ECN 7060, cours 13

William McCausland

2019-12-04

Un modèle de mélange I

- Les X_i sont iid, chaque X_i un mélange de deux gaussiens

$$F(x_i|\theta) = p\Phi\left(\frac{x_i - \mu_1}{\sigma_1}\right) + (1-p)\Phi\left(\frac{x_i - \mu_2}{\sigma_2}\right)$$

- Le vecteur de paramètres est $\theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2)$.
- Irregularité I : paramètres non-identifiés
  - (label switching)

    $$f(X|\theta) = f(X|\theta')$$

    où

    $$\theta' = (1-p, \mu_2, \mu_1, \sigma_2, \sigma_1)$$

  - (non-identification sous l'hypothèse $p = 1$)

    $$f(X|(1, \mu_1, \mu_2, \sigma_1, \sigma_2))$$

    ne dépend pas de μ_2, σ_2.

La question d'identification (ponctuelle)

- Identification I : θ_0 est la vrai valeur

$$\theta \neq \theta_0 \Rightarrow f(\cdot|\theta) \neq f(\cdot|\theta_0).$$

- Sinon, θ et θ_0 sont observationnellement équivalent.
- Identification II :

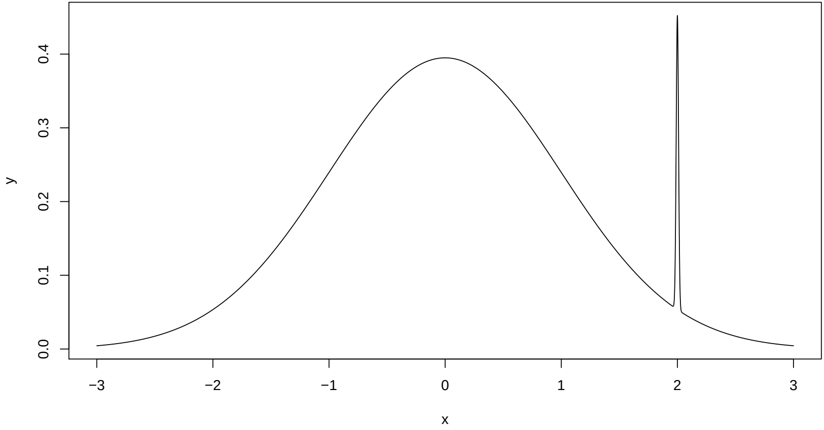
$$\theta_1 \neq \theta_2 \Rightarrow f(\cdot|\theta_1) \neq f(\cdot|\theta_2).$$

- C'est plus fort, mais on peut le vérifier.

Un mélange de deux gaussiens

- Ici, $p = 0.01$, $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = 1$, $\sigma_2 = 0.01$.

```
x = seq(-3, 3, by=0.001)
y = 0.01*dnorm(x, 2, 0.01) + 0.99*dnorm(x, 0, 1)
plot(x, y, type='l')
```



Un modèle de mélange II

- Rappel : $\theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2)$
- Irregularité II : la vraisemblance n'est pas bornée
  - $x = (x_1, \ldots, x_n)$ est arbitraire. Soit $\bar{x} = n^{-1} \sum x_i$.
  - Soit $\theta(\epsilon) = (n^{-1}, x_1, \bar{x}, \epsilon, s)$.
  - $f(x_1|\theta(\epsilon)) = n^{-1} \frac{1}{\sqrt{2\pi\epsilon}} + (1 - n^{-1}) \frac{1}{\sqrt{2\pi s}} e^{-(x_1 - \bar{x})/2s^2}$
  - $\lim_{\epsilon \downarrow 0} f(x_2, \ldots, x_n|\theta(\epsilon)) \neq 0$
  - $\lim_{\epsilon \downarrow 0} f(x|\theta(\epsilon)) = \infty$.
  - D'autres chemins où la vraisemblance croit sans borne
    - $p \in (0, 1)$ arbitraire
    - d'autres choix de μ_2, σ_2
    - échange d'index
  - D'autres modèles de mélange
- Implications pour la loi *a posteriori*
  - Pour certaines lois *a priori*, la densité *a posteriori* est bornée.
  - Bornée ou non, si la densité *a priori* est propre, la densité *a posteriori* l'est aussi.
  - Même sinon, la densité *a posteriori* est souvent propre (mais il faut vérifier)
  - Même si la densité postérieure n'est pas bornée, la région où la densité postérieure est plus grande que

Un modèle Bernoulli

- $X_1, \ldots, X_n \sim \mathrm{iid}\,\mathrm{Bn}(p)$
- $R = \sum_{i=1}^n X_i$ est une statistique suffisante minimale pour p.
- Si $r = 0$,
  - $f(x|p) = (1-p)^n$,
  - $\hat{p}_{MV}(x) = 0$,
  - $\mathrm{Var}_p[\hat{p}_{MV}(X)] = 0$ quand $p = \hat{p}_{MV}(x)$.
- Irregularité :
  - $\hat{p}_{MV}(x)$ se trouve sur la frontière de $\Theta = [0, 1]$.
  - La dérivée (à droite) de $\log f(x|p)$ n'égale pas à zéro à l'estimation MV :

$$\left.\frac{\partial n \log(1-p)}{\partial p}\right|_{p=0} = -n/(1-p)|_{p=0} = -n.$$

- D'autres cas :
  - Modèles avec restrictions sur les paramètres

Un modèle uniform

- $X_1, \ldots, X_n \sim \mathrm{iid}\, U(0, \theta)$.
- $X_{(1)} = \max_i X_i$ est une statistique suffisante minimale pour θ.
- Ici,
  - $f(x|\theta) = \theta^{-n} 1_{[x_{(1)}, \infty)}(\theta)$.
  - $\hat{\theta}_{MV} = X_{(1)}$, la valeur minimale possible de θ
  - Pour $\theta > x_{(1)}$,
    - $\log f(x|\theta) = -n \log \theta$
    - $\frac{\partial \log f(X|\theta)}{\partial \theta} = -\frac{n}{\theta}$
    - $E_\theta[\frac{\partial \log f(X|\theta)}{\partial \theta}] = -\frac{n}{\theta} \neq 0$.
- Irregularités :
  - le support de X_i dépend de θ
  - on ne peut pas prendre la dérivé dans l'intégral

$$\frac{\partial}{\partial \theta} \int_0^\theta f(x_i|\theta)\, dx = 0 \quad \text{mais} \quad \int_0^\theta \frac{\partial}{\partial \theta} f(x_i|\theta)\, dx = \int_0^\theta \frac{-1}{\theta^2}\, dx = -\frac{1}{\theta}.$$

Information de Fisher, deux formes

- Ici, $X = (X_1, \ldots, X_n)$, $x = (x_1, \ldots, x_n)$.
- Deux dérivées de la log vraisemblance, si elles existent :

$$\frac{\partial \log f(x|\theta)}{\partial \theta^\top} = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta^\top}$$

$$\begin{aligned}
\frac{\partial^2 \log f(x|\theta)}{\partial \theta \partial \theta^\top} &= \frac{1}{f(x|\theta)} \frac{\partial^2 f(x|\theta)}{\partial \theta \partial \theta^\top} - \frac{1}{f(x|\theta)^2} \frac{\partial f(x|\theta)}{\partial \theta^\top} \frac{\partial f(x|\theta)}{\partial \theta} \\
&= \frac{1}{f(x|\theta)} \frac{\partial^2 f(x|\theta)}{\partial \theta \partial \theta^\top} - \frac{\partial \log f(x|\theta)}{\partial \theta^\top} \frac{\partial \log f(x|\theta)}{\partial \theta}
\end{aligned}$$

- Espérance des deux côtés, si on peut passer l'espérance après les dérivées :

$$E_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta^\top} \right] = -E_\theta \left[\frac{\partial \log f(X|\theta)}{\partial \theta^\top} \frac{\partial \log f(X|\theta)}{\partial \theta} \right] \equiv -I_n(\theta)$$

- Attention : existence des dérivées, changement de l'ordre.

Additivité de l'information de Fisher

On considère ici les modèles où les X_i sont iid et on peut échanger l'ordre de l'espérance et le gradient.

- Si les X_i sont indépendantes, les $\partial \log f(X_i|\theta)/\partial\theta$ le sont aussi.
- Si on peut changer l'ordre de l'espérance et le gradient, ils ont une expérance de 0.
- Alors

$$\begin{aligned}
I_n(\theta) &= E_\theta\left[\sum_{i=1}^n \frac{\partial \log f(X_i|\theta)}{\partial\theta^\top}\sum_{i=1}^n \frac{\partial \log f(X_i|\theta)}{\partial\theta}\right] \\
&= \sum_{i=1}^n E_\theta\left[\frac{\partial \log f(X_i|\theta)}{\partial\theta^\top}\frac{\partial \log f(X_i|\theta)}{\partial\theta}\right] \\
&\equiv nI(\theta).
\end{aligned}$$

Gradient de la log vraisemblance

- Soit $l(\theta; x) \equiv \sum_{i=1}^{n} \log f(x_i|\theta)$.
- Soit $\hat{\theta}$ l'estimateur MV, θ_0 la vraie valeur du paramètre.
- Expansion du gradient à $\hat{\theta}$

$$\frac{\partial l(\hat{\theta}; x)}{\partial \theta^\top} \approx \frac{\partial l(\theta_0; x)}{\partial \theta^\top} + \frac{\partial^2 l(\theta_0; x)}{\partial \theta \partial \theta^\top}(\hat{\theta} - \theta_0).$$

- Alors si le gradient à gauche est nulle

$$(\hat{\theta} - \theta_0) \approx -\left[\frac{\partial^2 l(\theta_0; x)}{\partial \theta \partial \theta^\top}\right]^{-1} \frac{\partial l(\theta_0; x)}{\partial \theta^\top}$$

- Notes : on a besoin de
  - l'existence d'un maximum intérieur de la vraisemblance
  - l'existence des dérivées
  - la négligeabilité du terme résiduel
  - la non-singularité de la matrice hessienne

Continuous mapping theorem

- Si $g(\cdot)$ est continu, X, $X_1, X_2, \ldots$ des vecteurs aléatoires,

$$X_n \to_p X \Rightarrow g(X_n) \to_p g(X),$$

$$X_n \to_{ps} X \Rightarrow g(X_n) \to_{ps} g(X),$$

$$X_n \to_d X \Rightarrow g(X_n) \to_d g(X).$$

- Notes
  - Le théorème de Slutsky est un cas spécial parce que $X_n \to_p c \Rightarrow X_n \to_d c$.
  - Slutsky : si $X_n \to_d X$ et $Y_n \to_p c$, $X_n + Y_n \to_d X + c$, $X_n Y_n \to_d cX$, $X_n / Y_n \to_d X/c$ si $c > 0$.
  - On peut relacher la continuité : g peut avoir un ensemble D de points de discontinuité avec $P(X \in D) = 0$.

Pour préparer une analyse asymptotique

- Pour préparer une analyse asymptotique, on peut écrire

$$(\hat{\theta} - \theta_0) \approx \left[-\frac{1}{n} \frac{\partial^2 l(\theta_0; x)}{\partial\theta\partial\theta^\top} \right]^{-1} \left[\frac{1}{n} \frac{\partial l(\theta_0; x)}{\partial\theta^\top} \right].$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \left[-\frac{1}{n} \frac{\partial^2 l(\theta_0; x)}{\partial\theta\partial\theta^\top} \right]^{-1} \left[\sqrt{n} \frac{1}{n} \frac{\partial l(\theta_0; x)}{\partial\theta^\top} \right].$$

Théorème limite central, loi de grand nombres pour le gradient

$$\sqrt{n}\frac{1}{n}\frac{\partial l(\theta_0; X)}{\partial \theta^\top} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(X_i|\theta_0)}{\partial \theta^\top}$$

- Pour la vraie valeur θ_0,

$$E_{\theta_0}\left[\frac{\partial \log f(X_i|\theta_0)}{\partial \theta^\top}\right] = 0, \qquad \mathrm{Var}_{\theta_0}\left[\frac{\partial \log f(X_i|\theta_0)}{\partial \theta^\top}\right] = I(\theta_0).$$

- Par une loi de grand nombres,

$$\frac{1}{n}\frac{\partial l(\theta_0; x)}{\partial \theta^\top} \to_p 0.$$

- Par un théorème central limite,

$$\sqrt{n}\frac{1}{n}\frac{\partial l(\theta_0; x)}{\partial \theta^\top} \to_d N(0, I(\theta_0)).$$

Notes sur le gradient

On a besoin de

- l'existence des dérivées, échange d'ordre (intégral, dérivée),
- une variance fini,
- X_i indépendants, identiquement distribués.

Loi de grand nombres pour la matrice hessienne et son inverse

$$\frac{1}{n}\frac{\partial^2 l(\theta_0; X)}{\partial\theta\partial\theta^\top} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i|\theta_0)}{\partial\theta\partial\theta^\top} \to_p I(\theta_0)$$

- Par le théorème « continuous mapping, »

$$\left[\frac{1}{n}\frac{\partial^2 l(\theta_0; x)}{\partial\theta\partial\theta^\top}\right]^{-1} \to_p I(\theta_0)^{-1}.$$

Combinaison des résultats

- Convergence de $(\hat{\theta} - \theta)$ en probabilité :

$$(\hat{\theta} - \theta_0) \approx \left[-\frac{1}{n} \frac{\partial^2 l(\theta_0; x)}{\partial \theta \partial \theta^\top} \right]^{-1} \left[\frac{1}{n} \frac{\partial l(\theta_0; x)}{\partial \theta^\top} \right] \to_p 0.$$

- Convergence de $\sqrt{n}(\hat{\theta} - \theta)$ en loi :

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \left[-\frac{1}{n} \frac{\partial^2 l(\theta_0; x)}{\partial \theta \partial \theta^\top} \right]^{-1} \left[\sqrt{n} \frac{1}{n} \frac{\partial l(\theta_0; x)}{\partial \theta^\top} \right] \to_d N(0, I(\theta_0)^{-1}).$$

- Remarquez que $I(\theta_0)^{-1}$ est la borne inférieure Cramer-Rao. Sous les conditions de régularité, $\hat{\theta}_{MV}$ est un estimateur asymptotiquement efficace de θ_0.

Distribution asymptotique de la statistique test LRT

- Développement quadratique de $l(\theta|x)$ autour de θ_0, évalué à $\hat{\theta}$:

$$l(\theta_0|x) = l(\hat{\theta}|x) + \frac{1}{2}(\hat{\theta} - \theta_0)^\top \frac{\partial^2 l(\tilde{\theta}; x)}{\partial\theta\partial\theta^\top}(\hat{\theta} - \theta_0) + \dots$$

- Sous l'hypothèse nulle $H_0 : \theta = \theta_0$,

$$\begin{aligned} -2\log\lambda(x) &= -2(l(\theta_0|x) - l(\hat{\theta}|x)) \\ &\to_d (\hat{\theta} - \theta_0)^\top \frac{\partial^2 l(\hat{\theta}; x)}{\partial\theta\partial\theta^\top}(\hat{\theta} - \theta_0) \to_d \chi^2_k. \end{aligned}$$