Big data in multi-block data analysis: An approach to paralleling classic PLS algorithm

Alba Martinez-Ruiz*
Universidad Católica de la Santísima Concepción, Concepción, Chile - amartine@ucsc.cl

Cristina Montañola-Sales
Universitat Politècnica de Catalunya BarcelonaTech, Barcelona Supercomputing Center, Barcelona, Spain - cristina.montanola@upc.edu

The management and analysis of large data sets have become a fundamental need in several scientific disciplines. Nowadays, the integration of knowledge from different domains is essential for task performance. Early, mathematicians and computer scientists explored methodologies to propose parallel matrix algorithms to optimize computing power and profit large computer systems such as clusters, clouds or supercomputers. With the software-hardware infrastructures increase, examining big data is gradually more feasible. This situation makes the analysis of large volumes of data a major challenge of investigating the performance, efficiency and effectiveness of statistical methods. Multi-block data analysis, consisting of a set of well-established methods, is properly positioned to meet these challenges, since the methods are based on the analysis of components. However, as far as we know, no investigations have addressed these issues yet in the scientific community.

In this communication, we present the parallelization process of the classic partial least squares (PLS) algorithm implemented using an explicit parallelization to distribute large data calculations across a computer cluster. This approach allows a finer-grained control over the job distribution within the computer infrastructure. PLS algorithm was implemented in R and its parallelization was done with the standard high performance library Message Passing Interface (MPI). First, we give a brief literature review on the strategies that are used to parallelize algorithms and how researchers deal with computational complications in the processing of big data sets. Second, we discuss some important aspects of the parallelization process of the PLS algorithm. This work contributes to the understanding of methodological aspects of multi-block methods for the analysis of big data sets, an emerging area of outstanding interest for scientific communities in the short term.

Keywords: Multi-block methods; big data; PLS; parallelization.