

Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Rappels #3.1 (statistique & maximum de vraisemblance)

Likelihood / Vraisemblance

Assume that $\{x_1, x_2, \dots, x_n\}$ are obtained from i.i.d. random variables X_1, X_2, \dots, X_n , with identical distribution F_θ , and density f_θ .

$$\mathcal{L}(\theta) = f_\theta(\mathbf{x}) = f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$$

$$\log \mathcal{L}(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$$

$\widehat{\theta}$ is a **maximum likelihood estimator** of parameter θ if

$$\widehat{\theta} \in \operatorname{argmax}\{\mathcal{L}(\theta)\} = \operatorname{argmax}\{\log \mathcal{L}(\theta)\}$$

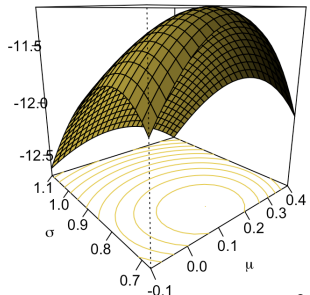
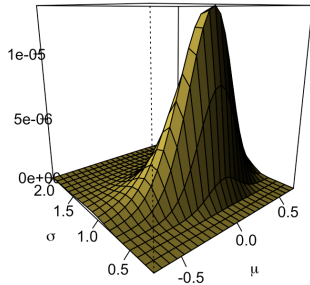
Likelihood / Vraisemblance

Given some sample $\{x_1, \dots, x_n\}$
from a $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$\mathcal{L}(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}\right)$$

$$\log \mathcal{L}(\mu, \sigma^2) = -\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)$$

Here $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.



Likelihood / Vraisemblance

Likelihood equations are, the first order condition

$$\left. \frac{\partial \log (\mathcal{L}(\theta; x_1, \dots, x_n))}{\partial \theta} \right|_{\theta=\widehat{\theta}} = 0$$

Second order condition is

$$\left. \frac{\partial^2 \log (\mathcal{L}(\theta; x_1, \dots, x_n))}{\partial \theta} \right|_{\theta=\widehat{\theta}} < 0$$

Example: if $X \sim \mathcal{P}(\lambda)$,

$$\log \mathcal{L}(\lambda; x_1, \dots, x_n) = \sum_{i=1}^n \log \left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right) = -n\lambda + n\bar{x} \log(\lambda) - \log \left(\prod_{i=1}^n x_i! \right)$$

$$\frac{\partial \log (\mathcal{L}(\lambda; x_1, \dots, x_n))}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}, \text{ so } \widehat{\lambda} = \bar{x}.$$

Likelihood / Vraisemblance

If $X \sim f_\theta$, $\mathbb{E}\left(\frac{d}{d\theta} \log f_\theta(X)\right) = 0$

Example: if $X \sim \mathcal{P}(\lambda)$,

$$\frac{d \log f_\lambda(X)}{d\lambda} = -1 + \frac{X}{\lambda}, \text{ so } \mathbb{E}\left(\frac{d}{d\lambda} \log f_\lambda(X)\right) = -1 + \frac{\mathbb{E}(X)}{\lambda} = 0$$

Fisher information associated with a density f_θ , with $\theta \in \mathbb{R}$ is

$$I(\theta) = \mathbb{E}\left(\frac{d}{d\theta} \log f_\theta(X)\right)^2 \text{ where } X \text{ has distribution } f_\theta,$$

$$I(\theta) = \text{Var}\left(\frac{d}{d\theta} \log f_\theta(X)\right) = -\mathbb{E}\left(\frac{d^2}{d\theta^2} \log f_\theta(X)\right).$$

For a sample of size n ,

$$I_n(\theta) = \mathbb{E}\left(\frac{\partial}{\partial \theta} \log \mathcal{L}(\theta, X_1, \dots, X_n)\right)^2 = nI(\theta)$$

Likelihood / Vraisemblance

Cramér-Rap bound: If $\widehat{\theta}$ is an unbiased estimator of θ , then

$$\text{Var}(\widehat{\theta}) \geq \frac{1}{nI(\theta)}$$

If that bound is attained, the estimator is said to be **efficient**.

An unbiased estimator $\widehat{\theta}$ is said to be **optimal** if it has the lowest variance among all unbiased estimators, see **bias**, **minimum variance unbiased estimator**

Example: if X has a Poisson distribution $\mathcal{P}(\theta)$,

$$\log f_{\theta}(x) = -\theta + x \log \theta - \log(x!) \text{ and } \frac{d^2}{d\theta^2} \log f_{\theta}(x) = -\frac{x}{\theta^2}$$

$$I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\theta^2} \log f_{\theta}(X)\right) = -\mathbb{E}\left(-\frac{X}{\theta^2}\right) = \frac{1}{\theta}$$

Example: if X has a binomial distribution $\mathcal{B}(n, \theta)$, $I(\theta) = \frac{n}{\theta(1-\theta)}$

Likelihood / Vraisemblance

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k, \left. \frac{\partial \log(\mathcal{L}(\boldsymbol{\theta}; x_1, \dots, x_n))}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$$

Second order condition is

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k, \left. \frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\theta}; x_1, \dots, x_n))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ is definite negative}$$

If $\boldsymbol{\theta} \in \mathbb{R}^k$, then Fisher information is the $k \times k$ matrix $I = [I_{i,j}]$ with

$$I_{i,j} = \mathbb{E} \left(\frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(X) \right).$$

i.e.

$$I(\boldsymbol{\theta}) = \mathbb{E} \left[\left(\frac{d}{d\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X) \right) \left(\frac{d}{d\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X) \right)^{\top} \right]$$

$$I(\boldsymbol{\theta}) = -\mathbb{E} \left(\frac{d^2}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}} \log f_{\boldsymbol{\theta}}(X) \right)$$

Likelihood / Vraisemblance

For a Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$, $l(\theta) = \frac{1}{\sigma^2}$

For a Gaussian distribution $\mathcal{N}(\mu, \theta)$, $l(\theta) = \frac{1}{2\theta^2}$

For a Gaussian distribution $\mathcal{N}(\theta)$, $l(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$

Likelihood / Vraisemblance

Let $\{x_1, \dots, x_n\}$ be a sample with distribution f_θ , where $\theta \in \Theta$. The maximum likelihood estimator $\widehat{\theta}_n$ of θ is

$$\widehat{\theta}_n \in \operatorname{argmax}\{\mathcal{L}(\theta; x_1, \dots, x_n), \theta \in \Theta\}.$$

Under some technical assumptions $\widehat{\theta}_n$ converges almost surely towards θ , $\widehat{\theta}_n \xrightarrow{a.s.} \theta$, as $n \rightarrow \infty$.

Under some technical assumptions $\widehat{\theta}_n$ is asymptotically efficient,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta)).$$

See [maximum likelihood estimation](#)

Optimization

Consider some Poisson model,

```
1 > set.seed(1)
2 > (y=rpois(10,3))
3 [1] 2 2 3 5 2 5 6 4 3 1
4 > mean(y)
5 [1] 3.3
6 > NLogL = function(lambda) -sum(log(dpois(y,lambda)))
7 > optim(fn = NLogL,par = 1)
8 $par
9 [1] 3.3
10
11 $value
12 [1] 18.59581
```

Likelihood / Vraisemblance

Consider a sample $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. from F_θ . Let

$$S_{n,\theta}(\mathbf{x}) = \frac{\partial \log \mathcal{L}(\theta; \mathbf{x})}{\partial \theta} = \sum_{i=1}^n S_{1,\theta}(x_i)$$

denote the **score function**. Then $S_{n,\theta}(\mathbf{X})$ is a random vector. Then

$$\mathbb{E}[S_{n,\theta}(\mathbf{X})] = \mathbf{0}$$

while

$$\text{Var}[S_{n,\theta}(\mathbf{X})] = I(\theta).$$

$$\frac{S_{n,\theta}(\mathbf{X})}{n} \xrightarrow{a.s.} 0 \quad \text{and} \quad \frac{S_{n,\theta}(\mathbf{X})}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_n(\theta)).$$

Likelihood / Vraisemblance

If θ is univariate, use Taylor approximation of S_n in the neighbourhood of θ_0 (the true value)

$$S_n(x) = S_n(\theta_0) + (x - \theta_0)S'_n(y) \text{ for some } y \in [x, \theta_0]$$

Set $x = \widehat{\theta}_n$, then

$$S_n(\widehat{\theta}_n) = 0 = S_n(\theta_0) + (\widehat{\theta}_n - \theta_0)S'_n(y) \text{ for some } y \in [\theta_0, \widehat{\theta}_n]$$

Hence, $\widehat{\theta}_n = \theta_0 - \frac{S_n(\theta_0)}{S'_n(y)}$ for $y \in [\theta_0, \widehat{\theta}_n]$. Hence,

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{S'_n(\widehat{\theta}_n^{(i)})},$$

from some starting value $\widehat{\theta}_n^{(0)}$ (hopefully well chosen).

Likelihood / Vraisemblance

Newton-Raphson:

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{S'_n(\widehat{\theta}_n^{(i)})},$$

where

$$S'_n(\widehat{\theta}_n^{(i)}) \sim \frac{S_n(\widehat{\theta}_n^{(i)} + h) - S_n(\widehat{\theta}_n^{(i)} - h)}{2h}$$

from some starting value $\widehat{\theta}_n^{(0)}$ (hopefully well chosen), and some small $h > 0$.

Fisher-Scoring technique:

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} + \frac{S_n(\widehat{\theta}_n^{(i)})}{nl(\widehat{\theta}_n^{(i)})},$$

again from some starting value.

Newton-Raphson

Consider some Poisson model, $S_1(\theta) = -1 + \frac{x}{\theta}$

```
1 > Sn = function(lambda) sum(-1+y/lambda)
2 > h = 1e-7
3 > dSn = function(lambda) (Sn(lambda+h)-Sn(lambda-h))
   /(2*h)
4 > L = rep(NA,10)
5 > L[1] = 1
6 > for(i in 1:9){
7 +   L[i+1] = L[i] - Sn(L[i])/dSn(L[i])
8 + }
9 > L
10 [1] 1.000 1.697 2.521 3.116 3.290 3.300 3.300 3.300
```

Scoring-Fisher

Consider some Poisson model, with Fisher information $I(\theta) = \frac{1}{\theta}$

```
1 > I = function(lambda) 1/lambda
2 > L = rep(NA,10)
3 > L[1] = 1
4 > for(i in 1:9){
5 +   L[i+1] = L[i] - Sn(L[i])/(length(y)*I(L[i]))
6 + }
7 > L
8 [1] 1.0 3.3 3.3 3.3 3.3 3.3 3.3 3.3 3.3 3.3
```