Approfondissement Modélisation Statistique - Projet

Aristote Koen
Thomas Chatrefou

Octobre 2019

1 Estimation par maximum de vraisemblance

1. $\mathbf{X}$ n'étant pas aléatoire, on a $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\theta, \sigma^2 I_n)$. La vraisemblance est donc donnée, pour $\theta \in \mathbb{R}^d$ par :

$$L_Y(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta)\right)$$

dont on déduit la log-vraisemblance

$$l(\theta) = -\frac{1}{2\sigma^2}\|Y - X\theta\|^2 + c$$

où $c \in \mathbb{R}$.

$$\nabla l(\theta) = \frac{1}{\sigma^2}X^T(Y - X\theta)$$

donc, comme $\det(X^T X) \neq 0$:

$$\nabla l(\hat{\theta}) = 0 \quad \Leftrightarrow \quad \hat{\theta} = (X^T X)^{-1} X^T Y$$

Enfin,

$$D^2 l(\theta) = -\frac{1}{\sigma^2} X^T X$$

est une matrice négative car $X^T X$ est positive au sens où $\forall v \in \mathbb{R}^d, v^T X^T X v \geq 0$.
Donc $\hat{\theta}_{\text{MLE}} = (X^T X)^{-1} X^T Y$

2. Estimer le MLE c'est maximiser $l(\theta)$, c'est-à-dire minimiser $\|Y - X\theta\|^2$. Ainsi estimer les MLE est équivalent à estimer le minimiseur de $\theta \mapsto \|Y - X\theta\|^2$, donc $\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{OLS}}$

3. En supposant X connu :

$$l(\theta) = -\frac{1}{2\sigma^2}\|Y - X\theta\|^2 + c = -\frac{1}{2\sigma^2}\left(\|Y\|^2 + \|X\|^2 - \hat{\theta}_{\text{MLE}}^T X^T X\theta - \theta^T X^T X \hat{\theta}_{\text{MLE}}\right)$$

donc par le théorème de factorisation, le MLE est exhaustif.
Calculons maintenant

$$\begin{aligned} \mathbb{E}_\theta\left[\hat{\theta}_{\text{MLE}}|X\right] &= \mathbb{E}_\theta\left[(X^T X)^{-1} X^T Y | X\right] \\ &= (X^T X)^{-1} X^T \mathbb{E}_\theta[Y] \end{aligned}$$

or $Y|X \sim \mathcal{N}(X\theta, \sigma^2 I_n)$, donc

$$\mathbb{E}_\theta\left[\hat{\theta}_{\text{MLE}}|X\right](X^T X)^{-1} X^T X\theta = \theta$$

Le MLE est donc sans biais.
En écrivant la vraisemblance sous la forme :

$$L_Y(\theta) = \exp\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\|Y\|^2 + \|X\theta\|^2 - 2\langle Y|X\theta\rangle\right)\right)$$

Le modèle appartient bien à la famille exponentielle. De plus,

$$\langle Y|X\theta\rangle = \left\langle XX^{-1}X^{T^{-1}}X^TY|X\theta\right\rangle = \langle X(X^TX)^{-1}X^TY|X\theta\rangle = \left\langle X\hat{\theta}_{\mathrm{MLE}}|X\theta\right\rangle$$

donc

$$L_Y(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}(\|Y\|^2 + \|X\theta\|^2)\right)\exp\left(\frac{1}{\sigma^2}\left\langle X\hat{\theta}_{\mathrm{MLE}}|\theta\right\rangle\right)$$

On en déduit que $X\hat{\theta}_{\mathrm{MLE}}$ est une statistique exhaustive complète et comme on peut vérifier les hypothèses de régularité, $X\hat{\theta}_{\mathrm{MLE}}$ est un estimateur efficace de l'espérance de la loi de Y qui est $X\theta$.
Calculons l'information de Fisher :

$$I_n = -\mathbb{E}_\theta\left[D^2 l(\theta)\right] = \frac{1}{\sigma^2}X^TX$$

or

$$\mathrm{Var}_\theta(\hat{\theta}_{\mathrm{MLE}}) = \left((X^TX)^{-1}X^T\right)\sigma^2 I_n\left((X^TX)^{-1}X^T\right)^T = \sigma^2(X^TX)^{-1} = I_n^{-1}$$

Donc le MLE est efficace.

4. D'après une proposition du cours, le risque quadratique est déterminé par :

$$\begin{aligned} R_X(\hat{\theta}_{\mathrm{MLE}},\theta) &= \mathrm{Tr}\left(\mathrm{Cov}(\hat{\theta}_{\mathrm{MLE}}|X)\right) \\ &= \mathrm{Tr}\left(\frac{\sigma^2}{n}\left(\frac{1}{n}X^TX\right)^{-1}\right) \\ R_X(\hat{\theta}_{\mathrm{MLE}},\theta) &= \frac{\sigma^2}{n}\mathrm{Tr}\,\hat{\Sigma}_n^{-1} \end{aligned}$$

5. On a :

$$\mathrm{Tr}\,\hat{\Sigma}_n^{-1} = \sum_{i=1}^{d}\frac{1}{\lambda_i(\hat{\Sigma}_n)} = d\int_{\mathbb{R}}\frac{1}{x}\hat{F}_n(dx)$$

donc

$$R_X(\hat{\theta}_{\mathrm{MLE}},\theta) = \frac{d\sigma^2}{n}\int_{\mathbb{R}}\frac{1}{x}\hat{F}_n(dx)$$

6. D'après la convergence faible de $\hat{F}_n$ vers F_γ, on a :

$$\begin{aligned} r(\gamma) &= \lim_{n\to\infty}\mathbb{E}_\theta\left[\frac{d\sigma^2}{n}\int_{\mathbb{R}}\frac{1}{x}\hat{F}_n(dx)\right] \\ &= \mathbb{E}_\theta\left[\lim_{n\to\infty}\frac{d\sigma^2}{n}\int_{\mathbb{R}}\frac{1}{x}\hat{F}_n(dx)\right] \\ &= \mathbb{E}_\theta\left[\frac{\gamma\sigma^2}{1-\gamma}\right] \\ r(\gamma) &= \frac{\gamma\sigma^2}{1-\gamma} \end{aligned}$$

7. Si $\gamma = 0$, alors $R_X(\hat{\theta}_{\text{MLE}}, \theta) \underset{n\to\infty}{\longrightarrow} 0$ et comme la convergence $\mathcal{L}^2$ implique la convergence en probabilité, on a $\hat{\theta}_{\text{MLE}}$ consistant.

8. On constate que plus n augmente (le nombre d'observations de chaque echantillon), plus le risque quadratique s'aligne avec $\dfrac{\sigma^2\gamma}{1-\gamma}$. De plus on observe que plus γ est petit - c'est à dire que le nombre de paramètres est petit comparé à n - et plus le risque est petit.

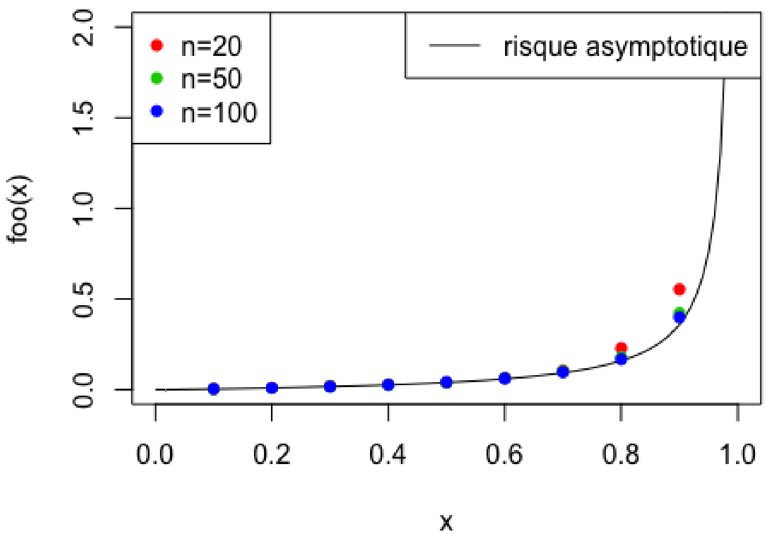

FIGURE 1 – Risque pour $\hat{\theta}_{\text{MLE}}$ en fonction de γ

9.

$$\begin{aligned}
\mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}}\|^2\right] &= \mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}} - \theta + \theta\|^2\right] \\
&= \mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}} - \theta\|^2\right] + 2\mathbb{E}_\theta\left[\left\langle \hat{\theta}_{\text{MLE}} - \theta | \theta \right\rangle\right] + \|\theta\|^2 \\
&= \mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}} - \theta\|^2\right] + 2\mathbb{E}_\theta\left[\left\langle \hat{\theta}_{\text{MLE}} | \theta \right\rangle\right] - 2\mathbb{E}_\theta\left[\langle \theta | \theta \rangle\right] + \|\theta\|^2 \\
&= \mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}} - \theta\|^2\right] + \|\theta\|^2
\end{aligned}$$

En passant à la limite :

$$\mathbb{E}_\theta\left[\|\hat{\theta}_{\text{MLE}}\|^2\right] = \frac{\sigma^2\gamma}{1-\gamma} + \|\theta\|^2$$

2 Régularisation de Tikhonov

11. Soit $f : \theta \mapsto -\dfrac{1}{n} l(\theta) + \dfrac{\lambda}{2\sigma^2}\|\theta\|^2$. Pour tout $\theta \in \mathbb{R}^d$:

$$\begin{aligned}\nabla f(\theta) &= -\frac{1}{n}\nabla l(\theta) + \frac{\lambda}{\sigma^2} \\ &= -\frac{1}{n}\left(\frac{-X^TX\theta + X^TY}{\sigma^2}\right) + \frac{\lambda}{\sigma^2}\theta \\ &= \frac{X^TX\theta}{n\sigma^2} + \frac{\lambda\theta}{\sigma^2} - \frac{X^TY}{n\sigma^2}\end{aligned}$$

Ainsi :

$$\nabla f(\theta) = 0 \iff \left(\frac{X^TX + n\lambda Id}{n\sigma^2}\right)\theta = \frac{X^TY}{n\sigma^2}$$

$\left(X^TX + n\lambda Id\right)$ est inversible car elle est définie positive par translation de son spectre. En effet, $\lambda > 0$ et X^TX est définie positive donc :

$$\theta = \left(X^TX + n\lambda Id\right)^{-1} X^TY$$

De plus, la Hessienne $D^2f(\theta) = \left(\frac{X^TX + n\lambda Id}{n\sigma^2}\right)$ est donc elle aussi définie positive pour les mêmes raisons et donc f admet un unique minimum global :

$$\hat{\theta}_\lambda = \left(X^TX + n\lambda Id\right)^{-1} X^TY$$

12. On a :

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_\lambda|X\right] &= \mathbb{E}\left[\left(X^TX + n\lambda Id\right)^{-1} X^TY|X\right] \\ &= \left(X^TX + +n\lambda Id\right)^{-1} X^T\mathbb{E}\left[Y|X\right] \\ &= \left(X^TX + +n\lambda Id\right)^{-1} X^TX\theta \\ &= \left(n\hat{\Sigma}_n + n\lambda Id\right)^{-1} n\hat{\Sigma}_n\theta \\ &= \left(\hat{\Sigma}_n + \lambda Id\right)^{-1} \hat{\Sigma}_n\theta\end{aligned}$$

Calculons maintenant la matrice de covariance :

$$\begin{aligned}\mathrm{Cov}_\theta(\hat{\theta}_\lambda|X) &= \left(X^TX + n\lambda I_d\right)^{-1} X^T\sigma^2 I_n\left(\left(X^TX + n\lambda I_d\right)^{-1} X^T\right)^T \\ &= \frac{\sigma^2}{n}(\hat{\Sigma}_n + \lambda I_d)^{-1}\hat{\Sigma}_n\left(\hat{\Sigma}_n + \lambda I_d\right)^{-1}\end{aligned}$$

car $\hat{\Sigma}_n + \lambda I_d$ est symétrique. Donc

$$\mathrm{Tr}\left(\mathrm{Cov}_\theta(\hat{\theta}_\lambda|X)\right) = \frac{d\sigma^2}{n}\int_{\mathbb{R}} \frac{x}{(x+\lambda)^2}\hat{F}_n(dx)$$

Comme $\lambda > 0$ et $x > 0$ presque sûrement, $\frac{1}{x} \geq \frac{x}{(x+\lambda)^2}$ et donc

$$R_X(\hat{\theta}_{\mathrm{MLE}}, \theta) \geq \mathrm{Tr}\left(\mathrm{Cov}_\theta(\hat{\theta}_\lambda|X)\right)$$

Posons $\bar{\theta} := \mathbb{E}_\theta\left[\hat{\theta}_\lambda|X\right]$. Alors

$$\bar{\theta} = \left(X^TX + n\lambda I_d\right)^{-1} X^TX\theta$$

soit

$$\bar{\theta} = \left(\hat{\Sigma}_n + \lambda I_d\right)^{-1}\hat{\Sigma}_n\theta$$

donc $\hat{\theta}_\lambda$ est biaisé.

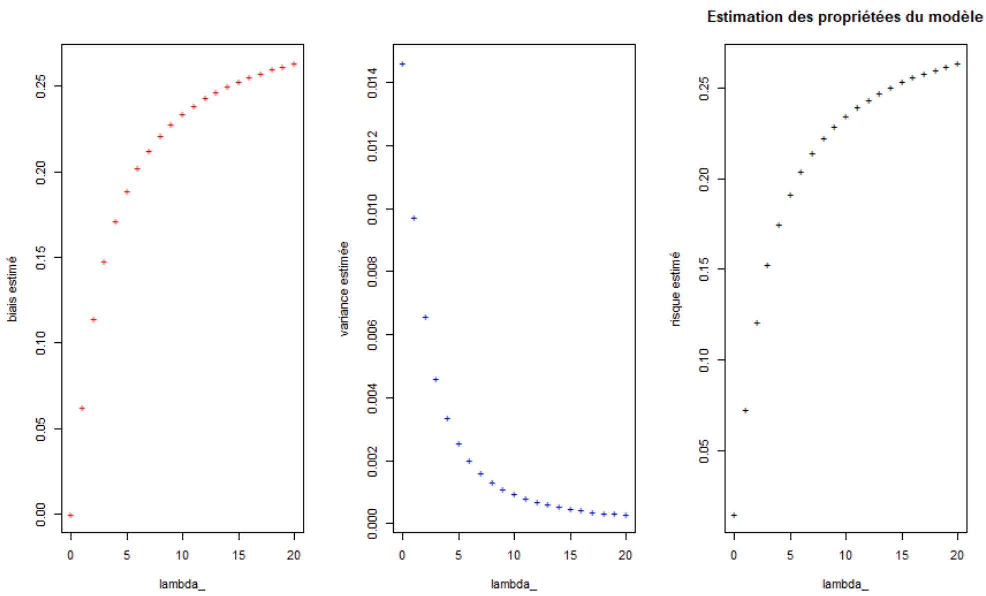

FIGURE 2 – Estimation des propriétés de l'estimateur régularisé

13. Sur la Figure 2 sont tracés respectivement de gauche à droite les estimateurs du biais, de la variance et du risque.

Plus on pénalise avec une valeur élevée de λ et plus on augmente le biais mais en réduisant considérablement la variance de l'estimateur.

14. Le risque pour $\hat{\theta}_\lambda$ est représenté sur la Figure 3.
On observe une augmentation linéaire du risque en fonction de gamma ce qui limite l'explosion de la variance.

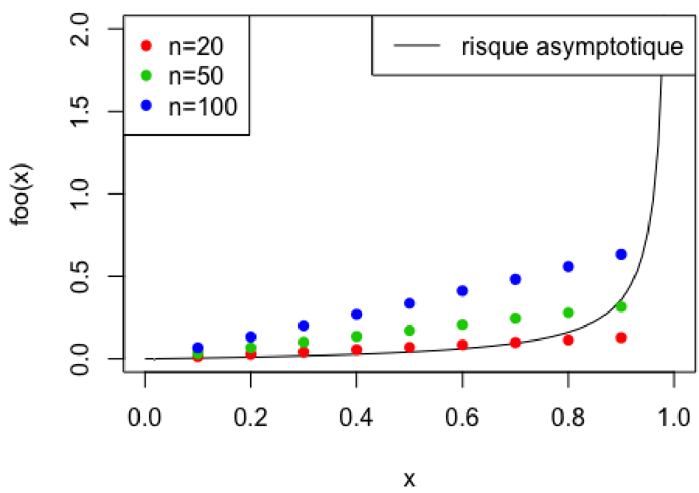

FIGURE 3 – Simulation avec $\hat{\theta}_\lambda$

15. Non, si gamma tend vers 1 il est plus intéressant de choisir un estimateur obtenu par régu-

larisation de Tikhonov afin d'avoir un estimateur avec une variance réduite. Mais si gamma est petit alors il vaut mieux garder un estimateur non biaisé.