ECN 7060, Cours 9

2020-11-04

Un paradoxe I

- Modèle : $X_i \sim \mathrm{iid}\, U(\theta, \theta+1)$
- Vraisemblance:

$$f(x|\theta) = \prod_{i=1}^{n} 1_{[\theta,\theta+1]}(x_i) = \prod_{i=1}^{n} 1_{[x_i-1,x_i]}(\theta) = 1_{[x_{(n)}-1,x_{(1)}]}(\theta),$$

  où $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ sont les statistiques d'ordre.
- Par le théorème de factorisation, $T(x) = (x_{(1)}, x_{(n)})$ est exhaustive pour θ.
- Vérification de minimalité par ratio :

$$\frac{f(x|\theta)}{f(y|\theta)} = \frac{1_{[x_{(n)}-1,x_{(1)}]}(\theta)}{1_{[y_{(n)}-1,y_{(1)}]}(\theta)}$$

  ne dépend pas de θ seulement si $x_{(1)} = y_{(1)}$ et $x_{(n)} = y_{(n)}$.

Un paradoxe II

- Une autre statistique exhaustive minimale est $T'(x) = (x_{(1)} + x_{(n)}, x_{(n)} - x_{(1)})$.
- Pourquoi? $T(x)$ et $T'(x)$ sont chacun une fonction de l'autre :

$$\begin{bmatrix} x_{(1)} + x_{(n)} \\ x_{(n)} - x_{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_{(1)} \\ x_{(n)} \end{bmatrix}.$$

- Le paradoxe : $x_{(1)} + x_{(n)}$ n'est pas exhaustive seule et $x_{(n)} - x_{(1)}$ est libre—sa distribution ne dépend pas de θ.
- Le concept de statistique complète est utile parce qu'une statistique complète et exhaustive est indépendante de n'importe quelle statistique libre.

Modèle binomial

- Modèle, $X_i \sim \mathrm{iid}\,\mathrm{Bn}(\theta)$, $\theta \in [0,1]$:

$$f(x_i|\theta) = \begin{cases} \theta & x_i = 1, \\ 1-\theta & x_i = 0. \end{cases}$$
$$= \theta^{x_i}(1-\theta)^{1-x_i}.$$

- Avec n observations, $x = (x_1, \ldots, x_n)$,

$$f(x|\theta) = \theta^{n_1}(1-\theta)^{n_0}$$

  où n_1 est le nombre de fois que $x_i = 1$, $n_0 = n - n_1$ est le nombre de fois que $x_i = 0$.

Une statistique suffisante

- Proposition : $T(x) = n_1$ est une statistique suffisante.
- Vérification par ratio :
  - $n_1 \sim \mathrm{Bi}(n, \theta)$,

$$q(T(x)|\theta) = \binom{n}{n_1}\theta^{n_1}(1-\theta)^{n-n_1}$$

  - $p(x|\theta) = \theta^{n_1}(1-\theta)^{n-n_1}$
  - $p(x|\theta)/q(T(x)|\theta) = 1/\binom{n}{n_1}$ ne dépend pas de θ.
- Vérification par factorisation :
  - $p(x|\theta) = g(T(x)|\theta)h(x)$ pour $g(T(x)|\theta) = \theta^{n_1}(1-\theta)^{n-n_1}$ et $h(x) = 1$.

Remarque sur le facteur $h(x)$

- Densité des données pour un modèle $\mathrm{Po}(\theta)$ (Poisson) :

$$\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

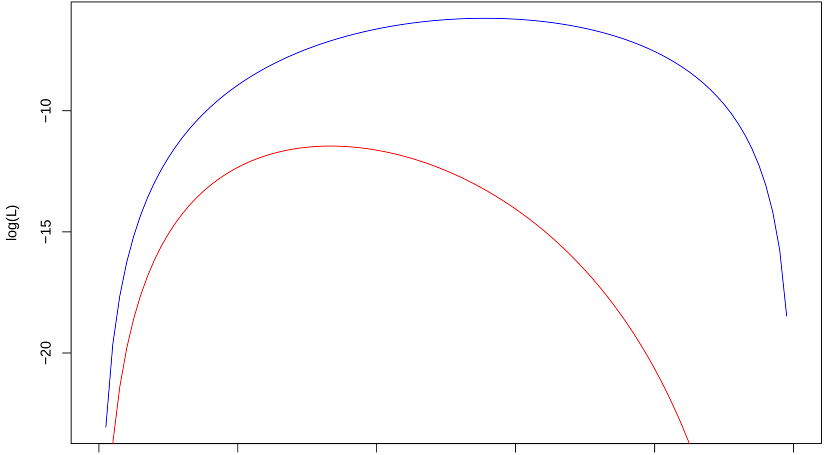
- $\sum_{i=1}^{n} x_i$ est une statistique suffisante minimale.
- Le facteur $h(x) = \left(\prod_{i=1}^{n} x_i!\right)^{-1}$ ne dépend pas de θ.

Minimalité de la statistique suffisante $T(x) = n_1$ dans le modèle binomial

- Proposition : $T(x) = n_1$ est une statistique suffisante minimale.
- Vérification par ratio de vraisemblances
  - Ratio de deux vraisemblances :
  $$\frac{f(x|\theta)}{f(y|\theta)} = \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}}.$$
  - Le ratio ne dépend pas de θ seulement si $\sum_i x_i = \sum_i y_i$ ou $T(x) = T(y)$.

Des log vraisemblances pour le modèle binomial

```
theta = seq(0, 1, by=0.01)
n0 = 4; n1 = 5; L = theta^n1 * (1-theta)^n0
plot(theta, log(L), col='blue', type='l')
n0 = 12; n1 = 6; L = theta^n1 * (1-theta)^n0
lines(theta, log(L), col='red')
```



Estimation de θ

- Par la méthode des moments :
  - $E[X_i] = \theta$ et $\frac{1}{n}\sum_{i=1}^n x_i = n_1/n$
  - La solution de $E[X_i] = \frac{1}{n}\sum_{i=1}^n x_i$ donne

$$\hat{\theta}_{MM} = n_1/n.$$

- Par la méthode de maximum de vraisemblance :
  - $\mathcal{L}(\theta; x) = \log L(\theta; x) = n_1 \log\theta + (n - n_1)\log(1-\theta)$
  - Deux dérivées par rapport à θ :

$$\frac{d\mathcal{L}(\theta; x)}{d\theta} = \frac{n_1}{\theta} - \frac{(n - n_1)}{1-\theta}$$

$$\frac{d^2\mathcal{L}(\theta; x)}{d\theta^2} = -\frac{n_1}{\theta^2} - \frac{(n - n_1)}{(1-\theta)^2}.$$

  - Deuxième toujours négative, première nulle pour $\theta = n_1/n$.
  - $\hat{\theta}_{ML} = n_1/n$.

La distribution de $\hat{\theta} = \hat{\theta}_{MM} = \hat{\theta}_{ML}$

- La distribution de $\hat{\theta} = T(X)/n$ vient de la distribution de X.
- Nous savons que $n\hat{\theta} = n_1 \sim \mathrm{Bi}(n, \theta)$.
- $E[\hat{\theta}] = n^{-1} \sum_{i=1}^{n} E[X_i] = \theta$.
- Puisque $E[X_i^2] = E[X_i] = \theta$, $\mathrm{Var}[X_i] = \theta(1-\theta)$ et
$$\mathrm{Var}[\hat{\theta}] = \theta(1-\theta)/n.$$
- $E[X_i^4] = \theta < \infty$ alors $\hat{\theta}$ converge à θ presque sûrement.
- $\sqrt{n}(\hat{\theta} - \theta)$ converge en loi à la loi $N(0, \theta(1-\theta))$.
- Notez la dépendance à θ partout.
- θ ici est inconnu mais fixe.

L'approche bayésienne

- Représenter l'incertitude concernant θ par une loi.
- Un modèle est une loi conjointe de θ et X.
- En pratique, le modèle est donné sous la forme $f(\theta)f(X|\theta)$.
- Une séparation entre l'apprentissage (automatique selon la règle de Bayes) et la prise des décisions.
- Au moment de prendre une décision, x est fixe (observé), θ est aléatoire, avec densité conditionnelle $f(\theta|x)$.

La loi beta

- La densité $\mathrm{Be}(\alpha, \beta)$ sur $[0, 1]$:

$$f(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- Moyenne et variance :

$$E[\theta] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

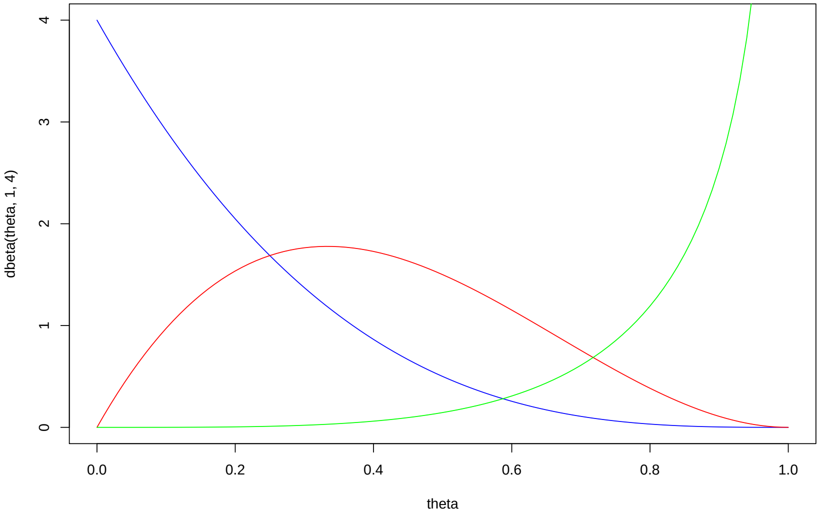
- Relation avec la loi gamma: si X et Y sont indépendantes, $X \sim \mathrm{Ga}(\alpha, \gamma)$ et $Y \sim \mathrm{Ga}(\beta, \gamma)$,

$$\frac{X}{X+Y} \sim \mathrm{Be}(\alpha, \beta).$$

- Remarquez la forme fonctionnelle en θ et sa ressemblance à la vraisemblance binomiale.

Des densités beta

```r
plot(theta, dbeta(theta, 1, 4), type='l', col='blue')
lines(theta, dbeta(theta, 2, 3), col='red')
lines(theta, dbeta(theta, 4.5, 0.5), col='green')
```



La loi conjointe de θ et x dans le modèle beta-binomial

- Si $\theta \sim \mathrm{Be}(\alpha, \beta)$, $x_i \sim \mathrm{iid}\,\mathrm{Bn}(\theta)$,

$$\begin{aligned} f(\theta, x) &= f(\theta) f(x|\theta) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{n_1}(1-\theta)^{n-n_1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}. \end{aligned}$$

- La densité postérieure de θ est proportionnelle à la densité conjointe :

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} \propto f(\theta, x) \propto \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}.$$

- $\theta|x \sim \mathrm{Be}(\alpha + n_1, \beta + n - n_1)$
- La densité postérieure normalisée est

$$f(\theta|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+n_1)\Gamma(\beta+n-n_1)} \theta^{n_1+\alpha-1}(1-\theta)^{n-n_1+\beta-1}.$$

Trois fonctions de perte pour l'analyse bayésienne

- Supposons que a est une action associée à l'estimation du paramètre θ.
- Trois fonctions de perte $L(\theta, a)$:
  1. Perte quadratique $L(\theta, a) = (a - \theta)^2$
  2. Perte valeur absolue $L(\theta, a) = |a - \theta|$
  3. Perte 0-1 $L_\epsilon(\theta, a) = 1 - 1_{[0,\epsilon]}(|a - \theta|)$

Trois estimateurs bayésiens de θ

1. La valeur $\hat{\theta}_1$ qui minimise $E[(\theta - \hat{\theta}_1)^2|x]$ est la moyenne postérieure.
2. La valeur $\hat{\theta}_2$ qui minimise $E[|\theta - \hat{\theta}_2||x]$ est la médiane postérieure.
3. La valeur $\hat{\theta}_3$ qui est la limite $(\epsilon \downarrow 0)$ de la valeur a qui minimise $E[1 - 1_{[0,\epsilon]}(|a - \theta|)|x]$ est le mode postérieur.

- Dans le modèle beta-binomial, si $\alpha + n_1,\ \beta + n - n_1 > 1$

$$\hat{\theta}_1 = E[\theta|x] = \frac{\alpha + n_1}{\alpha + \beta + n},$$

$$\hat{\theta}_2 = \frac{\alpha + n_1 - 1/3}{\alpha + \beta + n - 2/3},$$

$$\hat{\theta}_3 = \frac{\alpha + n_1 - 1}{\alpha + \beta + n - 2}.$$

Biais et variance dans le modèle binomial

- Calculs préliminaires ($X_i \sim \mathrm{iid}\,\mathrm{Bn}(\theta)$), $i = 1, \ldots, n$.
  - $E[X_i] = E[X_i^2] = \theta$, $\mathrm{Var}[X_i] = \theta - \theta^2 = \theta(1-\theta)$.
  - $n_1 = \sum_{i=1}^n X_i$, $E[n_1] = n\theta$, $\mathrm{Var}[n_1] = n\theta(1-\theta)$
- Propriétés de l'estimateur $\hat{\theta} = n_1/n$:
  - $E[\hat{\theta}] = \theta$, $\mathrm{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n}$, $\mathrm{Var}[\sqrt{n}(\hat{\theta} - \theta)] = \theta(1-\theta)$.
  - $\mathrm{biais}[\hat{\theta}] = E[\hat{\theta}] - \theta = 0$, $\mathrm{EQM}[\hat{\theta}] = \mathrm{Var}[\hat{\theta}] = \frac{\theta(1-\theta)}{n}$.
- Propriétés de l'estimateur $\hat{\theta}_1 = \frac{\alpha+n_1}{\alpha+\beta+n}$:
  - $E[\hat{\theta}_1] = \frac{\alpha+n\theta}{\alpha+\beta+n}$, $\mathrm{Var}[\hat{\theta}_1] = \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} < \mathrm{Var}[\hat{\theta}]$.
  - $\mathrm{biais}[\hat{\theta}_1] = E[\hat{\theta}_1] - \theta = \frac{\alpha(1-\theta)-\beta\theta}{\alpha+\beta+n}$,

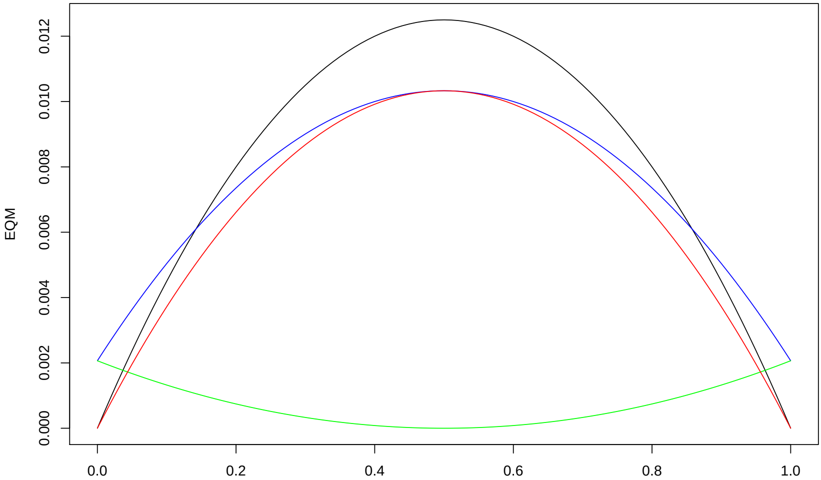
$$\mathrm{EQM}[\hat{\theta}_1] = \frac{n\theta(1-\theta) + [\alpha(1-\theta) - \beta\theta]^2}{(\alpha+\beta+n)^2}.$$

Illustration graphique I

```
theta = seq(0, 1, by=0.01)
n = 20; alpha = 1; beta = 1;
EQM = theta * (1-theta) / n
biais1 = (alpha*(1-theta) - beta*theta)/(alpha + beta + n)
var1 = n*theta*(1-theta)/(alpha + beta + n)^2
EQM1 = biais1^2 + var1
```

Illustration graphique II

```r
plot(theta, EQM, type='l')
lines(theta, EQM1, col='blue')
lines(theta, biais1^2, col='green')
lines(theta, var1, col='red')
```



Complétion du carré dans les modèles gaussiens I

- y_i scalaire, $i = 1, \ldots, n$

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\mu)^2 &= \sum_{i=1}^{n}((y_i-\bar{y})+(\bar{y}-\mu))^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2 + 2\sum_{i=1}^{n}(y_i-\bar{y})(\bar{y}-\mu) + n(\bar{y}-\mu)^2 \\
&= \sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2 \\
&= (n-1)S^2 + n(\bar{y}-\mu)^2
\end{aligned}$$

ou $\bar{y} \equiv n^{-1}\sum_{i=1}^{n} y_i$ et $S^2 \equiv (n-1)^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

Complétion du carré dans les modèles gaussiens II

- y est $n \times 1$, X est $n \times k$, β est $k \times 1$, $b = (X^\top X)^{-1} X^\top y$ existe.

$$
\begin{aligned}
u^\top u &\equiv (y - X\beta)^\top (y - X\beta) \\
&= (y - Xb + X(b - \beta))^\top (y - Xb + X(b - \beta)) \\
&= (y - Xb)^\top (y - Xb) + (b - \beta) X^\top X (b - \beta) \\
&\quad + 2(b - \beta)^\top X^\top (y - Xb) \\
&= (y - Xb)^\top (y - Xb) + (b - \beta) X^\top X (b - \beta)
\end{aligned}
$$

parce que $X^\top X b = X^\top y$

Complétion du carré dans les modèles gaussiens III

- y_i est $k \times 1$, $i = 1, \ldots, n$

$$\begin{aligned}
T(y) &= \sum_{i=1}^{n} (y_i - \mu)^\top H (y_i - \mu) \\
&= \sum_{i=1}^{n} ((y_i - \bar{y}) + (\bar{y} - \mu))^\top H ((y_i - \bar{y}) + (\bar{y} - \mu)) \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^\top H (y_i - \bar{y}) + n(\bar{y} - \mu)^\top H (\bar{y} - \mu) \\
&= \sum_{i=1}^{n} \mathrm{tr}[H(y_i - \bar{y})(y_i - \bar{y})^\top] + n(\bar{y} - \mu)^\top H (\bar{y} - \mu) \\
&= \mathrm{tr}\left[H \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^\top\right] + n(\bar{y} - \mu)^\top H (\bar{y} - \mu)
\end{aligned}$$