ML 03 Régularisation de Tikhonov

1 Problèmes linéaires mal posés

Avec moins d'observations que de fonctions de base ($M < N$), le système $\mathbf{Ax} = \mathbf{b}$ ne possède pas de solution unique. Même quand $M \geq N$, le système linéaire peut posséder une solution approchée préférable à la solution optimale. C'est en particulier vrai quand plusieurs observations sont très proches à un coefficient multiplicatif près (on parle de colinéarités). Par exemple, soit le système linéaire suivant :

$$\begin{pmatrix} 1 & 1 \\ 1 & 1.00001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.99 \end{pmatrix}$$

Sa solution est $\mathbf{x}^T = (1001, -1000)$. Cependant, la solution approchée $\mathbf{x}^T = (0.5, 0.5)$ semble préférable. En effet, la solution optimale a peu de chance de bien s'adapter à de nouvelles observations (par exemple, l'observation $(1, 2)$ serait projetée sur l'étiquette -999).

2 Ajout de contraintes de régularité

Ainsi, lorsqu'il faut choisir entre plusieurs solutions, il peut être efficace d'exprimer une préférence envers celles dont les coefficients (ou paramètres) ont de faibles valeurs. Cela consiste par exemple à minimiser $|x_1| + |x_2| + \ldots$ (aussi noté $\|\mathbf{x}\|_1$, la "norme 1") ou encore $x_1^2 + x_2^2 + \ldots$ (aussi noté $\|\mathbf{x}\|_2^2$, le carré de la "norme 2"). Dans ce dernier cas, il s'agit de résoudre un nouveau problème de minimisation :

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2 + \alpha \|\mathbf{x}|_2^2$$
$$avec \quad 0 \leq \alpha$$

C'est encore un problème de minimisation quadratique en $\mathbf{x}$ dont le minimum se découvre par annulation de la dérivée.

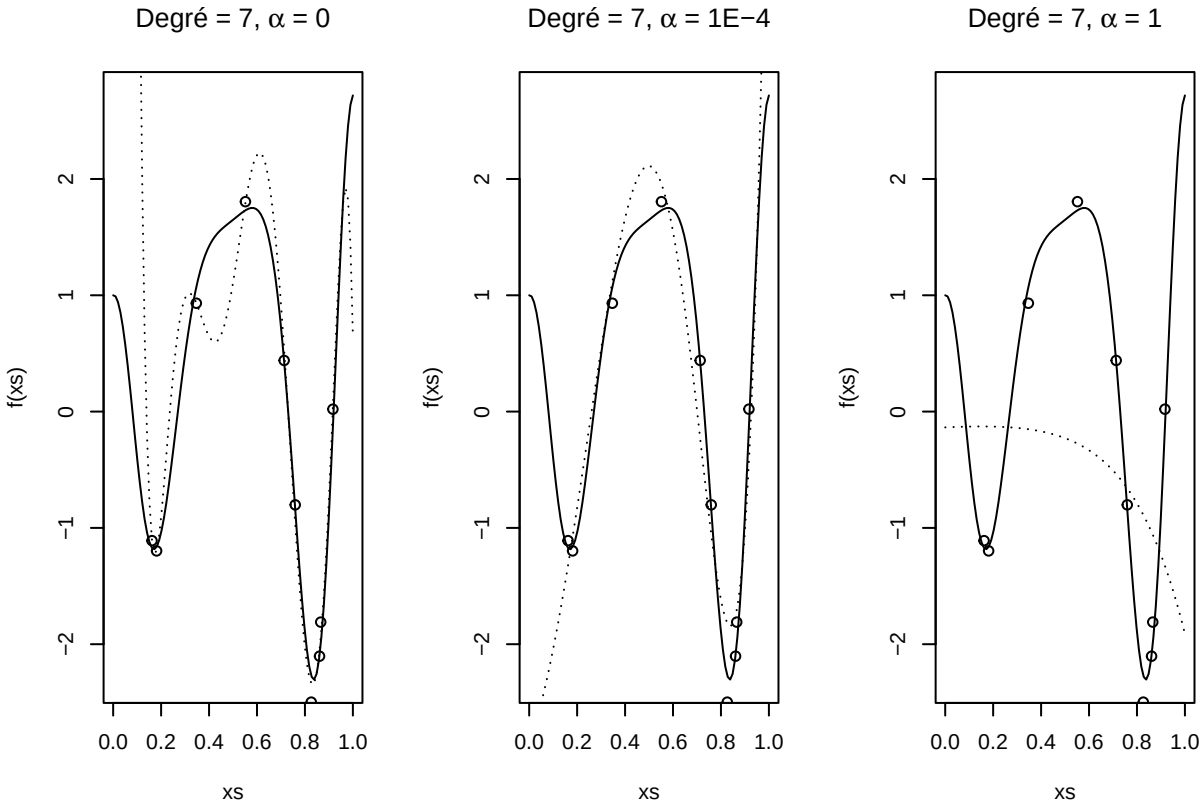
$$\mathbf{0} = 2\mathbf{A}^T\mathbf{Ax} - 2\mathbf{A}^T\mathbf{b} + 2\alpha\mathbf{x}$$
$$=$$
$$\left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I}_{n\times n}\right)\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

En pratique, il s'agit donc d'ajouter une petite valeur positive α aux éléments de la diagonale de la matrice de Gram. Cette approche porte plusieurs noms dont "régularisation de Tikhonov" ou "régression Ridge".

3 Visulation de l'effet sur les paramètres de différents niveaux de régularisation

Sur notre exemple synthétique, nous affichons la fonction génératrice, le jeu de donnée et le polynôme de degré au plus sept découvert par régression ridge avec une valeur de α égale soit à 0, soit à 10^{-4}, soit à 1.

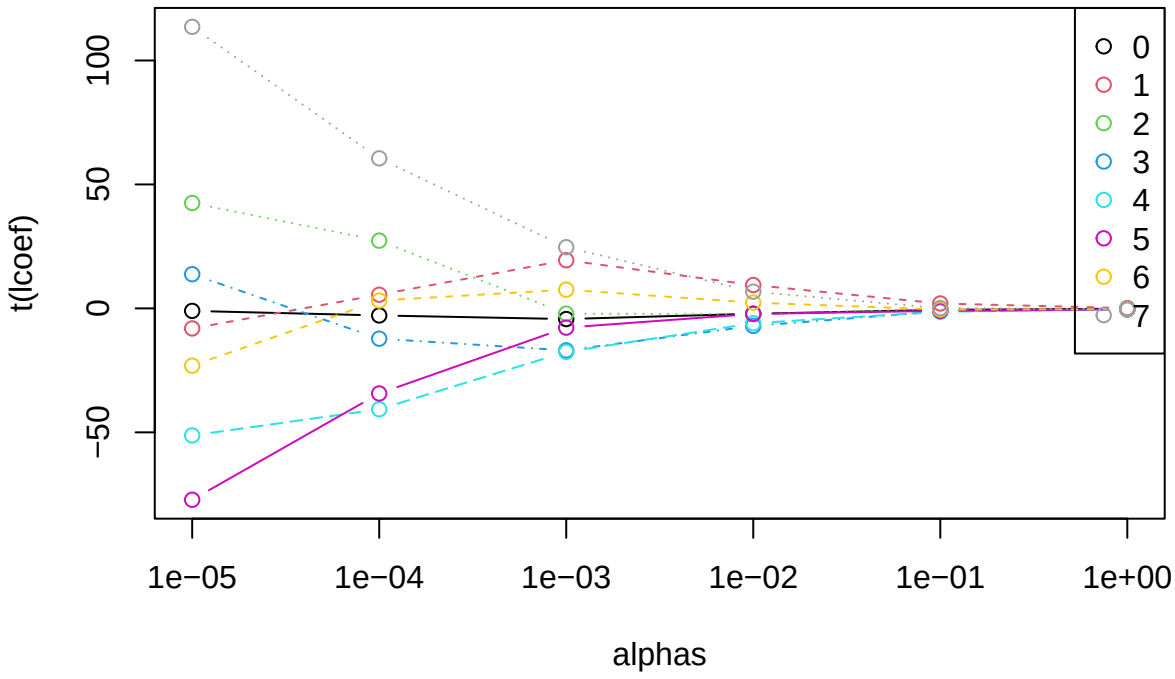
```
set.seed(1123)
# Image par f d'un échantillon uniforme sur l'intervalle [0,1], avec ajout d'un
# bruit gaussien de moyenne nulle et d'écart type 0.2
data = gendat(10,0.2)

par(mfrow=c(1,3))
coef <- ridge(0, data, 7)
plt(data,f,main=expression(paste(plain("Degré = "), 7, plain(", "), alpha,
                                 plain(" = 0"))))
pltpoly(coef)
coef <- ridge(1E-4, data, 7)
plt(data,f,main=expression(paste(plain("Degré = "), 7, plain(", "), alpha,
                                 plain(" = 1E-4"))))
pltpoly(coef)
coef <- ridge(1, data, 7)
plt(data,f,main=expression(paste(plain("Degré = "), 7, plain(", "), alpha,
                                 plain(" = 1"))))
pltpoly(coef)
```



Plus le coefficient de régularisation α est faible, moins il contraint les paramètres (c'est-à-dire, les coefficients du polynôme) à conserver de petites valeurs et plus le polynôme découvert peut être complexe, au risque de provoquer du sur-apprentissage et donc de limiter la capacité du modèle à bien prédire les étiquettes de nouvelles observations. A l'inverse, plus le coefficient de régularisation est élevé, plus le modèle découvert sera simple, au risque de sous-apprendre en ne modélisant pas convenablement les variations propres au processus qui a généré les observations.

Pour mieux visualiser l'effet du coefficient de régularisation ridge, nous affichons les valeurs des coefficients du polynôme découvert pour différentes valeurs de α. Plus α augmente, plus les coefficients du polynôme diminuent et tendent vers 0.

```
alphas <- c(1E-5, 1E-4, 1E-3, 1E-2, 1E-1, 1)
lcoef <- sapply(alphas, ridge, data, 7)
matplot(alphas, t(lcoef), type=c("b"), pch=1, col=1:8, log="x")
legend("topright", legend = 0:7, col=1:8, pch=1)
```



4 Régularisation et complexité

Essayons de mieux comprendre les raisons pour lesquelles la régularisation peut être efficace. Nous allons montrer qu'elle réduit la complexité d'un modèle prédictif. Dans ce contexte, qu'est-ce que la complexité ? C'est la sensibilité du modèle au bruit présent dans les données. Qu'est-ce que le bruit ? Il est possible de le définir après avoir fait l'hypothèse d'une famille de modèles qui auraient pu générer les données observées.

Prenons l'exemple d'un modèle linéaire : $F(\mathbf{X}) = \mathbf{W}^T\mathbf{X} + b$. Posons ensuite $\mathbf{X}^* = \mathbf{X} + \epsilon$ avec ϵ un faible bruit ($\|\epsilon\|$ est petit). $\mathbf{X}$ et $\mathbf{X}^*$ étant proches, une hypothèse de régularité demande que $F(\mathbf{X})$ et $F(\mathbf{X}^*)$ le soient aussi. La proximité de $F(\mathbf{X})$ et $F(\mathbf{X}^*)$ peut se mesurer avec $|F(\mathbf{X}) - F(\mathbf{X}^*)|$.

$$
\begin{aligned}
& |F(\mathbf{X}) - F(\mathbf{X}^*)| \\
= & \{F(\mathbf{X}) = \mathbf{W}^T\mathbf{X} + b\} \\
& |\mathbf{W}^T\mathbf{X} - \mathbf{W}^T\mathbf{X}^*| \\
= & \\
& |\mathbf{W}^T(\mathbf{X} - \mathbf{X}^*)| \\
= & \{\mathbf{X}^* = \mathbf{X} + \epsilon\} \\
& |\mathbf{W}^T\epsilon| \\
\leq & \{\text{Inégalité de Cauchy-Schwarz}\} \\
& \|\mathbf{W}\|_2\|\epsilon\|_2
\end{aligned}
$$

Donc, en pénalisant les grandes valeurs de $\|\mathbf{W}\|_2$, on réduit la sensibilité du modèle à de petites perturbations

dans les données observées. Autrement dit, par l'ajout de régularisation, les prédictions associées aux plus proches voisins de $\mathbf{X}$ tendent à être similaires à celle associée à $\mathbf{X}$.