Statistique pour la bio-informatique

Séance 6 - 12 Novembre 2003

Maximum de vraisemblance. Algorithmes stochastiques

1 Rappels de statistique : maximum de vraisemblance

1.1 Propriétés des estimateurs

Rappelons que l'un des objectifs de l'estimation statistique est de trouver une grandeur $\hat{\theta}(x_1, \ldots, x_n)$ dépendant des données $x_1, \ldots, x_n$ représentant avec le moins d'erreur possible un paramètre θ généralement présent dans un modèle probabiliste (par exemple, le taux d'une variable de loi exponentielle). La grandeur en question s'appelle une statistique, et la variable $\hat{\theta}$ obtenue en considérant le caractère aléatoire des données s'appelle un estimateur. Nous avons la relation

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

où les X_i ne sont plus des données mais des variables aléatoires tirées selon le modèle de référence dont on cherche à estimer le paramètre. Dans cette optique, un estimateur sans biais vérifie la relation

$$E[\hat{\theta}] = \theta\,.$$

Étant donné un modèle probabiliste aussi appelé une *structure statistique* dépendant du paramètre θ et un jeu de données $x_1, \ldots, x_n$, on appelle **vraisemblance** la fonction suivante

$$L(\theta) = \mathrm{P}(x_1, \ldots, x_n; \theta)$$

Cette notation abusive correspond à la loi des variables modélisées, lorsque le paramètre est θ. Par exemple, lorsque les données sont discrètes, la vraisemblance correspond à

$$L(\theta) = \mathrm{P}(X_1 = x_1; \ldots; X_n = x_n\,;\,\theta).$$

Lorsque les données sont à valeurs réelles, le vecteur $(X_1, \ldots, X_n)$ admet souvent une densité

$$L(\theta) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta).$$

La statistique de maximum de vraisemblance est

$$\hat{\theta}(x_1, \ldots, x_n) = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)$$

Cette statistique est très souvent utilisée pour l'estimation du fait de ses bonnes propriétés asymptotiques. En fait, on démontre dans de nombreux contextes que l'estimateur

- $\hat{\theta}$ est asymptotiquement **sans biais**

$$\lim_{n\to\infty} E[\hat{\theta}] = \theta\,.$$

- $\hat{\theta}$ est asymptotiquement **gaussien**

$$\frac{\hat{\theta} - E[\hat{\theta}]}{\sqrt{Var(\hat{\theta})}} \to \mathcal{N}(0,1), \quad \text{en loi.}$$

De plus, pour n inifini, il est de variance minimale parmi tous les estimateurs sans biais de θ (et d'erreur minimale parmi tous les autres) et atteint la borne de Cramer-Rao

$$Var(\hat{\theta}) \to \frac{1}{I(\theta)}$$

ou

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log L(\theta)\right)^2\right]$$

(Revoir les cours de stats math, ou de réseaux de neurones). Par ailleurs, l'estimateur de maximum de vraisemblance est souvent intuitif (relié au moindres carrés dans le cas gaussien, à des estimateurs de moments). Toutes ces propriétés font qu'il est très souvent utilisé.

Exercice. Calculer la statistique de maximum de vraisemblance du paramètre m dans le cas où les observations sont i.i.d. de loi $\mathcal{N}(m, \sigma^2)$. Calculer la borne de Cramer-Rao.

1.2 Données manquantes : l'algorithme EM

L'algorithme EM proposé par Dempster et ses collègues apparait dans de nombreuses situations en bio-informatique, dans des variantes qui n'indiquent pas toujours son origine. Il est très utile pour l'estimation de maximum de vraisemblance (maximum local) lorsqu'une partie des données est manquante. Les données manquantes correspondent à des variables que l'on ne peut pas observer. C'est le cas par exemple des séquences des ancêtres lorsque l'on tente de reconstruire l'évolution à partir de données moléculaires actuelles. Dans ce cas l'algorithme d'estimation s'appelle *algorithme de Felsenstein*. Parfois, les données manquantes correspondent à des données fictives, explicatives et introduites dans un soucis de pertinence de modélisation. C'est par exemple, le cas des algorithmes d'estimation dans les chaînes de Markov cachées (*algorithme de Baum-Welch*). Cette section propose donc un cadre unificateur pour plusieurs approches très classiques en bio-informatique.

On dispose d'un modèle probabiliste dont θ est le paramètre. La vraisemblance s'exprime de la manière suivante

$$L^c(\theta) = p(x, y; \theta).$$

Dans cette expression, $x = x_1, \ldots, x_n$ correspond à l'ensemble des données observées, dites aussi **incomplètes**, et $y = y_1, \ldots, y_m$ l'ensemble des données manquantes. Le couple (x, y) s'appelle les données **complètes**. La vraisemblance L^c est la **vraisemblance complète**. Afin de faire apparaître les données, nous noterons

$$L(\theta|\ x, y) = p(x, y; \theta)$$

et la **vraisemblance incomplète**

$$L(\theta|\ x) = L(\theta) = p(x; \theta) = \int p(x, y; \theta) dy.$$

De même, les log-vraisemblances seront notées

$$\ell(\theta|\ x, y) = \log p(x, y; \theta)$$

et

$$\ell(\theta|\ x) = \log p(x; \theta) = \log \int p(x, y; \theta) dy.$$

Proposition 1.1 *Soit θ_0 un paramètre quelconque. Nous avons*

$$\ell(\theta|\ x) = E[\ell(\theta|\ x, Y)|\ X = x; \theta_0] - E[\log p(y|\ x; \theta)|\ X = x; \theta_0]$$

Démonstration. Observons que

$$\ell(\theta|\ x) = \ell(\theta|\ x, y) - \log p(y|\ x; \theta)$$

et sachant $X = x$, nous pouvons prendre l'espérance conditionnelle des deux membres pour un θ_0 queconque.

Il y a bien deux valeurs différentes de θ dans l'équation précédente. L'une sert pour l'intégration (c'est θ_0) et l'autre est l'argument à optimiser. L'idée de l'algorithme EM est d'optimiser le premier terme de $\ell(\theta|\ x)$, et de compenser l'erreur faite de cette manière par une procédure de **ré-estimation**. Posons donc

$$Q(\theta;\theta_0) = E[\ell(\theta|\ x, Y) \mid X = x; \theta_0]$$

Le calcul de $Q(\theta;\theta_0)$ consiste en l'**étape E** de l'algorithme EM (E pour Expectation). Nous avons

$$Q(\theta;\theta_0) = \int \ell(\theta|\ x, y) p(y|\ x; \theta_0) dy$$

L'**étape M** consiste à maximiser $Q(\theta;\theta_0)$ en θ (M pour Maximization). Nous avons

$$\theta_1 = \mathrm{argmax}_\theta Q(\theta;\theta_0).$$

L'algorithme EM consiste à calculer de manière itérative

$$\theta_{n+1} = \mathrm{argmax}_\theta Q(\theta;\theta_n).$$

Le résultat important est la convergence de l'algorithme vers un maximum local de la vraisemblance.

Proposition 1.2 *La suite (θ_n) converge vers un maximum local de $\ell(\theta|\ x)$.*

Nous admettons en premier lieu le lemme suivant, dont la démonstration repose sur une inégalité de convexité (inégalité de Jensen).

Lemme 1.1 *Soient p, q deux densités de probabilité. L'entropie de Kullback de p et q est toujours positive*

$$K(p,q) = -\int \log\left(\frac{q(x)}{p(x)}\right) p(x) dx \geq 0$$

Démonstration. Par définition, nous avons $Q(\theta_{n+1},\theta_n) \geq Q(\theta_n;\theta_n)$. En prenant θ_n comme valeur pour θ_0, nous obtenons

$$\ell(\theta_{n+1}|\ x) - \ell(\theta_n|\ x) = Q(\theta_{n+1},\theta_n) - Q(\theta_n;\theta_n) + K_{n+1}$$

où

$$K_{n+1} = E\left[\log\left(\frac{p(Y|\ \theta_{n+1}, x)}{p(Y|\ \theta_n, x)}\right) \mid x; \theta_n\right]$$

Cette dernière grandeur est toujours positive, car elle correspond à l'entropie de Kullback des lois $p = p(\ \mid \theta_{n+1}, x)$ et $q = p(\ \mid \theta_n, x)$.

Un exercice : les données censurées.

On considère un échantillon de loi normale $\mathcal{E}(\theta)$ où σ^2. On suppose que l'échantillon est de taille $m + n$ et que m données ont été censurées. De ces données, la seule information conservée est qu'elles sont plus grandes qu'une constante c. Écrire les vraisemblances incomplètes et complètes. Décrire la procédure EM.

1.3 Statistique bayésienne. Méthodes MCMC

Dans cette section, nous faisons une brève présentation des méthodes d'estimation bayésiennes, en illustrant ces méthodes dans le cas de l'estimation d'une fréquence. Elles sont maintenant très utilisées en génétique, dans le contexte du coalescent de Kingman. Cette section peut être lue de manière séparée du texte.

En statistique bayésienne, on considère qu'une hypothèse donnée H possède une probabilité *a priori*, c'est à dire ne tenant pas compte des données. Cette notion, essentiellement liée à la subjectivité de l'hypothèse, est le point d'entrée principale des critiques adressées aux "bayésiens". Toutefois les méthodes bayésiennes sont particulièrement prisées lorsque l'on dispose d'informations à priori par exemple sous la forme d'un modèle bien établi, ou que le nombre de données est insuffisant pour une approche fréquentiste.

Notons $\mathrm{P}(H)$ la probabilité a priori de l'hypothèse H. Etant donné un nombre n d'observations rassemblées sous la notations D, la formule de Bayes montre que la probabilité de H est modifiée de la façon suivante

$$\mathrm{P}(H/D) = \frac{\mathrm{P}(H)\mathrm{P}(H/D)}{\mathrm{P}(D)}.$$

Cette probabilité est appelée *probabilité a posteriori* de l'hypothèse H.

Lorsque l'on doit choisir entre plusieurs hypothèses H_0, H_1, ..., H_r, la formule devient

$$\forall i = 0, \ldots, r, \quad \mathrm{P}(H_i/D) = \frac{\mathrm{P}(H_i)\mathrm{P}(H_i/D)}{\sum_{j=0}^{r} \mathrm{P}(H_j)\mathrm{P}(H_j/D)}.$$

Une stratégie raisonnable de décision est de choisir l'hypothèse dont la probabilité a posteriori est maximale (règle de Bayes). Notons que cela revient à maximiser le numérateur de la fraction précédente sans se préoccuper du dénominateur.

Dans cette démarche, il est tout à fait possible d'intégrer un continuum d'hypothèses en supposant par exemple que la loi de H admet une densité $p(h)$. Alors, nous parlerons de loi à posteriori

$$p(h/D) = \frac{p(h)p(D/h)}{\int p(D/h)p(h)dh}.$$

Dans ce cas, le maximum a posteriori h^* sera un estimateur naturel du paramètre h une fois observées les données. Toutefois les bayésiens préconisent souvent l'estimateur de moment

$$h^* = \int hp(h/D)dh.$$

Illustrons cette démarche avec un exemple classique. On dispose de n données iid de loi de Bernoulli de paramètre p et l'on cherche à estimer ce paramètre. Le fréquentiste estimera

$$\hat{p} = x/n$$

où x est le nombre de réalisations observées (de loi binomiale $\mathcal{B}(n,p)$). Dans la mesure où cet estimateur correspond à l'estimateur de maximum de vraisemblance, bien lui en prend lorsque le nombre de données est très grand.

Le bayésien formulera une hypothèse a priori sur p. Par exemple, la distribution de ce paramètre peut être la loi de Dirichlet de densité

$$\forall p \in (0,1), \quad f(p) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}(p - p^2)^{\alpha-1}$$

Ce choix est usuel, non pour des raisons de modélisation (biologique ou autre), mais parcequ'il permet le calcul de la loi a posteriori

$$f(p/x) = \frac{\Gamma(2\alpha + n)}{\Gamma(\alpha + x)\Gamma(\alpha + n - x)}(1-p)^{\alpha+n-x-1}p^{\alpha+x-1}$$

L'estimateur de maximum de vraisemblance a posteriori est

$$\hat{p} = \frac{\alpha + x - 1}{2\alpha + n - 2}$$

On voit donc que l'estimation de p est pour le bayésien un compromis entre ses a priori et les observations. Ainsi supposons que l'on observe 6 réalisations d'un événement dans une série de 20 répétitions d'une épreuve. Le fréquentiste estimera

$$\hat{p} = 0.3$$

Si $\alpha = 10$, le bayésien estimera

$$\hat{p} = 15/38 \approx 0.39$$

Cette estimation est plus proche de la moyenne de la loi de Dirichlet qui est égale à 0.5. Pour 600 réalisations sur un total de 20000 répétitions, les deux estimations seront quasiment identiques car dominées par l'information contenue dans les données (et non dans le modèle de Dirichlet).

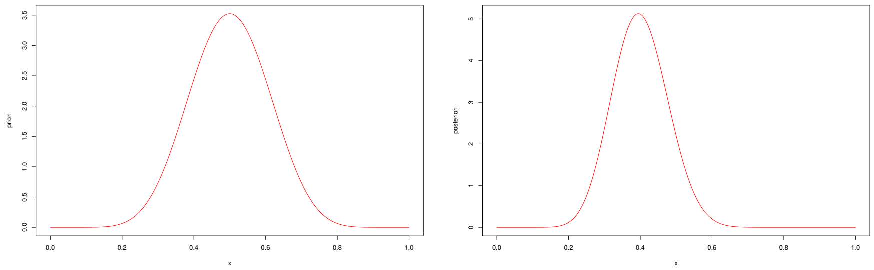

FIG. 1 – Distributions a priori (loi de Dirichlet) et a posteriori pour l'observation $x = 6$ dans une série de 20 répétitions.

La situation se complique pour le bayésien lorsqu'il va choisir le paramètre α. Dans ce cas, sa philosophie lui dicte une nouvelle loi a priori (hyper-prior). Considèrons par exemple la loi de Pascal de paramètre $\lambda = 1/2$. Nous avons alors

$$\forall p \in (0,1),\ \alpha \geq 0, \quad f(p,\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}(p-p^2)^{\alpha-1}\lambda^\alpha(1-\lambda)$$

Dans ce cas les estimateurs a posteriori deviennent plus difficile à calculer.

Lorsque calculer la loi a posteriori devient trop complexe, une stratégie consiste à simuler cette loi pour obtenir les estimateurs. Pour cela, une methode très utilisée est la simulation d'une chaîne de Markov dont la loi stationnaire est la loi cherchée. Cette technique s'appelle méthode de Monte Carlo par chaîne de Markov (MCMC). De telles chaînes ne sont pas difficiles à construire. Elles ne nécessitent souvent que le calcul du numérateur de la loi en question, évitant le (peut être) délicat calcul d'une intégrale au dénominateur

$$f(p,\alpha/x) \propto \lambda^\alpha \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} p^{x+\alpha-1}(1-p)^{n-x+\alpha-1}$$

L'algorithme est itératif. Dans une démarche typique, on part de paramètres initiaux $p = p_0$ et $\alpha = \alpha_0$ arbitraires, puis on fait une proposition pour changer les paramètres

$$p \to p' \quad \alpha \to \alpha'.$$

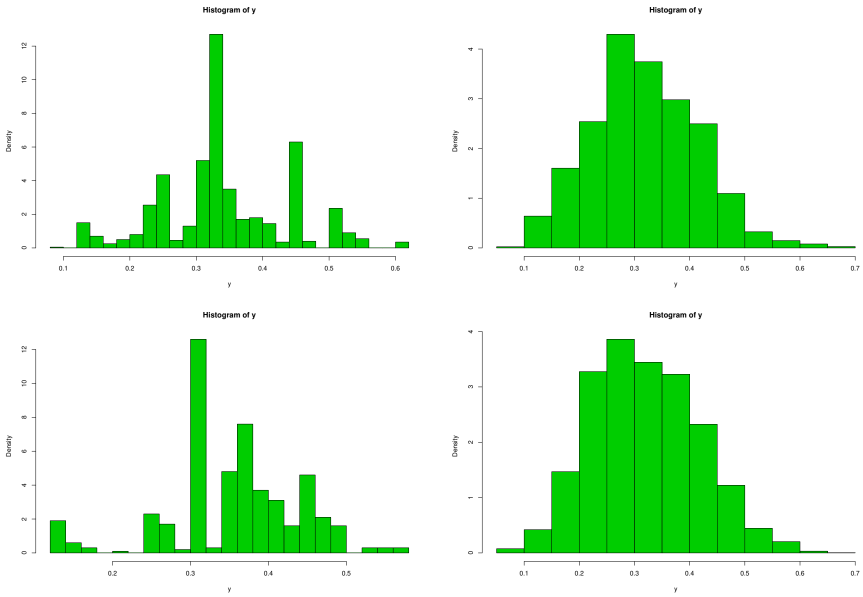

FIG. 2 – *De gauche à droite : (1) Equilibre non atteint ; (2) Equilibre atteint ; (3) Mauvais échantillonnage ; (4) changement de l'hyper-prior (peu d'effet).*

Les nouveaux paramètres sont pris par exemple uniformément dans un voisinage des paramètres courants. On calcule alors la probabilité de rejet des paramètres proposés

$$R = 1 - \min(1, \frac{f(p', \alpha'/x)}{f(p, \alpha/x)}).$$

Puis on réitère le processus jusqu'à ce que l'on juge que le régime stationnaire est établi. La chaîne de Markov ainsi créée est réversible, elle admet la loi souhaitée comme loi stationnaire.

Toutefois, des conditions classiques d'ergodicité (difficiles à vérifier au cas pas cas) sont en général nécessaires pour garantir la convergence de la méthode MCMC (Robert et Casella, 1999). De plus, il y a plusieurs difficultés liées à l'implémentation de la chaîne qui peuvent rendre difficile l'interprétation des résultats (Figure 2).

- La vitesse d'atteinte de l'équilibre est inconnue. Pour des problèmes complexes, des heures de calcul sont parfois nécessaires (*burn-in period*).
- Obtenir un échantillon iid de la loi stationnaire est un autre problème difficile à résoudre (*sampling period*). Les variables issues de la simulation d'une chaîne de Markov sont généralement corrélées.
- Le choix de la stratégie d'exploration de l'ensemble des paramètres influence fortement la vitesse de convergence. Il s'agit sans doute du problème le plus crucial

dans l'utilisation des algorithmes MCMC. L'art du progammeur de méthodes MCMC réside dans la définition d'un noyau de proposition bien adapté aux contraintes du problème qu'il souhaite traiter.

Exercice. Programmer la méthode précédente d'estimation d'une fréquence et estimer le temps de convergence numériquement pour différents tirages de loi binomiale.