Approaching Big Data in Statistics: Parallelization of Classic PLS Algorithm:
Cristina Montañola-Sales (a) (b), Alba Martínez-Ruiz (c)

(a) Universitat Politècnica de Catalunya (b) Barcelona Supercomputing Center (c) Universidad Católica de la Santísima Concepción, Chile

Challenges of Statistics for the Big Data era

Classical statistical methodologies have long demonstrated the powerful capabilities of statistics to analyse data for current policy or business analytics. However, in the era of Big Data how can we approach datasets of Teras or Petas with those techniques? We need new solutions to deal with increasing size in datasets and complex operations.

Distributed and parallel computing can provide a solution to perform complex mathematical operations by dividing the problem in small sizes to apply classical statistical methods.

The current statistical community relies in R as a preferred software. The reason is that R provides a good environment to run statistical methods widely used by the community. Moreover, it allows to publish and make available new packages with new developed techniques. Therefore, there is a need to include solutions in R to handle big data problems. Large problems can be overcome by distributing mathematical calculations across processors, so distributed datasets could be processed with current statistical instruments.

Currently, there are some R packages that can help on distributing R algorithms in High Performance Computing systems. The most important ones are Snow [1], Parallel [2], Rmpi [3] and pbR4 [4].

In this poster, we present our current work-in-progress to use pbXr for large multi-block data analysis. Specifically we propose an approach to parallelize the classic PLS algorithm [3][7]. This algorithm is useful in many applications in areas as diverse as sensometrics, information science and systems, marketing, strategic management and technological change.

High Performance Computing Systems for Big Data

To distribute and processing,
- Clusters of computers, a set of computer nodes interconnected with high-speed networks offering HPC at a relatively low-cost.
- Clouds, computers connected through a real-time communication network, typically the Internet, which provides dynamic and scalable resources as services to the end-user.
- Grids, computers connected through a real-time communication networks as clouds, which require more control by the end user.

Data Parallelism vs. Task parallelism

Big Data in Multi-Block Data Analysis

- Multi-block data analysis consisting of a set of well-established methods, is properly positioned to meet the big data analysis challenge, since the methods are based on the analysis of components.

PLS multi-block analysis
1. Estimation of standardized constructs using a PLS algorithm
2. Endogenous constructs are regressed on the exogenous constructs to estimate effects

To parallelize an algorithm
- Step 1: Choose J arbitrary initial vectors \( \mathbf{v}_j \), \( j = 1, \ldots, J \). To repeat until convergence, \( \mathbf{p} \).
- Step 2: Estimation
  - Step 3: Internal estimation
  - Step 4: Updating
  - Efficiency of a distributed processing system will depend on a proper mathematical design, factoring procedure and computer system.

References