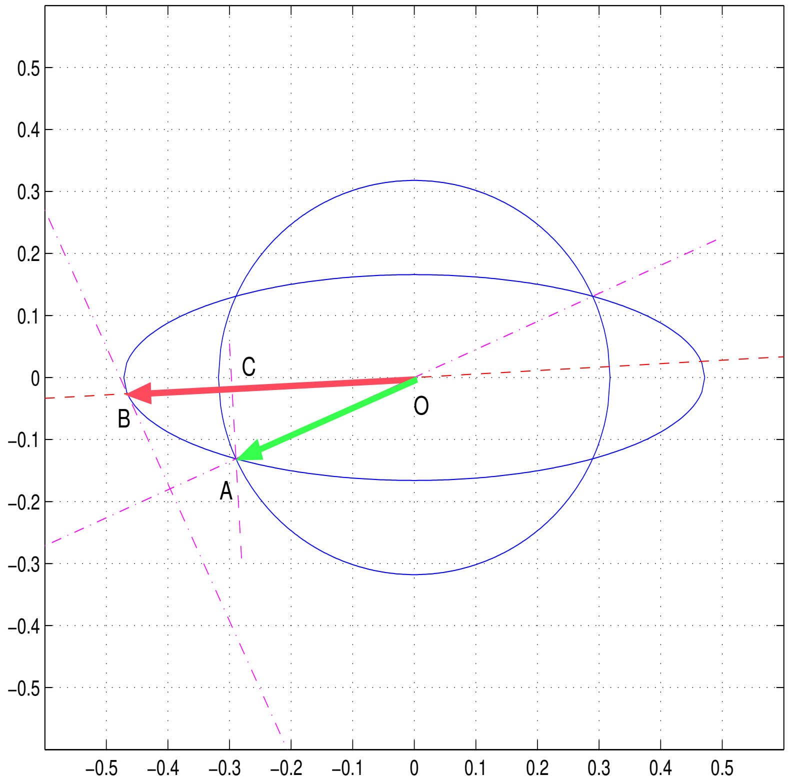
(G) Valeurs propres de $\mathrm{X}'\mathrm{X}$, de rang $p = 9$; et coefficients $< \mathrm{Y}; U_{l}. >$

(D) Pour $\kappa = 1, \cdots, 9$, on trace $l \mapsto \omega_{l,\kappa}$. Observons :

(i) valeur de $\omega_{1,\kappa}$; (ii) $\exists\, \omega_{l,\kappa} \leq 0$ et $\exists\, \omega_{l,\kappa} \geq 1$; (iii) $\exists\, l \geq \kappa$, $\omega_{l,\kappa} \geq 1$; (iv) pour $\kappa = 9$, PLS $\equiv$ OLS.

Propriétés de l'estimateur $\hat{\theta}^{\mathrm{PLS}}$ (suite)

▶ PLS vs OLS (shrinkage) : $\|\hat{\theta}^{\mathrm{PLS},\kappa}\| \leq \|\hat{\theta}^{\mathrm{OLS}}\|$ [Goutis (1996)]



Modèle canonique : $\mathrm{Y} \in \mathbb{R}^p, \mathrm{X} \in \mathbb{R}^{p\times p}$ diag, invers.

$$\mathcal{E} = \{s \in \mathbb{R}^p, s'\mathrm{X}^{-2}s = \|\mathrm{Y}'\mathrm{X}^{-2}\mathrm{Y}\|^2\}$$
$$\mathcal{C} = \{s \in \mathbb{R}^p, s's = \|\mathrm{Y}\|^2\}$$

$\overrightarrow{OA}$: coord. Y; Base. ell. $\mathrm{X}^{-1}\mathrm{Y} = \hat{\theta}^{\mathrm{OLS}}$

$\overrightarrow{OB} = \lambda\, \overrightarrow{OA} + \nu^{\perp}$, coord. $\mathbf{t}_1 = \mathrm{X}c \in \mathbb{R}^p$

$$\max_c | < \mathrm{X}c, \mathrm{Y} > | \Leftrightarrow \max | \overrightarrow{OB} \cdot \overrightarrow{OA} |$$
$$c'c = \|\hat{\theta}^{\mathrm{OLS}}\|^2 \Leftrightarrow B \in \mathcal{E}.$$

Rmq. : $\mathrm{Y}'\mathrm{Y} \leq \mathbf{t}_1'\mathbf{t}_1$.

$\overrightarrow{OC} = \mu\, \mathbf{t}_1 = \mathrm{X}\hat{\theta}^{\mathrm{PLS},1}$ d'où $\|\hat{\theta}^{\mathrm{PLS},1}\|^2 = \mu^2\ \mathbf{t}_1'\mathrm{X}^{-2}\mathbf{t}_1 = \mu^2\ \|\hat{\theta}^{\mathrm{OLS}}\|^2$,

$\| \overrightarrow{OC} \|^2 \leq \| \overrightarrow{OA} \|^2$ d'où $\mu^2\ \mathbf{t}_1'\mathbf{t}_1 \leq \mathrm{Y}'\mathrm{Y}$ puis $\mu^2 \leq 1$.

Notations

Y Vecteur de réponses $n \times 1$, à valeur dans $\{0,1\}^n$ $\mathrm{Y} = (\mathrm{Y}_1, \cdots, \mathrm{Y}_n)'$.

X Matrice de données $n \times p$, $\mathrm{X}_{i.}$ individu i $\mathrm{X}_{.j}$ covariable j.

Y dépend de X au travers du prédicteur linéaire $\theta_0 \mathbb{1}_n + \mathrm{X}\theta_{1:p}$:

$[\mathbb{1}_n \ \mathrm{X}]\,\theta = [\mathbb{1}_n \ \mathrm{X}^{\mathrm{red}}]\,\gamma$, $\mathrm{X}^{\mathrm{red}} \in \mathbb{R}^{n\times r}$ centrée, scalée en col., rang r $\gamma \in \mathbb{R}^{r+1}$.

Le paramètre γ est identifiable et

si $r = p$ *i.e.* X rang plein en col., $\theta \longleftrightarrow \gamma$, ex. : cas $n \geq p$ et $[\mathbb{1}_n \ \mathrm{X}]$ rang plein;

si $r < p$, $\theta \longrightarrow \gamma$, ex. : cas $n < p$ et $[\mathbb{1}_n \ \mathrm{X}]$ rang plein.

$\hookrightarrow$ Dans la suite de la section GLM,

$\mathrm{Y} \in \{0,1\}^n$, $\mathrm{Z}^{\mathrm{red}} = \left[\mathbb{1}_n \ \mathrm{X}^{\mathrm{red}}\right] \in \mathbb{R}^{n\times(r+1)}$, rang plein $(r+1)$

Inférence par Maximum de Vraisemblance, si existence et unicité :

- Solution de l'equation de vraisemblance $(\mathrm{Y}_k, \mathrm{X}_{k\cdot})_k$ indep.

$$Z^{\mathrm{red}\prime}(\mathrm{Y} - \Pi(\gamma)) = 0, \qquad \text{où } \Pi = (\Pi_1, \cdots, \Pi_n)'$$

avec

$$\Pi_i(\gamma) = [1 + \exp(-\eta_i(\gamma))]^{-1} \qquad \text{et} \qquad \eta_i(\gamma) = Z^{\mathrm{red}}_{i\cdot}\gamma.$$

- Algorithme Iteratively Reweighted Least Squares (IRLS) [Green (1995)]

Pour la valeur courante de la suite $\gamma^{(t)}$,

Mise à jour de la matrice de poids

$$W^{(t)} = \mathrm{diag}(W^{(t)}_{kk}, 1 \leq k \leq n), \qquad W^{(t)}_{kk} = \Pi_k(\gamma^{(t)})\left(1 - \Pi_k(\gamma^{(t)})\right).$$

Mise à jour de la pseudo-variable : $\psi^{(t)} = Z^{\mathrm{red}}\gamma^{(t)} + \left[W^{(t)}\right]^{-1}\left(\mathrm{Y} - \Pi(\gamma^{(t)})\right).$

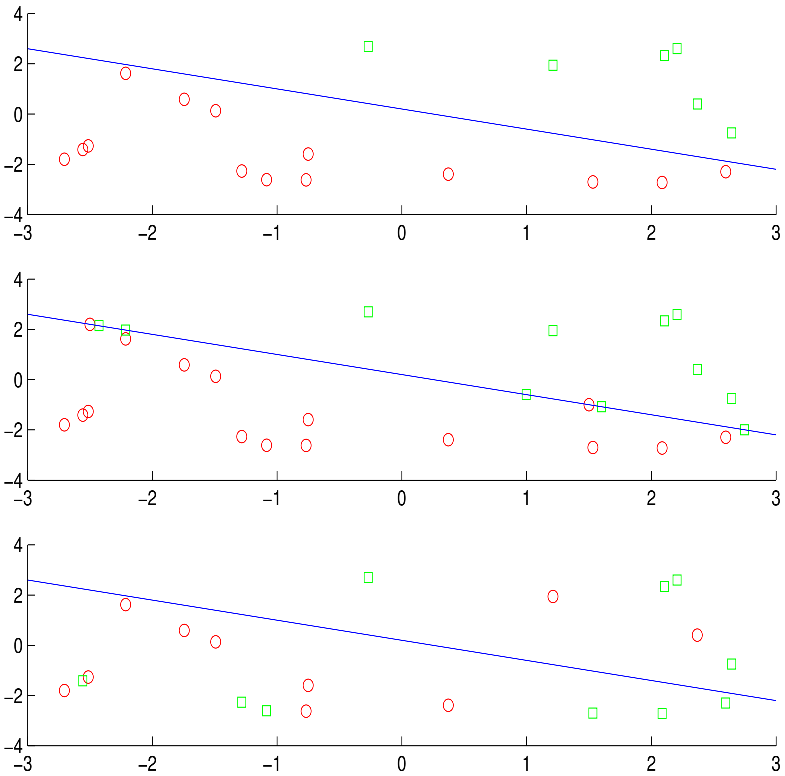
Régression pondérée : $\gamma^{(t+1)} = \left(Z^{\mathrm{red}\prime}W^{(t)}Z^{\mathrm{red}}\right)^{-1} Z^{\mathrm{red}\prime}W^{(t)}\psi^{(t)}.$

A convergence, $\hat{\gamma}^{\mathrm{ML}} = \gamma^{(\infty)}$.

Existence de l'estimateur du MV, $\hat{\gamma}^{\mathrm{ML}}$

$\mathrm{Y} \in \{0,1\}^n$, $\mathrm{Z}^{\mathrm{red}} \in \mathbb{R}^{n\times(r+1)}$, rang plein $(r+1)$

- Cas $n > (r+1)$ [Albert & Anderson (1984)]



$\mathrm{Y}_k = 0$; $\mathrm{Y}_k = 1$; Hyperplan $[1\ x_1\ x_2]\,\gamma = 0$.

Séparation complète : $\exists\gamma$ tq.

$$\mathrm{Z}^{\mathrm{red}}_{k\cdot}\gamma > 0 \text{ si } \mathrm{Y}_k = 1 \quad \mathrm{Z}^{\mathrm{red}}_{k\cdot}\gamma < 0 \text{ si } \mathrm{Y}_k = 0$$

$\hookrightarrow$ non existence.

Séparation quasi - complète : $\exists\gamma$ tq.

$$\mathrm{Z}^{\mathrm{red}}_{k\cdot}\gamma \geq 0 \text{ si } \mathrm{Y}_k = 1 \quad \mathrm{Z}^{\mathrm{red}}_{k\cdot}\gamma \leq 0 \text{ si } \mathrm{Y}_k = 0$$

$\hookrightarrow$ non existence.

Overlap: autres cas.

$\hookrightarrow$ existence et unicité.

- Cas $n = (r+1)$ cas $n << p$, en pratique

$$\forall k, \qquad \hat{\gamma}^{\mathrm{ML}}_k = \ln\left(\frac{\mathrm{Y}_k}{1-\mathrm{Y}_k}\right) \qquad \text{d'où } \|\hat{\gamma}^{\mathrm{ML}}\| = +\infty, \qquad \text{et } \hat{\gamma}^{\mathrm{ML}} \text{ n'existe pas.}$$

Comment étendre PLS aux GLM ?

On a vu :

(i) PLS, pour les modèles linéaires, approche 'moindres carrés';

(ii) Pour les GLM, approche 'maximum de vraisemblance'.

$\hookrightarrow$ Observer le cas 'modèle linéaire gaussien' car :

(i) inférence par moindres carrés $\Leftrightarrow$ inférence par maximum de vraisemblance,

(ii) un tel modèle est un cas particulier de GLM.

Côté technique :

(i) En entrée de PLS : $(\mathrm{Y}, \mathrm{Z}^{\mathrm{red}})$;

(ii) IRLS converge en une itération : $\hat{\gamma}^{\mathrm{ML}}$ est le vecteur des coeff. de

· Régression (pondérée) de Y sur $\mathrm{Z}^{\mathrm{red}}$

· Régression (pondérée) de la pseudo-variable $\psi^{(\infty)} = \mathrm{Y}$ à convergence de IRLS, sur $\mathrm{Z}^{\mathrm{red}}$.

$\hookrightarrow$ Appliquer PLS (pondéré), avec en entrée $(\psi^{(\infty)}, \mathrm{Z}^{\mathrm{red}})$. Mais $\cdots$

Ridge-PLS (Bernoulli) non restrictif

- INPUT : $\mathrm{Y} \in \{0,1\}^n$; $\mathrm{X} \in \mathbb{R}^{n\times p}$, centrée, scalée en col.; $\lambda > 0$; $\kappa > 0$ entier.

- ALGORITHME : soit $\mathrm{Z}^{\mathrm{red}} = [\mathbb{1}_n\ \mathrm{X}^{\mathrm{red}}]$, de rang $(r+1)$

Step A. Calcul de la pseudo-variable à convergence de l'algorithme *Iteratively Reweighted Ridge Least Squares*

(a) calcul itératif de l'estimateur de MV pénalisé

$$\hat{\gamma}^{\mathrm{RLS}} = \mathrm{argmax}_\gamma \left\{ l(\gamma) - \lambda \|\gamma_{2:(r+1)}\|^2 \right\};$$

(b) pseudo-variable à convergence

$$\psi^{(\infty)} = \mathrm{Z}^{\mathrm{red}}\hat{\gamma}^{\mathrm{RLS}} + [\mathrm{W}^{(\infty)}]^{-1}\left(\mathrm{Y} - \Pi(\hat{\gamma}^{\mathrm{RLS}})\right).$$

Step B. Algorithme PLS pondéré : κ composantes, en entrée $(\psi^{(\infty)}, \mathrm{Z}^{\mathrm{red}})$ et un poids $\mathrm{W}^{(\infty)}$.

- OUTPUT : $\hat{\gamma}^{\mathrm{PLS},\kappa} \in \mathbb{R}^{r+1}$, coeff. w.r.t. $\mathrm{Z}^{\mathrm{red}} \longrightarrow \hat{\theta}^{\mathrm{PLS},\kappa} \in \mathbb{R}^{p+1}$, coeff. w.r.t. $[\mathbb{1}_n\ \mathrm{X}]$.

- Rmq : λ choisi par critère BIC, en fin de Step A

$$\lambda = \mathrm{argmin}\left\{ -2\, l(\hat{\gamma}^{\mathrm{RLS}}) + \log n\ \mathrm{Tr}\left[\mathrm{Z}^{\mathrm{red}}\left(\mathrm{Z}^{\mathrm{red}\prime}\mathrm{W}^{(\infty)}\mathrm{Z}^{\mathrm{red}} + \lambda \mathrm{R}\right)^{-1}\mathrm{Z}^{\mathrm{red}\prime}\mathrm{W}^{(\infty)}\right]\right\}.$$

Utilisations de PLS dans les GLM (Bernoulli)

Etant donnés : $\mathrm{Y} \in \{0,1\}^n$; $\mathrm{X} \in \mathbb{R}^{n \times p}$, centrée, scalée en col.; $\kappa > 0$ entier.

- Méthode 1 [Nguyen & Rocke (2002)]
  - · Algorithme PLS avec (Y, X) en entrée; pour κ composantes.
  - · $\hat{\theta}$ défini comme l'estimateur du MV lorsque la matrice de design est $[\mathbb{1}_n \mathbf{t}_1 \cdots \mathbf{t}_\kappa]$.
  - STOP Input binaire; non existence de l'estimateur MV.

- Méthode 2 [Marx (1996)]
  - · Algorithme IRLS en remplaçant chaque régression LS par une régression PLS, avec $\kappa_{\max}$ composantes.
  - · A convergence, régression (sens ML) de Y sur κ composantes PLS obtenues à convergence de l'étape 1.
  - STOP Usage de $\kappa_{\max}$; non existence de l'estimateur MV.

- Méthode 3 [Ding & Gentleman (2003)]
  - · Critère de maximum de vraisemblance pénalisé, basé sur la pénalité de Firth.
  - · Max le critère $\leftrightarrow$ algo itératif de régression LS pondérée : substitution par de la régression PLS pondérée, à κ composantes.

Jeu de données initial $\longrightarrow$ Jeu de données étudié : pré-filtrage puis pré-sélection

$\hookrightarrow$ Gènes non informatifs car comportement 'uniforme' quelle que soit la classe de l'individu.

▶ Data set : subdivision en $\mathcal{L}$, *learning set* et $\mathcal{T}$, *test set*

▶ Pré-filtrage :

(i) Seuillage $\mathrm{X}_{ij} \in [100, 16000]$.

(ii) On ne garde que les gènes $\mathrm{X}_{.j}$ tq

$$\max_{k \in \mathcal{L}} \mathrm{X}_{kj} - \min_{k \in \mathcal{L}} \mathrm{X}_{kj} > 500 \qquad\qquad \frac{\max_{k \in \mathcal{L}} \mathrm{X}_{kj}}{\min_{k \in \mathcal{L}} \mathrm{X}_{kj}} > 5.$$

(iii) Transformation $\log_{10}$.

(iv) X centrée réduite en ligne.

D'où une matrice $\mathrm{X} \in \mathbb{R}^{n \times p_{\max}}$.

▶ Pré-sélection : les gènes sont ordonnés selon leur pouvoir 'discriminant' évalué par

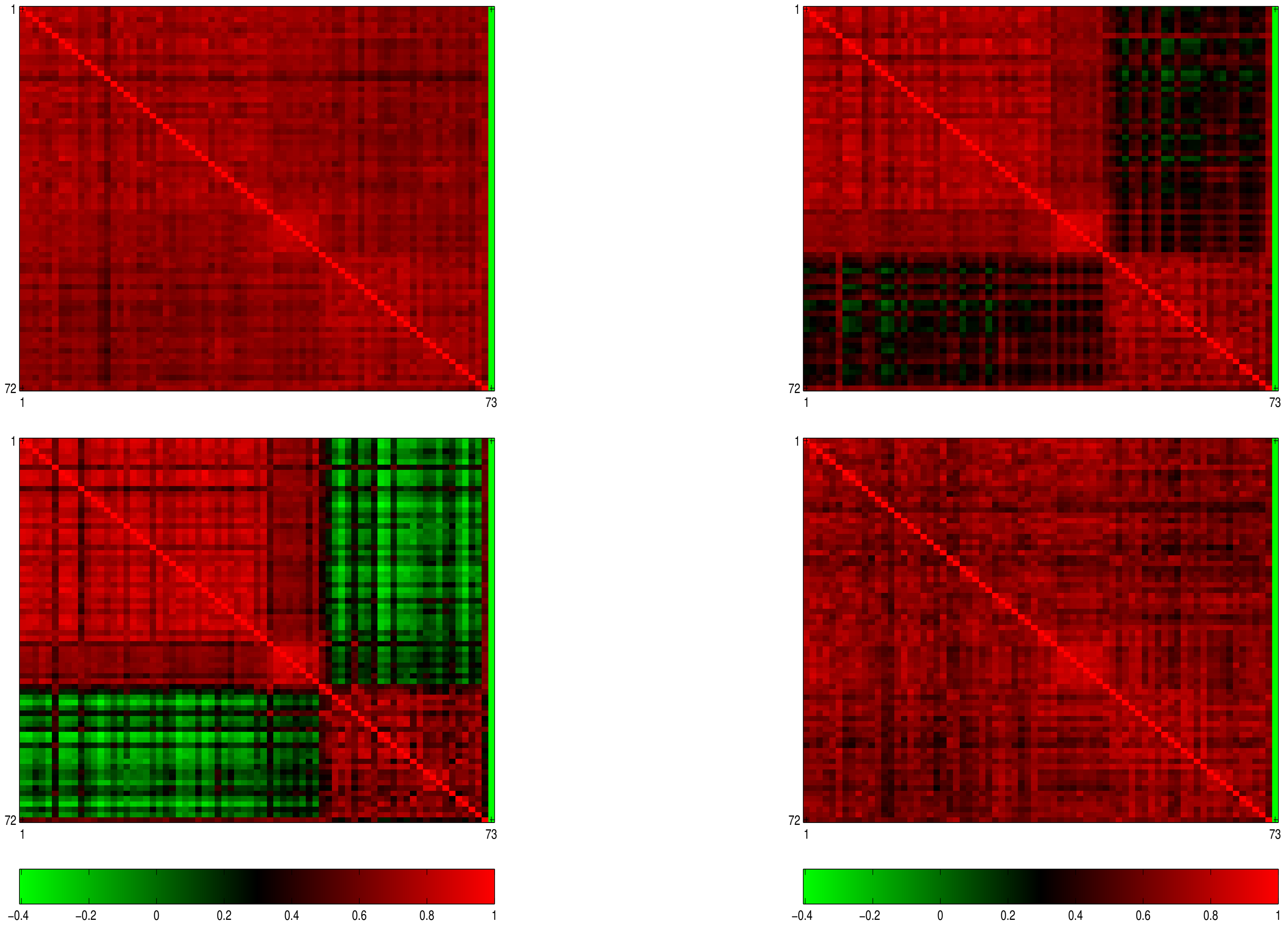
$$\frac{\sum_{k=1}^{nl} \sum_{i=1}^{2} (\mathrm{X}_{k,j} - m_j)^2 \mathbb{1}_{\mathrm{Y}_k = i}}{\sum_{k=1}^{nl} \sum_{i=1}^{2} (\mathrm{X}_{k,j} - m_{ij})^2 \mathbb{1}_{\mathrm{Y}_k = i}}, \qquad \text{sur les échantillons de } \mathcal{L}.$$

D'où une matrice $\mathrm{X} \in \mathbb{R}^{n \times p}$, $\quad p \leq p_{\max}$, avec $p_{\max}$ fonction de la subdivision $(\mathcal{L}, \mathcal{T})$.

Visualisation 'naïve' de l'effet de la pré-sélection

LEUKEMIA

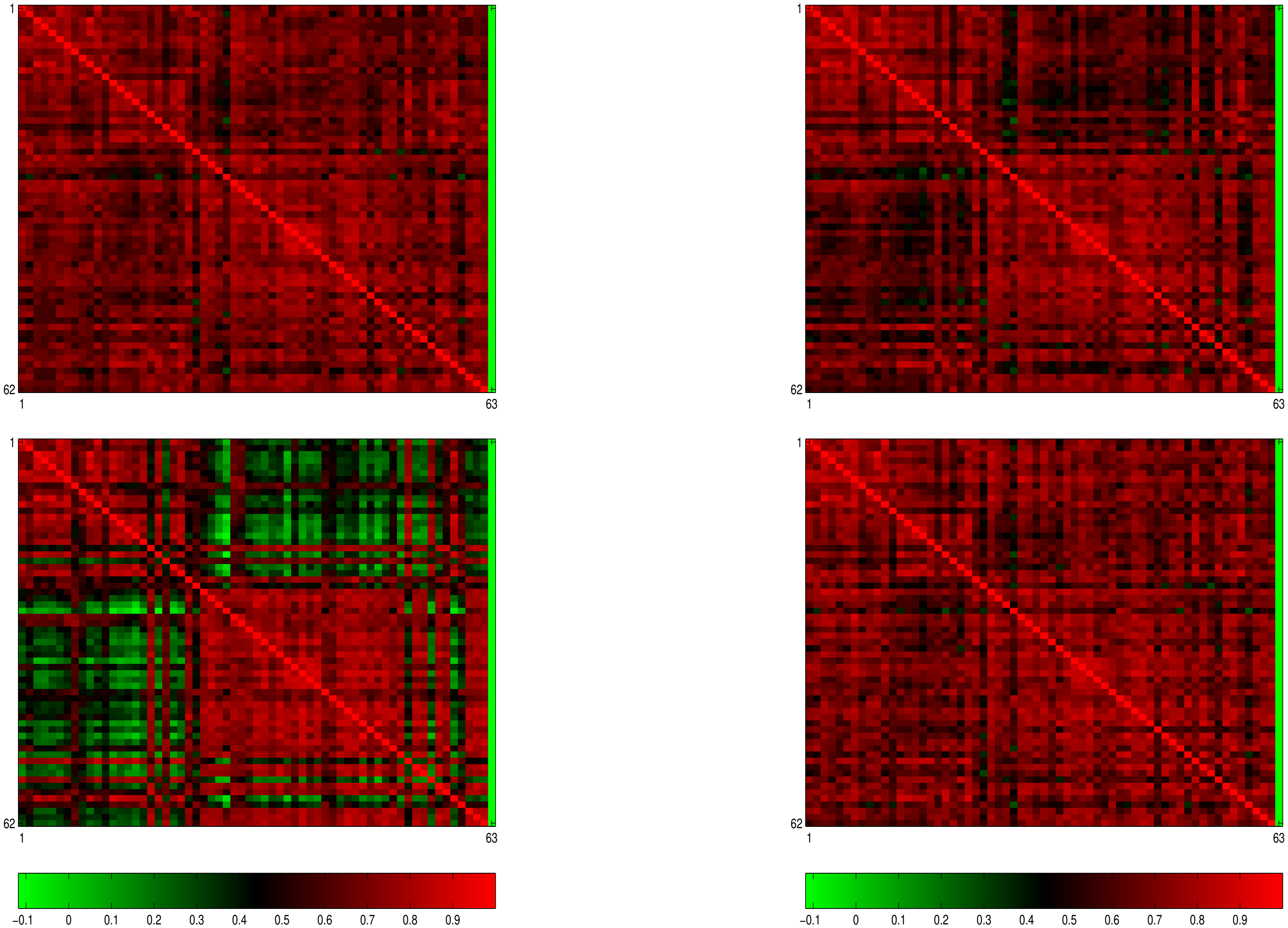
$\hookrightarrow$ Distance entre les 72 individus, représentés par un vecteur de $\mathbb{R}^p$: matrice de corrélation 72×72, lorsque $p = 3571$ [HG], $p = 500$ [HD], $p = 100$ [BG] et 100 au hasard [BD].



Visualisation 'naïve' de l'effet de la pré-sélection

COLON

$\hookrightarrow$ Distance entre les 62 individus, représentés par un vecteur de $\mathbb{R}^p$: matrice de corrélation 62×62, lorsque $p = 2000$ [HG], $p = 500$ [HD], $p = 100$ [BG] et 100 au hasard [BD].



Analyse Out of Sample

LEUKEMIA

sur la subdivision $(\mathcal{L}, \mathcal{T})$ de Golub (1999) : soit resp. 38 (27/11) ind. et 34 (20/14) ind.

- Résultats connus [Dudoit *et al.*(2000)] : 66 AML, 67 ALL sont difficiles à classer.

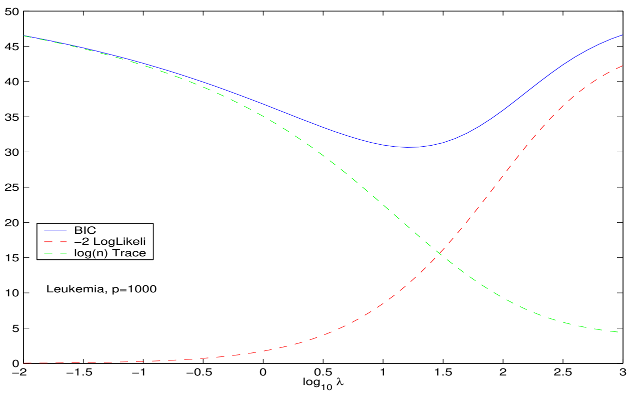
- Nombre d'ind. mal classés (parmi 34)

	R	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	RPLS6	DLDA	DQDA
50	1	1	1	1	1	1	1	1	1
100	1	1	1	2	3	2	2	1	4
300	3	3	1	2	2	1	1	2	1
500	3	3	2	3	3	3	2	2	1
1000	3	3	3	3	3	3	3	2	2

· Ridge - RPLS : ind. mal classés $54, 60, 66$ dans resp. $51, 51, 100\%$ des cas.

- λ minimisant le critère sur l'échelle $10^{-2:0.1:3}$

p	50	100	300	500	1000
λ_*	0.32	0.80	3.16	6.31	15.85



Analyse Leave One Out

LEUKEMIA

sur $\mathcal{L}$ de Golub (1999) : soit 38 (27/11) ind.

- Résultat connu [Dudoit *et al.*(2000)] : ind. 35 AML, $\mathcal{L}$ difficile à classer.

- Nombre d'ind. mal classés (parmi 38)

	R	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	RPLS6	DLDA	DQDA
50	0	1	1	0	1	1	1	1	1
100	1	1	0	0	1	1	0	1	1
300	2	0	1	1	0	1	1	1	1
500	2	0	0	0	0	0	0	1	2
1000	2	1	0	0	1	0	0	1	0

· Ridge - RPLS : ind. mal classés $12, 32, 35$ dans resp. $8, 14, 37\%$ des cas.

- λ minimisant le critère sur l'échelle $10^{-2:0.1:3}$

p	50	100	300	500	1000
λ_*	0.37	0.86	3.31	6.31	15.85

Analyse Resampling LEUKEMIA

100 subdivisions $(\mathcal{L}, \mathcal{T})$ du jeu de données en resp. 38 (27/11) ind. et 34 (20/14) ind.

- Nombre moyen (et std) d'ind. mal classés (parmi 34) $m_1^2 + \sigma_1^2 \leq m_2^2 + \sigma_2^2$.

	Ridge	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	DLDA	DQDA
50	1.52 (1.11)	1.24 (0.93)	1.36 (0.98)	1.43 (1.01)	1.40 (0.94)	1.40 (0.95)	2.01 (1.25)	3.27 (1.48)
100	1.35 (1.09)	1.18 (0.98)	1.24 (0.91)	1.32 (0.91)	1.34 (0.93)	1.33 (0.96)	1.96 (1.21)	3.20 (1.54)
300	1.62 (1.15)	1.12 (0.86)	1.15 (0.83)	1.10 (0.77)	1.09 (0.79)	1.08(0.80)	2.28 (1.52)	3.49 (1.45)
500	1.89 (1.21)	1.20 (0.97)	1.08 (0.79)	1.06 (0.79)	1.12 (0.78)	1.12 (0.77)	2.67 (1.63)	3.64 (1.49)
1000	2.83 (1.37)	1.45 (1.07)	1.27 (0.96)	1.14 (0.82)	1.39 (0.94)	1.24 (0.74)	3.44 (1.78)	3.82 (1.44)

- λ minimisant le critère sur l'échelle $10^{-2:0.1:3}$

p	50	100	300	500	1000
λ_*	0.38	0.94	3.76	7.30	18.42

Analyse Leave One Out COLON

sur 62 subdivisions $(\mathcal{L}, \mathcal{T})$ du data set tq. $|\mathcal{T}| = 1$.

- Résultat connu $N8, 12, 34$ et $T2, 30, 33, 36, 37$ ind. sont mal classés dans [Alon *et al.*, 2000] et $N8, 34, 36$ et $T30, 33, 36$ ind. sont mal classés dans [Furey *et al.*, 2000].

- Nombre d'ind. mal classés (parmi 62) xx syst. mal classés.

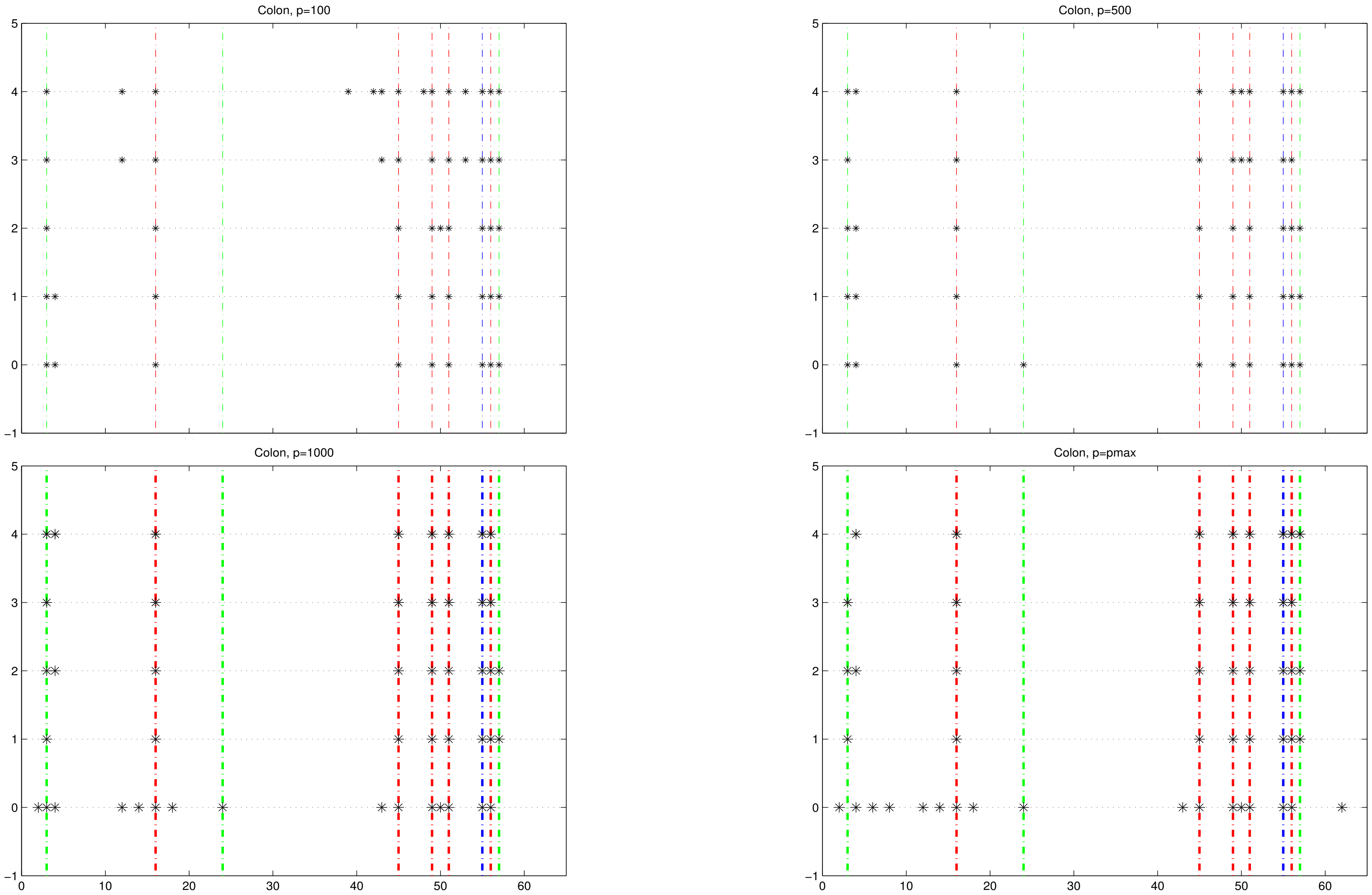
	R	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	RPLS6	DLDA	DQDA
100	9	9	9	11	14	14	12	17	17
300	9	8	8	10	11	12	12	17	21
500	10	9	9	8	10	12	11	18	22
1000	15	8	9	7	8	11	11	20	23
$p_{\max}$	17	8	9	7	8	12	9	22	25

- λ minimisant le critère sur l'échelle $10^{-2:0.1:3}$

p	100	300	500	1000	$p_{\max}$
λ_*	3.10	9.38	15.49	31.10	43.87

Analyse Leave One Out (Index des ind. mal classés) COLON

pour $p = 100, 500, 1000, p_{\max}$; pour Ridge et RidgePLS1 à RidgePLS4; Alon, Furey, les 2



Analyse Resampling COLON

100 subdivisions $(\mathcal{L}, \mathcal{T})$ du jeu de données en resp. 41 (14/27) ind. et 21 (8/13) ind.

- Nombre moyen (et std) d'ind. mal classés (parmi 21) $m_1^2 + \sigma_1^2$ min.

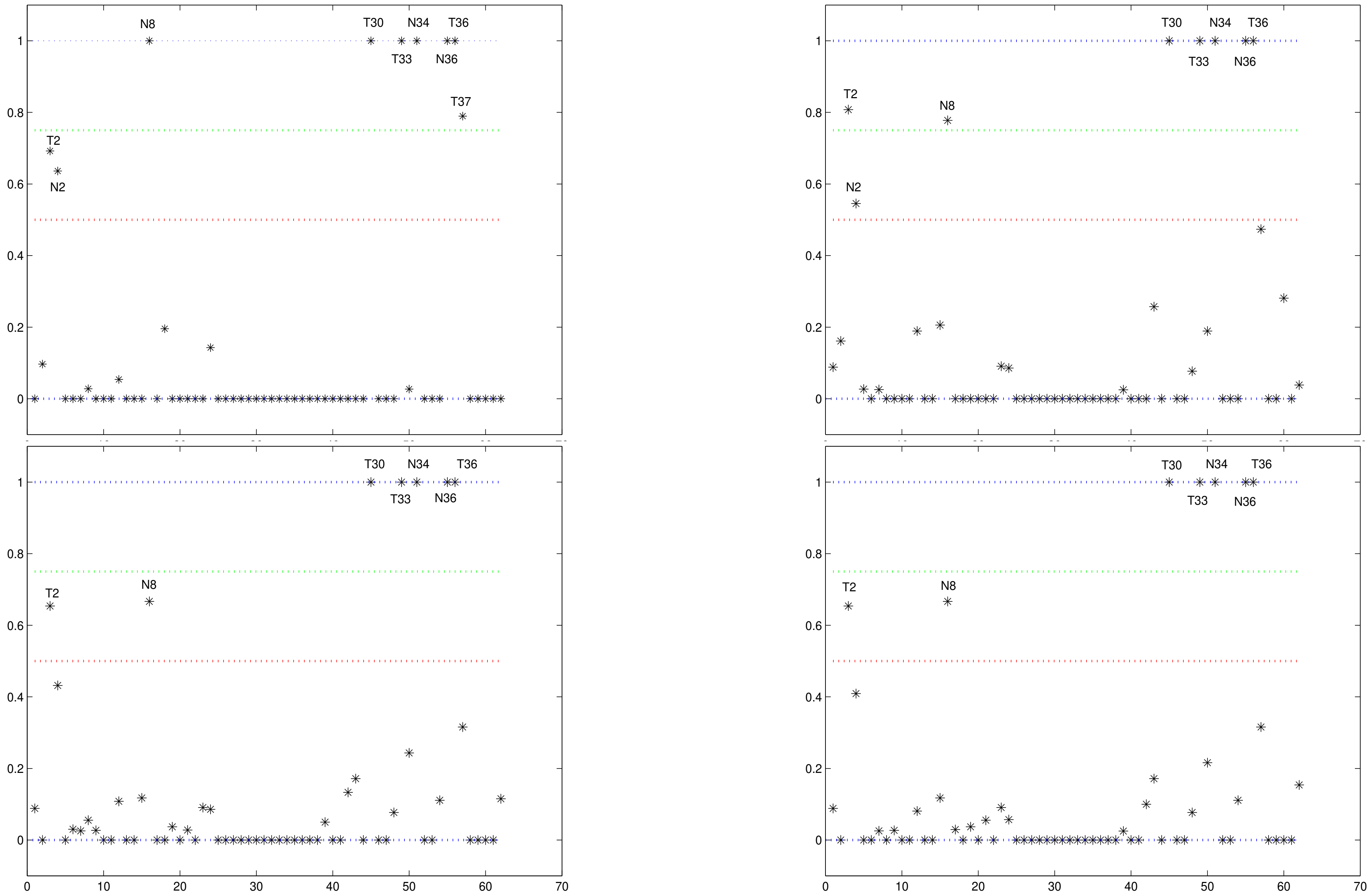
	Ridge	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	DLDA	DQDA
100	3.19 (1.19)	3.16 (1.22)	3.72 (1.36)	3.83 (1.36)	4.02 (1.39)	4.01 (1.40)	3.44 (1.42)	4.09 (1.38)
300	3.66 (1.27)	3.33 (1.31)	3.57 (1.34)	3.51 (1.17)	3.77 (1.23)	3.86 (1.13)	4.75 (1.68)	5.67 (1.66)
500	4.19 (3.40)	3.40 (1.36)	3.34 (1.27)	3.37 (1.30)	3.65 (1.23)	3.71 (1.11)	6.19 (1.94)	6.57 (1.74)
1000	6.56 (1.43)	3.48 (1.44)	3.08 (1.21)	3.10 (1.14)	3.40 (1.20)	3.54 (1.17)	7.42 (1.78)	7.40 (1.79)
$p_{\max}$	7.40 (1.13)	3.46 (1.44)	3.08 (1.24)	3.04 (1.17)	3.39 (1.17)	3.57 (1.13)	7.43 (1.84)	7.59 (1.78)

- λ minimisant le critère sur l'échelle $10^{-2:0.1:3}$

p	50	100	300	500	1000	$p_{\max}$
λ_*	1.32	2.85	9.39	17.11	49.82	66.62

Analyse Resampling ('proba' mal classé) COLON

pour $p = 100, 500, 1000, p_{\max}$; pour resp. RidgePLS1, RidgePLS2, RidgePLS3, RidgePLS3.



- Nous avons proposé
  - une méthode de réduction de dimension dans les GLM, adaptée au cadre statistique 'grande dimension' $n << p$, et au mauvais conditionnement (forte colinéarité) des régresseurs.
  - une extension de PLS aux GLM.
- Les applications considérées ont souligné l'intérêt de l'étape de réduction de dimension après une étape de régularisation.
- On pourra poursuivre ces travaux,
  - pour les étendre au cas de la classification multi-classe;
  - pour étudier l'influence de la technique de régularisation (choix de la pénalité).
  - pour mieux contrôler la technique PLS ...

Analyse Resampling LEUKEMIA

100 subdivisions $(\mathcal{L}, \mathcal{T})$ du jeu de données en resp. 38 (27/11) ind. et 34 (20/14) ind.

	RPCR1	RPCR2	RPCR3	RPCR4	RPCR5	RPCR6
50	1.73 (1.16)	1.74 (1.16)	1.70 (1.10)	1.80 (1.10)	1.80 (1.14)	1.84 (1.13)
100	1. 51 (1.00)	1.39 (0.99)	1.41 (0.91)	1.56 (1.02)	1.58 (1.04)	1.63 (0.99)
300	1.47 (1.11)	1.31 (0.97)	1.35 (0.95)	1.35 (0.89)	1.33 (0.88)	1.39 (0.82)
500	1.32 (0.92)	1.35 (0.98)	1.32 (0.95)	1.22 (0.82)	1.32 (0.89)	1.24 (0.82)
1000	1.36 (0.98)	1.31 (1.01)	1.26 (0.95)	1.30 (0.87)	1.21 (1.00)	1.16 (0.90)

	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	RPLS6
50	1.24 (0.93)	1.36 (0.98)	1.43 (1.01)	1.40 (0.94)	1.40 (0.95)	1.43 (0.97)
100	1.18 (0.98)	1.24 (0.91)	1.32 (0.91)	1.34 (0.93)	1.33 (0.96)	1.27 (0.89)
300	1.12 (0.86)	1.15 (0.83)	1.10 (0.77)	1.09 (0.79)	1.08(0.80)	1.12 (0.79)
500	1.20 (0.97)	1.08 (0.79)	1.06 (0.79)	1.12 (0.78)	1.12 (0.77)	1.13 (0.79)
1000	1.45 (1.07)	1.27 (0.96)	1.14 (0.82)	1.39 (0.94)	1.24 (0.74)	1.25 (0.77)

Analyse Resampling

COLON

100 subdivisions $(\mathcal{L}, \mathcal{T})$ du jeu de données en resp. 41 (14/27) ind. et 21 (8/13) ind.

	RPCR1	RPCR2	RPCR3	RPCR4	RPCR5	RPCR6
100	3.29 (1.34)	3.08 (1.33)	3.03 (1.33)	3.25(1.27)	3.33 (1.29)	3.41 (1.28)
300	3.38 (1.39)	3.27 (1.38)	3.05 (1.37)	3.09 (1.38)	3.01 (1.28)	3.05 (1.29)
500	3.63 (1.45)	3.40 (1.36)	3.12 (1.43)	3.05 (1.34)	3.02 (1.29)	2.89 (1.25)
1000	5.01 (1.58)	4.34 (1.80)	3.46 (1.66)	3.09 (1.37)	2.96 (1.35)	2.61 (1.27)
$p_{\max}$	5.75 (1.74)	4.96 (1.71)	3.64 (1.74)	3.25 (1.37)	3.02 (1.39)	2.58 (1.26)

	RPLS1	RPLS2	RPLS3	RPLS4	RPLS5	RPLS6
100	3.16 (1.22)	3.72 (1.36)	3.83 (1.36)	4.02 (1.39)	4.01 (1.40)	4.02 (1.44)
300	3.33 (1.31)	3.57 (1.34)	3.51 (1.17)	3.77 (1.23)	3.86 (1.13)	3.91 (1.06)
500	3.40 (1.36)	3.34 (1.27)	3.37 (1.30)	3.65 (1.23)	3.71 (1.11)	3.75 (1.10)
1000	3.48 (1.44)	3.08 (1.21)	3.10 (1.14)	3.40 (1.20)	3.54 (1.17)	3.50 (1.11)
$p_{\max}$	3.46 (1.44)	3.08 (1.24)	3.04 (1.17)	3.39 (1.17)	3.57 (1.13)	3.54 (1.10)