Correction feuille PC8 - modèle linéaire Gaussien - MAP433

1. Exercice 2 : Modèle de régression multiple

1. Par définition, l'estimateur des moindres carrés est donné par :

$$(\hat{\theta}_0, \hat{\theta})^\top \in \underset{(\theta_0,\theta)\in\mathbb{R}\times\mathbb{R}^k}{\operatorname{argmin}} \|y - \theta_0 e - X\theta\|_2 .$$

Alors $\hat{y} = \hat{\theta}_0 e + X\hat{\theta}$ est la projection orthogonale de y sur $\text{vect}(e, X^{(1)}, \cdots, X^{(k)})$ où $X^{(1)}, \cdots, X^{(k)}$ sont les vecteurs colonnes de X. En particulier, pour tout $\theta'_0 \in \mathbb{R}, \theta' \in \mathbb{R}^k$, on a

$$\left\langle y - \hat{y}, \theta'_0 e + X\theta' \right\rangle = 0.$$

En particulier, pour $\theta'_0 = 1, \theta' = 0$, on a $\left\langle y - \hat{y}, e \right\rangle = 0$ et comme $\bar{y} = n^{-1}\left\langle y, e \right\rangle$ (de même $\bar{\hat{y}} = n^{-1}\left\langle \hat{y}, e \right\rangle$), on a bien $\bar{y} = \bar{\hat{y}}$. De plus,

$$\bar{\hat{y}} = n^{-1}\left\langle \hat{y}, e \right\rangle = n^{-1}\left\langle \hat{\theta}_0 e + X\hat{\theta}, e \right\rangle = \hat{\theta}_0 + \bar{X}\hat{\theta}$$

où $\bar{X} = \left(\bar{X}^{(1)}, \cdots, \bar{X}^{(k)}\right)$.

2. $\bar{y}e$ est un élément de $\text{vect}(e, X^{(1)}, \cdots, X^{(k)})$. Comme $\hat{y}$ est le projeté orthogonal de y sur cet espace, on voit que $y - \hat{y}$ est orthogonal à $\bar{y}e - \hat{y}$. par Pythagore, on a

$$\|y - \bar{y}e\|_2^2 = \|y - \hat{y}\|_2^2 + \|\hat{y} - \bar{y}e\|_2^2 .$$

On a donc

$$R^2 = \frac{\|\hat{y} - \bar{y}e\|_2^2}{\|y - \bar{y}e\|_2^2} \leqslant 1.$$

   a) $R^2 = 1$ signifie que y est dans $\text{vect}(e, X^{(1)}, \cdots, X^{(k)})$ (modèle sans bruit). Donc les variables $e, X^{(1)}, \cdots, X^{(k)}$ "expliquent" très bien la sortie y.

   b) $R^2 = 0$ signifie que $\hat{y} = \bar{y}e$. Donc la variable qui explique le mieux y dans $e, X^{(1)}, \cdots, X^{(k)}$ est simplement e. Alors $X^{(1)}, \cdots, X^{(k)}$ sont des mauvaises variables pour expliquer ou prédire y.

3. Soit Proj l'opérateur de projection sur $\text{vect}(e, X^{(1)}, \cdots, X^{(k)})$. On a $Z\left(\hat{\theta}_0, \hat{\theta}\right)^\top = \text{Proj}(y)$ (pour alléger les notations, parfois, on ne distinguera pas θ de $\theta^\top$ dans la suite). On a pour tout $\theta'_0 \in \mathbb{R}, \theta' \in \mathbb{R}^k$, $\left\langle y - Z\left(\hat{\theta}_0, \hat{\theta}\right)^\top, Z\left(\theta'_0, \theta'\right)^\top \right\rangle = 0$. Par ailleurs,

$$\left\langle y - Z\left(\hat{\theta}_0, \hat{\theta}\right)^\top, Z\left(\theta'_0, \theta'\right)^\top \right\rangle = \left\langle Z^\top y - Z^\top Z\left(\hat{\theta}_0, \hat{\theta}\right)^\top, \left(\theta'_0, \theta'\right)^\top \right\rangle.$$

Donc $Z^\top y = Z^\top Z\big(\hat\theta_0, \hat\theta\big)^\top$. Comme la matrice carrée $Z^\top Z$ de taille $k+1$ est de rang $k+1$, elle est de rang plein donc inversible. Alors $\big(Z^\top Z\big)^{-1} Z^\top y = \big(\hat\theta_0, \hat\theta\big)^\top$.

On peux aussi voir que

$$(\hat\theta_0, \hat\theta)^\top \in \underset{\theta_0' \in \mathbb{R}, \theta' \mathbb{R}^k}{\operatorname{argmin}} \|y - \theta_0' e - X\theta'\|_2 .$$

Alors, $(\hat\theta_0, \hat\theta)^\top$ minimise la fonction convexe $F(u) = \|y - Zu\|_2^2$ sur $\mathbb{R}^{k+1}$. Alors $(\hat\theta_0, \hat\theta)^\top$ est solution de $F'(u) = 0$ càd $Z^\top(y - Zu) = 0$. Donc $\big(Z^\top Z\big)^{-1} Z^\top y = \big(\hat\theta_0, \hat\theta\big)^\top$.

La matrice de covariance de $\hat\Theta := (\hat\theta_0, \hat\theta)^\top$ est donnée par

$$\Sigma = \mathbb{E}\big[\big(\hat\Theta - \mathbb{E}\hat\Theta\big)\big(\hat\Theta - \mathbb{E}\hat\Theta\big)^\top\big].$$

L'espérance de $\hat\Theta$ est donnée par

$$\mathbb{E}\hat\Theta = \mathbb{E}\big(Z^\top Z\big)^{-1} Z^\top y = \big(Z^\top Z\big)^{-1} Z^\top Z\big(\theta_0, \theta\big)^\top = \big(\theta_0, \theta\big)^\top.$$

On en déduit que (étant donné que $\mathbb{E}\zeta\zeta^\top = \sigma^2 I_n$)

$$\Sigma = \mathbb{E}\big(Z^\top Z\big)^{-1} Z^\top \zeta\zeta^\top Z\big(Z^\top Z\big)^{-1} = \sigma^2\big(Z^\top Z\big)^{-1}.$$

Pour tout $j = 0, \dots, k$,

$$\operatorname{var}(\hat\theta_j) = \operatorname{var}\big(\big\langle e_j, \big(\hat\theta_0, \hat\theta\big)^\top\big\rangle\big) = \sigma^2 e_j^\top \big(Z^\top Z\big)^{-1} e_j = \sigma^2\big[\big(Z^\top Z\big)^{-1}\big]_{jj}.$$

4. On va montrer que $\hat\sigma^2 = (n-k)^{-1}\left\|y - Z\hat\Theta\right\|_2^2$ est un estimateur sans biais de σ^2. On rappel que Proj est l'opérateur de projection sur $\operatorname{vect}(e, X^{(1)}, \cdots, X^{(k)})$ et que $Z\hat\Theta = \operatorname{Proj}(y)$. Par ailleurs, $Z(\theta_0, \theta) \in \operatorname{vect}(e, X^{(1)}, \cdots, X^{(k)})$. On a donc :

$$\begin{aligned}
\mathbb{E}\hat\sigma^2 &= \frac{1}{n-k}\mathbb{E}\,\|(I_n - \operatorname{Proj})(y)\|_2^2 = \frac{1}{n-k}\mathbb{E}\,\|(I_n - \operatorname{Proj})(y)\|_2^2 \\
&= \frac{1}{n-k}\,\|(I_n - \operatorname{Proj})(\xi)\|_2^2 = \frac{1}{n-k}\mathbb{E}\operatorname{Tr}(I_n - \operatorname{Proj})\xi\xi^\top(I_n - \operatorname{Proj})^\top \\
&= \frac{\sigma^2}{n-k}\operatorname{Tr}(I_n - \operatorname{Proj}) = \sigma^2.
\end{aligned}$$

On a utilisé le fait que si $x \in \mathbb{R}^{k+1}$ alors $\|x\|_2^2 = \operatorname{Tr}(xx^\top)$ (où $xx^\top$ est une matrice $(k+1)\times(k+1)$), que la transposée d'une projection et que son carré sont égales à la projection et, finalement, que la trace d'une projection est égale à son rang.

La matrice de covariance de $(\hat\theta_0, \hat\theta)$ vaut $\Sigma = \sigma^2\big(Z^\top Z\big)^{-1}$ (cf. Question 3). Donc un estimateur de Σ est donné par $\hat\sigma^2\big(Z^\top Z\big)^{-1}$. De même un estimateur de la variance de $\hat\theta_j$ est donné par $\hat\sigma^2\big[\big(Z^\top Z\big)^{-1}\big]_{jj}$.

5. On n'a pas forcèment $\tilde{\theta} = \hat{\theta}$. Si Z est de rang $k+1$ et X est de rang k alors $\tilde{\theta} = \left(X^\top X\right)^{-1} X^\top y$ et $\hat{\theta}$ est la projection de $\left(Z^\top Z\right)^{-1} Z^\top y$ sur $\{0\} \times \mathbb{R}^k$. (Si Z n'est pas de rang $k+1$ ou X n'est pas de rang k, alors l'inversion correspond à l'inversion généralisée et $\tilde{\theta}$ (resp. $\hat{\theta}$) est défini à un élément près de $\ker(X^\top X)$ (resp. $\ker(Z^\top Z)$)). On peut alors trouver un exemple tel que $\tilde{\theta} \neq \hat{\theta}$. Il suffit de prendre $X = e_1$ où e_1 est le premier élément de la base canonique de $\mathbb{R}^n$. On a $\tilde{\theta} = y_1$ et

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta} \end{pmatrix} = \frac{1}{n-1} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ y_1 \end{pmatrix} = \begin{pmatrix} (n-1)^{-1} \sum_{i=2}^n y_i \\ y_1 - (n-1)^{-1} \sum_{i=2}^n y_i \end{pmatrix}.$$

   Alors $\hat{\theta} = y_1 - (n-1)^{-1} \sum_{i=2}^n y_i \neq y_1 = \tilde{\theta}$ dès que $\sum_{i=2}^n y_i \neq 0$.
6. On voit que $\hat{y} = X\tilde{\theta}$ est le projeté de y sur $\text{vect}(X^{(1)}, \cdots, X^{(k)})$. Si e est orthogonal à $\text{vect}(X^{(1)}, \ldots, X^{(k)})$ alors $\langle \hat{y}, e \rangle = 0$ et si $\langle y, e \rangle \neq 0$ alors on n'a pas $\langle e, y - \hat{y} \rangle = 0$ donc $\bar{y} \neq \bar{\hat{y}}$. Dans ce modèle R^2 n'a pas de sens.

2. Exercice 4 : modèle ANOVA = ANalysis Of VAriance

Le modèle ANOVA est un modèle de regression linéaire multiple $y = \sum_{i=1}^k x_i mi + \zeta$ dans lequel les k variables $x_1, \ldots, x_k$ prennent leurs valeurs uniquement dans $\{0, 1\}$. Ce type de modèle permet d'étudier l'effet de variables qualitatives x_i (appartenance ou non à une certaine classe $i\{1, \ldots, k\}$) sur une variable quantitative y.

Par example, si on veut étudier l'influence du secteur géographque (I modalités) et de la branche d'activité (J modalités) sur le salaire des ouvriers à partir de l'observation de ces salaires on aura $k = I \times J$ classes. Pour chaque classe $i \in \{1, \ldots, k\}$, on observe l salaires : $Y_{ij} = m_i + \zeta_{ij}, j = 1, \ldots, l$.

1. On peut écrire ce modèle comme un modèle de régression linéaire multiple $y = Xm + \zeta$ en posant :

$$y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1l} \\ Y_{21} \vdots \\ Y_{2l} \\ \vdots \\ Y_{kl} \end{pmatrix} ; X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} ; m = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} \text{ et } \zeta = \begin{pmatrix} \zeta_{11} \\ \vdots \\ \zeta_{1l} \\ \zeta_{21} \vdots \\ \zeta_{2l} \\ \vdots \\ \zeta_{kl} \end{pmatrix}.$$

   On a donc $y \in \mathbb{R}^{kl}$, $X \in \mathbb{R}^{kl \times k}$, $m \in \mathbb{R}^k$ et $\zeta \in \mathbb{R}^{kl}$. De plus,

$$B = X^\top X = lI_k.$$

   (l est le nombre d'observations par classe et k est le nombre de classes).

2. On considère la matrice $(k-1)\times k$ de différentiation discrète :

$$G = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ & \cdots & & \ddots & \ddots & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

On a bien $Gm = 0$ si et seulement si $m_1 = m_2 = \cdots = m_k$ (la dérivée est nulle en m si et seulement si le signal m est constant). Donc $\mathbf{H_0} : m \in \Theta_0$ où $\Theta_0 = \ker G$.

3. On est dans le cadre d'application du test de Fisher. On note par $\hat{\theta}$ l'estimateur des moindres carrés :

$$\hat{\theta} = B^{-1}X^\top y = (1/l)X^\top y = \begin{pmatrix} \frac{1}{l}\sum_{j=1}^{l} Y_{1j} \\ \vdots \\ \frac{1}{l}\sum_{j=1}^{l} Y_{kj} \end{pmatrix}.$$

(Chaque moyenne m_i est estimée par une moyenne empirique pour toutes les classes $i = 1, \ldots, k$). D'après la proposition 7.4,

$$G\hat{\theta} \sim N_{k-1}\big(Gm, (\sigma^2/l)GG^\top\big).$$

Par ailleurs, on a

$$GG^\top = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ & \vdots & & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

On note $D = (\sigma^2/l)GG^\top$. Sous $\mathbf{H_0}$, $D^{-1/2}G\hat{\theta} \sim N_{k-1}(0, I_{k-1})$. En particulier, $\left\|D^{-1/2}G\hat{\theta}\right\|_2^2 \sim \chi^2(k-1)$. Par ailleurs, σ^2 étant inconnu, on va l'estimer par

$$\hat{\sigma}^2 = \frac{\left\|y - X\hat{\theta}\right\|_2^2}{lk - k}.$$

On utilisera alors $\hat{D} = (\hat{\sigma}^2/l)GG^\top$ à la place de D. Pour notre problème, la statistique de Fisher est donnée par :

$$F = \frac{\left\|\hat{D}^{-1/2}G\hat{\theta}\right\|_2^2}{k-1} = \frac{l}{(k-1)\hat{\sigma}^2}\hat{\theta}^\top G^\top\big(GG^\top\big)^{-1}G\hat{\theta} = \frac{\left\|D^{-1/2}G\hat{\theta}\right\|_2^2/(k-1)}{\left\|\frac{y-X\hat{\theta}}{\sigma}\right\|_2^2/(lk-k)}$$

Par ailleurs, $\hat{\theta}$ et $y - X\hat{\theta}$ sont deux variables Gaussiennes indépendantes (on peut le voir en calculant la matrice de covariance du vecteur Gaussien $(y - X\hat{\theta}, \hat{\theta})$) donc $D^{-1/2}G\hat{\theta}$ et $y - X\hat{\theta}$ sont aussi indépendantes et Gaussiennes.

Alors F est distribuée selon une Fisher de paramètre $(k-1, lk-k)$. On rappelle que si $U \sim \chi^2(d_1)$ et $V \sim \chi(d_2)$ sont deux variables indépendantes alors $(U/d_1)/(V/d_2)$ est distribuée selon une Fisher de paramètre (d_1, d_2).

On construit finalement un test de niveau $\alpha \in (0,1)$ par :

$$\hat{t}_\alpha = \begin{cases} \mathbf{H_0} & \text{si } F > t_\alpha \\ \mathbf{H_1} & \text{sinon} \end{cases}$$

où $t_\alpha = q_{1-\alpha}(Fisher(k-1, lk-k)$.

3. Exercice 5 : Théorème de Gauss-Markov

1. Par définition, $\hat{\theta}$ minimise $F(u) = \|y - Xu\|_2^2$ donc $\hat{\theta} = \left(X^\top X\right)^\top X^\top y$. On remarque que $\text{rang}(X) = k$ donc $n \geqslant k$ et X est injective (donc $X^\top X$ est inversible : en effet, $X^\top X$ est symmétrique donc diagonalisable et si λ est une valeur propre de vecteur propre u alors $\|Xu\|_2^2 = \lambda \|u\|_2^2$, donc $\lambda \neq 0$ donc $X^\top X$ est inversible).

   On a donc $\mathbb{E}\hat{\theta} = \left(X^\top X\right)^{-1} X^\top \mathbb{E}y = \left(X^\top X\right)^{-1} X^\top X\theta = \theta$. Donc $\hat{\theta}$ est bien un estimateur sans biais. La matrice de covariance de $\hat{\theta}$ est donnée par

$$\Sigma := \mathbb{E}\big(\hat{\theta} - \mathbb{E}\hat{\theta}\big)\big(\hat{\theta} - \mathbb{E}\hat{\theta}\big)^\top = \left(X^\top X\right)^{-1} X^\top \mathbb{E}\zeta\zeta^\top X \left(X^\top X\right)^\top = \sigma^2 \left(X^\top X\right)^{-1}$$

2. On a $\mathbb{E}LY = LX\theta$. Pour que $\tilde{\theta} = LY$ soit sans biais, il faut et il suffit que $LX\theta = \theta$. Ceci étant vrai pour tout θ, on doit avoir $LX = I_k$.
3. $\Sigma = \mathbb{E}\big((\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^\top\big) = L\text{var}(Y)L^\top = \sigma^2 LL^\top$. Comme $LX = I_k$, on a :

$$\Delta X = LX - \left(X^\top X\right)^{-1} X^\top X = I_k - I_k = 0$$

   et la covariance de $\tilde{\theta}$ est donnée par :

   $\text{var}(\tilde{\theta}) = \text{var}(\Delta Y + \hat{\theta}) = \text{var}(\Delta Y) + \text{var}(\hat{\theta}) + \text{cov}(\hat{\theta}, \Delta Y) + \text{cov}(\Delta Y, \hat{\theta})$

   $\sigma^2 \Delta\Delta^\top + \text{var}(\hat{\theta}) + \text{cov}(\hat{\theta}, \Delta Y) + \text{cov}(\Delta Y, \hat{\theta})$.

   Par ailleurs, comme $\Delta X = 0$, on a $\mathbb{E}\Delta Y = 0$ et

$$\text{cov}(\Delta Y, \hat{\theta}) = \mathbb{E}\big[\Delta Y \hat{\theta}^\top\big] = \Delta\mathbb{E}\big[(X\theta + \zeta)\zeta^\top X (X^\top X)^{-1}\big] = 0$$

   car $\mathbb{E}\zeta\zeta^\top = \sigma^2 I_n$. De même $\text{cov}(\hat{\theta}, \Delta Y) = 0$. On en déduit que

$$\text{var}(\tilde{\theta}) = \text{var}(\hat{\theta}) + \sigma^2 \Delta\Delta^\top \succeq \text{var}(\hat{\theta}).$$

4. On a

$$\left\|\tilde{\theta} - \theta\right\|_2^2 = \sum_{j=1}^k (\tilde{\theta}_j - \theta_j)^2 = \sum_{j=1}^k e_j^\top (\tilde{\theta} - \theta)(\tilde{\theta} - \theta)^\top e_j$$

   alors

$$\mathbb{E}\left\|\tilde{\theta} - \theta\right\|_2^2 = \sum_{j=1}^k e_j \text{var}(\tilde{\theta}) e_j$$

de même $\mathbb{E}\left\|\hat{\theta}-\theta\right\|_2^2 = \sum_{j=1}^k e_j \mathrm{var}(\hat{\theta}) e_j$. Mais d'après 3., on a $\mathrm{var}(\tilde{\theta}) \succeq \mathrm{var}(\hat{\theta})$. Notamment, pour tout j, $e_j^\top \mathrm{var}(\tilde{\theta}) e_j \succeq e_j^\top \mathrm{var}(\hat{\theta}) e_j$. On a donc

$$\mathbb{E}\left\|\tilde{\theta}-\theta\right\|_2^2 \geqslant \mathbb{E}\left\|\hat{\theta}-\theta\right\|_2^2.$$

4. Exercice 6 : Régression Ridge

On peut voir la régression Ridge, comme une relaxation de la méthode MC dans le cas où les variables explicatives sont colinéaires (càd quand il y a de la redondance d'information dans les variables explicatives). Pour définir l'EMC de manière unique, on a besoin que $X^\top X$ soit inversible. Dans ce cas $\theta^{MC} = \left(X^\top X\right)^{-1} X^\top Y$. Comme $\ker(X^\top X) = \ker X$, on a voit que $X^\top X$ est inversible si et seulement si les colonnes de X sont linéairement indépendantes. D'un point de vue statistiques, des colonnes de X linéairement dépendantes signifie qu'il y a de la redondance d'information parmi les variables explicatives. Par ailleurs, quand $X^\top X$ est inversible mais que son conditionement (ratio plus grande valeur singulière sur plus petite valeur singulière) est grand alors un calcul effectif de l'EMC est difficile (car on doit trouver une solution au système $(X^\top X)x = X^\top y$). On va donc considérer, un estimateur qui "régularise" l'EMC ou "conditionne" la matrice de Gram $X^\top X$. Pour cela, on va inverser $X^\top X + \lambda I_k$ et ainsi considérer l'*estimateur Ridge*

$$\hat{\theta}_\lambda = \left(X^\top X + \lambda I_k\right)^{-1} X^\top Y.$$

Cet estimateur n'est plus sans biais mais il peut améliorer le risque quadratique de l'EMC. On peut voir ça comme un compromis biais/variance : on perd un peu sur l'espérance (= biais) mais on gagne sur la variance dans l'égalité

$$\mathbb{E}\left(\hat{\theta}_\lambda\right)^2 = \left(\mathbb{E}\hat{\theta}_\lambda - \mathbb{E}\theta\right)^2 + \mathrm{var}(\hat{\theta}_\lambda).$$

On doit aussi faire en sorte de bien choisir $\lambda > 0$. Ceci introduit le problème de la sélection de paramétre en statistique (et notamment la méthode de validation croisée).

1. Quand $k > n$, la matrice $X : \mathbb{R}^k \mapsto \mathbb{R}^n$ a un noyau et comme $\ker(X^\top X) = \ker X$, la matrice $X^\top X$ n'est plus inversible. On sait que l'EMC est défini comme solution de l'équation $X^\top X\hat{\theta} = X^\top Y$ qui admet une infinité de solution (un espace affine dirigé par $\ker(X^\top X)$). L'EMC n'est donc pas uniquement défini. On peut alors choisir parmi cet ensemble infini de solutions, une ayant certaines propriétés supplémentaires. On va chercher celle ayant une petite norme 2.
2. On introduit la fonction

   $$F(\theta) = \|Y - X\theta\|_2^2 + \lambda \|\theta\|_2^2, \quad \forall \theta \in \mathbb{R}^k.$$

   Cette fonction est strictement convexe et tend vers l'infini quand $\|\theta\|_2$ tend vers l'infini donc elle admet un unique minimum $\hat{\theta}_\lambda$ qui est solution de

l'équation $\Delta F(\hat{\theta}_\lambda) = 0$ càd $-2X^\top(Y - X\hat{\theta}_\lambda) + 2\lambda\theta = 0$. On a donc

$$\hat{\theta}_\lambda = \big(X^\top X + \lambda I_k\big)^{-1} X^\top Y.$$

3. Le biais de l'estimateur Ridge est donné par :

$$\mathbb{E}\hat{\theta}_\lambda = \big(X^\top X + \lambda I_k\big)^{-1} X^\top \theta$$

qui est différent de θ en général. Alors l'ER est en général un estimateur biaisé. La matrice de covariance est donnée par :

$$\begin{aligned}\mathrm{var}(\hat{\theta}_\lambda) &= (X^\top X + \lambda_k)^{-1} X^\top \mathbb{E}\zeta\zeta^\top X (X^\top X + \lambda_k)^{-1} \\ &= \sigma^2 (X^\top X + \lambda_k)^{-1} X^\top X (X^\top X + \lambda_k)^{-1}.\end{aligned}$$

4. Pour $k = 1$, on écrit $Y = X\theta + \zeta$ où X est un vecteur de $\mathbb{R}^n$. Dans ce cas $X^\top X = \|X\|_2^2$ alors l'EMC et l'ER sont donnés par :

$$\hat{\theta} = \hat{\theta}^{MC} = \frac{\langle X, Y\rangle}{\|X\|_2^2} \text{ et } \hat{\theta}_\lambda = \hat{\theta}^{ER} = \frac{\langle X, Y\rangle}{\|X\|_2^2 + \lambda}.$$

Le risque quadratique de l'EMC est

$$\begin{aligned}\mathbb{E}(\hat{\theta} - \theta)^2 = \mathrm{var}(\hat{\theta}) = \mathbb{E}\hat{\theta}^2 - (\mathbb{E}\hat{\theta})^2 = \frac{\mathbb{E}\langle X, Y\rangle^2}{\|X\|_2^4} - \theta^2 \\ = \frac{\mathbb{E}\langle X, X\theta + \zeta\rangle}{\|X\|_2^2} - \theta^2 = \frac{\sigma^2}{\|X\|_2^2}.\end{aligned}$$

La décomposition biais-variance du risque quadratique de l'ER donne :

$$\mathbb{E}\big(\hat{\theta}_\lambda - \theta\big)^2 = \big(\mathbb{E}\hat{\theta}_\lambda - \mathbb{E}\theta\big)^2 + \mathrm{var}(\hat{\theta}_\lambda) = \Big(\frac{\|X\|_2^2\,\theta}{\|X\|_2^2 + \lambda} - \theta\Big)^2 + \frac{\sigma^2\,\|X\|_2^2}{\big(\|X\|_2^2 + \lambda\big)^2}.$$

En posant $\mu = \lambda / \|X\|_2^2$, on est amené à chercher $\mu > 0$ tel que

$$\Big(\frac{1}{1+\mu} - 1\Big)^2 \theta^2 + \frac{\big(\sigma^2 / \|X\|_2^2\big)}{\big(1+\mu\big)^2} < \big(\sigma^2 / \|X\|_2^2\big) \tag{1}$$

càd $\mu\big(\theta^2 - \big(\sigma^2 / \|X\|_2^2\big)\big) < 2\big(\sigma^2 / \|X\|_2^2\big)$. Si $\theta^2 \|X\|_2^2 > \sigma^2$ alors pour tout λ tel que

$$\lambda < \frac{2\sigma^2 \|X\|_2^2}{\theta^2 \|X\|_2^2 - \sigma^2},$$

le risque quadratique de l'ER est moindre que celui de l'EMC. Quand $\theta^2 \|X\|_2^2 < \sigma^2$ alors pour tout $\lambda > 0$, le risque quadratique de l'ER est moindre que celui de l'EMC.

Le ratio θ^2/σ^2 (et en général pour tout k, $\|\theta\|_2^2/\sigma^2$) est appelé le "signal sur bruit". Quand il est grand ($\theta^2/\sigma^2 > \|X\|_2^{-2}$), il faut choisir λ assez petit et quand il est petit, l'ER est toujours meilleur (en terme de risque quadratique) que l'EMC pour n'importe quel λ.