"blockberry" : Un package expérimental sous R pour matrices structurées en blocs

Mohamed Hanafi et Gaston Sanchez

Oniris, Unité de Sensometrie et Chimiométrie,

Site de la Géraudière, rue de la Géraudière, BP 82225

Nantes 44322 Cedex 03, France

mohamed.hanafi@oniris-nantes.fr, gaston.stat@gmail.com

Mots clefs : Tableaux multiples, algèbre matricielle, R.

Une des questions majeures en statistique multivariée est l'étude des relations entre plusieurs blocs de données, aussi connue comme " Analyses des données multiblocs ou de tableaux multiples ». Cet ensemble de techniques [1] est utilisé dans divers domaine allant de la chimiométrie [2,3] à l'écologie [4] .

Après près d'un siècle de littérature consacrée aux développements des techniques multiblocs, il apparaît que : le cadre fourni par l'algèbre matricielle et les stratégies de factorisation matricielle [5] ne suffisent pas à décrire toute la richesse, la complexité et la manipulation des matrices structurées. Ceci explique en partie l'absence d'outils logiciels exhaustifs et spécifiquement dédiés aux analyses des données multiblocs. Des packages d'analyse de données tels que ADE 4 ou Factominer [7] permettent de réaliser des analyses multiblocs mais le nombre des méthodes disponibles est très limité.

La contribution de la présente communication est double : (i) proposer un cadre algébrique pour l'analyse des données impliquant une structuration des données en blocs. Il s'agit d'étendre l'algèbre matricielle classique pour le cas des matrices structurées en blocs. (ii) présenter un package R actuellement au stade expérimental ["***blockberry" package : https://github.com/gastonstat/blockberry.***] pour le stockage et la manipulation des matrices en bloc. Les intérêts pratiques de ce package pour une implémentation rapide des algorithmes à la base des méthodes d'analyse des données multiblocs seront démontrés sur la base d'exemples.

1 Algèbre des matrices structurée en blocs.

Définition. Une partition d'un entier **n** est une décomposition de **n** comme somme de **k** entiers tous positifs (nommés parties). **P (n,k)** désigne l'ensemble des toutes les partitions de l'entier **n** en **k** parties.

Définition. Une partition d'un couple d'entiers (n,p) est un couple de partitions (P(n,k), P(p,l)) formés d'une partition de l'entier **n** et d'une partition de l'entier **p**.

Définition . Une matrice structurées en blocs est un couple (X, (P(n,k),P(p,l))) où X est une matrice de dimension (n,p) et (P(n,k),P(p,l)) est un couple de partitions de sa dimension (n,p).

Définition. Par analogie avec la notion de dimension d'une matrice X, le couple de partitions (P(n,k),P(p,l)) est nommé ici bloc-dimension de X considérée comme une matrice structurée en blocs.

Peu étudiées dans littérature, les multiplications des matrices structurées, les multiplications des matrices structurées en blocs sont des notions très importantes car elles sont à la base des implémentations des algorithmes des méthodes multiblocs. Ces multiplications nécessitent de spécifier (i) les blocs qui doivent être multipliés, (ii) la règle de calcul utilisée pour le calcul du bloc résultat. La table 1, présente 3 exemples de multiplication possible entre matrices structurées en blocs par au moins 15 multiplications possibles.

Table 1. Noms et notations of quelques multiplications pour matrices structurées en blocs sont données (colonnes 1 et 2). La définition du bloc résultat est donnée dans la colonne 3. (*u) désigne le produit matriciel usuel (*h) désigne le produit matriciel d' Hadamard, (*s) désigne le produit d'un scalaire par une matrice.

Noms	Notations	definition du bloc
Produit (u,u)	$\mathbf{X} = \mathbf{A} *_{(u,u)} \mathbf{B}$	$\mathbf{X}_{ij} = \sum_{k=1} \mathbf{A}_{ik} *_u \mathbf{B}_{kj}$
Produit (u,h)	$\mathbf{X} = \mathbf{A} *_{(u,h)} \mathbf{B}$	$\mathbf{X}_{ij} = \sum_{k=1} \mathbf{A}_{ik} *_h \mathbf{B}_{kj}$
Produit (u,s)	$\mathbf{X} = \mathbf{A} *_{(u,s)} \mathbf{B}$	$\mathbf{X}_{ij} = \sum_{k=1} a_{ik} *_s \mathbf{B}_{kj}$

2. Un package R expérimental pour l'algèbre des bloc-matrices

Visant l'implémentation rapide des algorithmes multiblocs, le premier défi auquel nous avons dû faire face est de savoir comment traduire le concept matrice structurée en blocs dans R ?
R offre différentes solutions pour le stockage et la manipulation de données (vecteurs , matrices , des tableaux, des trames de données , les facteurs et les listes). Le problème est qu'il n'y a aucune structure de données conçu spécifiquement pour représenter le concept de matrice structurée en blocs. Pratiquement, tous les algorithmes à la base des méthodes statistiques mentionnées en [1,2] font recours aux matrices structurées en blocs et parfois même des tenseurs structurés en blocs.
D'un point de vue opérationnel, les concepts (bloc-dimension, matrices structurées en blocs, multiplications entre matrices structurées en blocs) sont traduits en différentes classes qui permettent le stockage des matrices structurées en blocs ainsi que nombreuses méthodes génériques pour les manipulations qui peuvent être utilisées pour le prototypage rapide des algorithmes. Ces classes étendent la fonctionnalité de R. La présente communication propose une description détaillée de l'ensemble de ces classes et leur utilisation pou mettre en œuvre des algorithmes multiblocs.
Ce travail a été réalisé dans le cadre du projet régional «Approches Intégratives du déterminisme structurel, génétique et écophysiologique de la qualité des fruits". Les auteurs remercient la Région « Pays de la Loire » pour son appui financier de ce travail.

References

[1] Hanafi, M., Kiers, H.A.L. (2006). Analysis of K sets of data, with differential emphasis on agreement between and within sets". Computational Statistics and Data Analysis. 51, 3, pp. 1491-1508,

[2] Hanafi, M., Mazerolles, G., Dufour, E., Qannari, E. M. (2006). Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques. Journal of Chemometrics. 20, (5), pp. 172-183.

[3] Kohler, A. Hanafi, M., Bertrand, D., Janbu A.O., Møretrø T., Naterstad, K., Qannari, E.M., Martens, H. (2007). Interpreting several types of measurements in bioscience. Modern concepts in biomedical vibrational spectroscopy, Lasch, P. ISUP, Inc. dba Blackwell Publishing England.

[4] Chessel, D., Hanafi. M. (1996) Analyses de la co-inertie de K nuages de points. Revue de Statistique Appliquée. XLVI, pp.35-60.

[5] Golub, G.H., Van Loan, C.F. (1996).. Matrix Computations. The Johns Hopkins University Press, Baltimore, MD.