Estimation pour le modèle M0

Estimateur du maximum de vraisemblance (EMV)

La vraisemblance de la suite d'observations $x_{1:n}$ sous le modèle $\mathbb{P}^\vartheta$ est

$$V(\vartheta) = \mathbb{P}^\vartheta(X_{1:n} = x_{1:n}).$$

L'estimateur du maximum de vraisemblance consiste à choisir la valeur de ϑ qui maximise $V(\vartheta)$, soit

$$\widehat{\vartheta} \leftarrow \max_\vartheta V(\vartheta).$$

Si la classe est M0, ϑ correspond aux poids $p = (p(x))_{x\in\mathcal{A}}$ et on peut tout calculer.

L'estimateur du maximum de vraisemblance de $(p(x))_{x\in\mathcal{A}}$ dans le modèle M0 pour la séquence $x_{1:n}$ est donné par

$$\widehat{p}_n(x) = \frac{N_n(x)}{n}, \quad x \in \mathcal{A}.$$

Conséquence : l'EMV est consistant. Quand $n \to \infty$,

$$\widehat{p}_n(x) \to p(x).$$

Deux prolongements :

- Mesurer la taille de l'erreur $|\widehat{p}_n(x) - p(x)|$.
- Généraliser ce résultat aux mots.

Taille de l'erreur pour le modèle M0

Résultat théorique : le théorème central limite
Outil : la variance (déjà vue)

Rappel :

$$\mathbb{E}(N_n(x)) = np(x), \quad \operatorname{var}(N_n(x)) = np(x)(1-p(x))$$

Donc

$$\mathbb{E}(R_n(x)) = p(x), \quad \operatorname{var}(R_n(x)) = p(x)\,(1-p(x))/n$$

Théorème central limite Quand n est grand, $R_n(x)$ ressemble à une variable aléatoire gaussienne de même moyenne et de même variance :

$$R_n(x) \approx \mathcal{N}(m_x, \sigma_x^2/n), \quad m_x := p(x),\ \sigma_x^2 := p(x)\,(1-p(x)).$$

Par exemple :

$$\mathbb{P}(R_n(x) \geqslant t) \approx \mathbb{P}(\mathcal{N}(m_x, \sigma_x^2/n) \geqslant t).$$

Rappel : $\mathcal{N}(m, \sigma^2) = m + \sigma\mathcal{N}(0,1)$. Et

$$\mathbb{P}(a \leqslant \mathcal{N}(0,1) \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x$$

Donc

$$\boxed{\mathbb{P}\left(R_n(x) \geqslant m_x + t\frac{\sigma_x}{\sqrt{n}}\right) \approx \mathbb{P}(\mathcal{N}(0,1) \geqslant t)}$$

Idem pour « inférieur ou égal »

Table de quantiles de la loi normale centrée réduite

$$\mathbb{P}(m - \sigma \leqslant \mathcal{N}(m, \sigma) \leqslant m + \sigma) = 68.26\%$$

$$\mathbb{P}(m - 2\sigma \leqslant \mathcal{N}(m, \sigma) \leqslant m + 2\sigma) = 95.44\%$$

$$\mathbb{P}(m - 3\sigma \leqslant \mathcal{N}(m, \sigma) \leqslant m + 3\sigma) = 99.74\%$$

À retenir :

Plus de ± 2 fois l'écart-type : 5%
Plus de ± 3 fois l'écart-type : 1%
Et pour M0, l'écart-type est en $1/\sqrt{n}$

Approximation : pour t positif pas trop petit,

$$\mathbb{P}(\mathcal{N}(0, 1) \geqslant t) \approx \exp(-t^2/2)$$

Donc, pour $t > p(x)$ et t pas trop proche de $p(x)$,

$$\mathbb{P}(R_n(x) \geqslant t) \approx \exp\left[-\frac{n(t - p(x))^2}{2p(x)(1 - p(x))}\right]$$

Idem pour $t < p(x)$

Fréquences des mots

L'ensemble de tous les mots $\mathcal{A}^*$ est la réunion des $\mathcal{A}^n$ pour tout $n \geqslant 1$. Pour des raisons techniques, on se donne aussi un mot vide noté $*$.

La longueur d'un mot $\mathtt{w} \in \mathcal{A}^*$ est $|\mathtt{w}| = n$ si $\mathtt{w} \in \mathcal{A}^n$. La longueur du mot vide est $|*| = 0$.

La loi des grands nombres ci-dessus affirme que, dans le modèle M0, $R_n(\mathtt{w}) \to p(\mathtt{w})$ pour tout mot $\mathtt{w}$ de longueur 1. En fait :

> Pour toute longueur $L \geqslant 1$ et tout mot $\mathtt{w} \in \mathcal{A}^*$ de longueur L, dans le modèle M0, quand n devient grand,
> $$R_n(\mathtt{w}) \to p(\mathtt{w}) \text{ avec } p(\mathtt{w}) = p(w_1) \cdots p(w_L).$$

La preuve de ce résultat est omise : sur le principe, on reprend la preuve du cas d'une lettre, donc on montre que

1) $\mathbb{E}(N_n(\mathtt{w})) = n\, p(\mathtt{w})$,
2) $\mathrm{var}(N_n(\mathtt{w}))$ se comporte comme un multiple de n,
3) on conclut par Bienaymé-Tchebychev.

Comme pour les lettres, on peut quantifier :

> $\mathrm{var}(N_n(\mathtt{w})) \sim n\sigma_{\mathtt{w}}^2$ et $\sigma_{\mathtt{w}}^2$ est calculable
> $R_n(\mathtt{w}) \approx p(\mathtt{w}) + \mathcal{N}(0,1)\sigma_{\mathtt{w}}/\sqrt{n}$

$$\mathbb{P}(N_n(\mathtt{w}) \geqslant np(\mathtt{w}) + 2\sigma_{\mathtt{w}}\sqrt{n}) \approx 2.28\%$$

Exemples : $\mathtt{w} = AA$, $\mathtt{w} = AT$. Calculer $\sigma_{\mathtt{w}}^2$.

Validation ou rejet de M0

On compte les lettres, on en déduit des valeurs plausibles de $p(x)$: on choisit le modèle M0 du maximum de vraisemblance. Mais on voudrait rejeter une séquence

$$AACCGGTTAACCTTGGCCAAAATTGG\ldots$$

Sou M0, le fréquence d'un mot $\mathtt{w} = w_1w_2$ doit vérifier

$$R_n(\mathtt{w}) \approx p(\mathtt{w}) = p(w_1)\,p(w_2), \quad R_n(w_1) \approx p(w_1), \quad R_n(w_2) \approx p(w_2),$$

donc

$$\frac{N_n(\mathtt{w})}{n} = R_n(\mathtt{w}) \approx R_n(w_1)\,R_n(w_2) = \frac{N_n(w_1)\,N_n(w_2)}{n^2}.$$

$$\boxed{\begin{array}{l} R_n(w_1w_2) \approx R_n(w_1)\,R_n(w_2) \\ \text{avec une erreur de l'ordre de } 1/\sqrt{n} \end{array}}$$

A contrario, si $n\,N_n(w_1w_2)$ est très différent de $N_n(w_1)\,N_n(w_2)$, le modèle M0 n'est pas pertinent !

Exemple : ADN d'*E. coli*. (Voir transparent.)

Que faire ? Réponse : les chaînes de Markov (à suivre).

Le modèle M1

À présent, les positions successives X_n ne sont plus indépendantes. On commence par le cas le plus simple : la distribution de X_n est influencée par la valeur de X_{n-1}. Les biologistes parlent de modèle M1, les mathématiciens de chaînes de Markov.

Définition Si $\mathbb{P}(B_2) \neq 0$, la probabilité conditionnelle de B_1 sachant B_2 est

$$\mathbb{P}(B_1|B_2) = \frac{\mathbb{P}(B_1 \cap B_2)}{\mathbb{P}(B_2)}.$$

Commentaire Cela correspond à l'intuition : considérons le cas où $B_1 =$ je suis en retard en cours, $B_2 =$ il neige. On peut penser que, si B_2, la circulation dans l'agglomération grenobloise devient plus difficile, donc B_1 a plus de chances d'être réalisé. Il vaudrait mieux évaluer B_1 par une probabilité éventuellement différente de $\mathbb{P}(B_1)$, qui rende compte du fait que B_2 est réalisé (il neige) : cette nouvelle valeur, c'est $\mathbb{P}(B_1|B_2)$.

Définition

La suite $X_{1:n}$ est une chaîne de Markov si, pour tout $1 \leqslant k \leqslant n-1$ et tout $x_{1:k+1}$,
$\mathbb{P}(X_{k+1} = x_{k+1}|X_{1:k} = x_{1:k}) = \mathbb{P}(X_{k+1} = x_{k+1}|X_k = x_k)$.

On parle aussi de mémoire à distance 1 : si on s'intéresse à la position $k+1$, on peut oublier les valeurs aux positions $1 : k-1$ et ne garder que la position k.

Un calcul facile montre alors que

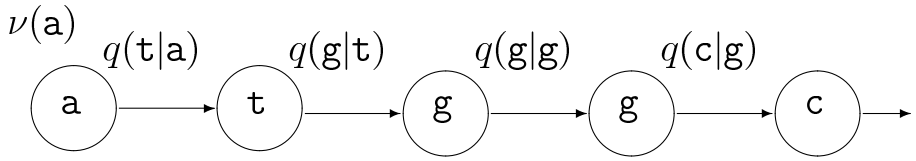
$$\boxed{\begin{aligned}&\mathbb{P}(X_{1:k} = x_{1:k}) = \mathbb{P}(X_1 = x_1)\times \\ &\\ &\times \mathbb{P}(X_2 = x_2|X_1 = x_1) \times \cdots \times \mathbb{P}(X_k = x_k|X_{k-1} = x_{k-1}).\end{aligned}}$$

On voit que la loi d'une chaîne de Markov (stationnaire, si on veut préciser) est décrite complètement dès qu'on connaît $\mathbb{P}(X_1 = x)$ pour tout $x \in \mathcal{A}$ et $\mathbb{P}(X_k = x'|X_{k-1} = x)$ pour tout $(x, x') \in \mathcal{A} \times \mathcal{A}$. On note

$$\boxed{\nu(x) = \mathbb{P}(X_1 = x), \quad q(x'|x) = \mathbb{P}(X_k = x'|X_{k-1} = x).}$$

On note aussi $q(x, x') = q(x'|x)$ (attention : l'ordre de x et x' change !).

Le modèle M1



Par exemple,

$$\mathbb{P}(X_{1:5} = \mathtt{atggc}) = \nu(\mathtt{a})\, q(\mathtt{t}|\mathtt{a})\, q(\mathtt{g}|\mathtt{t})\, q(\mathtt{g}|\mathtt{g})\, q(\mathtt{c}|\mathtt{g}).$$

Paramètres de la chaîne de Markov :

- Loi initiale ν : $\nu(x) \geqslant 0$ pour tout $x \in \mathcal{A}$ et $\sum_{x \in \mathcal{A}} \nu(x) = 1$.
- Matrice de transition q : $q(x, x') \geqslant 0$ pour tous x et $x' \in \mathcal{A}$ et, pour tout $x \in \mathcal{A}$,

$$\sum_{x' \in \mathcal{A}} q(x, x') = 1.$$

Donc $0 \leqslant \nu(x) \leqslant 1$ et $0 \leqslant q(x, x') \leqslant 1$.

$$\nu = (\nu(\mathtt{a}), \nu(\mathtt{c}), \nu(\mathtt{g}), \nu(\mathtt{t})),$$

et, dans l'ordre $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$,

$$q = \begin{pmatrix} q(\mathtt{a},\mathtt{a}) & q(\mathtt{a},\mathtt{c}) & q(\mathtt{a},\mathtt{g}) & q(\mathtt{a},\mathtt{t}) \\ q(\mathtt{c},\mathtt{a}) & q(\mathtt{c},\mathtt{c}) & q(\mathtt{c},\mathtt{g}) & q(\mathtt{c},\mathtt{t}) \\ q(\mathtt{g},\mathtt{a}) & q(\mathtt{g},\mathtt{c}) & q(\mathtt{g},\mathtt{g}) & q(\mathtt{g},\mathtt{t}) \\ q(\mathtt{t},\mathtt{a}) & q(\mathtt{t},\mathtt{c}) & q(\mathtt{t},\mathtt{g}) & q(\mathtt{t},\mathtt{t}) \end{pmatrix}.$$

Une chaîne de Markov (un modèle M1) est un processus sans mémoire (autre que celle de sa valeur actuelle). Rappel :

$$\mathbb{P}(X_{n+1} = x | X_{1:n} = x_{1:n}) = \mathbb{P}(X_{n+1} = x | X_n = x_n).$$

Cas ADN : M1 donne une meilleure approximation de la réalité que le modèle indépendant (M0). Mais en fait, bien sûr, les dépendances sont encore plus complexes.

On utilisera très vite des dépendances à m pas (modèles Mm). Le principe reste le même donc on va décrire M1.

Quelques remarques

• On peut utiliser les modèles pour une séquence génomique donnée. Alors $q(x, x')$ donne la probabilité que le site $n+1$ soit occupé par un x' sachant que le site n soit occupé par un x, i.e. n est un indice **spatial**.

On peut aussi utiliser n comme un indice **temporel.**

Donc X_n est le nucléotide en un site donné après n réplications de la molécule d'ADN et (par exemple) les sites évoluent indépendamment les uns des autres.

On peut penser qu'il y a eu beaucoup de réplications donc on s'intéressera à la distribution quand n devient grand.

Deux exemples classiques de modèles M1 d'évolution :

On s'intéresse à un site fixé et on suppose qu'il évolue indépendamment du reste de la séquence (ce qui est tout à fait faux, biologiquement ! !).

Jukes-Cantor Pour tous $x \neq x'$, $q_{\mathrm{JC}}(x, x') = p$ avec $0 \leqslant p \leqslant \frac{1}{3}$.

$$q_{\mathrm{JC}} = \begin{pmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{pmatrix}.$$

Le paramètre p dépend de l'échelle de temps considérée (voir plus loin).

Kimura Purines $\mathtt{a}, \mathtt{g}$ vs. pyrimidines $\mathtt{c}, \mathtt{t}$.

Pour chaque transition, probabilité u. Pour chaque transversion, probabilité v. Donc $0 \leqslant u + 2v \leqslant 1$. Dans l'ordre $\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}$,

$$q_{\mathrm{K}} = \begin{pmatrix} 1-u-2v & v & u & v \\ v & 1-u-2v & v & u \\ u & v & 1-u-2v & v \\ v & u & v & 1-u-2v \end{pmatrix}.$$

Même remarque que pour J-C.

Hasegawa, Kishino, Yano, autres...

Description du modèle M1

Lois = calcul matriciel !

Théorème
Dans un modèle M1, la distribution après n pas vaut νq^n.

Preuve :

$$\begin{aligned}\mathbb{P}(X_{n+1}=x) &= \sum_{x_{1:n}\in\mathcal{A}^n} \mathbb{P}(X_{1:n+1}=x_{1:n}x) \\ &= \sum_{x_{1:n}\in\mathcal{A}^n} \nu(x_1)\prod_{i=2}^{n} q(x_{i-1},x_i)\, q(x_n,x) \\ &= (\nu q^n)(x).\end{aligned}$$

Rappel : $(MN)(x,y) = \sum_z M(x,z)N(z,y)$.

Par exemple,

$$\mathbb{P}(X_4=\mathtt{g}|X_1=\mathtt{a}) = \sum_{x=\mathtt{a}}^{\mathtt{t}}\sum_{z=\mathtt{a}}^{\mathtt{t}} q(\mathtt{a},x)\, q(x,z)\, q(z,\mathtt{g}).$$

Problème : comment calculer la distribution de X_{101} ?
Additionner $4^{100} = 2^{200} \approx 10^{60}$ termes ? ? ?

Premier principe des processus M1
Convergence vers un équilibre (stochastique).

C'est-à-dire :

(1) Les νq^n varient avec n, on sent l'effet de l'âge.

(2) Chaque νq^n dépend de ν, on se souvient de son état initial.

(3) Mais tout ceci disparaît quand n devient grand, on finit par tout oublier. Convergence vers l'équilibre : si n est grand,

$$\mathbb{P}_\nu(X_n = x) \approx \pi(x).$$

Remarque : $\pi(x)$ est indépendant de n et de ν.

Que vaut π ? C'est un exemple de distribution stationnaire.

Distribution stationnaire : $\mu q = \mu$.

Si on converge, c'est vers une distribution stationnaire.

Le résultat

• Hypothèses techniques : chaîne de Markov finie, irréductible, apériodique.

• Alors la distribution stationnaire π existe et est unique et pour toute loi initiale ν,

$$\mathbb{P}_\nu(X_n = x) \to_{n\to\infty} \pi(x).$$

L'hypothèse d'apériodicité est satisfaite dès que $q(x,x) \neq 0$ pour au moins un $x \in \mathcal{A}$.

Remarque : Si on part de π, la loi de X_n est π pour tout n, donc calculs faciles.

Mais attention ! (X_n) n'est pas i.i.d. de loi π (voir ci-dessous).

Exemple : Jukes-Cantor

Loi après n pas ? Réponse : $(q_{\mathrm{JC}})^n$ est de type Jukes-Cantor, pour le paramètre

$$p_n = \tfrac{1}{4}\,(1-(1-4p)^n). \qquad (*)$$

Comme $p_n \to \frac{1}{4}$, on sait que $(q_{\mathrm{JC}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix}$ et

$$\nu(q_{\mathrm{JC}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix} = \pi.$$

Remarque : en fait, on n'avait pas besoin de faire le calcul de q^n, il suffisait de résoudre $\pi\, q = \pi$, ce qui, ici, est trivial.

En tous cas : si n est grand, $\mathbb{P}_\nu(X_n = x) \approx \pi(x) = .25$ pour tout $x \in \mathcal{A}$ et tout ν.

Caveat !

Sous le modèle M0, on aurait

$$\mathbb{P}(Y_n = x,\, Y_{n+1} = x') = \pi(x)\,\pi(x').$$

Ici, même pour n grand,

$$\mathbb{P}(X_n = x,\, X_{n+1} = x') \approx \pi(x)\, q(x, x').$$

Par exemple $\mathbb{P}(X_n = \mathtt{a},\, X_{n+1} = \mathtt{a}) \approx \frac{1}{4}\,(1-3p) \neq \frac{1}{4}\,\frac{1}{4}$.
Alors que $\mathbb{P}(X_n = \mathtt{a}) \approx \frac{1}{4}$ et $\mathbb{P}(X_{n+1} = \mathtt{a}) \approx \frac{1}{4}$.

Exemple : Kimura

Loi après n pas : encore de type Kimura, un peu compliqué.

Par contre, il est facile de savoir que

$$(q_{\mathrm{K}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 \end{pmatrix} = \pi,$$

et $\nu(q_{\mathrm{K}})^n \approx \begin{pmatrix} .25 & .25 & .25 & .25 \end{pmatrix}$.

Donc $(q_{\mathrm{JC}})^n \approx (q_{\mathrm{K}})^n$ mais (encore une fois), par exemple,

$$\mathbb{P}_{\mathrm{JC}}(X_n = \mathsf{a},\, X_{n+1} = \mathsf{a}) \approx \tfrac{1}{4}\,(1 - 3p)$$

et

$$\mathbb{P}_{\mathrm{K}}(X_n = \mathsf{a},\, X_{n+1} = \mathsf{a}) \approx \tfrac{1}{4}\,(1 - u - 2v)$$

donc les deux sont a priori différents.

Plus important : les deux modèles donnent des prévisions « macroscopiques » différentes.

Deuxième principe des processus M1
Convergence des fréquences empiriques.

Rappel : comptages $N_n(\mathtt{w}) = \sum_{i=1}^{n} \mathbf{1}(X_{i:i+\ell-1} = \mathtt{w})$ avec $\ell = |\mathtt{w}|$.

Et $R_n(\mathtt{w}) = N_n(\mathtt{w})/n$. Donc $R_n(\mathtt{w})$ est une variable aléatoire.

Le résultat (Loi des grands nombres)

• Hypothèses techniques : chaîne de Markov finie, irréductible, apériodique, π distribution stationnaire.
• Alors, il existe une fonction Π telle que, pour tout mot $\mathtt{w}$ et toute loi initiale ν,

$$R_n(\mathtt{w}) \to_{n\to\infty} \Pi(\mathtt{w}),$$

au moins aux sens où $\mathbb{E}_\nu(R_n(\mathtt{w})) \to \Pi(\mathtt{w})$ et, pour tout $\varepsilon > 0$,

$$\mathbb{P}_\nu(|R_n(\mathtt{w}) - \Pi(\mathtt{w})| \geqslant \varepsilon) \to_{n\to\infty} 0.$$

Ici,

$$\Pi(\mathtt{w}) = \mathbb{P}_\pi(X_{1:\ell} = \mathtt{w}) = \pi(w_1)\, q(w_1, w_2) \dots q(w_{\ell-1}, w_\ell).$$

Conséquence : si n est grand, $R_n(w_1) \approx \pi(w_1)$, mais aussi

$$R_n(w_1 w_2) \approx \pi(w_1)\, q(w_1, w_2),$$

et encore

$$R_n(w_1 w_2 w_3) \approx \pi(w_1)\, q(w_1, w_2)\, q(w_2, w_3) \dots$$

Estimation statistique dans le modèle M1

Comment estimer les paramètres $q(x, y)$ à partir d'une trajectoire $x_{1:n}$? Réponse : EMV !

$$v(q) = \nu(x_1) \prod_{i=2}^{n} q(x_{i-1}, x_i) = \nu(x_1) \prod_{x,y \in \mathcal{A}} q(x, y)^{N_{n-1}(x,y)}.$$

Donc

$$\log v(q) = \mathrm{C}^{\mathrm{te}} + \sum_{x,y \in \mathcal{A}} N_{n-1}(x, y) \log q(x, y).$$

Contraintes :

$$\sum_{y \in \mathcal{A}} q(x, y) = 1, \quad x \in \mathcal{A}.$$

Le résultat :

$$\widehat{q}(x, y) = N_{n-1}(x, y)/N_{n-1}(x).$$

On connaît une distribution $\widehat{\pi}$ stationnaire pour $\widehat{q}$, sans calcul :

$$\widehat{\pi}(x) = N_{n-1}(x)/(n-1).$$

En réalité, $\widehat{\pi}$ est seulement « presque » stationnaire car il y a un problème de comptage ± 1 au temps n.

Mais on fait comme si.

En pratique :

$$\widehat{q}(x, y) = N_n(x, y)/N_n(x), \quad \widehat{\pi}(x) = N_n(x)/n.$$

Avantage : estimation facile, on compte des mots de longueur 1 et 2.

Désavantage ou avantage ? Les fréquences des mots de longueur $\geqslant 3$ sont prédites par le modèle M1. Par exemple,

$$\begin{aligned} N_n(uvw) &= n\, R_n(uvw) \\ &\approx n\, \Pi(uvw) = n\, \pi(u)\, q(u,v)\, q(v,w) \\ &= n\, \Pi(uv)\, \Pi(vw)/\pi(v) \\ &\approx N_n(uv)\, N_n(vw)/N_n(v). \end{aligned}$$

Conséquence : si $N_n(uvw)$ est vraiment différent de

$$N_n(uv)\, N_n(vw)/N_n(v),$$

le modèle M1 est douteux.

Renvoi à la littérature : la procédure d'Arndt et al. pour résoudre le modèle d'évolution consiste à imposer l'égalité des fréquences

$$R_n(uvw) = R_n(uv)\, R_n(vw)/R_n(v).$$