Loi des EMV

Marie-Anne Poursat

M1 Bioinformatique Université Paris-Saclay

30 janvier 2024

Estimateur du maximum de vraisemblance

Modéle *non-linéaire* en les paramètres : *pas de calculs exacts*.

$\hookrightarrow$ l'étude asymptotique (n "grand") est la référence
$\hookrightarrow$ les calculs sont des *approximations asymptotiques* ($n \to \infty$)

$$\text{Modèle : } (Y_1, \ldots, Y_n) \text{ i.i.d. de densité } f_\theta, \theta = (\theta_1, \ldots, \theta_p)^T$$

L'EMV $\widehat{\theta}$ maximise $\log L(\theta) = \sum_{i=1}^n \log f_\theta(Y_i)$, solution des équations de vraisemblance :

$$G(\widehat{\theta})) = \frac{\partial \log L}{\partial \theta}(\widehat{\theta}) = 0$$

- Consistance (garantie d'un bon estimateur).
- Approximation de la loi de $\widehat{\theta}$?
  $\hookrightarrow$ indispensable pour calculer des intervalles de confiance ou des tests.
  Comment calculer la variance de $\widehat{\theta}$?

Information de Fisher

On appelle information de Fisher de l'échantillon $(Y_1, \ldots, Y_n)$ la quantité

$$I_n(\theta) = \mathrm{Var}\left[\frac{\partial \log L}{\partial \theta}(\theta)\right] = \mathrm{E}\left[\left(\frac{\partial \log L}{\partial \theta}(\theta)\right)^2\right]$$

- si θ est un réel, $I(\theta)$ est un nombre ;
- si θ est un vecteur, $I(\theta)$ est une matrice $p \times p$, définie positive (en $\widehat{\theta}$)).

On peut montrer :

$$I(\theta) = -\mathrm{E}\left[H(\theta)\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta_j^2} \log L(\theta)\right]_j$$

Interprétation géométrique : plus $I(\theta)$ est grande, meilleure est la localisation du maximum de la log-vraisemblance.

Calcul de I

Comment calculer I ? En général, on ne sait pas calculer l'espérance (dépend de θ inconnu, calcul analytique impossible), on l'estime par l'information de Fisher observée

$$\widehat{I} = -H(\widehat{\theta})$$

calculée au dernier pas de l'algorithme d'optimisation.

Exemple : échantillon de loi de Bernoulli.

Loi de l'EMV

Théorème ((généralisation du TLC))

Sous des hypothèses mathématiques de régularité du modèle ($\log L$ 2 fois dérivable, le support de la loi ne dépend pas de θ, $0 < I(\theta) < \infty$), dans le cas où θ est un réel,

1. *L'EMV $\widehat{\theta}$ est un estimateur consistant de θ,*
2. $\widehat{\text{s.e.}} = \text{s.e.}(\widehat{\theta}) \approx \sqrt{\dfrac{1}{\widehat{I}}}$
3. $\dfrac{\widehat{\theta} - \theta}{\widehat{\text{s.e.}}} \overset{\text{Loi}}{\rightsquigarrow} \mathcal{N}(0, 1)$

   Ce qui se traduit par une approximation de la loi de $\widehat{\theta}$:

$$\text{Loi}(\widehat{\theta}) \approx \mathcal{N}\left(\theta, \frac{1}{\widehat{I}}\right)$$

Les logiciels statistiques implémentent le calcul de l'écart-type (estimé) de l'EMV

IC de θ

On déduit de l'approximation de la loi de l'EMV l'intervalle de confiance de niveau (approché) $1-\alpha$ suivant :

$$\left(\widehat{\theta} - q_{1-\alpha/2}\,\widehat{\text{s.e.}},\quad \widehat{\theta} + q_{1-\alpha/2}\,\widehat{\text{s.e.}}\right)$$

avec $q_{1-\alpha/2}$ le quantile d'ordre $(1-\alpha/2)$ d'une loi $\mathcal{N}(0,1)$.

Modèle multi-paramètres

$\theta = (\theta_1, \ldots, \theta_p)^T$

$$I(\theta) = \begin{pmatrix} -E(H_{11}(\theta)) & -E(H_{12}(\theta)) & \ldots & -E(H_{1p}(\theta)) \\ -E(H_{12}(\theta)) & -E(H_{22}(\theta)) & \ldots & -E(H_{2p}(\theta)) \\ \vdots & \vdots & & \vdots \\ -E(H_{p1}(\theta)) & -E(H_{p2}(\theta)) & \ldots & -E(H_{pp}(\theta)) \end{pmatrix}$$

- $I(\theta)$ estimée par $\widehat{I}$.
- La variance de $\widehat{\theta}$ est estimée par $\widehat{I}^{-1}$, matrice p x p

Modèle multi-paramètres

Loi de l'EMV

Si $\theta \in \mathbb{R}^p$, on peut montrer que dans les modèles réguliers,

$$\widehat{\theta}^{ML} \sim \mathcal{N}\left(\theta, I^{-1}(\theta)\right)$$

Cette approximation est encore valide si $I(\theta)$ est estimée par $\widehat{I}$.

On peut donc calculer

$$\widehat{\mathrm{s.e.}}(\widehat{\theta}_j^{ML}) \approx \sqrt{\widehat{I}_{jj}}$$

et

$$IC(\theta_j) = \left[\widehat{\theta}_j^{ML} - q_{1-\alpha/2}\, \widehat{\mathrm{s.e.}}(\widehat{\theta}_j^{ML}); \quad \widehat{\theta}_j^{ML} + q_{1-\alpha/2}\, \widehat{\mathrm{s.e.}}(\widehat{\theta}_j^{ML})\right]$$

est un intervalle de confiance de niveau $1 - \alpha$.

Exemple : données

Deux séquences ADN alignées de longueur N
$\hookrightarrow$ *match* = positions qui présentent le même nucléotide A, C , G ou T (repérées par *)

```
*   *     *     * * *   * *                 *
G G A G A C T G T A G A C A G C T A A T ..... G A G
G A A C G C C C T A G C C A C G A G C C ..... G G C
1 2 3 4 5 6 7 8 9 .............................  N
```

Y = nombre de *matchs* consécutifs observés avant une position où les nucléotides diffèrent

- Observation : $n \leq N$ réalisations $Y_1, \ldots, Y_n$ de Y le long de l'alignement
- n est donc le nombre de positions de l'alignement qui ne sont pas des *match*
- Sur l'exemple : $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, $y_4 = 1$, $y_5 = 0$, $y_6 = 3$, etc...

Exemple : estimation

On suppose que $Y_1, \ldots, Y_n$ forme un échantillon indépendant et de même loi géométrique :

$$\mathrm{P}(Y = y) = (1-\theta)\theta^y, \quad y = 0, 1, 2, \ldots.$$

On a alors $\mathrm{E}(Y) = \dfrac{\theta}{1-\theta}$.

Paramètre : θ est la probabilité d'observer un *match*.

1. Déterminer la log-vraisemblance des observations.
2. Montrer que l'estimateur du maximum de vraisemblance de θ est $\widehat{\theta} = \frac{\overline{Y}}{1+\overline{Y}}$, où $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
3. Calculer l'information de Fisher des observations. En déduire une approximation de la loi de $\widehat{\theta}$.
4. Donner le code R qui permet de calculer un intervalle de confiance pour θ de niveau 95%.