Titre : Description et sélection de données en grande dimension

L'évolution des technologies actuelles permet de traiter un grand nombre d'expériences (ou de simulations) et d'envisager un nombre important de paramètres. Cette situation conduit à des matrices de grande, voire très grande, dimension et nécessite le développement de nouveaux outils pour évaluer et visualiser ces données et, le cas échéant, en réduire la dimension. L'évaluation de la qualité de l'information apportée par l'ensemble de points constituant une base de données ou un plan d'expériences peut se faire au travers de critères basés sur des calculs de distance, qui renseigneront sur l'uniformité de la répartition dans l'espace multidimensionnel. Parmi les méthodes de visualisation, l'Analyse en Composantes Curvilignes a l'avantage de projeter des données en grande dimension dans un espace bidimensionnel en préservant la topologie locale, ce qui peut aider à détecter des amas de points ou des zones lacunaires. La réduction de dimension s'appuie sur une sélection judicieuse de sous-ensembles de points ou de variables, via des algorithmes. Les performances de ces méthodes ont été évaluées sur des cas d'étude issus des études QSAR, de la spectroscopie et de la simulation numérique.

Mots clés : données en grande dimension ; plans d'expériences ; simulation numérique ; critères intrinsèques ; Analyse en Composantes Curvilignes ; algorithme WSP

Title: Description and selection of high-dimensional data

Technological progress has now made many experiments (or simulations) possible, along with taking into account a large number of parameters, which result in (very) high-dimensional matrix requiring the development of new tools to assess and visualize the data and, if necessary, to reduce the dimension. The quality of the information provided by all points of a database or an experimental design can be assessed using criteria based on distances that will inform about the uniformity of repartition in a multidimensional space. Among the visualization methods, Curvilinear Component Analysis has the advantage of projecting high-dimensional data in a two-dimensional space with respect to the local topology. This also enables the detection of clusters of points or gaps. The dimensional reduction is based on a judicious selection of subsets of points or variables, via accurate algorithms. The performance of these methods was assessed on case studies of QSAR, spectroscopy and numeric simulation.

Keywords: high-dimensional data ; experimental designs ; numerical simulation ; intrinsic criteria ; Curvilinear Component Analysis ; WSP algorithm