Optimisation à finalité statistique

Salim Lardjane

Université de Bretagne-Sud

Cours 7 - Applications Statistiques : Maximum de Vraisemblance

Applications statistiques

On va illustrer l'utilisation des algorithmes d'optimisation vus plus haut sur des problèmes statistiques relativement classiques. On introduira de plus de *nouveaux algorithmes* particulièrement utilisés en Statistique.

Maximum de vraisemblance standard

L'estimation par maximum de vraisemblance consiste à estimer un paramètre (ou plusieurs) en maximisant la *fonction de vraisemblance*, celle-ci étant la densité de probabilité conjointe des observations, vue comme fonction du ou des paramètres.

Ainsi, étant donné un échantillon $x_1, \ldots, x_n$ issue de la densité $f(x; \theta)$ dépendant du seul paramètre θ, la fonction de vraisemblance s'écrit

$$\mathcal{L}(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdots f(x_n; \theta)$$

Le *principe* du maximum de vraisemblance implique qu'on adopte comme estimation de θ la valeur $\widehat{\theta}$ qui *maximise* cette vraisemblance.

Maximum de vraisemblance standard

Lorsque la fonction de vraisemblance est différentiable à l'ordre deux en θ, alors $\widehat{\theta}$ est solution de

$$\left\{ \begin{array}{rcl} \frac{\partial \mathcal{L}}{\partial \theta}(\widehat{\theta}) & = & 0 \\ \frac{\partial^2 \mathcal{L}}{\partial \theta^2}(\widehat{\theta}) & < & 0 \end{array} \right.$$

Lorsque les deux conditions précédentes sont vérifiées, $\widehat{\theta}$ correspond à un *maximum local* de la fonction de vraisemblance.

La valeur de $\widehat{\theta}$ correspondant au plus grand de ces maxima est l'estimation requise de θ.

Maximum de vraisemblance standard

En pratique, il est souvent plus simple de travailler avec le logarithme de la fonction de vraisemblance, appelé *log-vraisemblance*, qui prend ses maxima aux même valeurs de θ.

L'estimateur du maximum de vraisemblance possède un certain nombre de bonnes propriétés statistiques comme la convergence, la normalité asymptotique, ...etc.

Dans la suite, on s'intéressera surtout aux *méthodes de calcul* des estimations des paramètres par maximum de vraisemblance.

Maximum de vraisemblance standard

Dans certains cas, l'équation au premier ordre admet une solution *explicite*. Toutefois, le cas le plus fréquent en pratique est celui où elle doit être résolue en ayant recours à un algorithme d'optimisation.

Exemple univarié

Intéressons-nous par exemple à l'estimation des paramètres d'une *loi de Poisson tronquée* (typiquement utilisée en Biologie et en Fiabilité) de densité (fonction de masse)

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!(1-e^{-\theta})}, \quad x = 1, 2, \ldots$$

La log-vraisemblance s'écrit

$$\begin{aligned}\ell(\theta) &= -n\theta + \log(\theta)\sum_i x_i - \sum_i \log(x_i!) \\ &\quad -n\log(1-e^{-\theta})\end{aligned}$$

d'où

$$\frac{d\ell}{d\theta}(\theta) = -n + \frac{\sum_i x_i}{\theta} - \frac{ne^{-\theta}}{1-e^{-\theta}}$$

Exemple univarié

Si on pose $d\ell/d\theta(\theta) = 0$, l'équation *n'admet pas de solution explicite en* θ.

Or on peut déterminer de façon relativement directe la dérivée seconde de ℓ par rapport à θ.

On a aura donc recours préférentiellement à l'*algorithme de Newton-Raphson*. On a de fait

$$\frac{d^2\ell}{d\theta^2}(\theta) = -\frac{\sum_i x_i}{\theta^2} + \frac{ne^{-\theta}}{(1 - e^{-\theta})^2}$$

En combinant les expression des dérivées première et seconde, on peut utiliser la formule de mise à jour de l'algorithme de Newton-Raphson et obtenir une estimation de θ.

Exemple univarié

Exercice 6. Le faire pour les *données groupées* suivantes :

x	f_x
1	1486
2	694
3	195
4	37
5	10
6	1

où, pour chaque valeur de x, f_x est la *fréquence observée*.

Exemple univarié

Comme valeur initiale du paramètre, on pourra prendre la moyenne des observations, qui est de 1.5118.

Sous R, la méthode de Newton-Raphson est implémentée par la fonction **nlm()** (non linear minimization). Comme il s'agit d'une méthode de minimisation, on l'appliquera à $-\ell$.

Plusieurs critères d'arrêt sont autorisés, un code étant retourné par la procédure, indiquant le critère utilisé.

Algorithme de Fisher

Une modification de l'algorithme de Newton-Raphson très utilisée en Statistique est l'*algorithme du score* de Fisher.

Cette méthode a pour formule de mise à jour

$$\theta^{i+1} = \theta^i - \left[\mathbb{E}\left(\frac{d^2\ell}{d\theta^2}(\theta^i)\right)\right]^{-1}\left(\frac{df}{d\theta}(\theta_i)\right)$$

On retrouve essentiellement la formule de mise à jour de Newton-Raphson où le terme $d^2\ell/d\theta^2$ est *remplacé par son espérance*.

Algorithme de Fisher

Dans l'exemple précédent, on a

$$\frac{d^2\ell}{d\theta^2}(\theta) = -\frac{\sum_i x_i}{\theta^2} + \frac{ne^{-\theta}}{(1-e^{-\theta})^2}$$

l'espérance $\mathbb{E}[\frac{d^2\ell}{d\theta^2}(\theta)]$ peut donc être déterminée à partir de $\mathbb{E}(x_i)$. Or,

$$\mathbb{E}(x_i) = \sum_x x \, \frac{e^{-\theta}\theta^x}{x!(1-e^{-\theta})} = \frac{\theta}{1-e^{-\theta}}$$

d'où

$$\mathbb{E}\left(\frac{d^2\ell}{d\theta^2}(\theta)\right) = -\frac{n}{\theta(1-e^{-\theta})} + \frac{ne^{-\theta}}{(1-e^{-\theta})^2}$$

Algorithme de Fisher

L'algorithme du score de Fisher ne semble pas être implémenté sous R mais il est utilisée par exemple par SAS/Insight pour déterminer des estimations par maximum de vraisemblance pour les modèles linéaires généralisés.

De façon générale, on considère que l'algorithme du score de Fisher est plus stable que l'algorithme de newton-Raphson et qu'il est parfois plus rapide.

Algorithme de Fisher

Notons que pour des problèmes *univariés*, tels que celui de l'exemple précédent, on peut également avoir recours à des algorithmes de recherche directe de type Fibonacci, section dorée ou interpolation quadratique.

Une combinaison des deux dernières approches est implémentée sous R par la fonction **optimize()**.

Algorithme de Fisher

Exercice 7. Appliquer cette dernière fonction aux données précédentes.

Exemple multivarié

La méthode du maximum de vraisemblance est également souvent utilisée pour *l'estimation simultanée de paramètres* en Statistique multivariée.

Intéressons-nous par exemple à l'estimation des trois paramètres de la loi gamma de densité

$$f(x;\alpha,\sigma,p) = \frac{1}{\sigma\,\Gamma(p)}\left(\frac{x-\alpha}{\sigma}\right)^{p-1} \text{exp}\left\{-\left(\frac{x-\alpha}{\sigma}\right)\right\}$$

où $\alpha \leq x \leq \infty$, $\sigma > 0$ et $p > 2$ (cette loi intervient notamment en Fiabilité).

Exemple multivarié

La log-vraisemblance correspondante s'écrit

$$\begin{aligned}\ell(\alpha,\sigma,p) &= -np\log(\sigma) - n\log\Gamma(p) \\ &\quad +(p-1)\sum_i \log(x_i-\alpha) \\ &\quad -\sum_i \frac{x_i-\alpha}{\sigma}\end{aligned}$$

Les *équations de vraisemblance* sont obtenues en égalant les dérivées de cette fonction par rapport à α, σ et p à zéro, ce qui donne les conditions du premier ordre :

$$\left\{\begin{array}{rcl}\frac{\partial\ell}{\partial\alpha} &=& -(p-1)\sum_i (x_i-\alpha)^{-1} + \frac{n}{\sigma} = 0 \\ \frac{\partial\ell}{\partial\sigma} &=& -\frac{np}{\sigma} + \sum_i \frac{x_i-\alpha}{\sigma^2} = 0 \\ \frac{\partial\ell}{\partial p} &=& -n\log(\sigma) - n\frac{d}{dp}\log\Gamma(p) + \sum_i \log(x_i-\alpha) \\ &=& 0\end{array}\right.$$

Exemple multivarié

Ces trois équations simultanées *ne peuvent être résolues explicitement* ; on doit donc avoir recours à des algorithmes d'optimisation pour résoudre le problème.

Un problème supplémentaire est que les paramètres sont sujets à des contraintes : celles-ci sont cependant assez simples et peuvent être traitées par une *reparamérisation* de la forme

$$\left\{ \begin{array}{lcl} \sigma & = & u^2 \\ p & = & 2 + v^2 \\ \alpha & = & x_{min} - w^2 \end{array} \right.$$

où x_{min} désigne la valeur minimale de l'échantillon.

Exemple multivarié

On se ramène ainsi à un *problème d'optimisation sans contraintes*.

Exemple multivarié

Exercice 8. Générer cent réalisations issues de la loi précédente pour $\alpha = 5$, $\sigma = 1$ et $p = 6$.

Cela peut être réalisé sous R à l'aide de la fonction

rgamma(100,shape=p,scale=σ)

combinée avec une translation de $\alpha = 5$.

Utiliser ensuite l'algorithme du simplexe de Nelder-Mead pour maximiser la log-vraisemblance en partant des valeurs initiales $\alpha = x_{min} - 0.01^2$, $\sigma = 4$ et $p = 3$, ce qui correspond à $w = 0.01$, $u = 2$ et $v = 1$.

Exemple multivarié

Pour cet algorithme, les dérivées premières et secondes de la fonction de log-vraisemblance *ne sont pas requises*.

On minimise directement moins la log-vraisemblance, ce qui constitue un avantage lorsque les dérivées sont difficiles à calculer, comme c'est le cas ici.