Inférence d'un modèle SBM par une approche de type tensor train

M.A. Abouabdallah[1,2], A. Franc[1,2], O. Coulaud[3], N. Peyrard[4]

[1] *BIOGECO, INRAE, Univ. Bordeaux, 33612 Cestas, France*
[2] *Pleiade, EPC INRIA–INRAE–CNRS, Univ Bordeaux 33405 Talence France*
[3] *HiePACS - INRIA Bordeaux - Sud-Ouest, 33405 Talence, France*
[4] *INRAE, UR MIAT, F-31320 Castanet-Tolosan, France*

Un modèle graphique est un modèle statistique qui décrit une loi ou distribution jointe entre possiblement un très grand nombre n de variables aléatoires comme un produit de facteurs, chacun n'impliquant qu'un petit nombre de ces variables. Nous nous intéressons à l'inférence d'un type particulier de modèle graphique : les modèles SBM pondérés (Weighted Stochastic Block Model, [4]) sur un tableau de distances. Etant donné n individus et B classes, un tel modèle est défini par deux paramètres : un vecteur $\alpha = (\alpha_1, \ldots, \alpha_k)$ où α_k est la probabilité qu'un individu choisi au hasard soit dans la classe k, et une matrice $B \times B$ de coefficients $\lambda_{kk'}$ tels que la distribution de la distance entre deux individus appartenant respectivement à la classe k et k' suive une loi de Poisson de paramètre $\lambda_{kk'}$. Les classes des individus sont les variables latentes du modèle. L'inférence des paramètres par maximum de vraisemblance connaissant une réalisation des distances est un problème difficile qui se résoud classiquement par un algorithme de type EM. La difficulté vient de ce que, dans l'étape E, il faut calculer les marginales unaires et binaires d'une distribution à n variables. Pour cela, dans [3] les auteurs ont développé une approche variationnelle de type champ moyen. Cette approche dite VEM est plus rapide et moins précise que l'approche MCEM. Ici, nous présentons une approche où l'approximation champ moyen de la loi jointe conditionnelle des états latents est remplacée par une approximation de rang faible de type tensor train (TT, voir [6, 7] pour la décomposition de type TT). L'approximation champ moyen est l'approximation TT de rang 1 avec la distance de Kullback-Leibler. Aussi, même si nous travaillons avec la norme de Frobenius, nous espérons une meilleure précision.

A chaque loi jointe de n variables aléatoires discrètes chacune prenant ses valeurs dans un ensemble Λ à B valeurs nous pouvons associer un tenseur à n modes où la dimension de chaque mode est B. Nous avons choisi le format TT d'approximation de ce tenseur car il se prête facilement à la marginalisation. Dans l'approximation champ moyen, ou de rang un, chaque coefficient du tenseur est le produit de ses marginales unaires. Dans l'approche TT, il est un produit de matrices, chacune ne dépendant que d'un mode (une variable du modèle graphique). On peut ainsi écrire, pour une approximation de rang r d'un tenseur à trois modes dont les indices sont associés aux variables (z_i, z_j, z_k) : $\mathbf{A}[z_i, z_j, z_k] \approx u(z_i).G(z_j).v(z_k)$, où $u(z_i)$ est un vecteur $1 \times r$, $G(z_j)$ une matrice $r \times r$ et $v(z_k)$ un vecteur $r \times 1$. La distributivité de la multiplication par rapport à l'addition dans l'anneau des matrices permet de calculer facilement la fonction de partition et les marginales unaires d'un tenseur en format TT, de la même façon que pour l'approche champ-moyen. La difficulté est d'écrire la loi jointe d'un modèle graphique dans le format TT, ou une approximation : vues ses dimensions, B^n coefficients, le tenseur ne rentre pas en mémoire et il est impossible d'appliquer l'algorithme TT-SVD d'Oseledets dès que n est significatif. Récemment, dans [5], les auteurs ont réussi le tour de force de produire exactement (résultat algébrique) la décomposition TT de la loi jointe d'un modèle graphique à partir de la décomposition TT de chacun de ses facteurs, qui est elle accessible (peu de variables). Le prix à payer est que cela implique la manipulation de matrices de taille immense (r^n). Le deuxième tour de force de ce groupe est d'avoir construit une librairie de calcul matriciel où toutes les opérations élémentaires

(somme, produits, produit de Hadamard, etc ...) peuvent être réalisés dans le format TT, sans écrire les matrices en mémoire. Ces éléments permettent un progrès significatif dans l'inférence des modèles graphiques.

L'approche de Novikov & al. permet d'écrire la loi jointe d'un modèle graphique à n variables comme $\Psi(z_1, \ldots, z_n) \approx A_1(z_1) \times \ldots \times A_n(z_n)$ où les $A_i(z_i)$ sont des matrices. Si on note $B_i = \sum_{z_i} A_i(z_i)$, alors une marginale $m-$aire se calcule comme un produit de matrices B_i (les sommes sur les autres variables) et $A_i(z_i)$ (les variables des marginales à calculer). Notre contribution est alors de montrer que toutes les marginales $m-$aires peuvent être calculées à partir des matrices C_{ij} où $C_{ij} = B_i B_{i+1} \ldots B_{j-1} B_j$. Nous proposons pour cela une méthode de programmation dynamique, en écrivant un arbre de dépendance des calculs intermédiaires où un nœud est une matrice C_{ij} et une arête un produit matriciel. Un parcours de ce graphe livre le calcul de toutes les matrices C_{ij}, en $n(n-1)/2$ produits matriciels. Les marginales $m-$aires se calculent alors par différents assemblages.

Dans un modèle SBM, les facteurs de la loi conditionnelle des variables latentes sont unaires ou binaires. L'approximation TT des facteurs se réduit alors à une simple SVD, exacte, ou approchée au rang r. Nous présenterons la mise en œuvre du calcul des marginales binaires cette loi conditionnelle. Puis nous montrerons comment utiliser ce calcul pour développer un algorithme EM approché, que nous appelons TT-EM, intégrant cette approximation des marginales dans l'étape E de l'algorithme EM. Nous présenterons la boite python associée que nous développons, en cours de finalisation. Dans un premier temps, nous évaluerons la qualité de l'inférence du TT-EM sur des jeux de données simulées comme réalisation d'un modèle SBM, en comparant notamment au VEM. Puis nous appliquerons la méthode à un jeu de données test composé d'un tableau de distances moléculaires entre 1500 arbres de Guyane [2, 1].

Références

[1] A. M. Abouabdallah, N. Peyrard, and A. Franc. Evaluating the adequacy between morphological-based and molecular-based inventories at high taxonomic level. *Submitted*, 2021.

[2] H. Caron, J.-F. Molino, D. Sabatier, P. Léger, P. Chaumeil, C. Scotti-Saintagne, J.-M. Frigério, I. Scotti, A. Franc, and R. J. Petit. Chloroplast DNA variation in a hyperdiverse tropical tree community. *Ecology and Evolution*, 9(8) :4897–4905, 2019.

[3] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18 :173–183, 2008.

[4] C. Lee and D.J. Wilkinson. A review of Stochastic Block Models and extensions for graph clustering. *Applied Network Science*, 4(122), 2019.

[5] A. Novikov, A. Rodomanov, A. Osokin, and D. Vetrov. Putting MRFs on a Tensor Train. In *Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, 2014.*, 2014.

[6] I. V. Oseledets. A new tensor decomposition. *Dokl. Math.*, 80(1) :495–496, 2009.

[7] I. V. Oseledets. Tensor-Train decomposition. *SIAM J. Comput.*, 33 :2295–2317, 2011.