TP1 “MODÉLISER L’ALÉA” CHAÎNES DE MARKOV CACHÉES

Clément Riu - Louis Trezzini

24 avril 2017

2 Crabes de Waldon

Question 1. On visualise l’histogramme des données ainsi que la gaussienne la plus proche afin d’évaluer si les données suivent une loi normale.

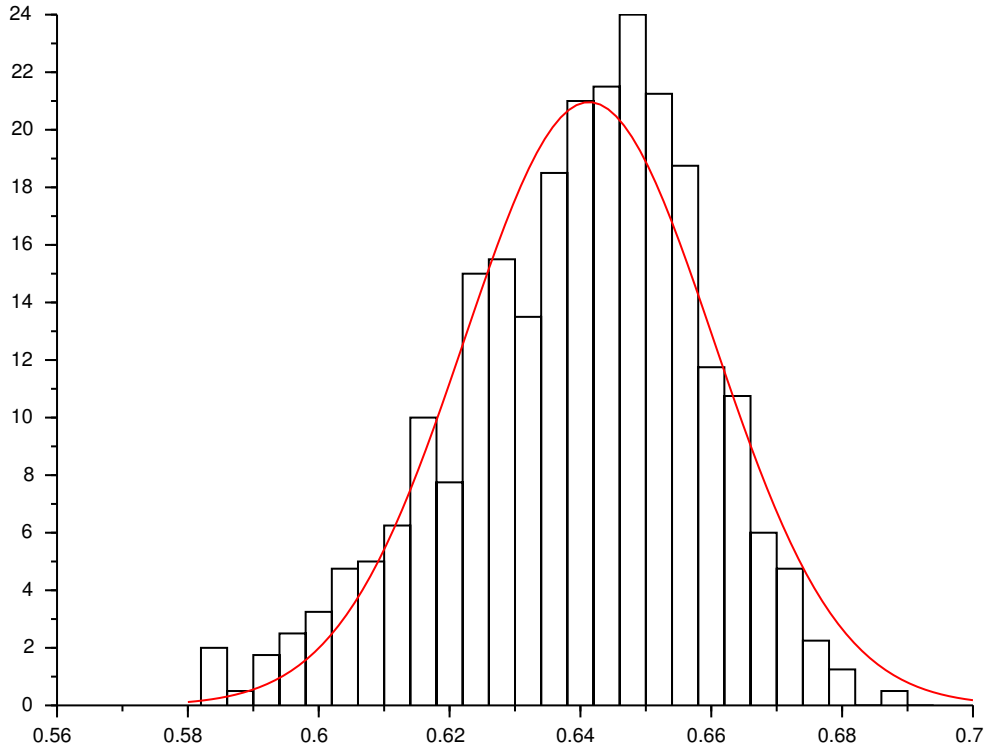

FIGURE 1 – Loi empirique des données et gaussienne la plus proche.

Il y a un décallage entre la courbe théorique et les données empiriques, la loi normale n’explique pas bien les données.

Question subsidiaire 1. Lorsqu’on applique un test du χ^2 aux données pour tester la normalité des données, le test est négatif : une loi normale ne permet pas d’expliquer les données.

Question 2.

Question 3. L'application de l'algorithme EM renvoie la courbe suivante :

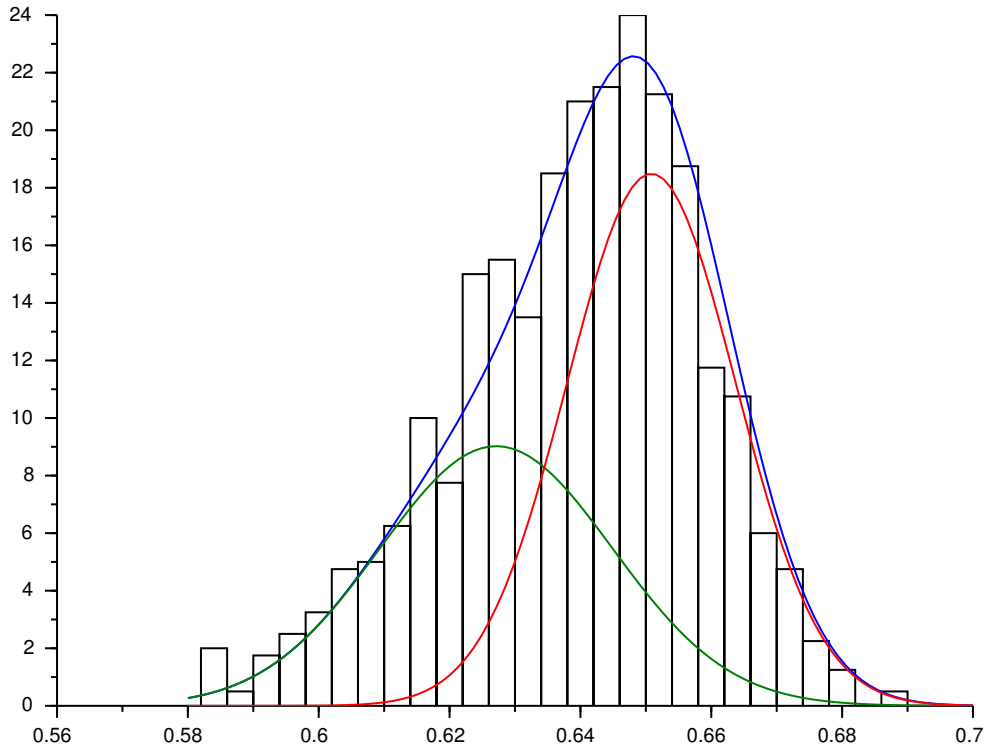

FIGURE 2 – Loi empirique des données et les gaussiennes calculées par l'algorithme EM.

On remarque que la courbe en cyan suit bien les données empiriques, cette loi semble donc plus adéquate que la première proposition.

Question subsidiaire 2. On obtient dans le cas de 3 populations le résultat suivant :

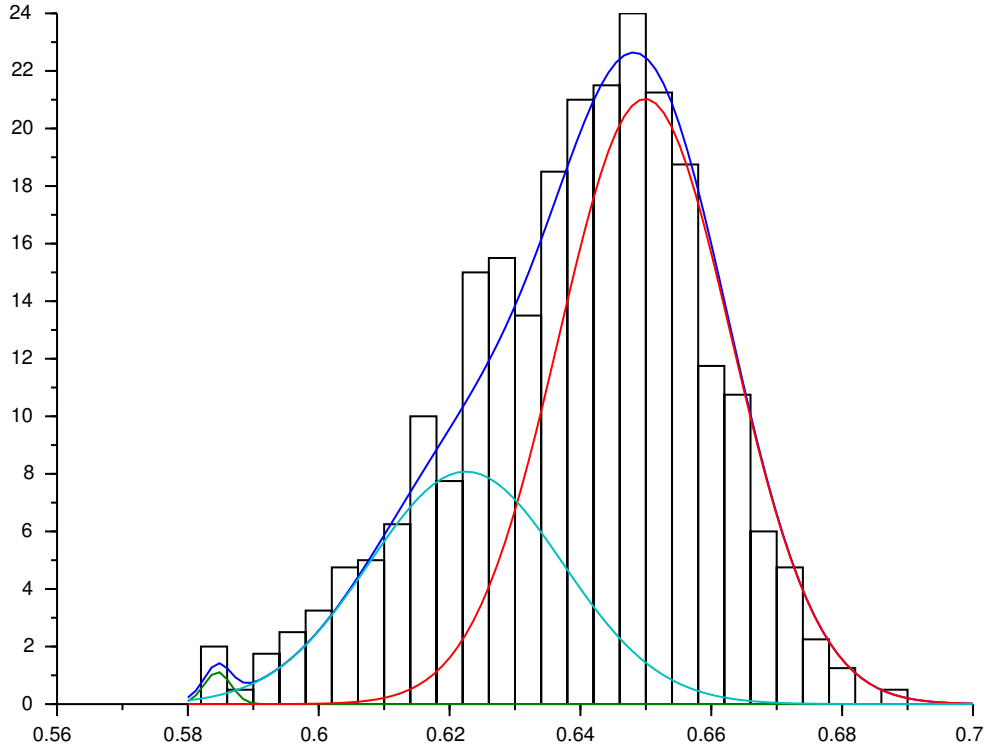

FIGURE 3 – Loi empirique des données et les gaussiennes calculées par l'algorithme EM.

On observe que les crabes de faibles tailles font partis d'une population à part.

3 Recherche de zones homogènes dans l'ADN

Question 4. On applique le calcul aux données restreintes avec les données initiales suivantes :

$$a = \begin{pmatrix} 0.99 & 0.01 \\ 0.03 & 0.97 \end{pmatrix}$$
$$b = \begin{pmatrix} 0.2697410 & 0.2084444 & 0.1983422 & 0.3234723 \\ 0.2463460 & 0.2475527 & 0.2982972 & 0.2078041 \end{pmatrix}$$
$$\pi_0 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

On obtient alors les résultats finaux suivant :

$$a = \begin{pmatrix} 0.9977068 & 0.0022932 \\ 0 & 1 \end{pmatrix}$$
$$b = \begin{pmatrix} 0.2576651 & 0.203861 & 0.2155776 & 0.3228962 \\ 0.1938853 & 0.2570135 & 0.3833224 & 0.1657789 \end{pmatrix}$$

On observe alors les régions codantes et non-codantes :

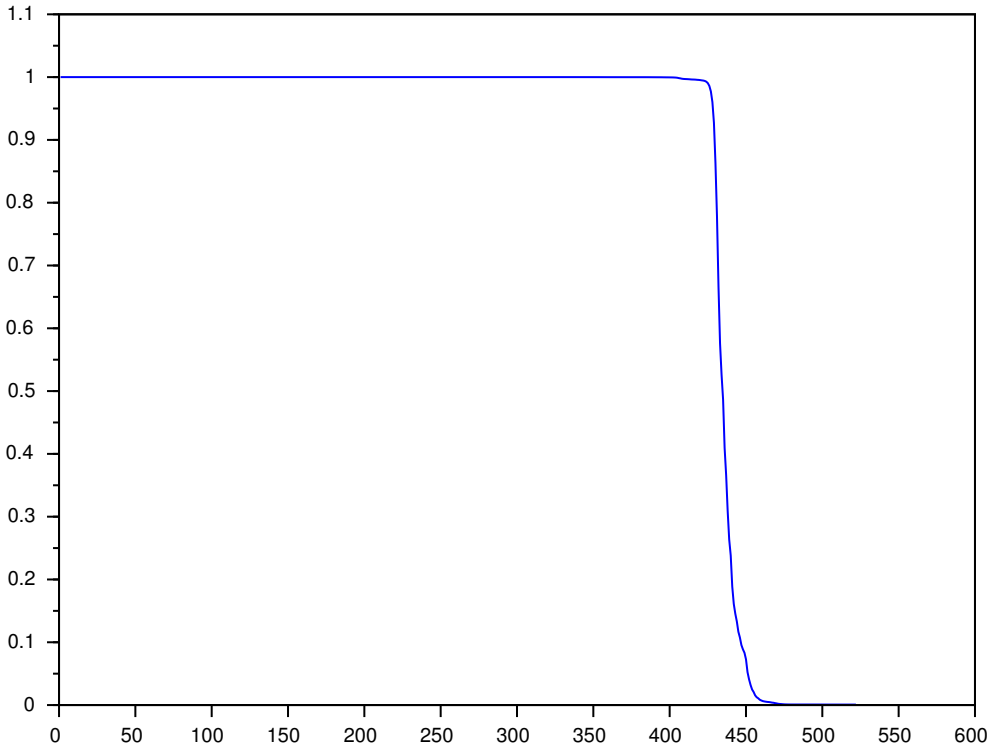

FIGURE 4 – Les zone homogènes de la séquence d'ADN.

Les matrices a et b évoluent comme suit :

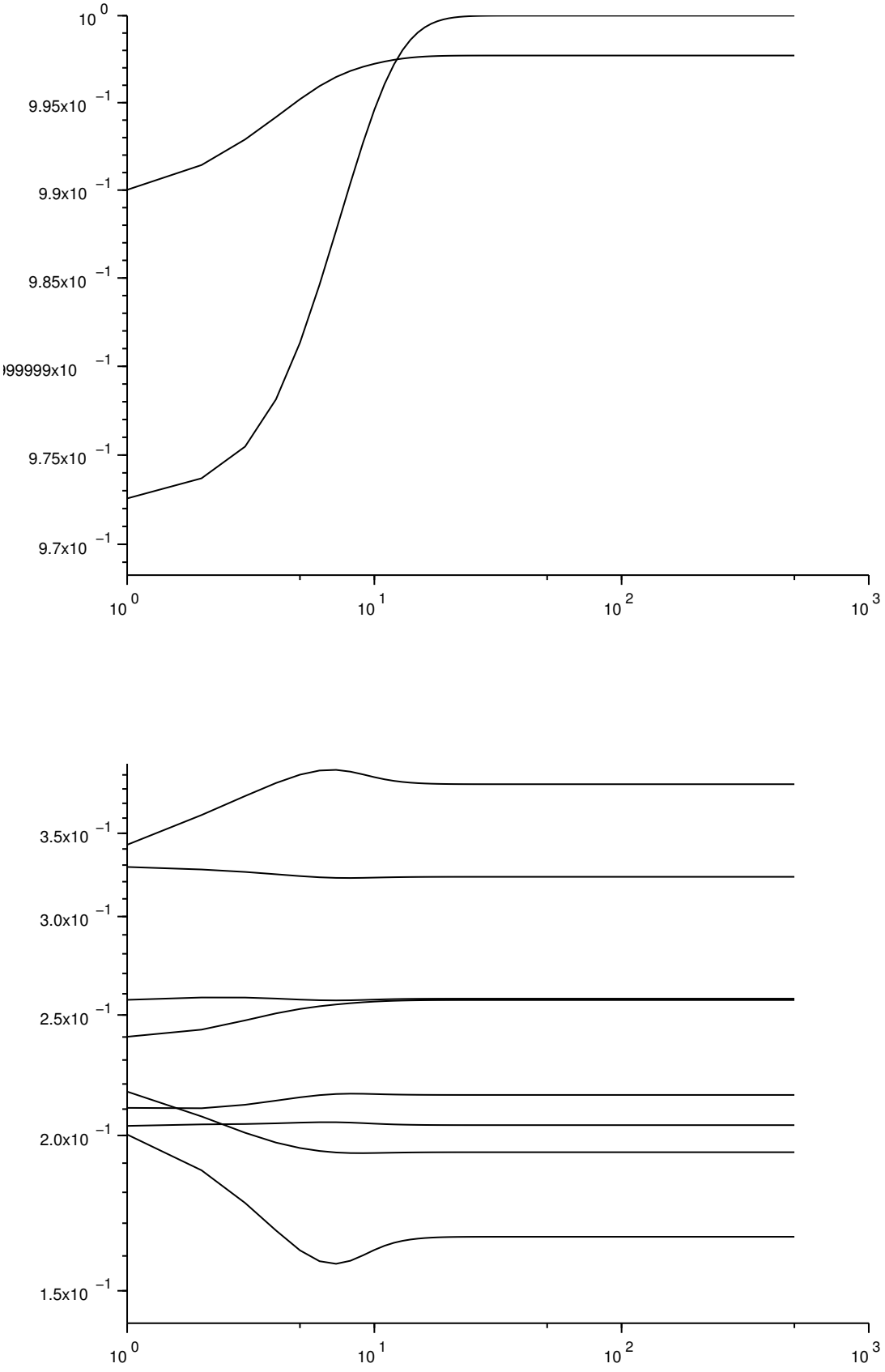

FIGURE 5 – Évolution en échelle log-log des coefficients de la matrice a (en haut) et de la matrice b (en bas).

On remarque une convergence relativement rapide de l'algorithme, comme attendu d'après les résultats du livre *Modèles aléatoires* de J.-F. Delmas et B. Jourdain.

Les données restreintes donnant des résultats satisfaisants, on observe le résultat sur la chaine complète :

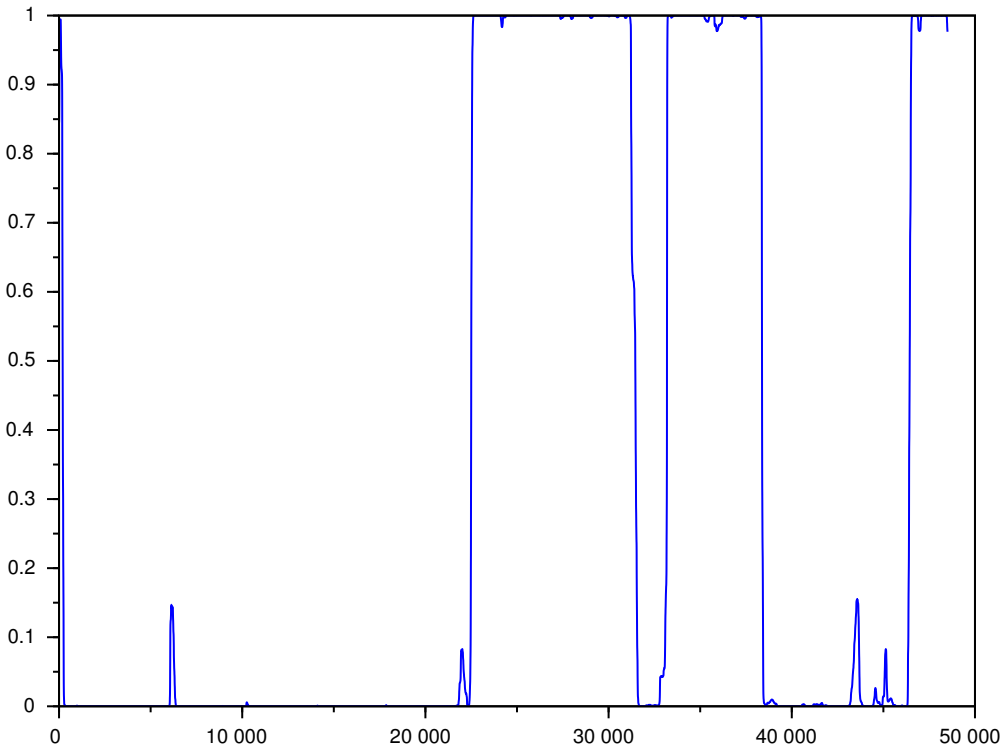

FIGURE 6 – Les zone homogènes de la séquence d'ADN.

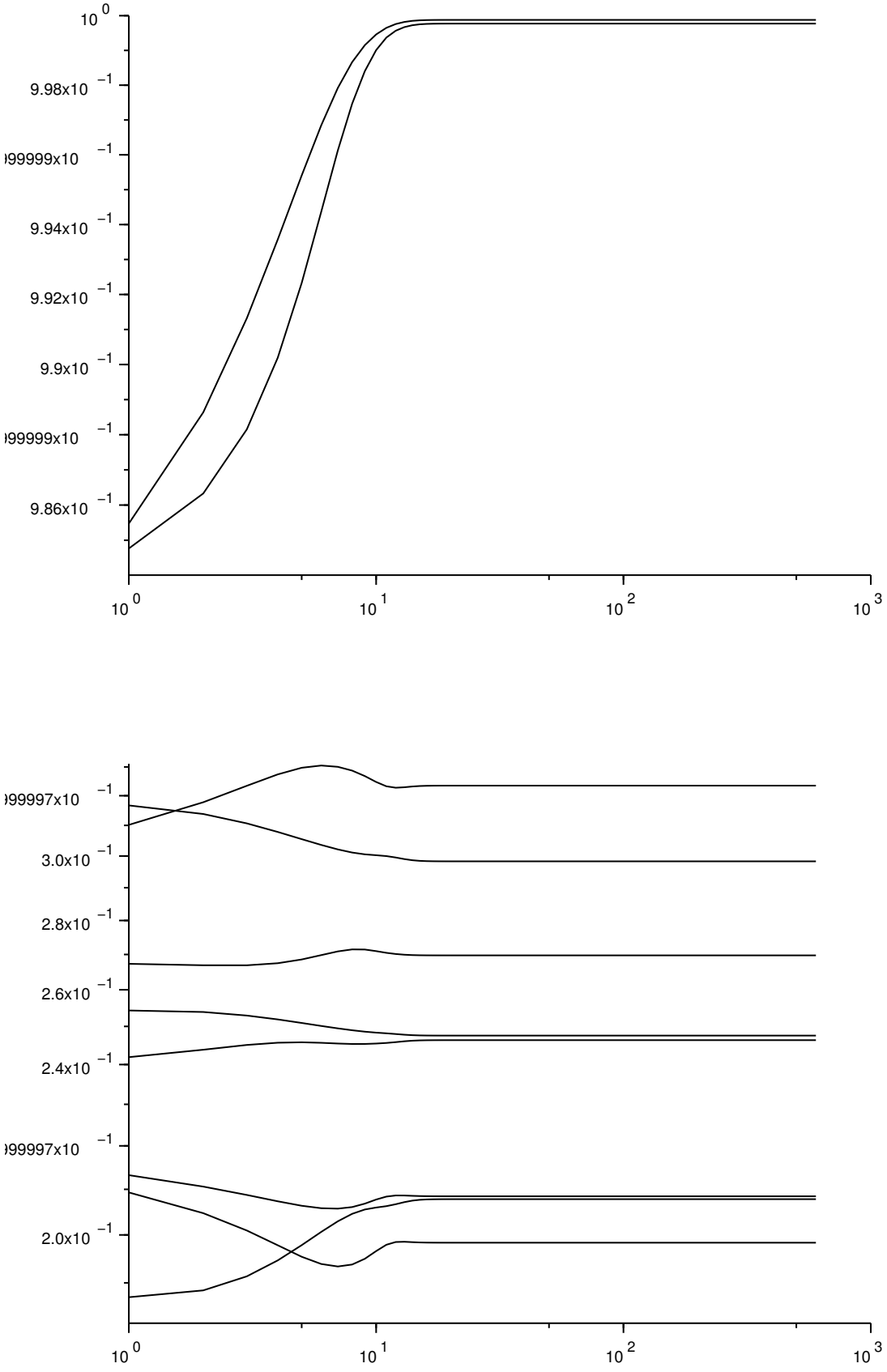

FIGURE 7 – Évolution en échelle log-log des coefficients de la matrice a (en haut) et de la matrice b (en bas).

En comparant aux résultat de la page 159 on conclue que le résultat est satisfaisant.