QÜESTIIÓ, vol. 21, 1 i 2, p. 37-58, 1997

REMARQUES SUR LE MAXIMUM DE VRAISEMBLANCE

CATHERINE HUBER*
Université de Paris V

MIKHAIL NIKULIN**
Université de Bordeaux II

Les bonnes et les mauvaises propriétés de la méthode du maximum de vraisemblance sont examinées dans cet article. En ce qui concerne les bonnes, elles sont emboîtées dans le cadre plus général de celles des $M-$estimateurs. En ce qui concerne les mauvaises, l'absence de robustesse dans les cas les plus usuels d'hypothèse gaussienne et la difficulté rencontrée dans la définition même de la vraisemblance sont les deux traits essentiels, avec celui de la qualité asymptotique de l'optimalité de la méthode, illustrée par un exemple paradoxal de jeu où la stratégie du maximum de vraisemblance n'est pas admissible.

Some remarks on the maximum likelihood

Keywords: Vraisemblance, $M-$estimateurs, paradoxe.

* C. Huber. Université de Paris V, 45 rue des Saints-Pères 75 270 Paris cedex 06

** M. Nikulin. Université de Bordeaux II, 146 rue Léo Saignat 33 800 Bordeaux

– Article rebut el juliol de 1995.

– Acceptat el març de 1996.

1. INTRODUCTION

La méthode du maximum de vraisemblance est à la fois l'une des plus utilisées et des plus controversées en statistique. Elle a en effet un attrait à la fois intuitif, parce que la vraisemblance semble bien contenir toute l'information fournie par les observations, et théorique, à cause des bonnes propriétés asymptotiques des estimateurs correspondants sous certaines conditions de régularité. Cependant, cette méthode a plusieurs inconvénients qui ont été mis en évidence en particulier par L. Le Cam [1975] et P. J. Huber [1980], mais aussi par d'autre auteurs tels que J. Berkson [1980], R. R. Bahadur [1958], D. Basu [1980], T. S. Ferguson [1982], M. Goldstein et J. V. Howard [1991], S. Dharmadhikari et K. Joag-Dev [1985], V. G. Voinov et M. S. Nikulin [1993] par exemple. Comme cette méthode est universellement applicable, on a tendance à l'utiliser parfois sans précautions, ce qui conduit à des résultats désastreux. Un des exemples les plus célèbres est celui de L. Le Cam qui exhibe un estimateur «du maximum de vraisemblance» qui converge vers deux fois la valeur du paramètre qu'il est censé estimer, simplement parce que le nombre des paramètres inconnus nuisibles croît à la même vitesse que la taille de l'échantillon [L. Le Cam, 1979]. La vraisemblance elle-même pose des problèmes lorsqu'il y a plusieurs versions de la densité de probabilité sur laquelle repose cette vraisemblance. De plus, chaque fois qu'il n'y a pas de propriété de convexité, il est difficile, voire impossible de trouver un maximum global. Même dans ce cas particulièrement favorable, le maximum de la vraisemblance peut être, et il l'est souvent, si «plat» que l'estimateur correspondant est numériquement très mal défini. B. Efron [1982] nous paraît avoir là-dessus un point de vue très sain: si l'on est suffisamment attentif aux conditions dans lesquels on applique la méthode, on n'aura pas de déboire majeur. Il n'en reste pas moins que deux défauts en quelque sorte intrinsèques subsistent:

1. L'absence de définition correcte d'un estimateur du maximum de vraisemblance lorsqu'il peut y avoir une ambiguïté sur la version de la densité à employer, ou bien lorsque le modèle n'est pas dominé. La tentative de F. W. Scholz [1980] pour donner une définition unifiée du maximum de vraisemblance est une réponse à cette question. Elle n'est malheureusement pas très connue et n'a pas eu l'impact auquel on aurait pu s'attendre.
2. Le manque de robustesse des estimateurs du maximum de vraisemblance les plus employés: ceux qui sont relatifs au modèle gaussien. Le plongement des estimateurs M-V dans un ensemble plus général, celui des M-estimateurs, par P. J. Huber [1980] traite de cette question car il permet de corriger l'absence de robustesse de certains estimateurs M-V.

Cet exposé se déroule en trois parties. Nous allons voir successivement:

- Les propriétés des estimateurs du maximum de vraisemblance considérés comme un cas particulier des M-estimateurs, sous des conditions de régularité

du modèle statistique supposé dominé. Trois propriétés essentielles sont considérées: la consistance, la normalité asymptotique et la sensibilité de ces estimateurs à des écarts par rapport au modèle. Ce point de vue sur le maximum de vraisemblance est celui qui est issu des études de robustesse [P.J. Huber, 1981] et il met en évidence l'instabilité des estimateurs usuels du maximum de vraisemblance.

- Un essai de théorie unifiée du maximum de vraisemblance, que le modèle soit ou non dominé. Cette partie de l'exposé est fondée sur un article assez peu connu de F.W. Scholz qui, bien conscient des ambiguïtés qui subsistent dans la définition des estimateurs du maximum de vraisemblance, défauts qui ont pour conséquence de mauvaises propriétés de ces estimateurs, a fait une très intéressante tentative de définition unifiée du maximum de vraisemblance.
- Quelques paradoxes du maximum de vraisemblance. On peut trouver de tels exemples en grand nombre chez L. Le Cam [1979], mais aussi dans les articles cités plus haut.

2. MAXIMUM DE VRAISEMBLANCE GÉNÉRALISÉ

Soit $X_1, X_2, \ldots, X_n$ un échantillon d'une variable aléatoire X gouvernée par une probabilité $P_\theta \in H$, où H est un ensemble de mesures de probabilité absolument continues par rapport à une mesure μ:

$$H = \{P_t, t \in \Theta \subset \mathbb{R}^p\}$$

On note f la densité de P par rapport à μ , soit

$$f(.,t) = \frac{dP_t}{d\mu}$$

et ℓ la dérivée par rapport à t du logarithme de f, soit

$$\ell(x,t) = \frac{\partial Log\, f(x,t)}{\partial t}.$$

Alors l'estimateur du maximum de vraisemblance de θ est $\hat{\theta}$ défini par l'équation

$$\sum_{i=1}^{n} \ell(x_i, \hat{\theta}) = 0$$

2.1. Définition d'un M-estimateur

Un estimateur du maximum de vraisemblance généralisé est solution d'une équation analogue, où la fonction ℓ est remplacée par une fonction ψ vérifiant de bonnes conditions de régularité . Ces conditions de régularité sont destinées à permettre à l'estimateur d'avoir toutes les propriétés des bons estimateurs du maximum de vraisemblance, c'est à dire consistance, et normalité asymptotique, plus celle d'être peu sensible à des écarts des observations par rapport au modèle. L'équation de la vraisemblance est alors remplacée par

$$\sum_{i=1}^{n} \psi(x_i, \hat{\theta}) = 0$$

Pour être plus rigoureux, on ne devrait pas parler d'*un* estimateur du maximum de vraisemblance, mais d'une suite $\{T_n\}_{n\in\mathbb{N}}$ d'estimateurs. Cette suite $\{T_n\}_{n\in\mathbb{N}}$ peut être considérée comme la restriction à

$$\mathcal{F}_n = \{\text{lois empiriques d'ordre n}\}$$

de T, fonctionnelle définie sur $\mathcal{F}$ par

$$\int \psi(x, T(F))dF(x) = 0$$

Exemples:

1. Modèle de translation: l'équation s'écrit dans ce cas

$$\int \psi(x - T(F))dF(x) = 0.$$

   Si la fonction ψ est l'identité, l'estimateur obtenu est la moyenne et si c'est la valeur absolue, la médiane.

2. Modèle de régression linéaire: l'équation correspondante lorsque Y, variable aléatoire réelle est égale à une combinaison linéaire des composantes d'un vecteur x de dimension p plus une variable aléatoire S, l'erreur, de moyenne nulle, s'écrit

$$\int \psi(y - \langle\theta, x\rangle)x_j dF(x) = 0$$

$$j = 1, 2, \ldots, p$$

2.2. Propriétés des M-estimateurs

La fonction $\psi(x,t)$ dépend des deux variables, éventuellement multidimensionnelles, t et x, désignant respectivement le paramètre et la variable. On définit:

$$m(\theta,t) = \int \psi(x,t) dP_\theta(x).$$

Theorem 1 (Consistance) *Si l'application $t \mapsto m(\theta,t)$ est nulle pour $t = \theta$ et décroissante dans un voisinage de θ, alors $\exists$ une suite $\{T_n\}_{n\in\mathbb{N}}$ de solutions de l'équation*

$$t \mapsto \int \psi(x,t) dF_n(x) = 0$$

qui converge presque sûrement vers θ.

Theorem 2 (Normalité asymptotique) *Sous les hypothèses du théorème 1 et si de plus, l'application $t \mapsto m(\theta,t)$ est continuement dérivable, de dérivée $\frac{\partial\psi(x,t)}{\partial t}$ notée $\psi'(x,t)$ telle que:*

- *$t \mapsto \psi'(x,t)$ soit continue, uniformément en x*
- *$0 < \int \psi^2(x,\theta) dP_\theta(x) = d^2(\theta) < \infty$*
- *$0 < \int \psi'(x,\theta) dP_\theta(x) = c(\theta) < \infty$*

Alors, si $\{T_n\}_{n\in\mathbb{N}}$ est une suite de solutions de

$$\int \psi(x,T_n) dF_n(x) = 0$$

on a

$$\sqrt{n}(T_n - \theta) \quad \xrightarrow{\mathcal{L}} \quad N(0, \frac{d^2(\theta)}{c^2(\theta)})$$

Démonstration: Par définition de T_n, $0 = \psi_n(x,T_n) = \psi_n(x,\theta) + (T_n - \theta)\psi'(x,\theta_n)$ où θ_n tend presque sûrement vers θ. On peut donc écrire:

$$\begin{array}{rclcll} \psi'(x,\theta_n) & = & [\psi'(x,\theta_n) - \psi'(x,\theta)] & + & [\psi'_n(x,\theta) - E_\theta\psi'(X,\theta)] & + & E_\theta\psi'(X,\theta) \\ & = & A_n & + & B_n & + & c(\theta) \end{array}$$

Or A_n tend presque sûrement vers 0 à cause de la continuité en t de ψ', qui est uniforme en x, et B_n tend aussi presque sûrement vers 0 à cause de la loi forte des grands nombres. Donc

$$\sqrt{n}(T_n - \theta) = \frac{-\sqrt{n}\psi_n(x,\theta)}{A_n + B_n + c(\theta)}$$

qui, d'après le théorème limite centrale appliqué à $n\psi_n = \sum_{i=1}^{n} \psi(X_i,\theta)$ est asymptotiquement normal de moyenne nulle et de variance $\frac{d^2(\theta)}{c^2(\theta)}$.

La sensibilité d'un estimateur à des écarts par rapport au modèle peut être mesurée par l'impact sur cet estimateur du retrait d'une petite masse ε de probabilité qui serait mise au point *x*, soit $\varepsilon\delta_x$. La courbe de sensibilité décrit, en fonction de *x*, la limite de cet impact lorsque ε tend vers 0. Cette courbe s'exprime simplement en fonction de ψ pour un M-estimateur comme le montre le théorème suivant (pour la démonstration et pour plus de détails, voir P. J. Huber [1980] ou C. Huber-Carol [1985]).

■

Theorem 3 (Sensibilité) *Soit F une fonction de répartition sur* $(\mathbb{R},\mathcal{B})$*, et soit Y une variable aléatoire réelle définie sur* $(\mathbb{R},\mathcal{B})$*. Si, pour tout t réel, l'équation* $\int \psi(y,t)\, dF(y) = 0$ *définit sans ambiguïté une fonctionnelle T sur*

$$\mathcal{G} = \{G = (1-\varepsilon)F + \varepsilon\delta_x, x \in \mathbb{R}, 0 \leq \varepsilon \leq \varepsilon_0\}$$

pour un ε_0 *de* $[0;1]$ *et si, de plus,* $t \mapsto \psi(y,t)$ *est continue dérivable, de dérivée* $\psi'(y,t)$ *telle que* $|\psi'(y,t)| < g(y)$ *pour une fonction g de* $L^2(F)$*, alors la suite* $\{T_n\}_{n\in\mathbb{N}}$ *a pour courbe d' influence:*

$$CI_{T,F}(x) = \frac{-\psi(x,T(F))}{\int \psi'(y,T(F))dF(y)}$$

3. THÉORIE UNIFIÉE DU MAXIMUM DE VRAISEMBLANCE

3.1. Cas discret

C'est dans le cas discret que la définition d'un estimateur du maximum de vraisemblance ne pose en général pas de problème. En effet, si $X \sim P_\theta$, où P_θ est une loi discrète dont le paramètre θ appartient à un ensemble Θ, *un* estimateur du maximum de vraisemblance est défini comme

$$\hat{\theta} = \arg\max_{\theta\in\Theta} P(X = x|\, \theta)$$

3.2. Cas d'un modèle dominé

Prenons un modèle dominé par une mesure μ, par exemple la mesure de Lebesgue. Soit $P_\theta \ll \mu, \theta \in \Theta$. On considère *une* version $f_\theta = \frac{dP_\theta}{d\mu}$ de la densité de P_θ par rapport à μ et la vraisemblance de l'observation x devient $V(\theta) = f(x|\theta)$. Aussi la définition utilisée dans le cas discret ne peut elle plus être employée sans le risque d'avoir une ambiguïté: *l'estimateur du maximum de vraisemblance dépendra de la version choisie pour la densité*. Au lieu de prendre la valeur de la densité au point x observé, on va considérer un voisinage $N(x)$ de ce point et, en notant $V(N(x))$ le volume de ce petit voisinage de x, on aura

$$P(X \in V(x)| \quad \theta) \simeq f(x|\theta)V(N(x))$$

3.3. Cas d'un modèle non dominé

Considérons la situation non paramétrique suivante: la famille de lois est celle des lois symétriques par rapport à 0. Le paramètre est un paramètre de translation. Dans ce cas, il n'y a pas de mesure dominante σ-finie. Il n'y a donc pas de densités à comparer. Dans ce cas, deux tentatives ont été faites pour généraliser la méthode du maximum de vraisemblance: l'une par Kiefer et Wolfowitz en1956, l'autre par Kalbfleish et Prentice en 1980. F. W. Scholz en propose une troisième qui est plus satisfaisante.

1. Généralisation de Kiefer et Wolfowitz On prend les probabilités de la famille deux par deux. A ce moment-là, le couple $\{P,Q\}$ est toujours dominé par la mesure somme $P+Q$. Cependant, comme ci-dessus, la version des dérivées de Radon-Nicodym n'étant pas spécifiée, à chaque choix va correspondre une solution différente.

   Exemple:
   Soit $\varphi(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$ la densité de la loi normale $N(0,1)$ et
   $$\varphi*(x) = \begin{cases} \varphi(x), & \text{si} \quad x \neq 1, \\ 10, & \text{si} \quad x = 1. \end{cases}$$
   Soit le modèle $\{P_\theta = f_\theta(x).\mu = \varphi(x-\theta).\mu; \theta \in \mathbb{R}\}$ où μ est la mesure de Lebesgue, c'est à dire que $P_\theta = N(\theta,1)$. Supposons que l'on ait une seule observation x. Alors, l'estimateur de θ sera x avec la première version et $x-1$ avec la deuxième, et cela pour tout x. Si maintenant on a n observations, $x_1, x_2, \ldots, x_n$, ce sera $x_{(i)} - 1$ pour l'une des statistiques d'ordre $x_{(i)} - 1$. On voit donc que cette généralisation a les mêmes défauts que le cas continu: l'estimateur dépend de la spécification choisie pour la version de la densité, *à moins que la version de la densité ne soit, a priori, une donnée du problème*.

2. Généralisation de Kalbfleish et Prentice Cette généralisation a été proposée par Kalbfleish et Prentice dans le contexte des durées de survie et pour l'estimateur de Kaplan-Meier de la fonction de répartition. Elle consiste à transformer les données de la manière suivante:

   - Discrétiser les données: elles sont de toute faç on toujours discrètes, à cause de la limitation de la précision des mesures, de l'arrondi.
   - Faire un maximum de vraisemblance sur la multinomiale correspondante.
   - Espérer que le maximum de vraisemblance discret va converger vers une limite quand l'erreur d'arrondi tend vers 0. Ce dernier point n'est pas encore élucidé (cf Van der Laan, 1995): deux points restent obscurs, le premier est: y-a-t-il une limite ?, et le deuxième: cette limite, quand elle existe, dépend elle du groupement ?

3. Théorie unifiée de F. W. Scholz Cette définition unifiée résulte du mariage des deux idées précédentes: d'une part, on prend les probabilités par couples, et, d'autre part, on considère un voisinage de l'observation x. On définit les quantités suivantes:

$\mathcal{X}$	un espace métrique de métrique d ,
$\mathcal{B}$	la sigma-algèbre des boréliens de $\mathcal{X}$,
$\mathcal{P}$	une famille de probabilités sur $(\mathcal{X}, \mathcal{B})$,
Alors, $\forall x \in \mathcal{X}$,	on définit

Définition 1

$$\mathcal{N}_x = \{N_x : x \in \overset{o}{N}_x, N_x \in \mathcal{B}\}$$

et $D(N_x)$ est le diamètre de N_x.

Définition 2 ***(Dominance en un point x)*** *Soit $P, Q, \in \mathcal{P}$. On dit que $P \overset{x}{\geq} Q$ si*

$$\liminf_{\varepsilon \to 0} \{\frac{P(N_x)}{Q(N_x)} : N_x \in \mathcal{N}_x, D(N_x) \leq \varepsilon\} \geq 1,$$

où, par convention $\frac{0}{0} = 1$.

Notation: On notera la quantité qui intervient à gauche dans la formule précédente $\underline{\lim} \frac{P(N_x)}{Q(N_x)}$.

□

Définition 3 (Equivalence en un point x) *On dit que P et Q sont équivalentes en x si* $P \overset{x}{\geq} Q$ *et* $Q \overset{x}{\geq} P$ *et on note* $P \overset{x}{=} Q$.

Commentaires:

- La relation $\overset{x}{\geq}$ est réflexive $P \overset{x}{\geq} P \quad \forall P$.
- Elle est transitive: $P \overset{x}{\geq} Q$ et $Q \overset{x}{\geq} R$ impliquent que $P \overset{x}{\geq} R$.
- Si on définit $\{P\}_x$ comme la classe d'équivalence de P pour x, soit

  $$\{P\}_x = \{Q \in \mathcal{P} : Q \overset{x}{=} P\}$$

  et $\mathcal{P}_x$ comme l'ensemble de ces classes, pour P variant dans $\mathcal{P}$, alors $\{\mathcal{P}_x, \overset{x}{\geq}\}$ est un ensemble partiellement ordonné.
- $P \overset{x}{=} Q \Leftrightarrow lim_{D(N_x)\to 0} \frac{P(N_x)}{Q(N_x)}$ existe et est égale à 1.

Définition 4 (Maximum de vraisemblance) *La statistique* $P_0 \in \mathcal{P}$ *est un estimateur du maximum de vraisemblance par rapport à x et à* $\mathcal{P}$ *si* $\forall Q \in \mathcal{P}$ *tel que* $Q \overset{x}{\leq} P_0$ *alors* $Q \overset{x}{=} P_0$ *c'est à dire que* P_0 *est M-V si et seulement si l'une des deux conditions équivalentes suivantes est réalisée:*

1. *Il n'existe pas de loi Q dans* $\mathcal{P} - \{P_0\}_x$ *telle que* $Q \overset{x}{\geq} P_0$
2. $$\liminf \frac{Q(N_x)}{P_0(N_x)} < 1, \qquad \forall Q \in \mathcal{P} - \mathcal{P}_0.$$

L'existence d'un estimateur M-V n'est pas garantie par cette définition

4. UN PARADOXE DE LA VRAISEMBLANCE

L'exemple suivant est considéré par Fraser(1984), Wolpert(1988) et Joshi (1989) comme mettant en question les principes de la vraisemblance. On considère, dans une urne, 6 boules numérotées respectivement k, k, 4k +1, 4k +2, 4k + 3 et 4k + 4. Il s'agit d'estimer le paramètre $\theta \in \mathbb{N}$, qui vaut k, après avoir tiré au hasard une boule numéroté $S = s$. La vraisemblance $V(s, \theta)$ vaut:

$$\begin{aligned} V(s,\theta) &= 2/6 \text{ si } s=\theta \\ &= 1/6 \text{ si } s=4\theta+1 \\ &= 1/6 \text{ si } s=4\theta+2 \\ &= 1/6 \text{ si } s=4\theta+3 \\ &= 1/6 \text{ si } s=4\theta+4 \end{aligned}$$

Par conséquent, comme $max_\theta V(s,\theta) = 1/3$ pour $\hat{\theta} = s$, l'estimateur M-V de θ est

$$u_L(s) = s.$$

Mais en fait, pour une valeur observée s de S , il n'y a que deux valeurs possibles de θ:

θ	=	s	si l'on a tiré l'une des deux boules numérotées θ
θ	=	$\left[\frac{s-1}{4}\right]$	si l'on a tiré l'une des quatre autres boules numérotées $4\theta+i$ pour $1 \leq i \leq 4$.

On peut écrire:

$$\begin{aligned} V(\theta,s) &= 0 \quad \text{si } \theta \notin \{s; \left[\tfrac{s-1}{4}\right]\} \\ &= \tfrac{1}{3} \quad \text{si } \theta = s \\ &= \tfrac{2}{3} \quad \text{si } \theta = \left[\tfrac{s-1}{4}\right] \end{aligned}$$

On peut donc envisager une deuxième stratégie

$$u_C(s) = \left[\frac{s-1}{4}\right].$$

Le paradoxe provient alors du fait que la première stratégie u_L, qui est celle du maximum de vraisemblance, donne la bonne réponse vaut 1/3 alors que la deuxième stratégie donne la bonne réponse avec la probabilité 2/3, si θ est strictement positif. Dans le cas où θ est nul, cette dernière probabilité vaut même 1. Si l'on considère le tableau croisé de S et θ, on obtient le tableau suivant:

		0	1	2	3	4	5	6	7	8	9	10	11	12
	0	2	1	1	1	1								
	1		2				1	1	1	1				
θ	2			2										
	3				2						1	1	1	1
	4					2								

On peut encore accroître cette disparité en généralisant l'exemple précédent de la manière suivante: on a dans une urne 99 boules marquées k, et 99^2 boules marquées

$99_k^2 + i$, i variant de 1 à 99^2. Si les deux estimateurs de k sont

$$\begin{aligned} u_C(s) &= \left[\frac{s-1}{99^2}\right] \\ u_L(s) &= s \end{aligned}$$

la probabilité que u_C soit correct est supérieure ou égale à $0,99$ alors que celle de u_L est de $0,01$. Autrement dit, la vraisemblance de u_C est 99 fois plus élevée que celle de u_L.

Les principes en cause sont les suivants:

- LP (Likelihood Principle):
  Ce principe est celui selon lequel, toute l'information sur le paramètre θ est contenue dans la fonction de vraisemblance de θ étant donnée l'observation s.

  Ce principe n'est pas remis en cause par ce paradoxe.

- LPF (Likelihood Preference Principle):
  Lorqu'on fait de l'inférence statistique, on doit toujours préférer les valeurs de θ qui correspondent aux valeurs les plus grandes de la vraisemblance.

  C'est ce principe qui est remis en cause par cet exemple.

- CPP (Confidence Preference Principle):
  Si deux règles A et B donnent des intervalles de confiance pour θ de même taille, et si, pour tout θ, la probabilité que A soit correct est supérieure ou égale à celle de B, et, pour au moins un θ, strictement supérieure, on doit préférer A à B.
  Ce principe donne la préférence à la stratégie u_C dans notre exemple. Cet exemple met donc en évidence une contradiction entre les deux principes LPF et CPP.

En fait, il semble que *aucun* de ces principes ne soit bon. On peut s'en rendre compte en cherchant une loi a priori sur $\Theta = \mathbb{N}$ qui mène , d'un point de vue bayésien, à la préférence de l'une ou de l'autre des deux stratégies u_C et u_L.
Soit la loi a priori sur $\mathbb{N}$

$$p_i = P(\theta = i), i \in \mathbb{N}.$$

Alors

$$\frac{P(\theta = u_C|s)}{P(\theta = u_L|s)} = \frac{q(s)}{1-q(s)} = \frac{p_{u_C}}{2p_{u_L}}$$

pour $s \geq 0$.

Pour que cela conduise au choix de u_L il faut que

$$p_s \geq \frac{1}{2} p_{\left[\frac{s-1}{4}\right]}$$

Cependant, il y a beaucoup de lois a priori pour lesquelles u_C est préférable, par exemple:

$$p_i = \frac{1}{2^{i+1}}.$$

On peut se poser les questions suivantes:

- Existe-t-il une loi a priori qui conduirait à u_L quel que soit s ?
  La condition à remplir est

  $$p_s \geq (1/2) p_{[s-1/4]}$$

  pour $s = 1, 2, \ldots$ qui entraînerait que

  $$\sum_{s=1}^{\infty} p_s \geq 2 \sum_{s=0}^{\infty} p_s = 2$$

  ce qui est impossible.

- Y a-t-il seulement un nombre fini de valeurs de s pour lesquelles u_C est préférable ?
  Si c'était le cas, il existerait n au-delà duquel on aurait:

  $$\sum_{s=n}^{\infty} p_s \geq 2 \sum_{s=n}^{\infty} p_s$$

  ce qui est impossible, sauf si le support de θ n'était pas infini.

Il y a deux arguments contre le principe LPP:

- Aucune loi a priori sur $\Theta = \mathbb{N}$ ne conduit à la stratégie u_L pour tout s.
- Pour toute loi a priori sur $\Theta = \mathbb{N}$, il y a une infinité de valeurs de s pour lesquelles u_C est préférable.

Commentaires:

- Si le support de θ n'est plus $\Theta = \mathbb{N}$ mais $\Theta = \{0, 1, 2, \ldots, N\}$, où N est fini fixé, tout change.

Exemple: $0 \leq \theta \leq 4$

Le tableau suivant donne les probabilités de ne pas se tromper pour N=4. C'est la stratégie u_C qui est choisie si le critère est maximin.

		θ					Minimum
		0	1	2	3	4	
Règle	C	1	$2/3$	$2/3$	$2/3$	$2/3$	$2/3$
Règle	L	$1/3$	$1/3$	1	1	1	$1/3$

- Une autre possibilité consiste à utiliser la règle mixte M_4: choisir u_C avec la probabilité $2/3$ et u_L avec la probabilité $1/3$ lorsque l'observation s est comprise entre 1 et 4, bornes comprises. Dans les autres cas, c'est à dire $s = 0$ ou $s = 5$, les deux stratégies u_C et u_L sont de toute façon, identiques. On obtient alors le tableau suivant:

		θ					Minimum
		0	1	2	3	4	
Règle	C	1	$2/3$	$2/3$	$2/3$	$2/3$	$2/3$
Règle	L	$1/3$	1	1	1	1	
Règle	M_4	$7/9$	$7/9$	$7/9$	$7/9$	$7/9$	$7/9$

5. UNE GÉNÉRALISATION PRATIQUE DU MAXIMUM DE VRAISEMBLANCE

Une autre approche, qui peut être considérée comme une généralisation de la méthode du maximum de vraisemblance, a été proposée par Weiss et Wolfowitz (1966), (1974), voir aussi Blyth (1983), Weiss (1986), Voinov et Nikulin (1993) etc. On fait ici quelques remarques concernant cette approche.

Soit $X = (X_1, \ldots, X_n)^T$ un échantillon, $X_i \sim f(x;\theta), \theta \in \Theta \subseteq \mathbb{R}^1$. Notons $L(X;\theta)$ la fonction de vraisemblance de X:

$$L(X;\theta) = \prod_{i=1}^{n} f(X_i, \theta),$$

et soit r un nombre positif, $r \in \mathbb{R}^1_+$. Dans ce cas la statistique

$$T_r = T_r(X) = \arg\max_t \int_{t-r}^{t+r} L(X;\theta)d\theta, \quad T_r \in \Theta. \tag{1}$$

s'appelle *l'estimateur du maximum de probabilité* pour θ. On remarque que en réalité la relation (1) détermine une famille $\{T_r, r > 0\}$ des estimateurs T_r pour θ, ce qui nous permet de choisir l'estimateur optimal dans un sens ou un autre. Il est clair que

$$T_0 = \lim_{r \to 0} T_r = \hat{\theta}_n, \tag{2}$$

où $\hat{\theta}_n$ est l'estimateur de maximum de vraisemlence, à condition que le maximum globale dans (1) soit un point de continuité de $L(X;\theta)$. Si $L(X;\theta)$ est discontinue dans ce cas la relation (2) peut être considérée comme une définition de l'estimateur du maximum de vraisemblance. Cette convention nous permet d'inclure T_0 dans la classe $\{T_r\}$ et par conséquent de considérer, suivant en cela C. Blyth, la méthode du maximum de probabilité comme une généralisation de la méthode du maximum de vraisemblance.

Exemple 1

Soit $X = (X_1, X_2, \ldots, X_n)^T$ un échantillon,

$$X_i \sim f(x;\theta), \qquad \theta \in \Theta = \mathbb{R}^1, \qquad f(x;\theta) = \begin{cases} \exp -(x-\theta) & x \geq \theta, \\ 0, & x < \theta. \end{cases} \tag{3}$$

Il est bien connu que la statistique $X_{(1)} = \min(X_1, \ldots, X_n)$ est l'estimateur du maximum de vraisemblance pour θ. Pour construire un estimateur du maximum de probablité il nous faut trouver n'importe quelle statistique T_r pour laquelle

$$\int_{T_r - r}^{T_r - r} L(X;\theta) d\theta$$

atteint son maximum global. Blyth a montré (1983) que

$$T_r = X_{(1)} - \text{ r.} \tag{4}$$

Par des calculs directs on peut montrer que le le risque quadratique

$$E(T_r - \theta)^2 = r^2 - \frac{2r}{n} + \frac{2}{n^2}$$

de l'estimateur T_r atteint son minimum quand $r = 1/n$,

$$E(T_{1/n} - \theta)^2 = \min_r E(T_r - \theta)^2 = \frac{1}{n^2}. \tag{5}$$

On remarque que

$$E\,(X_{(1)}-\theta)^2=\frac{2}{n^2} \tag{6}$$

et donc de ce point de vue l'estimateur du maximum de probabilité est mielleur que l'estimateur du maximum de vraisemblance, qui est superefficace dans ce modèle et par conséquent $X_{(1)}$ est inadmissible par rapport à la fonction de perte quadratique.

On compare maintenant ces deux méthodes par rapport à la fonction de perte de Laplace. On trouve facilement que on minimise le risk

$$\min E\mid T_r-\theta\mid=\min_r\left\{\frac{1}{n}(nr-1+2e^{-nr})\right\}=\frac{0.693}{n}$$

si on choisit $r=(\ln 2)/\mathrm{n}$. Donc l'estimateur $T_{(\ln 2)/n}$ est le meilleur dans la classe $\{T_r\}$ par rapport à la fonction de perte de Laplace. On trouve en même temps que

$$E\mid X_{(1)}-\theta\mid=\frac{1}{n},$$

et on en tire de nouveau que l'estimateur du maximum de probabilité est mielleur que l'estimateur du maximum de vraisemblance et donc l'estimateur de maximum de vraisemblance est inadmissible par rapport à la fonction de perte de Laplace. $X_{(1)}$ est superefficace puisque sa variance est plus petite que la borne de Crámer-Rao. On peut trouver d'autre exemples intéressants on peut trouver chez Blyth (1983), Voinov et Nikulin (1993, 1996).

6. MÉTHODE DU MAXIMUM DE VRAISEMLANCE ET LA MÉTHODE DES MOMENTS

Soit $X=(X_1,X_2,\ldots,X_n)^T$ un échantillon, et d'après l'hypothèse

$$H_0: X_i\sim f(x;\theta),\qquad \theta=(\theta_1,\ldots,\theta_s)^T\in\Theta\subseteq\mathbb{R}^s,$$

$$f(x;\theta)=h(x)\exp\left\{\sum_{k=1}^{s}\theta_k x^k+v(\theta)\right\},\qquad x\in\mathcal{X}\subseteq\mathbb{R}^1, \tag{1}$$

où $\mathcal{X}$ est un ensemble borelien dans $\mathbb{R}^1$,

$$\mathcal{X}=\{x: f(x;\theta)>0\}\text{ pour tout }\theta\in\Theta.$$

La famille (1) est tres riche, on y trouve, par exemple, la famille des lois normales $N(\mu,\sigma^2)$:

$$f(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],\ |x|<\infty,\ \Theta=\{\theta=(\mu,\sigma^2)^T:\mid\mu\mid<\infty,\sigma>0\},$$

et la famille des lois de Poisson $P(\theta)$:

$$f(x;\theta) = \frac{\theta^x}{x!}e^{-\theta}, \quad \lambda \in \{0.1....\}, \quad \theta \in \Theta =]0,\infty[,$$

appartiennent à (1) avec $s = 2$ et $s = 1$ respectivement.
Supposons que

1) Le support $\mathcal{X}$ ne dépend pas de θ;

2) Le Hessien

$$H_v(\theta) = -\left\| \frac{\partial^2}{\partial\theta_i\partial\theta_j} v(\theta) \right\|_{(s\times s)} \tag{2}$$

de la fonction $v(\theta)$ soit défini positif sur Θ.

3) Le moment $a_s = \mathbf{E}X_1^s$ d'ordre s existe.

Dans ce modèle H_0

$$U_n = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i^3, \ldots, \sum_{i=1}^n X_i^s \right)^T \tag{3}$$

est la statistique exhaustive minimale. De (1) on trouve que

$$-grad v(\theta) = a(\theta) = (a_1(\theta), a_2(\theta), \ldots, a_s(\theta))^T, \tag{4}$$

et donc la statistique

$$T_n = \frac{1}{n}U_n \tag{5}$$

est le meilleur estimateur sans bias (**MVUE**, minimum variance unbiased estimateur, voir per exemple Voinov et Nikulin, 1993) pour $a(\theta)$, car

$$E_\theta T_n \equiv a(\theta), \qquad \theta \in \Theta, \tag{6}$$

et cette relation nous permet d'une façon unique d'obtenir l'estimateur θ_n^* pour θ par la méthode des moments de l'équation

$$T_n = a(\theta) \tag{7}$$

en fonctions de la statistique exhaustive U_n.

Par ailleurs, les conditions 1)-3) nous garantissent l'existence de l'estimateur de maximum de vraisemblance $\hat{\theta}_n$, qui est la racine unique (!) de la même équation

$$T_n = a(\theta).$$

On en déduit donc que pour la famille exponentielle (1) les méthodes de maximum de vraisemblance et des moments nous amènent au même estimateur, $\hat{\theta}_n = \theta_n^*$, de θ.

Exemple 1

On connait bien que pour la loi normale $N(\mu, \sigma^2)$ l'estimateur de maximum de vraisemblance $\hat{\theta}_n$ pour $\theta = (\mu, \sigma^2)^T$ est $\hat{\theta}_n = (\bar{X}_n, s_n^2)^T$, où

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \tag{8}$$

et donc $\hat{\theta}_n = \theta_n^*$.

Exemple 2

Soit $X = (X_1, \ldots, X_n)^T$ un échantillon, et d'apres l'hypothese H_0:

$$X_i \sim f(x;\theta) = \frac{\theta^x}{x!}e^{-\theta}, x \in \mathcal{X} = \{0, 1, 2, \ldots\}.$$

On connait bien que dans ce cas

$$\hat{\theta}_n = \theta_n^* = \bar{X}_n.$$

Enfin on note que cette remarque nous permet de considérer la méthode de vraisemblance comme un cas particulier de la méthode des moments, voir Huber et Nikulin (1993), Greenwood et Nikulin (1996).

RECONNAISSANCE

Les auteurs sont reconnus à un arbitre et à C.M. Cuadras par leur outiles commentaires.

7. REFERENCES

[1] **Bahadur, R.R.** (1958). «Examples of inconsistency of maximum likelihood estimates». *Sankhyā*, **20**, pp 207–210.

[2] **Barnett, V.D.** (1966). «Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots». *Biometrika*, **53, 1 and 2**, 151–165.

[3] **Basu, D.** (1954). *An inconsistency of the method of maximum likelihood*.

[4] **Berger J.O., Wolpert R.L.** (1988). *The likelihood principle*, Hayward : Institute of mathematical statistics.

[5] **Berkson, J.** (1980). «Minimum chi-square, not maximum likelihood». *AS*, **8, 3**, 457–487.

[6] **Blyth C.R.** (1982). *Maximum probability estimation in small samples*, in A. Festschrift for Erik Lehmann, (Bickel and Doksum ed.), Wadsworth, pp 83–86.

[7] **Dharmadhikari, S.** and **Joag-Dev, K.** (1985). «Examples of non unique maximum likelihood estimators». *The American Satistician*, **39, 3**, 199–200.

[8] **Efron, B.** (1980). «Discussion of a paper by Joseph Berkson on Minimum chi-square». *AS*, **8, 3**, 457–487.

[9] **Evans B., Fraser D.A.S.** and **Monette J.** (1986). «On principles and arguments to likelihood». *Canadian Journal of Statistics*, **14**, pp 181–199.

[10] **Ferguson, T.S.** (1982). «An inconsistent maximum likelihood Estimate». *JASA*, **77, 380**.

[11] **Fraser D.A.S., Monette G.** and **Ng K.W.** (1984), *Marginalization, likelihood and structural models*, in Multivariate Analysis VI, (ed. P.R. Krishnaiah), pp 209-217, Amsterdam : North Holland.

[12] **Goldstein, M.** and **Howard, J.V.** (1991). «A likelihood paradox». *JRSS*, **53, 3**, 619–628.

[13] **Huber, P.J.** (1980). *Robustness theory*.

[14] **Huber-Carol, C.** (1985). *Théorie de l'inférence statistique robuste*. Springer-Verlag.

[15] **Huber C., Nikulin M.** (1993). *Applications statistiques des transformations des variables aléatoiress*, Preprint, U.F.R. M.I.2S, Université Bordeaux 2.

[16] **Joshi V.M.** (1989). «A counter-example against the likelihood principle». *JRSS B*, **51**, 215–216.

[17] **Kalbfleish J.D.** and **Prentice R.L.** (1980). *The statistical Analysis of failure time data*, Wiley, New York.

[18] **Kempthorne, O.** (1989). «The fate worse than death and other curiosities and stupidities». *The American statistician*, **43, 3**.

[19] **Kiefer, J.** and **Wolfowitz, J.** (1956). «Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters». *AMS*, **27**, 887–906.

[20] **Le Cam, L.** (1979). «Maximum likelihood: an introduction». *Lecture notes*, **18**.

[21] **Lehmann, E.L.** (1983). *Theory of point Estimation*. Wiley, N.Y.

[22] **Norton, H.W.** (1956). «One likelihood adjustment may be inadequate». *Biometrics*, 79–81.

[23] **Richards, F.S.** (1961). *A method of maximum likelihood estimation*.

[24] **Russell, A.B.** and **Samaniego, F.J.** (1983). «Maximum Likelihood estimation for a discrete multivariate shock model». *JASA*, **78, 382**, 445–448.

[25] **Scholz, F.W.** (1980). «Towards a unified definition of maximum likelihood». *The Canadian Journal of Statistics*, **8, 2**, 193–203.

[26] **Van der Laan, M.** (1995). *Efficient and Inefficient Estimation in Semiparametric Models*. Mathematische Instituut van de Universiteit. Utrecht, 189 p.

[27] **Voinov, V.G.** et **Nikulin, M.S.** (1993). *Unbiased estimators and their applications*. Vol 263, Kluwer Academic Publishers, Dordrecht, Boston, London.

[28] **Weiss L., Wolfowitz J.** (1966). «Generalized maximum likelihood estimators». *Theory of Probability and its Applications*, **11, 1**, 68–93.

[29] **Weiss L., Wolfowitz J.** (1974). *Maximum Probability Estimators and Related Topics*, Lectures Notes in Mathematics, Springer-Verlag.

[30] **Weiss L.** (1986). «A note on small-sample maximum probability estimation». *Statistics and Probabiliy letters*, **4**, 109–111.

ENGLISH SUMMARY

SOME REMARKS ON THE MAXIMUM LIKELIHOOD

CATHERINE HUBER*
Université de Paris V

MIKHAIL NIKULIN**
Université de Bordeaux II

Some paradoxes on the maximum likelihood principle are presented and commented. We consider the properties of the maximum likelihood estimators as a particular case of the $M-$estimators. We propose and unified theory which includes non-dominated models. Several examples are given.

Keywords: $M-$estimators, likelihood paradoxes, unified maximum likelihood theory.

* C. Huber. Université de Paris V, 45 rue des Saints-Pères 75 270 Paris cedex 06

** M. Nikulin. Université de Bordeaux II, 146 rue Léo Saignat 33 800 Bordeaux

– Received July 1995.

– Accepted March 1996.

The maximum likelihood method is both one of the mostly used and one of the most controversial method in statistical estimation. It is intuitively appealing because the likelihood appears to capture the whole available and useful information contained in the data, and it has, too, a theoretical justification, via its excellent asymptotic properties under regularities conditions on the underlying model, which turn out very often to be even optimality properties.

Nevertheless, this method has several which drawbacks have been pointed out by L. Le Cam [1979] and P. J. Huber [1981], but also, among others, by J. Berkson [1980], R. R. Bahadur [1958], D. Basu [1980], T. S. Ferguson [1982], M. Goldstein and J. V. Howard [1991], S. Dharmadhikari and K. Joag-Dev [1985], V. G. Voinov and M. S. Nikulin [1993] among others. As this method is universally applicable, there is a tendeny to use it sometimes without much care, which leads to disastrous results. One of the most famous examples is the one of Le Cam which exhibits a maximum likelihood estimates which is not consistent (as the number n of observations grows to infinity, it converges to twice the value of the parameter it is supposed to estimate, because the number of nuisance parameters involved in the model increases at speed n) [L. Le Cam, 1979]. The likelihood itself is a source of problems when there are several versions of the probability density. Moreover, convexity properties are needed in order to get a global maximum, otherwise it is difficult or even impossible to get an overall maximum. And even though the estimate is unique, in that very favourable case, it happens very often that the likelihood curve is very flat in a neighbourhood of the maximum so that a big change in the estimate results in a very tiny modification of the likelihood, which results in numerical problems and lack of reliability of the evaluation of the estimate.

It is thus necessary, as is recommended B. Efron [1980], to be rather careful when using this method, and, in that case, no major disaster is to be feared of. But, anyway two intrinsic drawbacks are to be taken into account:

1. First, the lack for a definition of a maximum likelihood estimate when there is an ambiguity upon the version of the density which has to be in use, or else when the model is not dominated. F. W. Scholz [1980] tentatively defined a unified version of the maximum likelihood. It seems that it is not well known and that this interesting idea did not spread as much as it should have done.

2. Second, the lack of robutness of the most usual maximum likelihood estimates, those related to the gaussian models. Imbedding those estimates in the more general $M-$estimates [P.J. Huber, 1981], is a way of correcting the lack of stability of certain M-L estimates.

This paper treats of three different topics:

- The properties of the maximum likelihood estimates considered as a special case of M estimates under regularity conditions of the statistical model: consistency, asymptotic normality and sensitivity.

  An M estimate is a solution of equation

  $$\sum_{i=1}^{n} \varphi(x_i, \hat{\theta}) = 0$$

  where $\varphi(x,t)$ can be chosen to be equal to $\ell(x,t) = \frac{\partial Log f(x,t)}{\partial t}$, which is the special case of an M-L estimate regularity properties on φ ensures consistency, asymptotic normality and also robutness through the concept of sensitivity.

- A tentative unified theory of the maximum likelihood based on F.W. Scholz approach.

- A paradox arising from a misure of maximum likelihood is given on a finite probability space.