Exploration de la puissance de l'approche par graphe disjoint de l'analyse des données compositionnelles

Jules BERTIN ; Emmanuel CURIS
Laboratoire de biomathématiques, UR 7537 BioSTM, faculté de pharmacie de Paris, université Paris Cité

De plus en plus d'études sont fondées sur des données compositionnelles : données décrivant la composition d'une population. Ces données peuvent résulter soit de la nature même des données, soit, plus fréquemment, d'une étape de normalisation des données (typiquement en transcriptomique ou en métabolomique).

Les données compositionnelles sont caractérisées par l'existence d'une forte contrainte : la somme des variables associées à chaque constituant de la population est nécessairement égale à 1 (100 %) — ou tout autre valeur, arbitraire, correspondant à la population dans son ensemble (par exemple, la masse totale d'ARN ou de protéines traitées ; la surface totale de l'image analysée…). Cette contrainte induit des corrélations naturelles entre ces variables. Par ailleurs, on ne peut s'intéresser pour elles qu'aux variations relatives de ces variables entre elles et non à leur variation absolue.

Plusieurs méthodes ont été proposées pour traiter ce type de données. Parmi elles, la méthode des graphes disjoints permet de classer les différents constituants d'une composition en groupes évoluant de façon similaire en fonction d'un facteur externe [1]. Pour réaliser ce classement, cette méthode contrôle le risque de voir apparaître à tort des graphes disjoints, sur la base donc d'un test « H0 : le graphe est connexe » (un seul groupe) contre « H1 : le graphe est disjoint » (deux groupes ou plus).

Si, dans cette approche, le risque de première espèce, α, est bien définie, les choses se gâtent lorsque l'on aborde la notion de puissance. En effet, contrairement aux tests usuels qui font intervenir des hypothèses continues et que l'on peut ordonner, ici les hypothèses sont discrètes et il n'est pas évident de les ordonner de façon naturelle.

Ce travail présente les questions que soulève cette difficulté. Dans un premier temps, nous discuterons de la notion de puissance pour des hypothèses fondées sur des structures de graphes, conduisant à l'introduction de nouveaux indicateurs pour affiner la notion de puissance. Ces notions seront ensuite illustrées au travers de l'étude de graphe à 3 sommets (compositions ternaires).

Références

[1] Emmanuel CURIS, Cindie COURTIN, Pierre Alexis GEOFFROY, Jean-Louis LAPLANCHE, Bruno SAUBAMÉA, Cynthia MARIE-CLAIRE. « Determination of sets of covariating gene expression using graph analysis on pairwise expression ratios », *Bioinformatics*, vol. 35 n° 2, janvier 2019, p. 258-265.