Soutenance d'HDR de Véronique Cariou (unité StatSC, ONIRIS, INRAe, Nantes)

Véronique Cariou, maître de conférences au sein de l'unité Statistique, Sensométrie et Chimiométrie, soutiendra son Habilitation à Diriger des Recherches (HDR) intitulée «Approches factorielles et classificatoires pour l'analyse des données multiblocs. Applications en sensométrie et en chimiométrie.» / «Multibloc data analysis with factorial and clustering methods applied to sensometrics and chemometrics.»

mardi 30 juin 2020 à 15h30 en visio.

Jury :

Rapporteurs :
– M. Hervé ABDI, Professeur, Université du Texas
– M. Mohamed NADIF, Professeur, Université Paris-Descartes
– M. Arthur TENENHAUS, Professeur, Centrale-Supélec

– Mme Pascale KUNTZ-COSPEREC, Professeur, Polytech'Nantes
– M. Serge RUDAZ, Professeur, Université de Genève
– M. Pascal SCHLICH, Directeur de Recherche, CSGA
– Mme Evelyne VIGNEAU, Professeur, Oniris

Résumé. Les travaux de recherche, présentés dans le cadre de mon projet d'Habilitation à Diriger des Recherches, s'inscrivent en analyse des données multiblocs avec une optique de réduction de la dimensionnalité par des méthodes factorielles et classificatoires à base de composantes latentes. Ces méthodes répondent à des problématiques rencontrées en sensométrie (traitement des données issues d'évaluation sensorielle), chimiométrie (traitement des données issues de capteurs) et pour l'analyse des données –omiques. Les développements récents en évaluation sensorielle et les avancées analytiques ont en effet généré un intérêt croissant pour la mise en relation de plusieurs tableaux, appariés sur un voire deux modes, dans une optique d'une meilleure compréhension des phénomènes biologiques et biochimiques. Dans ce contexte, mes travaux se sont concentrés sur deux structures de données : (i) multibloc, où un ensemble d'observations est décrit par différents blocs de variables et (ii) trois-voies, où un ensemble d'observations est décrit par un même ensemble de variables mesurées suivant un troisième paramètre, comme le temps. Les méthodes proposées s'articulent autour de : (i) une approche géométrique avec la détermination de composantes latentes pour les sorties graphiques, l'aide à l'interprétation, la prédiction etc.(ii) une approche tensorielle pour la présentation des critères à optimiser et (iii) des algorithmes séquentiels de détermination successives des composantes, à l'instar de la régression PLS, et des algorithmes de type moindres carrés alternés. Ces développements sont pour partie intégrés dans les packages R ClustVarLV et ClustBlock. Mes contributions méthodologiques s'appuient sur deux méthodes antérieurement proposées à StatSC : (i) l'Analyse en Composantes Communes et Poids Spécifiques (renommée ComDim par la suite) et (ii) la Classification autour de Variables Latentes, CLV. Du point de vue factoriel, mes travaux ont porté sur l'extension de ComDim au cadre supervisé, avec PComDim et à la situation où il existe un graphe orienté de relations entre les blocs, avec PathComDim. Dans le cadre classificatoire, je me suis intéressée à l'extension de CLV pour le partitionnement d'un ensemble de tableaux de données avec CLUSTATIS et à la segmentation d'un cube de données suivant l'un de ces modes avec CLV3W. Ces stratégies présentent des analogies avec les approches Clusterwise. Enfin, l'intérêt de l'analyse des données multiblocs est illustré par

quelques applications, que cela soit l'identification de bio-marqueurs en lien avec la croissance pondérale de nourrissons dans un contexte de prématurité, la structuration d'un lexique des termes d'odeurs du vin ou encore la segmentation de consommateurs sur la base d'émotions suscitées par des odeurs de cafés.

Mots-clés : données multiblocs, données trois-voies, composantes latentes, régression PLS, sensométrie, chimiométrie

Summary :

The research work, presented for my Habitation à Diriger des Recherches project, fits within the scope of multiblock data analysis with the aim of a dimensionality reduction by means of factorial and clustering strategies based on latent components. These methods address some issues encountered in sensometrics (analysis of data from sensory evaluation), in chemometrics (analysis of data acquired with sensors) and for the analysis of –omics datasets. Recent developments in sensory evaluation and analytical techniques have more and more led to consider the simultaneous analysis of several data tables, coupled by one or even two modes, providing a more comprehensive view for a better understanding of biological and biochemical phenomena. In this context, my work focuses on two data structures: (i) multiblock, where a set of observations is described by different blocks of variables and (ii) three-way, where a set of observations is described by the same set of variables measured according to a third parameter, such as time or occasions. The proposed methods aim at : (i) providing geometrical tools with latent components for interpretation, graphical display, prediction etc. (ii) adopting a tensorial approach for the associated criteria to be optimized and (iii) implementing sequential algorithms for the determination of the various components, as in PLS regression, and alternating least squares algorithms. Part of these developments are integrated into the R packages ClustVarLV and ClustBlock. Two methods, originated from StatSC, are the cornerstone of my methodological contributions: (i) the Common Components and Specific Weights Analysis (renamed ComDim thereafter) and (ii) the Clustering around Latent Variables, CLV. From a factorial point of view, my work focuses on the extension of ComDim to the supervised framework, with PComDim and to the situation of a path diagram, with PathComDim. In the clustering framework, I propose the extension of CLV for the partitioning of a set of data tables with CLUSTATIS and for the segmentation of an array according to one of its mode with CLV3W. The underlying strategy is closely associated with Clusterwise approaches. Finally, the contributions to multibloc analysis are illustrated by different applications either for the identification of lipidomic biomarkers related to the weight growth of infants in a context of prematurity, for the structuration of a wine odor terms' lexicon or for the consumer' segmentation on the basis of emotional data in a sensory context.

Keywords : multiblock data, three-way data, latent components, PLS regression, sensometrics, chemometrics