Épreuve de mathématiques – 25 janvier 2007

Notes de cours et calculettes sont autorisées.

Exercice 1. *De séquences écrites dans l'alphabet* {t,c,a,g}*, on ne garde que l'information « purine »* ({a,g}*) ou « pyrimidine »* ({t,c}*). Dans cet alphabet, on compte combien de fois apparaît chaque mot de deux lettres sur des segments homologues issus de deux bactéries, Rickettsia akari et Rickettsia prowozeski.*

$$N_{Ra} = \begin{pmatrix} 110 & 90 \\ 140 & 60 \end{pmatrix}, \quad N_{Rp} = \begin{pmatrix} 70 & 30 \\ 60 & 40 \end{pmatrix}$$

On admettra que les modèles markoviens (d'ordre 1) ajustés sur ces deux observations ont des log-vraisemblances égales à $L_{\pi_{Ra}} = -259.800$ *et* $L_{\pi_{Rp}} = -128.388$*. Pour tester si c'est le même modèle ou non qui gère ces deux séquences, on considérera les deux observations comme provenant d'un même échantillon, on estimera la matrice de transition commune,* π_R*, on calculera la log-vraisemblance de ce modèle commun, et on fera le test.*

k	1	2	3	4	5	6	7	8
u	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51
v	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.54

TABLE 1 – Table donnant le seuil u tel que $\mathbb{P}(T_k \geq u) = 0.05$ et v tel que $\mathbb{P}(T_k \geq v) = 0.01$ lorsque $T_k \sim \chi^2(k)$.

Solution.

Estimation des paramètres du modèle commun. Si l'on considère que les deux observations proviennent d'un même échantillon, nous avons la matrice suivante pour les fréquences de transition des mots purine et pyrimidine (il suffit de sommer les deux matrices de l'énoncé) :

$$N_R = \begin{pmatrix} 180 & 110 \\ 200 & 100 \end{pmatrix},$$

où l'élément $N_R(i,j)$ indique le nombre de fois où l'on a observé une transition du mot i vers le mot j (par exemple $N_R(1,1)$ représente le nombre de fois où l'on a observé la transition purine vers purine sur la séquence).

L'estimateur du maximum de vraisemblance de la matrice de transition associée à N_R est défini par $\widehat{\pi}_R(a,b) = \frac{N_{a,b}}{N_{a\cdot}}$, ce qui donne

$$\widehat{\pi}_R = \begin{pmatrix} 180/300 & 110/300 \\ 200/300 & 100/300 \end{pmatrix} = \begin{pmatrix} 3/5 & 2/5 \\ 2/3 & 1/3 \end{pmatrix}.$$

Au passage, vous pouvez calculer la loi stationnaire μ associée à la matrice $\widehat{\pi}_R$ pour vous entraîner : il est facile de voir que $\mu = \begin{pmatrix}5/8 & 3/8\end{pmatrix}$, en vérifiant que $\mu\widehat{\pi} = \pi$.

Pour pouvoir calculer précisément la vraisemblance, il faut également disposer d'un estimateur de la loi initiale μ_0. Celui maximisant la vraisemblance est donné, dans le cas d'un modèle d'ordre 1, par $\widehat{\mu}_0(a) = \frac{N_R(a)}{\ell}$, où ℓ est la longueur de la chaîne. On obtient $\widehat{\mu}_0(\text{pur}) = 300/600 = 1/2$ et $\widehat{\mu}_0(\text{pyr}) = 300/600 = 1/2$. Ainsi, même sans connaître le premier élément de la séquence, on sait que la probabilité de voir apparaître l'un ou l'autre des deux mots possibles est la même. La contribution à la vraisemblance du premier élément sera donc 1/2.

Calcul de la vraisemblance du modèle commun. Par définition,

$$\begin{aligned} L_r = \log\left(\mathbb{P}(X_0)\prod_{k=1}^{\ell}\mathbb{P}(X_k|X_{k-1})\right) &= \log\left(\mathbb{P}(X_0)\prod_{i,j\in\{\text{pur,pyr}\}}\pi_R(i,j)^{N_R(i,j)}\right) \\ &= \log\mu(X_0) + \sum_{i,j\in\{\text{pur,pyr}\}} N_R(i,j)\log\pi_R(i,j) \end{aligned}$$

Soit, en remplaçant par les estimateurs,

$$\begin{aligned} L_R &= \log\left(\frac{1}{2}\right) + 180\log\left(\frac{3}{5}\right) + 110\log\left(\frac{2}{5}\right) + 200\log\left(\frac{2}{3}\right) + 100\log\left(\frac{1}{3}\right) \\ &= -384.88. \end{aligned}$$

Remarque 1. *Lorsque l'on ne connaît pas le début de la séquence, on peut également négliger le premier terme. C'est visiblement ce qui est fait pour le calcul de L_{Ra} et L_{Rp} données dans l'énoncé. Si on fait de même pour L_R, histoire d'avoir des choses comparables, on trouve -383.69 (en ôtant le terme $\log(1/2)$).*

Test du modèle le mieux adapté. Il s'agit de comparer la vraisemblance $L_R = -383.69$ du modèle commun avec celle du modèle à deux régimes (un par bactérie), pour lequel on somme les vraisemblances : posons $L_{Ra+Rp} = L_{Ra} + L_{Rp} = -259.8 - 128.388 = -388.188$. On peut considérer que ce sont les mêmes modèles qui gèrent les deux séquences si la vraisemblance du modèle commun et la somme des vraisemblances sur les modèles séparés sont suffisamment proches, ce qui constitue l'hypothèse H_0. Sous cette hypothèse, la statistique $T = 2D$ où D est la différence des log-vraisemblances L_R et L_{Ra+Rp} suit une loi du χ^2 à $k - h$ degrés de liberté, où k est le nombre de paramètres du modèle à deux régimes et h le nombre de paramètres pour le modèle commun. On a $2 \times 2 = 4$ paramètres libres pour le modèle à 2 régimes et 2 paramètres libres pour le modèle commun (1 par ligne de la matrice de transition – l'autre se déduisant par complémentarité) [1].

La valeur observée de la statistique de test $T \sim \chi^2_2$ est $t_{obs} = 9$. On se situe au-delà du seuil pour un niveau de 0.01, donc on ne peut pas considérer que ces deux séquences soient issues d'un même modèle.

1. Pour calculer le nombre de paramètres libres d'un modèle de Markov d'ordre m défini sur un alphabet à s éléments, vous pouvez utiliser la formule $(s-1) \times s^m$.

Exercice 2. *On place un singe orang-outan devant un clavier d'ordinateur comportant trois touches :* `a`, `b` *et* `c`. *Une étude préalable montre que l'on peut accepter l'hypothèse selon laquelle il tape ces trois touches avec même probabilité* $1/3$ *et que les lettres sont indépendantes.*

Quelle est la probabilité de voir apparaître en une position donnée le mot `abc` *? Même question pour le mot* `aaa`.

Quelle est l'espérance du nombre de lettres jusqu'à l'apparition du mot `abc` *? Même question pour le mot* `aaa`. *Pourquoi ces espérances diffèrent-elles ?*

On pourra utiliser les formules suivantes

$$\begin{pmatrix} 1/3 & -1/3 & 0 \\ -2/3 & 1 & -1/3 \\ -2/3 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 27 & 9 & 3 \\ 24 & 9 & 3 \\ 18 & 6 & 3 \end{pmatrix} \; ; \; \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 15 & 9 & 3 \\ 12 & 9 & 3 \\ 9 & 6 & 3 \end{pmatrix}$$

Solution. La séquence X_n générée par le singe suit un modèle $M0$ où la probabilité d'apparition d'une lettre est une loi uniforme sur l'alphabet {`a`,`b`,`c`}.

En régime stationnaire, la probabilité de voir apparaître les mots `aaa` ou `abc` est la même ! C'est tout simplement

$$\mathbb{P}_\mu(\texttt{aaa}) = \mathbb{P}_\mu(X_0 = \texttt{a})\mathbb{P}(X_1 = \texttt{a}|X_0 = \texttt{a})\mathbb{P}(X_2 = \texttt{a}|X_1 = \texttt{a}) = \frac{1}{3^3},$$

idem pour `abc`.

En revanche, le calcul de l'espérance du nombre de lettres jusqu'à l'apparition des mots `aaa` ou `abc` ne donne pas la même chose. Pour le montrer, on utilise la méthode de Fu :

Mot `aaa`. Considérons tout d'abord le mot `aaa`, auquel on associe la chaîne de Markov Y_n d'espace $\{\emptyset, \texttt{a}, \texttt{aa}, \texttt{aaa}\}$, c'est-à-dire que les états de Y_n sont les *préfixes* du mot recherché. On peut décrire le graphe associé à cette chaîne comme suit :

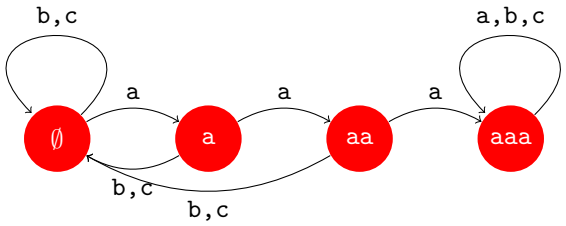

FIGURE 1 – Graphe de transition pour la chaîne Y

Les flèches indiquent l'apparition d'une lettre placée en fin de mot, et les états concernent le préfixe courant, d'où l'état absorbant en fin de graphe. La matrice de transition associée est

$$\Pi_1 = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \\ 2/3 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Considérons maintenant la matrice N_1 à laquelle on a ôté ligne et colonne correspondant à l'état absorbant, c'est-à-dire,

$$N_1 = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 2/3 & 0 & 1/3 \\ 2/3 & 0 & 0 \end{pmatrix}.$$

Un résultat du cours indique que l'espérance du nombre de fois où la chaîne Y, partant d'un état u, passe par l'état v avant d'être absorbée (c'est-à-dire avant d'atteindre le motif `aaa`) est donnée par le potentiel R_1 défini par

$$R_1(u,v) = (I - N_1)^{-1}(u,v).$$

La matrice R_1 est justement celle donnée dans l'énoncé :

$$R_1 = (I - N_1)^{-1} = \begin{pmatrix} 1/3 & -1/3 & 0 \\ -2/3 & 1 & -1/3 \\ -2/3 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 27 & 9 & 3 \\ 24 & 9 & 3 \\ 18 & 6 & 3 \end{pmatrix}.$$

Ainsi, l'espérance *du nombre de pas*[2] avant absorption, partant de u, est $\sum_v R_1(u,v)$, c'est-à-dire la somme par ligne de cette matrice : si l'on part d'un mot quelconque $\emptyset$ autre que `a` ou `aa`, le nombre de lettres à parcourir avant d'atteindre le motif `aaa` est $27 + 9 + 3 = 39$.

Mot `abc`. On fait un raisonnement similaire pour le mot `abc`. Le graphe associé à la chaîne Z_n d'espace $\{\emptyset, \mathtt{a}, \mathtt{ab}, \mathtt{abc}\}$ est :

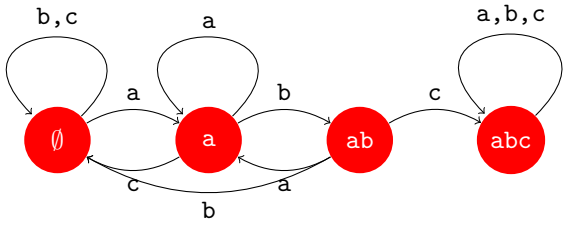

FIGURE 2 – Graphe de transition pour la chaîne Z

La matrice de transition associée est

$$\Pi_2 = \begin{pmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

La matrice N_2 à laquelle on a ôté ligne et colonne correspondant à l'état absorbant est

$$N_2 = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \end{pmatrix},$$

2. ce qui revient au nombre de lettres

et on a

$$R_2 = (I - N_2)^{-1} = \begin{pmatrix} 15 & 9 & 3 \\ 12 & 9 & 3 \\ 9 & 6 & 3 \end{pmatrix}.$$

Ainsi, si l'on part d'un mot quelconque $\emptyset$ autre que `a` ou `ab`, le nombre de lettres à parcourir avant d'atteindre le motif `abc` est $15 + 9 + 3 = 27$: on trouve bien des résultats différents, ce qui est logique étant donné que `aaa` est *auto-recouvrant* contrairement à `abc`.

Exercice 3. *On considère une séquence X modélisée par une chaîne de Markov cachée $M1M0$ à trois régimes, notés $R1$, $R2$ et $R3$. À la matrice de transition π_0 entre ces régimes on associe la matrice T (elle aussi 3×3), de terme général $T(t,u) = \log(\pi_0(t,u))$. Aux probabilités $\mu_u(a) = \mathbb{P}(X_i = a \,|\, S_i = u)$, on associe $V(u,a) = \log(\mu_u(a))$* [3]

$$T = \begin{pmatrix} 3.0 & -1.0 & -1.0 \\ 1.6 & 2.5 & 1.6 \\ -1.0 & -1.0 & 3.0 \end{pmatrix} \qquad V = \begin{pmatrix} 1.0 & 1.5 & 2.0 & 2.5 \\ 2.5 & 2.0 & 1.5 & 1.0 \\ 1.9 & 1.9 & 1.9 & 1.9 \end{pmatrix}$$

(indiqué dans l'ordre `t,c,a,g`*)*

On notera $Z_k(v)$ le meilleur score sur $[1 \dots k]$ tel que $S_k = v$. En initialisant par $Z_1(u) = V(u,a)$ (on aurait aussi pu faire intervenir la loi stationnaire de π_0.), annotez les séquences `ctagac` *et* `ctagacc`*.*

Solution. Il s'agit simplement de faire tourner l'algorithme de Viterbi que vous avez dû voir en cours avec les mêmes notations. Il permet de déterminer les régimes les plus probables dans un modèle de Markov cachée (cela concerne donc la partie « cachée » du modèle).

Pour chaque position k, le meilleur score $[1 \dots k]$ tel que $S_k = v$ sur vaut

$$Z_k(v) = \max_u \big(Z_{k-1}(u) + T(u,v) + V(v,a) \big), \tag{1}$$

où on a initialisé $Z_1(u)$ à $V(u,a)$.

Annotons par exemple la séquence `ctagacc` : on obtient le tableau des $Z_i(v)$ où les flèches indiquent pour quel régime u le maximum dans (1) a été atteint :

`c`		`t`		`a`		`g`		`a`		`c`		`c`
1.5	→	5.5		11.1	→	16.1	→	**21.6**	→	26.1	→	30.6
2.0	→	**7.0**	→	10.5	→	14.5	→	18.0		22.6	→	27.6
1.9	→	6.4	→	**11.7**	→	**16.6**	→	21.5	→	**26.6**	→	**31.3**

La première colonne est tout simplement la colonne de V correspondant à la lettre `c` (c'est-à-dire la 2e : les états sont dans l'ordre `t,c,a,g`).

3. On peut ajouter une constante arbitraire à T et/ou à V sans changer la solution – ce que nous avons fait ici. Notons que dans cet exemple les 'contrastes' entre les régimes ont été choisis extrêmes ($\mu_{R1} = (3.2\%, 8.7\%, 23.7\%, 64, 4\%)$, par exemple), ce qui permet sur une séquence très courte d'obtenir des changements de régime)

Voyons comment construire la suivante. La formule est

$$Z_2(v) = \max_u \big(Z_1(u) + T(u,v) + V(v,\mathtt{t})\big),$$

puisque la lettre courante est `t`. Lorsque $v = 1$ (premier régime), on a

$$\begin{aligned} Z_2(1) &= \max_u \big(Z_1(u) + T(u,1)\big) + V(1,\mathtt{t}) \\ &= \max_u \big(Z_1(1) + T(1,1); Z_1(2) + T(2,1); Z_1(3) + T(3,1)\big) + 1.0 \\ &= \max_u \big(1.5 + 3; 2.0 + 1.6; 1.9 - 1\big) + 1.0 = 4.5 + 1 = 5.5 \end{aligned}$$

Le maximum a été atteint pour $u = 1$, on fait donc une flèche partant du régime 1 vers le régime courant $v = 1$.

Lorsque $v = 2$,

$$Z_2(2) = \max_u \big(Z_1(u) + T(u,2)\big) + V(2,\mathtt{t}) = 4.5 + 2.5 = 7,$$

où le maximum est atteint en u égal 2. On fait donc une flèche depuis le régime $u = 2$ vers le $v = 2$.

Pour $v = 3$

$$Z_2(3) = \max_u \big(Z_1(u) + T(u,3)\big) + V(3,\mathtt{t}) = 4.9 + 1.9 = 6.8.$$

Cette fois, c'est de $u = 3$ que le maximum est atteint, d'où la flèche du régime 3 vers le $v = 3$.

Et ainsi de suite ! Dans chaque colonne, on a indiqué en gras le maximum, c'est à dire le régime que l'algorithme attribuerait à cette position si l'annotation s'arrêtait là. Remontant les flèches depuis cette position, on trouve l'annotation de tout le début de la séquence. Remarquons qu'en un seul pas, on peut remettre en cause toute cette annotation : à la $5^{\text{ème}}$ lettre, l'annotation optimale est $R2R2R1R1R1$, à la $6^{\text{ème}}$, elle devient $R3R3R3R3R3R3$.