Mesure de l'information - Fisher

La **théorie de l'Information** résulte initialement des travaux de Ronald Aylmer Fisher. Celui-ci, statisticien, définit formellement l'information comme égale à la valeur moyenne du carré de la dérivée du logarithme de la loi de probabilité étudiée.

$$\mathcal{I}(\theta) = \mathrm{E}\left\{ \left[\frac{\partial}{\partial\theta} \ln f(X;\theta) \right]^2 \middle| \theta \right\}$$

À partir de l'inégalité de Cramer, on déduit que la valeur d'une telle *information* est proportionnelle à la faible variabilité des conclusions résultantes. En termes simples, *moins* une observation est probable, *plus* son observation est porteuse d'information. Par exemple, lorsque le journaliste commence le journal télévisé par la phrase "Bonsoir", ce mot, qui présente une forte probabilité, n'apporte que peu d'information. En revanche, si la première phrase est, par exemple "La France a peur", sa faible probabilité fera que l'auditeur apprendra qu'il s'est passé quelque chose, et, partant, sera plus à l'écoute.

Information de Fisher

L'**information de Fisher** est une notion de statistique introduite par R.A. Fisher qui quantifie l'information relative à un paramètre contenue dans une distribution.

Soit $f(x;\theta)$ la distribution de vraisemblance d'une grandeur x (qui peut être multidimensionnelle), paramétrée par θ. La technique d'estimation de θ par le maximum de vraisemblance, introduite par Fisher consiste à choisir la valeur maximisant la vraisemblance des observations X :

$$E\left[\frac{\partial \log f(X;\theta)}{\partial\theta}|\theta\right] = 0$$

.

L'information de Fisher est quant à elle définie comme la variance associée à ce maximum :

$$I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2 |\theta\right]$$

.

Formulation discrète

Les différentes observations x_i nous permettent d'échantillonner la fonction de densité de probabilité $f(x;\theta)$. Selon le théorème de Bayes, en l'absence d'a priori sur θ on a

$P(\theta/X) \propto P(X/\theta)$ Si les observations sont dé-corrélées, la valeur la plus probable $\hat{\theta}$nous est donnée par le maximum de

$$\prod_i P(x_i / \theta),$$

qui est aussi le maximum de

$$\lambda(\theta) = \sum_i \log P(x_i / \theta).$$

Le passage en logarithme permet de transformer le produit en somme, ce qui nous autorise à trouver le maximum par dérivation :

$$\sum_i \left[\frac{\partial}{\partial \theta} \log P(x_i/\theta) \right]_{\theta=\hat{\theta}} = 0.$$

Cette somme correspond pour un nombre d'observations suffisamment élevé à l'espérance mathématique. La résolution de cette équation permet de trouver un estimateur de θ à partir du jeu de paramètre au sens du maximum de vraisemblance. Maintenant, la question est de quantifier la précision de notre estimation. On cherche donc à estimer la forme de la distribution de probabilité de θ autour de la valeur donnée par l'estimateur $\hat{\theta}$. À partir d'un développement limité à l'ordre 2, comme le terme linéaire est nul au maximum, on obtient :

$$\lambda(\theta) = \lambda(\hat{\theta}) - \frac{\theta^2}{2} I(\hat{\theta}) + o(\theta^2)$$

où $I(\hat{\theta})$est l'information de Fisher relative à θ au point de maximum de vraisemblance. Ceci signifie que la distribution est en première approximation une gaussienne de variance $1/I(\hat{\theta})$:

$$P(\theta/X) \propto \exp\left(-\frac{\theta^2}{2} I(\hat{\theta}) \right)$$

Cette variance est appelé la borne de Cramer-Rao et constitue la meilleure précision d'estimation atteignable en absence d'a priori.

Formulation multiparamétrique

Dans le cas où la distribution de probabilité dépend de plusieurs paramètres, θ n'est plus un scalaire mais un vecteur $\vec{\theta} = (\theta_1, \theta_2, \cdots)$. La recherche du maximum de vraisemblance ne se résume donc non pas à une seule équation mais à un système :

$$E\left[\frac{\partial}{\partial\theta_i}\log f(X;\vec{\theta})\right]=0,\qquad \forall i$$

on dérive vis à vis des différentes composantes de $\vec{\theta}$. Enfin, l'information de Fisher n'est plus définie comme une variance scalaire mais comme une matrice de covariance :

$$I(\theta_i,\theta_j)=E\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\vec{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\vec{\theta})\right)\right].$$

Cette matrice est couramment appelée la métrique d'information de Fisher. En effet, le passage de l'espace des observations à l'espace des paramètres est un changement de système de coordonnées. Dans la base des paramètres, avec comme produit scalaire la covariance, cette matrice est la métrique.

L'inverse de cette matrice permet quant à elle de déterminer les bornes de Cramer-Rao, i.e. les covariances relatives aux estimations conjointes des différents paramètres à partir des observations : en effet, le fait que tous les paramètres soient à estimer simultanément rend l'estimation plus difficile. Ce phénomène est une manifestation de ce qui est parfois appelé le « fléau de la dimension ». C'est pour cette raison que l'on utilise quand on le peut des a priori sur les paramètres (méthode d'estimation du maximum a posteriori). Ainsi, on restreint l'incertitude sur chacun des paramètres, ce qui limite l'impact sur l'estimation conjointe.

Information apportée par une statistique [modifier]

De la même façon que l'on à définit l'information de Fisher pour le vecteur des observations X on peut définir l'information de Fisher contenue dans une statistique S(X):

$$I_S(\theta)=\mathbb{E}_\theta\left[(\nabla_\theta\log f_S(S;\theta))\cdot(\nabla_\theta\log f_S(S;\theta))'\right].$$

Cette définition est exactement la même que celle de l'information de Fisher pour X pour un modèle multiparamétrique on remplace juste la densité de X par celle de S(X) la statistique S. Deux théorèmes illustrent l'intérêt de cette notion:

- Pour une statistique exhaustive on a $I_S(\theta)=I(\theta)$ ce qui permet de voir une statistique exhaustive comme une statistique comprenant toute l'information du modèle. L'on a aussi la réciproque à savoir que si $I_S(\theta)=I(\theta)$ alors S est exhaustif bien que cette caractérisation est rarement utilisée dans ce sens la définition grâce au critère de factorisation des statistiques exhaustives étant souvent plus maniable.

- Quelle que soit la statistique S, $I_S(\theta)\leq I(\theta)$ avec un cas d'égalité uniquement pour des statistiques exhaustives. On ne peut donc récupérer plus d'information que celle contenue dans une statistique exhaustive. Ceci explique en grande partie l'intérêt des statistiques exhaustives pour l'estimation. La relation d'ordre est ici la relation d'ordre partielle sur les matrices symétriques à savoir qu'une matrice $A\leq B$ si B-A est une matrice symétrique positive.