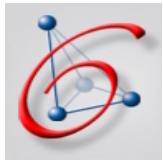
Master 1 Informatique
Éléments de statistique inférentielle

Faicel Chamroukhi
Maître de Conférences
UTLN, LSIS UMR CNRS 7296




email: chamroukhi@univ-tln.fr
web: chamroukhi.univ-tln.fr

2014/2015

Plan I

Méthode du maximum de vraisemblance (Maximum Likelihood)

- Définition de la fonction de vraisemblance
- Maximum de vraisemblance
- Propriétés
- Cas gaussien

Méthode du maximum de vraisemblance I

Introduction

Introduite par le statisticien Fischer en 1922

la méthode du maximum de vraisemblance est devenue la méthode générale la plus importante de l'estimation d'un point de vue théorique.

Son plus grand atout réside dans le fait que certaines propriétés très générale associées à cette procédure peuvent être dérivées et, dans le cas de grands échantillons, ce sont des propriétés optimales en fonction des critères d'absence de biais, de variance minimale, de consistance et d'efficacité.

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

Soit $f(x;\theta)$ la densité de probabilité d'une v.a X à densité où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $x = (x_1, \ldots, x_n)$ un échantillon d'observations des variables aléatoires $(X_1, \ldots, X_n)$. La *vraisemblance* du paramètre θ pour l'échantillon x est donnée par la densité jointe de x et se note ainsi :

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \quad (1)$$

Fonction de vraisemblance

Définition : **Fonction de vraisemblance**

Soit $f(x;\theta)$ la densité de probabilité d'une v.a X à densité où θ est le paramètre (vrai paramètre) à estimer (Nous prenons ici le cas simple d'un seul paramètre). Soit $x = (x_1, \ldots, x_n)$ un échantillon d'observations des variables aléatoires $(X_1, \ldots, X_n)$. La *vraisemblance* du paramètre θ pour l'échantillon x est donnée par la densité jointe de x et se note ainsi :

$$L(\theta; x) = L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta). \quad (1)$$

Vraisemblance pour un échantillon *i.i.d.*

Pour le cas *i.i.d.*, la fonction de vraisemblance est donnée par

$$\begin{aligned} L(\theta; x) = L(\theta; x_1, \ldots, x_n) &= f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta). \end{aligned} \quad (2)$$

Fonction de vraisemblance

Dans le cas où les X_i sont des v.a discrètes, on a

$$\begin{aligned} L(\theta; x) &= P(x_1; \theta) P(x_2; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta). \end{aligned} \tag{3}$$

⚠ Remarque : On peut aussi rencontrer la notation $L(\theta)$ de la vraisemblance de θ au lieu de $L(\theta; x_1, \ldots, x_n)$ (notamment dans ce cours :)).

$\rightarrow$ pour des valeurs d'échantillon données, la fonction de vraisemblance est seulement fonction du paramètre θ.

Maximum de vraisemblance

Définition : **Définition du maximum de vraisemblance**

Maximum de vraisemblance. L'estimation de θ par la méthode du maximum de vraisemblance consiste à choisir, comme estimation de θ, la valeur de θ qui maximise la fonction de vraisemblance $L(\theta)$.

En effet, en choisissant une valeur de θ qui maximise L (ou $\ln L$), cela revient à dire que, parmi les valeurs possible de θ, nous prenons la valeur qui rend le plus probable que possible l'évènement que les les valeurs de l'échantillon observé $(x_1, \ldots, x_n)$ viennent de la population de densité $f(x; \theta)$.

Maximum de vraisemblance : Cas d'un seul paramètre θ

maximum(s) d'une fonction

Mathématiquement, le maximum d'une fonction $L(\theta)$ correspond à la valeur de θ pour laquelle la dérivée de L par rapport à θ est nulle :

$$\frac{\mathrm{d}\, L(\theta)}{\mathrm{d}\, \theta} = 0$$

qui permet d'identifier les extrema de $L(\theta)$ (mais ne permet pas de savoir lesquels parmi ces extrema sont des maxima (que nous recherchons) ou bien des minima (qui ne nous intéressent pas). Il faut donc, après que les solutions de l'équation aient été trouvées, sélectionner celles qui correspondent à des maxima. Un maximum vérifie la dérivée seconde par rapport à θ est négative :

$$\frac{\mathrm{d}^2\, L(\theta)}{\mathrm{d}^2\, \theta} < 0$$

Maximum de vraisemblance : Cas d'un seul paramètre θ

Estimateur du Maximum de vraisemblance (MV)

L'*estimateur du maximum de vraisemblance* (MV) de θ, noté $\hat{\theta}$, à partir des valeurs de l'échantillon $(x_1, \cdots, x_n)$ peut être déterminé à partir de

$$\frac{\mathrm{d}\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{4}$$

$$\frac{\mathrm{d}^2\, L(\theta; x_1, \ldots, x_n)}{\mathrm{d}^2\,\theta}|_{\theta=\hat{\theta}} < 0 \tag{5}$$

il faut donc sélectionner parmi les solutions de la première équation celles qui vérifient cette deuxième équation.

Maximum de vraisemblance

⚠ Remarque :

Bien que la plupart des vraisemblances soient différentiables, les solutions de l'équation de vraisemblance (4) ne s'expriment pas toujours par des formes analytiques.

$\Rightarrow$ On a souvent recours à des méthodes d'optimisations numériques pour identifier les maxima de la fonction de vraisemblance (par exemple comme en régression Logistique, mélange de densités, modèles de Markov cachés, etc)

$\Rightarrow$ par exemple la montée de gradient, l'algorithme de Newton Raphson, l'algorithme EM, etc.

Maximum de vraisemblance

le logarithme est une fonction monotone et la fonction de vraisemblance étant positive

$\Rightarrow$ la fonction de vraisemblance atteint donc son maximum pour la même valeur que son logarithme

Fonction de **log-vraisemblance**

Maximiser la fonction de vraisemblance revient à maximiser son logarithme. Le logarithme de la fonction de vraisemblance s'appelle *log-vraisemblance*.

Manipuler le log de la fonction de vraisemblance au lieu de la vraisemblance elle même vient aussi du fait que, comme cette dernière s'écrit souvent comme produit de densités (de probabilités dans le cas discrets), cela peut résulter en des valeurs très faibles qui peuvent dans certains cas dépasser la précision de calculateurs. Ainsi, traiter des logarithmes revient plutôt à sommer et donc d'éviter des problèmes numériques.

Maximum de vraisemblance

L'équation de vraisemblance devient donc :

$$\frac{\mathrm{d} \ln L(\theta; x_1, \ldots, x_n)}{\mathrm{d}\,\theta}|_{\theta=\hat{\theta}} = 0. \tag{6}$$

Dans le cas où cette fonction est concave et admet donc une seule racine, l'estimateur du maximum de vraisemblance correspond à cette racine et on parle de **maximum global**.

Cependant, la fonction de vraisemblance peut avoir plus d'un maximum (maxima). Dans ce cas, on parle de **maxima locaux**

⚠ Remarque : (lorsque tous les maxima ont été identifiés, seul le plus grand d'entre-eux doit être retenu).

Maximum de vraisemblance : Vraisemblance multivariée I

Plusieurs densités admettent plus d'un paramètre. Par exemple l'estimation d'une densité normale monodimensionnelle nécessite l'estimation de la moyenne μ et de la variance σ^2.

Fonction de log-vraisemblance :

$$\ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)$$

et les estimateurs de MV de $\theta_j, j = 1, \cdots, m$, sont obtenus en résolvant simultanément le système d'équations de vraisemblance

$$\frac{\mathrm{d} \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}\, \theta_j}|_{\theta_j = \hat{\theta}_j} = 0 \quad \text{pour } j = 1, \ldots, m \qquad (7)$$

Maximum de vraisemblance : Vraisemblance multivariée II

et comme pour le cas univarié, mais dans ce cas multivarié c'est plus complexe, il faut en plus que au moins une des dérivées partielles secondes de L soit strictement négative pour au moins un j

$$\frac{\mathrm{d}^2 \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}^2 \theta_j}\Big|_{\theta_j = \hat{\theta}_j} < 0 \quad \text{pour au moins un } j$$

et le déterminant de la matrice des dérivées partielles secondes de L soit strictement positif :

$$\left|\frac{\mathrm{d}^2 \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n)}{\mathrm{d}^2 \theta_j}\right|_{\theta_j = \hat{\theta}_j} > 0$$

Cette dernière condition est en général difficile à vérifier, même dans les cas simples.

Propriétés du maximum de vraisemblance

Soit $\hat{\theta}$ la valeur de l'estimateur du maximum de vraisemblance $\hat{\Theta}$ de θ estimée à partir de l'échantillon $(x_1, \cdots, x_n)$ de taille n

Convergence

L'estimateur obtenu par la méthode du maximum de vraisemblance est convergent : $\lim_{n\to\infty} \mathbb{P}(|\hat{\Theta}_n - \theta| > \epsilon) = 0 \qquad \forall\, \epsilon > 0.$

Absence de biais et efficacité asymptotiques

Quand n tend vers l'infini on a :

$$\lim_{n\to\infty} \mathbb{E}[\hat{\Theta}] = \theta \quad : \text{asymptotiquement sans biais} \tag{8}$$

$$\lim_{n\to\infty} \text{var}[\hat{\Theta}] = \frac{1}{n\mathbb{E}\left[\left(\frac{\partial f(X;\theta)}{\partial \theta}\right)^2\right]} = \frac{1}{\mathcal{I}_n(\theta)} = \text{CRLB} \ : \text{asymptotiquement efficac}$$

Des résultats analogues sont obtenus lorsque X est une v.a. discrète.

Propriétés du maximum de vraisemblance

Normalité asymptotique

La distribution de $\hat{\Theta}$ tend vers une distribution normale lorsque n devient grand. L'EMV est donc *asymptotiquement normal*.

$$\sqrt{n}\,(\hat{\Theta} - \theta) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}\left(0, \mathcal{I}_n(\theta)^{-1}\right). \tag{9}$$

Invariance

On peut montrer que, si $\hat{\Theta}$ est l'EMV de θ, alors l'EMV d'une fonction bijective différentiable de θ, soit $g(\theta)$, est $g(\hat{\Theta})$.

$\Rightarrow$ Cette importante propriété d'invariance implique que, par exemple, si $\hat{\sigma}$ est l'EMV de l'écart type σ pour une distribution donnée, alors l'EMV de la variance σ^2 est $\hat{\sigma}^2$.

Maximum de vraisemblance dans le cas Gaussien

Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires réelles issues d'une population de densité normale $\mathcal{N}(\mu, \sigma^2)$, alors

1. la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de l'espérance μ
2. la variance empirique *corrigée* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$ est un estimateur sans biais de la variance σ^2

et on a

$$\begin{aligned} \bar{X} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \\ \frac{(n-1)S^2}{\sigma^2} &\sim \chi^2_{n-1} \end{aligned}$$

Loi de student

Soit Z une v.a de loi normale centrée et réduite et soit Uune v.a indépendante de Z et distribuée suivant la loi du χ^2 à n degrés de liberté. Par définition la variable

$$T = \frac{Z}{\sqrt{U/n}}$$

suit une loi de Student à n degrés de liberté.

Loi du χ^2

Soient $X_1, \ldots, X_n$, n v.a. indépendantes suivant des lois normales de moyennes respectives μ_i et d'écart-type σ_i ; $Y_i = \frac{X_i - \mu_i}{\sigma_i}$ leurs variables centrées et réduites, alors par définition la variable X , telle que

$$X := \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

suit une loi du χ^2 à n degrés de liberté.

Maximum de vraisemblance dans le cas Gaussien

On peut donc remarquer que $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ suit une loi de student de paramètre $n-1$.

Cela vient du fait que $(\bar{X}-\mu) \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ donc $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$ et comme S^2 suit une loi de χ^2, en remplaçant σ par son estimateur S on a alors la loi de student t_{n-1}.