Rapport TP1 : Chaînes de Markov cachées

Jean-Christophe Corvisier
Mohammed Amine Kheldouni

Avril 2017

Question 1

La modélisation des données avec une loi normale n'est pas satisfaisante car comme nous pouvons le voir sur les courbes, la superposition pose problème : on observe un pic sur les données des crabes en dehors de la plage parcourue par la loi gaussienne théorique, de même que quelques valeurs trop petites par rapport à celle obtenu avec la loi gaussienne théorique.

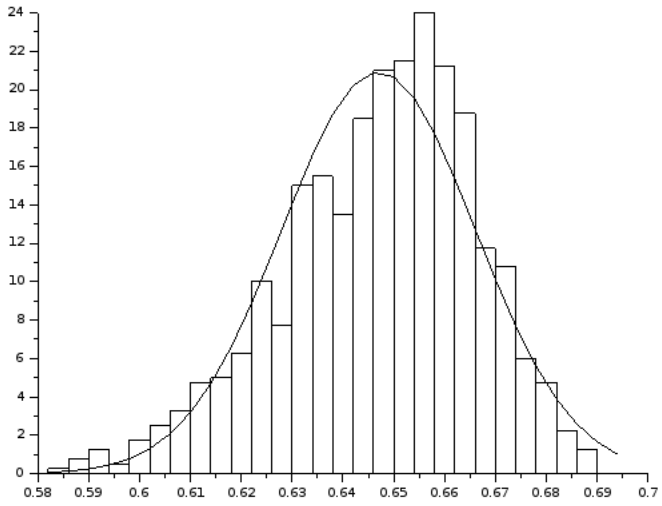

FIGURE 1 – Approximation de l'histogramme des données par une loi normale

Question subsidiaire

On teste dans cette question la normalité de notre variable aléatoire X associée aux données sur les crabes répartis sur chacun des intervalles.

On pose d'abord clairement les hypothèses du test d'adéquation du χ^2 :

$$\mathcal{H}_0 = \{X \sim N(\mu, \sigma)\}$$
$$\mathcal{H}_1 = \{X \text{ n'est pas distribuée normalement}\}$$

On prend une moyenne égale à la moyenne du des observations sur les crabes donc : $\mu = 0.647$ et de même, une variance égale à la variance sur les observations de : $\sigma^2 = 3.64 \times 10^{-4}$.
On considère également un risque d'erreur pour ce test du χ^2 de $\alpha = 5\%$. On calcule ensuite les fréquences théorique des crabes dans les intervalles de discrétisation donnés :

$Intervalle$	$Nombre$	$Nombretheorique$
$[0.580, 0.584[$	1	0.25
$[0.584, 0.588[$	3	0.51
$[0.588, 0.592[$	5	1
$[0.592, 0.596[$	2	1.78
$[0.596, 0.600[$	7	3.12
$[0.600, 0.604[$	10	5.22
$[0.604, 0.608[$	13	8.36
$[0.608, 0.612[$	19	12.82
$[0.612, 0.616[$	20	18.8
$[0.616, 0.620[$	25	26.4
$[0.620, 0.624[$	40	35.49
$[0.624, 0.628[$	31	45.66
$[0.628, 0.632[$	60	56.2
$[0.632, 0.636[$	62	66.25
$[0.636, 0.640[$	54	74.7
$[0.640, 0.644[$	74	80.68
$[0.644, 0.648[$	84	83.4
$[0.648, 0.652[$	86	82.46
$[0.652, 0.656[$	96	78.1
$[0.656, 0.660[$	85	70.75
$[0.660, 0.664[$	75	61.36
$[0.664, 0.668[$	47	51
$[0.668, 0.672[$	43	40.47
$[0.672, 0.676[$	24	30.78
$[0.676, 0.680[$	19	22.41
$[0.680, 0.684[$	9	15.6
$[0.684, 0.688[$	5	10.4
$[0.688, 0.692[$	0	6.64
$[0.692, 0.696[$	1	4.1

Une fois le seuil d'admissibilité de l'hypothèse $\mathcal{H}_0$ fixé, et les fréquences théoriques d'apparition dans les intervalles calculées grâce à la fonction de répartition de la loi normale, on se donne une statistique de test qui suit la loi du χ^2 comme suit :

$$\mathcal{S} = \sum_{i=1}^{29} \frac{(X_{observe,i} - X_{theorique,i})^2}{X_{theorique,i}} = \sum_{i=1}^{29} \frac{(Nombre_{observe,i} - Nombre_{theorique,i})^2}{Nombre_{theorique,i}}$$

$$\Rightarrow \mathcal{S} = 84,67$$

En cherchant dans la table du χ^2 pour un seuil d'admissibilité de 5% et un un degré de liberté de 28, on obtient la valeur $\chi^2_{5\%,28} = 41,34$.
Finalement, comme $\mathcal{S} = 84,67 > \chi^2_{5\%,28}$, l'hypothèse $\mathcal{H}_0$ est rejetée et donc nos données X ne suivent pas forcément la loi normale de paramètres $(\mu = 0.647,\ \sigma^2 = 3.64 \times 10^{-4})$.

Question 2

On rappelle que dans cette section, on veut maximiser la fonctionnelle $Q(\theta, \theta')$. Pour cela, on récrit cette dernière comme une somme d'une fonction de π notée A_0 :

$$A_0 \ : \ \pi \to \sum_{i \in \mathcal{I}} \pi_i^* log(\pi_i)$$

et des fonctions A_j pour $j \in \mathcal{I}$:

$$A_j \ : \ (\mu_j, \sigma_j) \to \sum_{j \in \mathcal{I}} \sum_{k=1}^{N} \rho'_{j,k} log(f_{\mu_j,\sigma_j}(y_k))$$

Comme on sait que les π^* forment une loi de probabilité, l'application du Lemme 5.2.8 du livre " Modèles aléatoires : Applications aux sciences de l'ingénieur et du vivant" de Benjamin Jourdain et Jean-François Delmas, nous donne donc que le maximum de A_0 est atteint pour la valeur $\pi = \pi^*$. Ainsi, A_0 est maximal pour $\pi = \pi^*$.

Maximalité de A_j
D'autre part, les fonctions A_j sont dérivables en μ_j et σ_j et le calcul des dérivées partielles se fait comme suit :

$$\frac{\partial A_j}{\partial \mu_j} = \sum_{k=1}^{N} \rho'_{j,k} \frac{(y_k - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial A_j}{\partial \sigma_j} = -\sum_{k=1}^{N} \rho'_{j,k} \frac{1}{\sigma_j} [1 - \frac{(y_k - \mu_j)^2}{\sigma_j^2}]$$

La recherche de point critique (pour lequel les dérivées partielles s'annulent) fournit un vecteur de paramètres (μ^*, σ^*) :

$$\forall j \in \mathcal{I}, \ \ \mu_j^* = \frac{\sum_{k=1}^{N} \rho'_{j,k} y_k}{\sum_{k=1}^{N} \rho'_{j,k}}$$

$$\forall j \in \mathcal{I}, \quad (\sigma_j^*)^2 = \frac{\sum_{k=1} \rho'_{j,k}(y_k - \mu_j^*)^2}{\sum_{k=1}^{N} \rho'_{j,k}}$$

Ensuite, un simple argument sur la concavité de la fonction, basé sur sa matrice Hessienne, prouve que le point critique obtenu est en fait un maximum de la fonction (et même l'unique).

Question 3

On applique dans cette question ci l'algorithme E.M. maximisant la fonctionnelle étudiée dans la section précédente $Q(\theta, \theta')$.
On effectue donc un calcul itératif des valeurs moyennes et des variances des loi normales que l'on désire mélanger pour obtenir une meilleure approximation de la loi des données traitées sur les crabes.
Pour cela, on forme un vecteur ρ selon les densités gaussiennes actualisées et le vecteur de probabilité π comme suit :

$$\rho'_{i,k} = \frac{\pi'_i f_{\mu'_i, \sigma'_i}(y_k)}{f_{\theta'}(y_k)}$$

Un premier résultat avec deux populations (donc un mélange de deux loi gaussiennes) nous fournit la figure (2) ci-dessous :

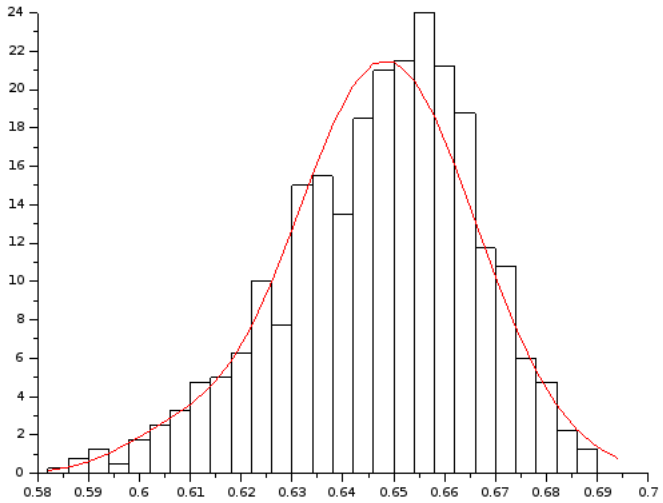

FIGURE 2 – Tracé du mélange de deux populations gaussiennes par rapport à l'histogramme des données

On converge avec ce schéma à deux population vers les valeurs suivantes :

$$\pi = (0.043, 0.957) \qquad \mu_1 = 0.6 \qquad \mu_2 = 0.65 \qquad \sigma_1^2 = 0.00011 \qquad \sigma_2^2 = 0.00031$$

Question subsidiaire

En généralisant le modèle du mélange de loi normales à trois population cette fois ci, on obtient une courbe plus fidèle à l'histogramme des données (cf. Figure 3).
Les résultats sont les suivants :

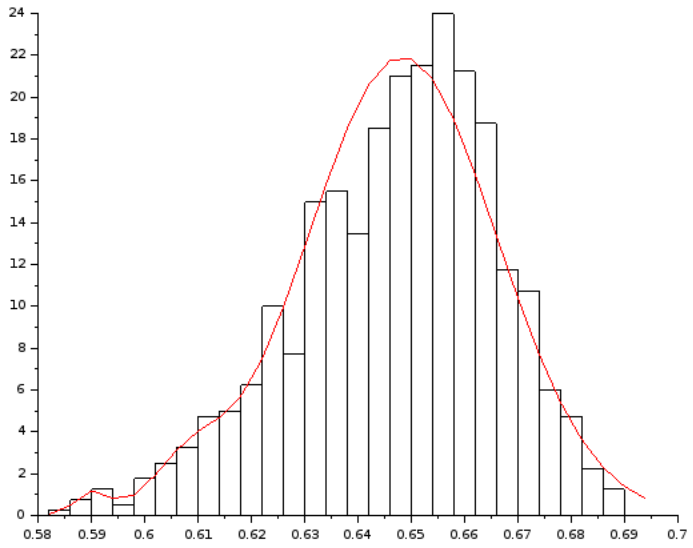

FIGURE 3 – Tracé du mélange de trois populations gaussiennes par rapport à l'histogramme des données

$$\mu_1 = 0.6 \quad \mu_2 = 0.65 \quad \mu_3 = 0.65 \quad \sigma_1^2 = 0.000045 \quad \sigma_2^2 = 0.00028 \quad \sigma_3^2 = 0.00022$$

Question 4

Pour la séquence d' ADN de test (seqlambda2.txt), on utilise le programme initialisé avec les valeurs suivantes, qui se trouvent être les valeurs utilisées lors de la génération de la séquence, et ce pour 1000 itérations :

$$a = \begin{pmatrix} 0.99 & 0.03 \\ 0.03 & 0.97 \end{pmatrix}$$

$$b = \begin{pmatrix} 0.2697410 & 0.2084444 & 0.1983422 & 0.3234723 \\ 0.2463460 & 0.2475527 & 0.2982972 & 0.2078041 \end{pmatrix}$$

$$\pi_0^0 = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix}$$

On trouve finalement à la fin des itérations les résultats suivants :

$$a_{11} = 0.9977, \qquad a_{22} = 1.$$

On obtient aussi une matrice b comme suit :

$$b = \begin{pmatrix} 0.258 & 0.204 & 0.216 & 0.323 \\ 0.194 & 0.257 & 0.383 & 0.166 \end{pmatrix}$$

Bien que les valeurs trouvées ne correspondent exactement à celles utilisées pour la génération de cette séquence d'ADN, elles sont néanmoins assez proches. De plus, on peut observer une convergence des coefficients pour a et b, comme le montre les figures 4,5 et 6.

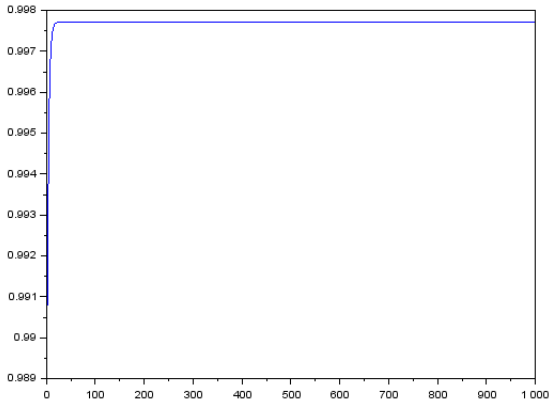

FIGURE 4 – évolution de a_{11} en fonction du nombre d'itération

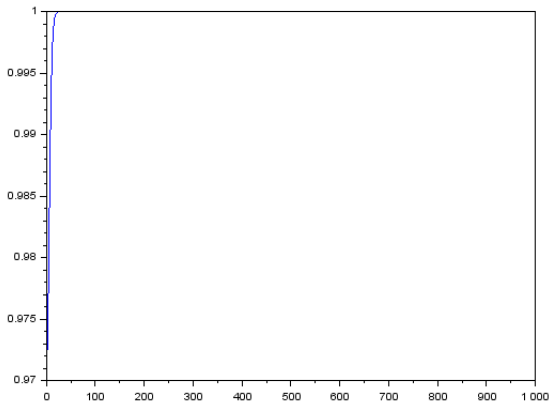

FIGURE 5 – évolution de a_{22} en fonction du nombre d'itération

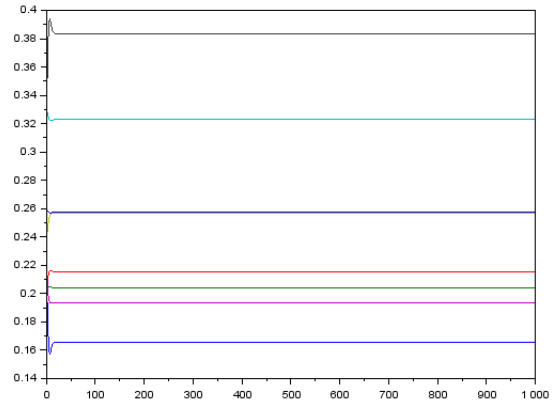

FIGURE 6 – évolution des coefficients de b en fonction du nombre d'itération

On peut en effet observer la convergence des coefficients diagonaux de la matrice a : au bout de 50 itérations les deux coefficients atteignent déjà leurs valeurs finales. De même les coefficients de b convergent au bout de 50 itérations. On en conclut donc que la statégie d'optimisation est efficace, puisqu'en moins de 100 itérations les valeurs de a et b convergent.

Les valeurs de $n \to \mathbb{P}(S_n = +1|Y_1^{N_0})$ et $n \to \mathbb{P}(S_n = -1|Y_1^{N_0})$sont représentées sur les figures 7 et 8.

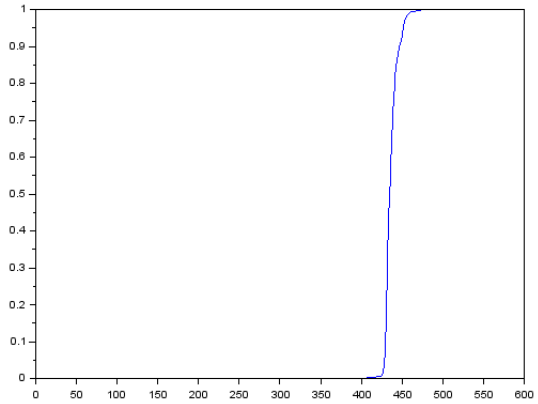

FIGURE 7 – Valeur de la probabilité d'être dans l'état +1 selon le nucléotide considéré pour la séquence d'ADN de test

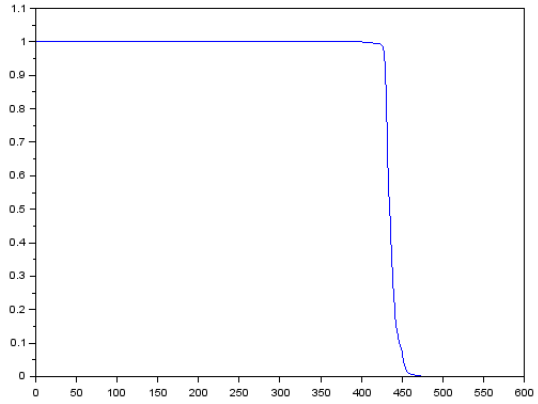

FIGURE 8 – Valeur de la probabilité d'être dans l'état -1 selon le nucléotide considéré pour la séquence d'ADN de test

On observe sur le petit échantillon d'ADN (sequence seqlambda2.text) deux régions distinctes : une première région (les 450 premiers nucléotides) qui comprend des nucléotides étant presque sûrement dans l'état +1. Et une deuxième région (les nucléotides situées entre la 450 ème et la dernière position) qui eux sont presque sûrement dans l'état -1, ce qui est conforme aux prédictions attendus comme précisé dans le mail d'information qui nous a été envoyé.

Concernant l'échantillon d'ADN réel, on exécute le même programme, avec des valeurs d'initialisations pour a et b et pi_0^0 identiques à celles utilisées pour la séquence de tests, et ce pour 100 itérations (le programme est plus gourmand en calcul du fait de la longueur de la séquence considérée).

$$a_{11} = 0.999772, a_{22} = 0.9998785$$

$$b = \begin{pmatrix} 0.269741 & 0.2084444 & 0.1983422 & 0.3234723, \\ 0.246346 & 0.2475527 & 0.2982972 & 0.2078041 \end{pmatrix}$$

La convergence de notre algorithme a bien lieu, comme le montre les figures 9, 10 et 11. Au bout de 20 itérations, les coefficients atteignent déjà leur valeur finale et ne varie plus.

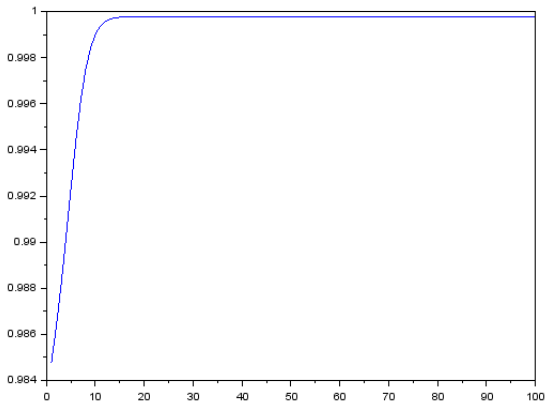

FIGURE 9 – évolution de la valeur du coefficient a_{11} en fonction du nombre d'itération pour le brin d'ADN réel

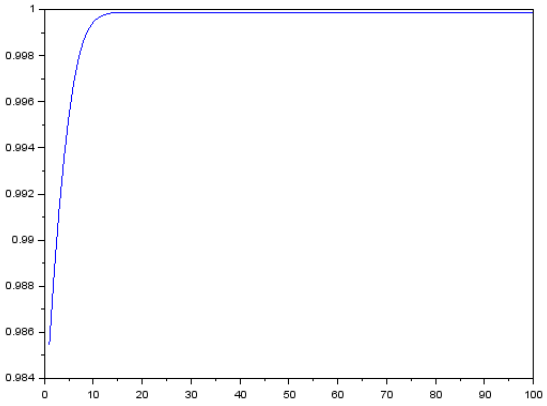

FIGURE 10 – évolution de la valore du coefficient a_{22} en fonction du nombre d'itération pour le brin d'ADN réel

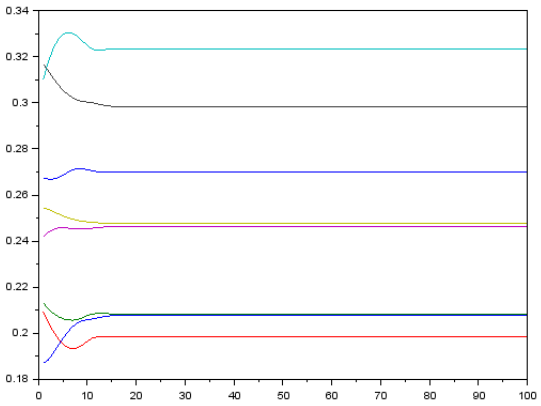

FIGURE 11 – évolution des valeur des coefficients de b en fonction du nombre d'itération pour le brin d'ADN réel

Enfin Les valeurs de $n \to \mathbb{P}(S_n = +1|Y_1^{N_0})$ et $n \to \mathbb{P}(S_n = -1|Y_1^{N_0})$ obtenues pour la séquence d'ADN réelle sont représentées sur les figures 12 et 13.

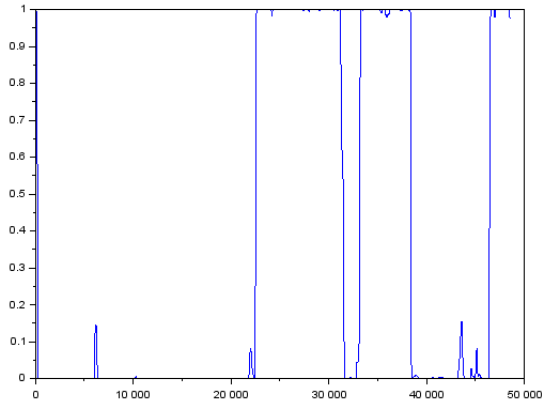

FIGURE 12 – Valeur de la probabilité d'être dans l'état +1 selon le nucléotide considéré pour la séquence d'ADN réelle

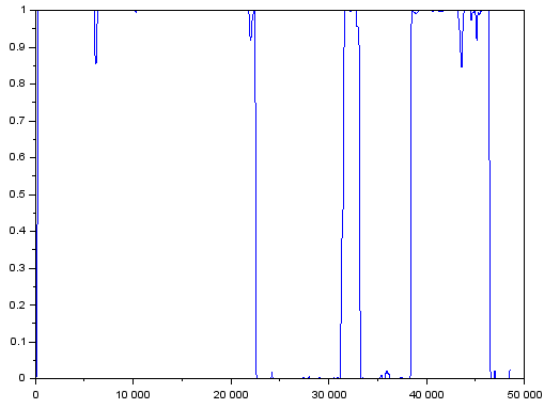

FIGURE 13 – Valeur de la probabilité d'être dans l'état -1 selon le nucléotide considéré pour la séquence d'ADN réelle

Ces deux figures nous permettent donc de mettre en évidence six importantes zones homogènes : les nucléotides dont les positions varient environ entre 22 000 et 31 000, entre 33 000 et 39 000 et entre 47000 et 50000 sont presque sûrement dans l'état + (ils sont codants). Pour les nucléotides dont les positions ne sont pas dans ces intervalles, ceux-ci sont presque sûrement dans l'état -1 (c'est à dire que le nucléotide qui leur est apparié est codant).