TD 6 - Régression Linéaire, Pénalisation.

Exercice 6.1. Soient X variable aléatoire à valeur dans $\mathbb{R}^p$ et Y variable aléatoire à valeurs dans $\mathbb{R}$. Soient N couples indépendants d'observations $(x_i, y_i)_{1\leq i\leq N}$. On s'intéresse à la régression linéaire de Y sachant X:

$$Y = \beta_0 + \beta^t X + \xi, \text{ avec } \xi \sim \mathcal{N}(0, \sigma^2).$$

On suppose que $\overline{x} = 0$ et $\overline{y} = 0$. On note x_{ij}, la j-ème composante des variables explicatives de l'échantillon i, x_i. Notons:

$$A = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \dots & x_{Np} \end{pmatrix}.$$

0. Comment peut-on s'assurer que $\overline{x} = 0$ et $\overline{y} = 0$?

1. Montrer qu'alors on peut prendre $\beta_0 = 0$. C'est ce que nous considèrerons pour la suite.

2. On suppose que $p > N$. La régression linéaire classique n'est alors pas possible (pourquoi ?). On utilise une pénalisation ℓ^2 et on met en oeuvre la régression ridge. Expliquer pourquoi celle-ci permet de résoudre le problème et rappeler la formule de l'estimateur ridge $\hat{\beta}^{ridge}$ pour β.

3. a) On suppose que A est de rang r. Montrer qu'il existe $V \in \mathcal{M}_p$ orthogonale, $V = (V_1, V_2)$ avec $V_1 \in \mathcal{M}_{p,r}$, $V_2 \in \mathcal{M}_{p,p-r}$, $D = \text{diag}(d_1, \dots, d_r)$, avec $d_1 \geq d_2 \geq \dots \geq d_r > 0$ telles que:

$$V_1^T A^T A V_1 = D \text{ et } A V_2 = 0 .$$

b) En déduire qu'il existe $U \in \mathcal{M}_{N,N}(\mathbb{R})$ orthogonale ($U^T U = I_p$), $\Sigma \in \mathcal{M}_{N,p}(\mathbb{R})$, matrice diagonale $\Sigma = \text{Diag}(d_1, \dots, d_r, 0, \dots, 0)$, avec $d_1 \geq d_2 \geq \dots \geq d_r > 0$, et $V \in \mathcal{M}_{p,p}$ orthogonale telles que:

$$A = U \Sigma V^T.$$

On parle de décomposition en valeurs singulières.

4. Soit λ le coefficient de pénalisation pour la régression ridge, et u_j, $1 \leq j \leq r$, les r premières colonnes de la matrice U.

a) Montrer qu'alors:

$$A\hat{\beta}^{ridge} = \sum_{j=1}^{r} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y .$$

Comparer à ce que l'on obtiendrait pour la régression linéaire classique dans le cas où le rang A est égal à p.

b) Soit v_j, $1 \leq j \leq p$ les vecteurs colonnes de V. Calculer la variance empirique sur l'échantillon du vecteur aléatoire obtenu par projection orthogonale de X sur v_j. En déduire une interprétation du résultat obtenu à la question a).

c) Si à partir des données, on souhaite construire un modèle linéaire avec un seul paramètre, quelle serait votre suggestion en vous inspirant des questions a) et b) ? Et avec q paramètres, $q < r$?

Solution: 0. On peut toujours réaliser un changement de variables et prendre $\tilde{x}_i = x_i - \overline{x}$ et $\tilde{y}_i = y_i - \overline{y}$. On a bien: $\overline{\tilde{x}} = 0$ et $\overline{\tilde{y}} = 0$.

1. $R(\beta_0, \beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta^T x_i\right)^2$

Au minimum: $\nabla R\left(\hat{\beta}_0, \hat{\beta}\right) = 0$. Or nous, avons:

$$\frac{\partial R}{\partial \beta_0}(\beta_0, \beta) = -2 \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i\right)$$

Donc

$$\frac{\partial R}{\partial \beta_0}\left(\hat{\beta}_0, \hat{\beta}\right) = 0 \Leftrightarrow \hat{\beta}_0 = \frac{1}{N} \sum_{i=1}^{N} y_i - \beta^T x_i = \bar{y} - \beta^\top \bar{x} = 0$$

.

2. On suppose que $p > N$. A n'est pas injective, et il existe une infinité de solutions. On utilise la régression ridge. On minimise $J(\beta)$, avec

$$J(\beta) = \|A\beta - y\|^2 + \lambda \|\beta\|^2, \lambda > 0$$

(La norme est la norme L^2.)

On prend le gradient:

$$\nabla J(\beta) = 2A^\top (A\beta - y) + 2\lambda\beta$$

et donc

$$\nabla J(\hat{\beta}) = 0 \Longleftrightarrow \left(A^\top A + \lambda Id\right) = A^\top y$$

Or pour $\lambda > 0$, $\left(A^\top A + \lambda Id\right)$ est toujours inversible. En effet:

si $\left(A^\top A + \lambda Id\right) x = 0$, alors $x^\top \left(A^\top A + \lambda Id\right) x = 0$, soit $\|Ax\|^2 + \lambda \|x\|^2 = 0$ et donc $x = 0$.

Donc $\left(A^\top A + \lambda Id\right)$ est injective, et donc inversible comme elle est carrée.

L'estimateur ridge $\hat{\beta}^{ridge}$ de β est donc donné par:

$$\hat{\beta}^{ridge} = \left(A^T A + \lambda I_d\right)^{-1} A^T y \ .$$

3. a) A^TA est symétrique, semi-définie positive. D'après le théorème de décomposition spectrale, $\exists Q \in \mathcal{M}_p$, orthogonale: $Q^TQ = I_p$ et $\Delta \in \mathcal{M}_p$, diagonale:

$$\Delta = \begin{pmatrix} d_1 & & (0) \\ & \ddots & \\ (0) & & d_p \end{pmatrix}$$

avec $d_1 \geq d_2 \cdots \geq d_p \geq 0$, telles que $Q^TA^TAQ = \Delta$.

Si A est de rang r: $d_1 \geqslant d_2 >, \ldots \geqslant d_r > 0$ et $d_{r+1} = \cdots = d_p = 0$.

Donc:

$$\Delta = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)$$

avec D matrice diagonale de rang r.

On pose: $Q = (Q_1|Q_2)$, avec $Q_1 \in \mathcal{M}_{pr}$, $Q_2 \in \mathcal{M}_{p(p-r)}$.

On a alors:

$$\begin{pmatrix} Q_1^\top \\ Q_2^\top \end{pmatrix} A^\top A \, (Q_1, Q_2) = \begin{pmatrix} Q_1^\top A^\top AQ_1 & Q_1^\top A^\top AQ_2 \\ Q_2^\top A^\top AQ_1 & Q_2^\top A^\top AQ_2 \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$$

Donc $Q_1^\top A^\top AQ_1 = D$ et $Q_2^\top A^\top AQ_2 = 0$.

Donc pour tout z, $z^TQ_2^\top A^\top AQ_2z = 0$, soit $\|AQ_2z\|^2 = 0$ et donc $AQ_2 = 0$, ce qui termine la démonstration.

b) En posant:

$$W_1 = D^{-1/2}Q_1^TA^T, W_1 \in \mathcal{M}_{rN},$$

on a:

$$W_1AQ_1 = D^{1/2}$$

et

$$W_1W_1^T = D^{-1/2}Q_1^TA^TAQ_1D^{-1/2} = Id$$

comme $Q_1^TA^TAQ_1 = D$.

Par complétion de base orthonormée, on peut choisir $W_2 \in \mathcal{M}_{(N-r)N}$ telle que

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \in \mathcal{M}_N$$

et soit orthogonale.

On a alors:

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} A(Q_1, Q_2) = \begin{pmatrix} W_1AQ_1 & W_1Aq_2 \\ W_2AQ_1 & W_2AQ_2 \end{pmatrix} = \begin{pmatrix} D^{1/2} & 0 \\ 0 & 0 \end{pmatrix}$$

en utilisant le fait que $AQ_2 = 0$ et $W_2AQ_1 = W_2W_1^TD^{1/2} = 0$ (comme $W_1 = D^{-1/2}Q_1^TA^T$, $AQ_1 = W_1^TD^{1/2}$ et $W_2W_1^T = 0$ par orthogonalité).

En posant:

$U = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}^T$, $V = Q$, et $\Sigma \in \mathcal{M}_{Np}$: $\Sigma = \begin{pmatrix} D^{1/2} & 0 \\ 0 & 0 \end{pmatrix}$.

nous obtenons bien: $A = U\,\Sigma\,V^T$.

4. La fonction de prévision est donnée par :

$$\hat{y} := A\hat{\beta}^{ridge} = A\left(A^\top A + \lambda I_d\right)^{-1} A^\top y.$$

Or:

$$A^\top A = V\Sigma^T U^\top U\Sigma V^T = V\Sigma^T\Sigma V^T.$$

Notons que Σ est diagonale mais rectangle. En revanche, $\Sigma^T\Sigma$ est diagonale carrée dans $\mathcal{M}_p$:

$$\Sigma^T\Sigma = \begin{pmatrix} d_1^2 & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & d_r^2 & \ddots & & \vdots \\ \vdots & & \ddots & 0 & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & 0 \end{pmatrix}$$

et

$$\begin{aligned} \left(A^TA + \lambda I_d\right)^{-1} &= \left(V\Sigma^T\Sigma V^T + \lambda VV^T\right)^{-1} \\ &= \left(V(\Sigma^T\Sigma + \lambda I_d)V^T\right)^{-1} \\ &= V\left(\Sigma^T\Sigma + \lambda I_d\right)^{-1}V^T \\ &= V \begin{pmatrix} \frac{1}{d_1^2+\lambda} & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & \frac{1}{d_1^2+\lambda} & \ddots & & \vdots \\ \vdots & & \ddots & \frac{1}{\lambda} & & \\ & & & & \ddots & 0 \\ 0 & & \cdots & & 0 & \frac{1}{\lambda} \end{pmatrix} V^T \end{aligned}$$

D'où:

$$A\hat{\beta}^{ridge} = U\Sigma V^T V \begin{pmatrix} \frac{1}{d_1^2+\lambda} & 0 & \cdots & & 0 \\ 0 & \ddots & & & \\ & & \frac{1}{d_1^2+\lambda} & \ddots & \vdots \\ \vdots & & \ddots & \frac{1}{\lambda} & \\ & & & \ddots & 0 \\ 0 & & \cdots & 0 & \frac{1}{\lambda} \end{pmatrix} V^T V \Sigma^T U^T y$$

$$= U\Sigma \begin{pmatrix} \frac{1}{d_1^2+\lambda} & 0 & \cdots & & 0 \\ 0 & \ddots & & & \\ & & \frac{1}{d_1^2+\lambda} & \ddots & \vdots \\ \vdots & & \ddots & \frac{1}{\lambda} & \\ & & & \ddots & 0 \\ 0 & & \cdots & 0 & \frac{1}{\lambda} \end{pmatrix} \Sigma^T U^T y$$

$$= U \begin{pmatrix} \frac{d_1^2}{d_1^2+\lambda} & 0 & \cdots & & 0 \\ 0 & \ddots & & & \\ & & \frac{d_r^2}{d_1^2+\lambda} & \ddots & \vdots \\ \vdots & & \ddots & 0 & \\ & & & \ddots & 0 \\ 0 & & \cdots & 0 & 0 \end{pmatrix} U^T y$$

et finalement:

$$A\hat{\beta}^{ridge} = \sum_{j=1}^{r} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y \ .$$

Si le rang A est égal à p, on obtient:

$$A\hat{\beta}^{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y \ ,$$

et dans le cas de la régression linéaire classique (équivalent à $\lambda = 0$):

$$A\hat{\beta}^{ridge} = \sum_{j=1}^{p} u_j u_j^T y \ .$$

b) On pose $z_i = v_1^t x_i$, et nous avons $\bar{z} = 0$ (comme $\bar{x} = 0$).

La variance empirique s'écrit:

$$\begin{aligned} S_z^2 &= \frac{1}{N}\sum_{i=1}^{N} z_i^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} v_1^T x_i x_i^T v_1 \\ &= \frac{1}{N} v_1^T \left(\sum_{i=1}^{N} x_i x_i^T\right) v_1 \\ &= \frac{1}{N} v_1^T A^T A v_1 \\ &= \frac{1}{N} v_1^\top V \Sigma^\top \Sigma V^\top v_1 \end{aligned}$$

Or

$$V^T v_1 = \begin{pmatrix} 1 & (0) \\ (0) & (0) \end{pmatrix}$$

et donc finalement: $S_z^2 = \frac{d_1^2}{N}$.

En notant:

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Nous avons $AV = U\Sigma$, soit $AVe_1 = U\Sigma e_1$ et $Av_1 = d_1 u_1$.

Cela correspond aux projections des x_i sur le vecteur v_1, qui est appelé le premier vecteur principal, de variance $\frac{d_1^2}{N}$. $u_j^T y$ correspond à la coordonnée de la projection de y selon les u_j et $\sum_{j=1}^{p} u_j u_j^T y$ correspond à la projection de y sur les u_j. On parle de projection selon les composantes principales. Pour des détails sur cette interprétation, nous renvoyons au chapitre ultérieur du cours sur l'Analyse en Composantes Princpales (Section 5.1 du poly de cours).

Pour la régression ridge: les coordonnées de la projection sont multipliées par le coefficient $\frac{d_j^2}{\lambda + d_j^2}$ pour la j-ème composante. Ce coefficient est plus petit que 1, toutes les coordonnées sont donc réduites, mais on a:

$$\frac{\frac{d_j^2}{\lambda + d_j^2}}{\frac{d_k^2}{\lambda + d_k^2}} = \frac{\lambda d_j^2 + d_j^2 d_k^2}{\lambda d_k^2 + d_j^2 d_k^2} > 1 \text{ si } d_j^2 > d_k^2.$$

Les coordonnées sont donc d'autant plus réduites que la variance de la composante est faible: on privilégie les composantes de plus grande variance.

c) Cette partie est appelé, régression sur composantes principales (principal component regression).

Pour un x_0, on prédit y_0. Avec un seul paramètre, la meilleur régression est obtenue en considérant le projeté de x_0 selon le vecteur principal, soit:

$$y_1 = \theta_1 (v_1, x_0) .$$

Pour estimer θ_1, on minimise:

$$\|y - \theta_1 A v_1\|^2$$

soit :

$$\hat{\theta}_1 = \frac{(y, Av_1)}{(Av_1, Av_1)} = \frac{v_1^T A^T y}{d_1^2}.$$

A q paramètres:

$$y_0 = \sum_{i=1}^{9} \theta_i (v_i, x_0) .$$

Les v_i sont orthogonaux, donc:

$$\hat{\theta}_i = \frac{(y, Av_i)}{d_i^2} .$$

Exercice 6.2. Soient X variable aléatoire à valeur dans $\mathbb{R}^p$ et Y variable aléatoire à valeurs dans $\mathbb{R}$. On considère le modèle de régression linéaire:

$$Y = \beta_0 + \beta^t X + \xi, \text{ avec } \xi \sim \mathcal{N}(0, \sigma^2),$$

avec $\beta = (\beta_1, \ldots, \beta_p)^T$. On suppose σ^2 connu et on s'intéresse désormais à l'estimation Bayésienne des paramètres $\gamma = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

On suppose que la distribution de X ne dépend pas de θ.

Pour tout $0 \leq i \leq p$, on choisit comme prior pour β_i une loi de Laplace de paramètre $(0, \tau)$.

On rappelle que si $Z \sim$ Laplace(μ, τ) la densité est donnée par:

$$p(z|\mu, \tau) = \frac{1}{2\tau} \exp\left(-\frac{|z - \mu|}{\tau}\right)$$

1. Soit un échantillon de données $(x_i, y_i)_{1 \leq i \leq N}$. Donner à une constante multiplicative près la loi a posteriori de γ.

2. Si on prend comme estimateur pour β le maximum a posteriori (c'est à dire le maximum de la loi a posteriori), que retrouve-t-on ?

Solution:

1. On écrit tout d'abord la vraisemblance:

$$p\left(y|\beta_0,\cdots,\beta_p\right) = e^{-\frac{1}{2\sigma^2}\sum_{1=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2}$$

Pour le prior:

$$\begin{aligned} p\left(\beta_0,\cdots,\beta_p\right) &\propto e^{-\frac{|\beta_0|}{\tau}} e^{-\frac{|\beta_1|}{\tau}} \ldots e^{-\frac{|\beta_p|}{\tau}} \\ &\propto e^{-\frac{1}{\tau}\sum_{i=1}^{p}|\beta_i|} \end{aligned}$$

En utilisant la formule de Bayes, nous avons alors:

$$p\left(\beta_0,\cdots,\beta_p|y\right) = e^{\left(-\frac{1}{2\sigma^2}\sum_{1=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2 - \frac{1}{\tau}\sum_{i=0}^{p}|\beta_i|\right)} .$$

2. Le maximum a posteriori s'obtient en maximisant $p\left(\beta_0,\cdots,\beta_p|y\right)$, ce qui est équivalent à minimiser:

$$J\left(\beta_0,\cdots,\beta_p\right) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2 + \frac{1}{\tau}\sum_{i=0}^{p}|\beta_i|$$

ou

$$\tilde{J}\left(\beta_0,\cdots,\beta_p\right) = \sum_{i=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2 + \frac{2\sigma^2}{\tau}\sum_{i=0}^{p}|\beta_i| \ .$$

C'est donc l'estimateur du lasso pour $\lambda = \frac{2\sigma^2}{\tau}$!

Pour aller plus loin...

Exercice 6.3. Soient X variable aléatoire à valeur dans $\mathbb{R}^p$ et Y variable aléatoire à valeurs dans $\mathbb{R}$. On suppose un modèle de régression général de la forme:

$$Y = f(X) + \epsilon, \text{ avec } \epsilon \sim \mathcal{N}(0,\sigma^2) \text{ et } \epsilon \text{ indépendante de } f(X).$$

On appelle $\hat{f}$ l'estimateur du modèle: c'est une statistique sur les échantillons aléatoires d'observations $(X_i,Y_i)_{1\leq i\leq N}$.

Soit un point $x_0 \in \mathbb{R}^p$. On s'intéresse à l'espérance de l'erreur de prédiction au carré en x_0, noté $\mathcal{E}(x_0)$:

$$\mathcal{E}(x_0) = \mathbb{E}\left[\left(Y-\hat{f}(x_0)\right)^2 | X = x_0\right] \ .$$

1. Écrire une décomposition biais-variance pour $\mathcal{E}(x_0)$.

On suppose pour la suite que les $(x_i)_{1\leq i\leq N}$ sont fixés.

2. Exprimer $\mathcal{E}(x_0)$ en fonction des valeurs inconnus $f(x_i)$:

a) Dans le cas du modèle des k-plus proches voisins. Commenter sur l'effet de k.

b) Dans le cas de la régression linéaire. Pour celle-ci, estimer $\mathbb{E}(\mathcal{E}(x_0))$ sur l'échantillon. Commenter sur l'effet du nombre de paramètres.

Solution: 1.

$$\begin{aligned}\mathcal{E}(x_0) &= \mathbb{E}\left[\left(Y - \hat{f}\left(x_0\right)\right)^2 | X = x_0\right] \\ &= \mathbb{E}\left[\left(f\left(x_0\right) + \varepsilon - \hat{f}\left(x_0\right)\right)^2\right] \\ &= E\left[\varepsilon^2 + 2\varepsilon\left(f(x) - \hat{f}\left(x_0\right)\right) + \left(f\left(x_0\right) - \hat{f}\left(x_0\right)\right)^2\right]\end{aligned}$$

Or $\mathbb{E}\left(\varepsilon\left(f(x) - \hat{f}\left(x_0\right)\right)\right) = 0$ et $\mathbb{E}(\varepsilon^2) = \sigma^2$, donc:

$$\begin{aligned}\mathcal{E}(x_0) &= \sigma^2 + \mathbb{E}\left(\left(f\left(x_0\right) - \hat{f}\left(x_0\right)\right)^2\right) \\ &= \sigma^2 + \mathbb{E}\left[\left(f\left(x_0\right) - \mathbb{E}\left(\hat{f}\left(x_0\right)\right)\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left(\hat{f}\left(x_0\right)\right) - \hat{f}\left(x_0\right)\right)^2\right]\end{aligned}$$

Mais $\left(f\left(x_0\right) - \mathbb{E}\left(\hat{f}\left(x_0\right)\right)\right)^2$ est une constante, et:

$$\mathbb{E}\left[\left(\mathbb{E}\left(\hat{f}\left(x_0\right)\right) - \hat{f}\left(x_0\right)\right)^2\right] = \mathbb{V}\left(\hat{f}\left(x_0\right)\right)$$

donc finalement:

$$\mathcal{E}(x_0) = \sigma^2 + \left(f\left(x_0\right) - \mathbb{E}\left(\hat{f}\left(x_0\right)\right)\right)^2 + \mathbb{V}\left(\hat{f}\left(x_0\right)\right)$$

où dans le membre de droite, le premier terme est l'erreur irréductible, le deuxième terme le biais, et le dernier terme la variance.

2. Les $(x_i)_{1 \leq i \leq N}$ étant fixés, l'aléatoire vient seulement des y_i.

a) Dans le cas du modèle des k-plus proches voisins, désignons par $\mathcal{V}_k(x_0)$ l'ensemble des indices des k plus proches voisins pour le point x_0:

$$\hat{f}\left(x_0\right) = \frac{1}{k}\sum_{l \in \mathcal{V}_k(x_0)} y_l = \frac{1}{k}\sum_{l \in \mathcal{V}_k(x_0)} \left(f\left(x_l\right) + \varepsilon_l\right)$$

et donc:

$$\mathbb{E}\left(\hat{f}\left(x_0\right)\right) = \frac{1}{k}\sum_{l \in \mathcal{V}_k(x_0)} f\left(x_l\right)$$

et

$$\hat{f}\left(x_0\right) - \mathbb{E}\left(\hat{f}\left(x_0\right)\right) = \frac{1}{k}\sum_{l \in \mathcal{V}_k(x_0)} \varepsilon_l$$

Donc

$$\mathcal{E}(x_0) = \sigma^2 + \left(f(x_0) - \frac{1}{k} \sum_{l \in \mathcal{V}_k(x_0)} \varepsilon_l \right)^2 + \frac{\sigma^2}{k} \ .$$

Quand k augmente, le terme de variance diminue, mais le biais augmente car on choisit des points de plus en plus loin de x_0. A l'inverse quand k est petit, le biais est potentiellement faible (les points choisis étant proches de x_0) mais la variance est plus grande. Cela indique que la complexité du modèle augmente quand k diminue.

b) Dans le cas de la régression linéaire à p paramètres:

$$\hat{f}_p(x) = \hat{\beta}^\tau x = x^t \hat{\beta}$$

Donc:

$$\hat{f}_p\left(x_0\right) = x_0^T \left(A^T A\right)^{-1} A^T y$$

avec

$$y = \begin{pmatrix} f_p(x_1) + \epsilon_1 \\ \vdots \\ f_p(x_N) + \epsilon_N \end{pmatrix}.$$

$$\mathbb{E}\left(\hat{f}_p\left(x_0\right)\right) = x_0^T \left(A^\top A\right)^{-1} A^\top \begin{pmatrix} f_p\left(x_1\right) \\ \vdots \\ f_p\left(x_N\right) \end{pmatrix}$$

et

$$\begin{aligned} \mathbb{V}\left(\hat{f}_p\left(x_0\right)\right) &= x_0^\top \left(A^\top A\right)^{-1} A^\top \mathbb{V}(y) A \left(A^\top A\right)^{-1} x_0 \\ &= \sigma^2 \quad x_0^\top \left(A^\top A\right)^{-1} x_0 \end{aligned}$$

D'où:

$$\mathcal{E}\left(x_0\right) = \sigma^2 + \left(f\left(x_0\right) - x_0^T \left(A^\top A\right)^{-1} A^\top \begin{pmatrix} f_p\left(x_1\right) \\ \vdots \\ f_p\left(x_N\right) \end{pmatrix} \right) + \sigma^2 x_0^\top \left(A^T A\right)^{-1} x_0.$$

$$\begin{aligned} \mathbb{E}\left(\mathcal{E}\left(x_0\right)\right) &\simeq \frac{1}{N} \sum \mathcal{E}\left(x_i\right) \\ &\simeq \sigma^2 + \frac{1}{N} \sum_{i=1}^{N} \left(f\left(x_i\right) - x_i^\top \left(A^T A\right)^{-1} A^T \begin{pmatrix} f_p(x_1) \\ \vdots \\ f_p\left(x_N\right) \end{pmatrix} \right) + \frac{\sigma^2}{N} \sum_{i=1}^{N} x_i^\top \left(A^\top A\right)^{-1} x_i \end{aligned}$$

Mais

$$\sum_{i=1}^{N} x_i^\top \left(A^\top A\right)^{-1} x_i = Tr\left(A \left(A^T A\right)^{-1} A^T\right) = rang(A) = p$$

$A \left(A^T A\right)^{-1} A^T$ correspond à une matrice de projection orthogonale, et si $(A^T A)$ est inversible, on sait qu'elle est de rang p.

D'où:

$$\mathbb{E}\left(\mathcal{E}\left(x_0\right)\right) \simeq \sigma^2 + \frac{1}{N}\sum_{i=1}^{N}\left(f\left(x_i\right) - x_i^{\top}\left(A^T A\right)^{-1} A^T \begin{pmatrix} f_p(x_1) \\ \vdots \\ f_p\left(x_N\right) \end{pmatrix}\right) + \frac{\sigma^2}{N}p$$

Ici, la variance augmente avec le nombre de paramètres p, ce qui est bien caractéristique d'un modèle dont la complexité augmente, et correspond bien à ce que l'on sait: lorsqu'on augmente le nombre de paramètres d'un modèle, on réduit potentiellement le biais, mais on augmente la variance. On parle d'over-fitting ou surapprentissage.