Statistical learning for OTU identification and biodiversity characterization

Abouabdallah Mohamed Anwar
Directeur : Olivier Coulaud, Alain Franc, Nathalie Peyrard.

Inrae, UMR BioGeCo, Pierroton & EPC Inrae/Inria Pleiade Talence, France

March 31, 2022

Dans le cadre de mes travaux de thèse, j'étudie les modèles à blocs stochastiques qui sont des modèles de plus en plus populaire dans l'apprentissage statistique. J'ai utilisé ces modèles pour du barcoding sur des données issues d'une parcelle botanique. J'ai ensuite adopté une approche basée sur l'algèbre tensorielle afin d'améliorer les performances de ces modèles.

1 Introduction :

Les espèces sont la première unité de mesure de la biodiversité, l'une des manières de les retrouver est le barcoding qui repose sur l'utilisation de leurs séquences ADN. A partir de ces dernières, on peut les retrouver dans un échantillon pour lesquels on dispose des séquences et les distances entre séquences en procédant par une classification non supervisée (clustering) pour regrouper des individus aux séquences très proches qu'on appel OTUs (Operational Taxonomic Unit), c'est-à-dire regrouper n individus en Q classes à partir d'une matrice de distances deux à deux entre individus (pairwaise distance). Parmi les modèles de classification non supervisée disponibles, il y a les modèles graphiques où chaque sommet serait considéré comme une variable aléatoire et le modèle graphique décrit une loi ou distribution jointe entre ces sommets comme un produit de facteurs, chacun n'impliquant qu'un petit nombre de ces variables.

2 Modèle :

Nous nous intéresserons à un type particulier de modèles graphiques : Les modèles à bloc stochastiques (stochastic block model SBM) et nous souhaitons trouver $Z = (Z_1, \ldots, Z_n)$ le vecteur des classes des individus.

Paramètres :

- Un vecteur $\alpha = (\alpha_1, \ldots, \alpha_Q)$ les proportions des individus dans les classes.
- Une matrice de connectivité $\in \mathbb{R}^{Q \times Q}$ entre les classes construite selon le modèle : Pour un SBM Poisson (respectivement SBM Gaussien): Λ de coefficients $\lambda_{kk'}$ (respectivement μ de coefficients $\mu_{kk'}$) le paramètre d'une loi de Poisson (respectivement gaussienne)d'avoir une distance d entre deux individus appartenant respectivement à la classe k et $k^{'}$

Hypothèses :

- H_1 : Les distances sont indépendantes sachant les individus.
- H_2 : Les variables $z_{i,i=1,\ldots,n}$ sont iid sur $\{1, \ldots Q\}$.

3 Comparaison entre la classification botanique et moléculaire

Nous avons utilisé ces deux types de modèles pour faire des classifications non supervisées sur des tableaux de

distances entre séquences d'arbres issus d'une parcelle botanique.
Nous avons comparé les résultats issus des modèles SBM avec la classification ascendante hiérarchique avec trois critères d'agrégation ainsi que la classification botanique sur différents niveaux taxonomiques via un indice basé sur l'information mutuelle.

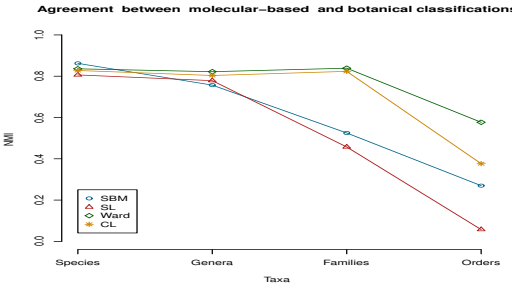

Figure 1: *Evolution de l'information mutuelle par rapport au niveau taxonomique[2]*

4 Approche de type tensor train pour inférer des modèles SBM

Dans un second temps, nous nous sommes intéressés à la méthode d'inférence des paramètres des modèles SBM : Les classes des individus étant les variables latentes du modèle, alors l'inférence des paramètres se fait par maximum de vraisemblance. Une manière naturelle pour le faire est de procéder par l'algorithme expectation-maximization (EM).
La difficulté de cet algorithme vient de l'étape E qui nécessite le calcul de Q^{n-1} sommes. Afin de palier à cette difficulté, il existe deux familles d'approche : Les approches basées sur le champs moyen (VEM) qui sont rapides et les approches de Monte-Carlo (MCMC) qui sont plus précises.
Nous présentons une approche où l'approximation champ moyen est remplacée par une approximation de rang faible de type tensor train TT[3]. Notre but est de remplacer l'approximation champ moyen qui est une approximation TT de rang 1 avec la distance de Kullback-Leibler par une approche où on peut faire varier le rang entre 1 et Q. Nous pensons qu'elle devrait être plus précise au sens de la norme de Frobenius.
Nous avons choisi ce format car il se prête facilement à la marginalisation et pour une distribution de probabilité jointe, chaque coefficient de $\mathbf{\Psi}$, le tenseur de cette loi peut s'exprimer comme un produit de matrices où chacune ne dépend que d'un mode c'est à dire d'une seule variable du modèle SBM. Ainsi :

$$\forall z_1, ..., z_n, \mathbf{\Psi}[z_1, ..., z_n] = A_1[z_1].A_2[z_2]...A_{n-1}[z_{n-1}].A_n[z_n]$$

où les $A_i[z_i]$ représentent les "cores" de Ψ (matrices). Grâce à cela, le calcul des marginales s'exprime très facilement parce que les variables z_i sont séparées. Dans Novikov et al[4], il a été démontré qu'on peut exprimer la décomposition TT de la loi jointe d'un modèle graphique à partir des TT formats de chacun des facteurs. Soit Ψ une fonction qui exprime un modèle graphique,

$$\Psi[z_1, \ldots, z_n] = \Pi_{\ell=1,n}\psi_{A_\ell}(Z)$$

Dans l'approche de Novikov on calcul la décomposition par une SVD des facteurs :

$$\psi_{ij} = G_i^{i,j}[z_i]G_j^{i,j}[z_j]$$

On rajoute les dimensions non essentielles comme scalaires et matrices identités :

$$\psi_{ij}[z_i, z_j] = \underbrace{G_1^{i,j} \times \ldots \times G_{i-1}^{i,j}}_{1} \underbrace{G_i^{i,j}[z_i]}_{R^{1\times Q}} \underbrace{G_{i+1}^{i,j} \times \ldots \times G_{j-1}^{i,j}}_{\mathbb{1}_Q} \underbrace{G_j^{i,j}[z_j]}_{R^{Q\times 1}} \underbrace{G_{j+1}^{i,j} \times \ldots \times G_n^{i,j}}_{1}$$

Puis faire la construction des TT-matrices $\mathbf{A_i}[\mathbf{z_i}]$ de la sorte :

$$\mathbf{A_i}[\mathbf{z_i}] = G_i^{1,2}[z_2] \cdot G_i^{1,3}[z_2] \cdot \ldots \cdot G_i^{n-1,n}[z_2]$$

$B_i = \sum_{z_i} A_i[z_i]$
Les marginales s'expriment comme produits de TT-matrices

$$p_{i,j}(z_i, z_j) = B_1 \times \ldots \times B_{i-1}A_i[z_i]B_{i+1} \times \ldots \times B_{j-1}A_j[z_j]B_{j+1} \times \ldots \times B_n$$

Les TT-matrices $A_i[z_i]$ sont de TT-rang 1 et les TT-matrices B_i sont de TT-rang Q. Ainsi, l'un des problèmes qui apparaissent lors du produit au dessus est l'augmentation exponentielle des TT-rang. Afin de comprendre cela, nous avons développé un dictionnaire qui fait le lien entre les structures algébriques et informatiques des TT-matrices.
De plus, pour diminuer l'augmentation du TT-rang nous faisons une opération qui s'appelle le rounding qui consiste à faire autant de QR et d'orthogonalisation que de cores qui compose les TT-matrices pour calculer une approximation de la TT-matrice avec un TT-rang plus bas. Cela nous a poussé à adopter plusieurs approches comme la fusion des cores, faire un rounding avec un TT-rang fixé et adopter des approches scalables pour le calcul des marginales.

Enfin, cette approche du calcul des marginales binaires d'une loi jointe peut s'étendre au delà du SBM, et s'appliquer à tout modèle graphique.

References

[1] J.J. Daudin et al, *A mixture model for random graphs*, Stat Comput (2008)

[2] MA. Abouabdallah, A. Peyrard N, A. Franc, *Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study.*, Mol Ecol Resour, (2022) 276–287.

[3] I. Oseledets, *Tensor-Train decomposition*, SIAM J. Comput.(2011)

[4] A. Novikov, A. Rodomanov, A. Osokin, D. Vetrov, *Putting MRFs on a Tensor Train*. Proceedings of the 31 st ICML, Beijing, China, 2014