Une famille de matrices sparses pour une modélisation multi-échelle par blocs

Camille Brunet*, Thomas Villman**
Vincent Vigneron*

*IBISC, Université d'Evry Val d'Essonnes
40 rue du Pelvoux - 91020 Evry Courcouronne - France
camille.brunet@ibisc.univ-evry.fr
**University of applied sciences Mittweida
Technikumplatz 17 - 09648 Mittweida - Germany

Résumé. La sériation est une technique d'analyse de données qui ordonne les observations directement à partir de leur tableau de valeurs afin de révéler une structure intrinsèque à ces données. Une telle approche présente de nombreux avantages de visualisation mais dès lors que les données sont bruitées ou que les groupes se superposent, la visualisation de toute structure devient difficile. Pour faire face à ces problèmes, nous introduisons de la parcimonie dans les données à travers une famille de matrices indicatrices de voisins communs. Celles-ci sont ordonnées selon un algorithme de type *branch and bound* et la matrice révélant la meilleure structure au sens de "diagonale par blocs" est sélectionnée au moyen d'un critère dérivé des problématiques de compression de données. Cet outil de partitionnement identifie des sous-ensembles de données relatifs aux *clusters* tout en écartant celles qui sont bruitées ou extrêmes ce qui permet de visualiser la structure globale intrinsèque aux données. Cependant, une trop grande sparsité des données amène parfois à l'éviction de données sous-représentées; nous proposons à cet effet, une approche multi-échelle combinant différents niveaux de sparsité dans une même visualisation.

1 Introduction

La notion de complexité des données peut s'appréhender de différentes manières dans la littérature. Elle peut en effet être liée à la nature et à l'hétérogénéité des données, à leur dimension plus ou moins grande, à leur qualité (données très bruitées ou non) ou encore à leur structure globale qui peut être telle que la distinction de groupes dans les données est rendue difficile. Les outils statistiques d'analyse de données ont pour objectif d'extraire de l'information. Ils cherchent à explorer des données multidimensionnelles pour les synthétiser d'une part et permettre de structurer l'information contenue dans ces données d'autre part. Dans la littérature, il existe plusieurs approches associées à cette démarche telles les méthodes factorielles (Pearson (1901); Jolliffe (1986)), les approches par classification non supervisée –