Approche en Réseaux de tableaux de données

Mohamed Hanafi[1], Julien Boccard[2] et Serge Rudaz[2]
[1] Oniris, INRAE, StatSC, 44300 Nantes, France
[2]School of Pharmaceutical Sciences, University of Geneva
Mohamed.Hanafi@oniris-nantes.fr, Benoit.Jaillais@inra.fr,
Julien.Boccard@unige.ch, Serge.Rudaz@unige.ch

Keywords: Données multiblocs, NetPCA.

1 Introduction

Les études qui consistent à caractériser un ensemble d'échantillons selon différents procédés : chimiques, physiques, structuraux (plusieurs techniques spectroscopiques par exemple, infrarouge, fluorescence, etc…) sont nombreuses. Ces acquisitions multiples traduisent une volonté de l'expérimentateur de prendre en compte le mieux possible la complexité des échantillons étudiés dans le but de mieux cerner leurs propriétés à différentes échelles. Les acquisitions multiples conduisent à des données multivariées consignées dans différents tableaux (données multiblocs) qui se trouvent appariés par les individus (Fig. 1.1), par les variables (Fig. 1.2), ou par les deux à la fois (Fig. 1.3). On entend ici par appariement de deux tableaux par les individus (ou par les variables) le fait que les entrées représentantes des lignes c'est-à-dire les individus (ou que les entrées des variables) des deux tableaux sont identiques en nombre et en nature.

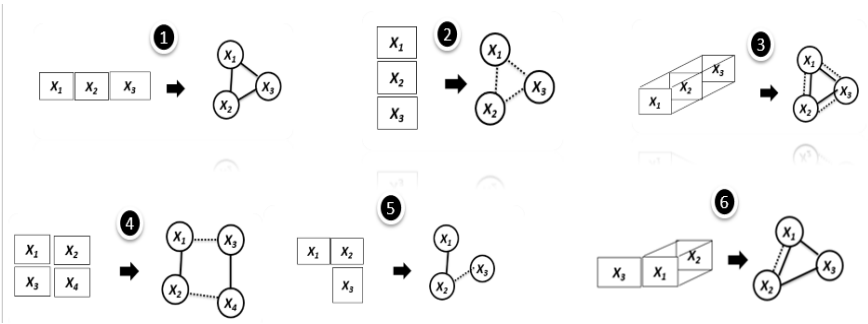

Figure 1. Polymorphismes des données multi-blocs versus réseaux entre tableaux de données.

.

2 Motivation

Les travaux en chimiométrie relatifs à l'analyse des données représentées dans la Figure 1 se sont développés d'une manière significative lors des trente dernières années conduisant à un paysage de méthodes à la fois riche et varié.

Ces méthodes supposent un appariement total de l'ensemble des tableaux à analyser. Cette condition peut s'avérer très restrictive et rendre l'application de ces méthodes impossible dans de nombreuses situations. On cite à titre d'exemple, les données structurées en groupes d'individus et en groupes de variables (Fig. 1.4), et les données organisées sous d'autres formes (Fig. 1.5 et 1.6).

Contrairement aux tableaux représentés dans la Figure 1 (1, 2 et 3), les tableaux représentés dans la Figure 1 (4, 5 et 6) ne sont pas totalement appariées (ne se partagent pas tous la même entrée), ces tableaux sont dit partiellement appariés.

3. Contribution

Pour décrire une collection de tableaux appariés partiellement ou totalement par les lignes ou par les colonnes, la présente communication introduit la notion de réseaux de tableaux [1,2]. Il s'agit d'une collection de tableaux dont la relation d'appariement entre tableaux est modélisée par des graphes. Comme illustrée dans la figure 1, la notion de réseaux entre tableaux semble convenir à décrire différentes formes de de données multi-blocs. Ainsi, cette notion est générique, elle hérite cette généricité de la notion de graphe qui constitue son fondement. La notion de modèle d'un réseau entre tableaux de données consiste en une décomposition de chaque bloc (sommets) formant le réseau entre tableaux (figure 2). Cette décomposition rappelle la Décomposition en Valeurs Singulières d'une matrice qui est au cœur de l'Analyse en Composantes Principales.

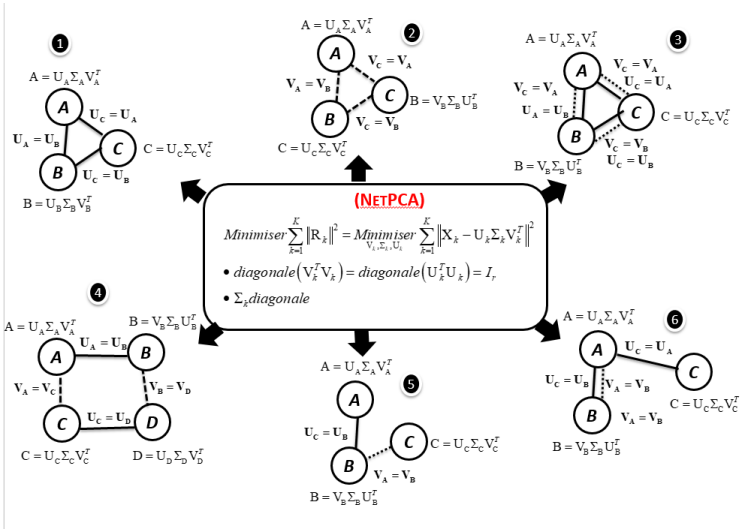

Figure 2. Modèle d'analyse d'un réseau entre tableaux

L'estimation du modèle de réseaux entre tableaux peut se faire par la minimisation des sommes des carrées des résidus. Ce problème peut être résolu par un algorithme de type moindres carrés alternés: NetPCA. L'application de l'approche en réseaux est illustrée sur des données métabolomiques acquises dans le cadre d'une étude sur la déficience rénale chronique [1] Les données exploitables se présentent sous forme d'un tableau structuré en trois groupes d'individus (Stade de la maladie) et deux groupes de variables (métabolites identifiées ou annotés). En plus de la double partition du tableau de données, les blocs ne sont pas du même ordre. Le nombre d'entrée des blocs associés au troisième groupe d'individus est de 3 alors que pour tous les autres blocs le nombre d'entrée est de 2. Cette hétérogénéité d'ordre des blocs confère aux données à analyser le statut de données complexes dont une modélisation en réseaux entre tableaux est nécessaire pour pouvoir envisager sa réduction de la dimensionnalité.

4. References

[1] Codesido S, Hanafi M, Gagnebin Y, González-Ruiz V, Rudaz S, Boccard J. (2021). Network principal component analysis: a versatile tool for the investigation of multigroup and multiblock datasets. Bioinformatics. 9;37(9):1297-1303.

[2] Boccard J, Schvartz D, Codesido S , Hanafi M, Gagnebin Y, Ponte B, Jourdan F, Rudaz S. (2022). Gaining Insights Into Metabolic Networks Using Chemometrics and Bioinformatics: Chronic Kidney Disease as a Clinical Model. Frontiers in Molecular Biosciences