MÉTHODE DU MAXIMUM DE VRAISEMBLANCE (H3)
(19 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

Il existe trois méthodes générales d'**estimation** d'un **modèle statistique** ou d'une **structure statistique** :

(a) la **méthode des moments**, ou la **méthode des caractéristiques**, qui procède d'une même démarche ;

(b) les méthodes de minimisation de distances ou de normes (méthodes par projection) (cf **méthodes à distance minimale**, **méthode de moindre norme**) ;

(c) les **méthodes non paramétriques** : méthodes par noyau (cf **méthode du noyau**) ou par **pénalisation** ;

(d) la **méthode du maximum de vraisemblance**, qui suppose définie la notion de **vraisemblance**, ie celle de **modèle dominé**.

Toutes ces méthodes ont été adaptées à des **situations statistiques** variées.

C'est sur la méthode du maximum de vraisemblance (cf **estimateur à distance minimum**, **estimateur du maximum de vraisemblance**), ou sur des méthodes dérivées, que de nombreux **estimateurs** ou **tests** sont fondés lorsque les autres méthodes usuelles ne sont pas aisément applicables.

(i) On considère un **modèle statistique** paramétré $(\Omega, \mathcal{T}, P_\theta)_{\theta \in \Theta}$, un **espace d'observation** $(\mathcal{X}, \mathcal{B})$ et une **va** (eg un **échantillon** ou une **statistique**) $X : \Omega \mapsto \mathcal{X}$. On suppose que $(P_\theta^X)_{\theta \in \Theta}$ (**famille** des lois images des P_θ par X) est une **famille de lois dominée** uniformément par une **mesure positive** (σ-finie) μ indépendante du paramètre $\theta \in \Theta$.

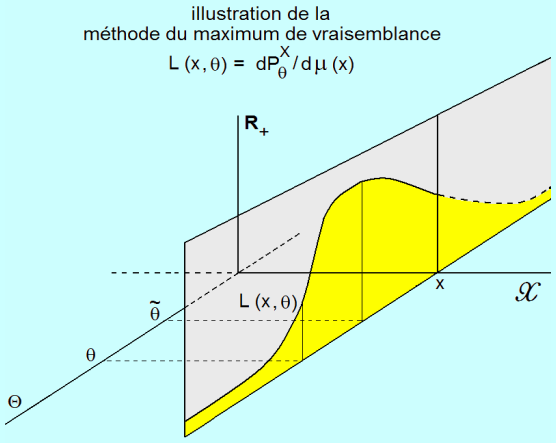
On note :

(1) $f(x, \theta)$ ou $L(x, \theta) = dP_\theta^X(x) / d\mu(x)$

la **fonction de vraisemblance** (**dérivée de NIKODYM-RADON** de P_θ^X pr à μ) associée au modèle image ainsi défini.

On appelle **méthode du maximum de vraisemblance** (sous forme paramétrée) (R.A. FISHER) (ou méthode du mv) la méthode d'**estimation** (ponctuelle) de θ basée sur la valeur $\tilde{\theta}$ (si elle existe) qui maximise f, ie qui est tq (cf schéma ci-après) :

(1) $f(x, \tilde{\theta}) \geq f(x, \theta)$, μ-p.p., $\forall\, \theta \in \Theta$.



On appelle encore **méthode du maximum de vraisemblance** la méthode qui maximise toute fonction mesurable φ strictement croissante pr à la fonction f (cf **fonction numérique**, **application mesurable**). Ainsi, si φ = Log, on définit la **Log-vraisemblance** selon L (x, θ) = Log f (x, θ), μ-p.p..

En général, $\tilde{\theta}$ définit un estimateur t : $\mathcal{X} \mapsto \Theta$ de θ selon $\tilde{\theta}$ = t (X). On appelle **estimateur du maximum de vraisemblance** (mv) de θ l'application (supposée **mesurable**) t o X : $\Omega \mapsto \Theta$ tq t (X (ω)) = $\tilde{\theta}$. On note aussi $\tilde{\theta}$ (x) au lieu de t (x), avec x = X (ω) (**observation** de X).

(ii) On montre que :

(a) **équation estimante**. Si $\Theta \in \mathcal{O}(\mathbf{R}^Q)$ (**ouvert** de $\mathbf{R}^Q$), si f > 0 (strictement) sur $\mathcal{X}$ x Θ et si la vraisemblance (application partielle) $\theta \mapsto$ f (x, θ) est de classe C^1 (P_θ-p.s., $\forall\ \theta \in \Theta$), alors l'estimateur du mv $\tilde{\theta}$ est solution de l'équation estimante suivante, appelée **équation de vraisemblance** (condition nécessaire du premier ordre pour un extrémum) :

$$D_2 f (x, \theta) \text{ ou } \mathrm{Grad}_\theta f (x, \theta) = 0, \ \mu\text{-p.p..},$$

(3) ou

$$D_2 L (x, \theta) \text{ ou } \mathrm{Grad}_\theta L (x, \theta) = 0, \ \mu\text{-p.p..}$$

Plus généralement, on appelle **équation de φ-vraisemblance** l'équation :

(4) $D_2 \varphi (f (x, \theta)) = 0,$

dans laquelle φ est une fonction strictement croissante.

En particulier, on appelle **équation de Log-vraisemblance**, ou parfois même encore **équation de vraisemblance**, l'équation :

(5) $D_2 \text{ Log } f(x, \theta) = 0$.

Chacune de ces équations est donc une équation estimante particulière ;

(b) **invariance** pr aux mesures de base. Si $(P_\theta^X)_{\theta \in \Theta}$ est une **famille de lois homogène** et si les **densités** associées $f_\theta = f(., \theta)$, $\forall\ \theta \in \Theta$, sont strictement positives, alors la méthode du mv est invariante par changement de **mesure** dominante μ . Autrement dit, si μ est remplacée par une autre mesure ν tq $\mu << \nu$, l'estimateur $\tilde{\theta}$ est inchangé ;

(c) **critère de factorisation**. Si S = s (X) est une **statistique exhaustive** pour θ, l'estimateur du mv t o X est fonction de s (ou de S) ;

(d) **invariance par changement de paramètre** (régulier). La méthode du mv est invariante par transformation « régulière » dans l'**ensemble** Θ des valeurs de θ : en effet, si $\tilde{\theta}(x)$ est l'estimateur du mv de θ, si g désigne une **bijection** dans Θ, et si l'on veut estimer $g(\theta)$ (au lieu de θ), alors $g(\tilde{\theta}(x))$ est l'estimateur du mv de $g(\theta)$ (cf **invariance du maximum de vraisemblance**) ;

(e) **biais**. L'estimateur du mv est en général biaisé (cf **estimateur sans biais**) :

(6) $E_\theta (t \text{ o } X)$ ou $E_\theta \tilde{\theta}(X) \neq \theta, \qquad \forall\ \theta \in \Theta$;

où E_θ désigne l'espérance de t o X calculée avec P_θ .

(f) **convergences stochastiques**. On suppose que le **modèle image** $(\mathcal{X}, \mathcal{B}, (P_\theta^X)_{\theta \in \Theta})$ est un **modèle d'échantillonnage** (fini) consistant en N tirages iid, avec $\mathcal{X} = \mathcal{X}_0$, $\mathcal{B} = \mathcal{B}_0^{\otimes N}$ et $P_\theta^X = (P_\theta^\xi)^{\otimes N}$ $(\forall\ \theta \in \Theta)$, où $\xi : \Omega \mapsto \mathcal{X}_0$ est une **va** donnée et X = $(X_1, ..., X_N)$ un N-**échantillon iid** issu de la **variable parente** ξ.

On suppose aussi que $\Theta \in \mathcal{P}(\mathbf{R}^Q)$ (paramètre vectoriel réel), que $(P_\theta^\xi)_{\theta \in \Theta}$ est une **famille de lois identifiable** (ie que $\theta \mapsto P_\theta^\xi$ est une **application injective**) et que la vraisemblance $\theta \mapsto f(x, \theta)$ est strictement positive et de classe C^1 (P_θ-p.s., $\forall\ \theta \in \Theta$).

Enfin, on note $f_\xi(., \theta)$ ou $f_\xi = dP_\theta^\xi / d\mu_0$ la **densité de probabilité** de P_θ^ξ pr à une **mesure σ-finie** μ_0 donnée sur $\mathcal{B}_0$ tq $\mu_0^{\otimes N} = \mu$.

Dans ce cadre, l'équation de Log-vraisemblance s'écrit :

(7) $D_2 f(x, \theta) = (\partial / \partial \theta) \text{ Log } (\Pi_{n=1}^N f_\xi(X_n, \theta)) = \Sigma_{n=1}^N D_2 \text{ Log } f_\xi(X_n, \theta) = 0$.

Alors, $\forall\ \theta \in \Theta$, il existe une suite $(\tilde{\theta}_N(X))_{N \in \mathbf{N}^*}$ de racines de l'équation (7) qui converge P_θ-p.s. vers θ, ie (cf **convergence presque sûre**) :

(8) $\tilde{\theta}_N(X)$ ou $t_N(X) \to \theta$, P_θ-p.s., $\forall\, \theta \in \Theta$.

De plus, cette suite converge en loi vers θ (cf **convergence en loi**) ;

(g) la méthode du mv est liée à la notion d'**entropie**. En effet, si le modèle statistique $(\mathcal{X}, \mathcal{B}, (P_\theta^X)_{\theta \in \Theta})$ est dominé par une **mesure positive** μ et si sa vraisemblance s'écrit $f(., \theta) = dP_\theta^X / d\mu$, l'entropie du modèle s'écrit :

(9) $H(P_\theta^X) = -\int f(., \theta) \operatorname{Log} f(., \theta)\, d\mu = -\int \operatorname{Log} f(., \theta)\, dP_\theta^X$.

L'**entropie croisée**, ou **co-entropie**, ou encore **entropie relative**, des lois P_θ^X et P_τ^X s'écrit $\forall\, (\theta, \tau) \in \Theta \times \Theta$:

(10) $H(\theta, \tau) = H(P_\theta^X, P_\tau^X) = -\int f(., \theta) \operatorname{Log} f(., \tau)\, d\mu = -\int \operatorname{Log} f(., \tau)\, dP_\theta^X$.

L'estimateur du mv vérifie alors :

(11) $H(\theta^*, \tilde{\theta}_N) = \inf_{\theta \in \Theta} H(\theta^*, \theta)$,

où θ^* désigne la **vraie valeur** (inconnue) du paramètre.

(iii) La méthode du mv peut être définie dans un contexte d'**estimation non paramétrique** (ou non paramétrée).

Ainsi, avec un modèle image $(\mathcal{X}, \mathcal{B}, \mathcal{P}^\xi)$, on peut chercher à estimer la densité f de la loi générique $P^\xi \in \mathcal{P}^\xi$ au vu d'un N-**échantillon iid** X issu de ξ. Le nombre de « paramètres » du modèle est alors « infini » puisque f (dérivée de NIKODYM-RADON de P^ξ) appartient à un espace fonctionnel.

L'**estimateur du mv** de P^ξ n'est autre que la **loi empirique** $P_N = N^{-1} \sum_{n=1}^N \delta(X_n)$ associée à X (cf aussi **statistique naturelle**, **fonction de répartition empirique**).

Mais l'estimateur du mv de la densité f elle-même n'est pas défini. Des méthodes possibles ont été définies.

Ainsi, dans la **méthode des cribles** ou la **méthode du maximum de vraisemblance pénalisé(e)** (cf **pénalisation**), la Log-vraisemblance du modèle est modifiée (« pénalisée ») selon :

(12) $\sum_{n=1}^N \operatorname{Log} f(X_n) - \psi(f)$,

où ψ représente une **pénalisation**, ou **pénalité**, qui peut être définie eg pr à une notion de **courbure** selon :

(13) $\psi(f) = \lambda \cdot \int ([(f)^{1/2}]'')^2 \, dx,$

où λ est un **hyper-paramètre** appelé **paramètre de pénalisation** ou **paramètre de lissage** et g'' désigne la **dérivée** seconde de g.