Project title / Job position title:
Ph.D. Fellowship in Applied Mathematics – Applied Statistics

Area of knowledge:
Physical Sciences, Mathematics and Engineering Panel

Group disciplines:
Theoretical and Applied Mathematics, IT

Research Project / Research Group Description:

Background:
Traditionally, the area of Statistics that deals with large amount of data is the field of “High-dimensional Statistics”, which studies data whose dimension is larger than dimensions considered in classical multivariate analysis (Bühlmann and van de Geer (2011). Problems in biology, engineering, medical research, epidemiology, demography, environmental sciences, genetics, etc. are dealing with enormous amounts of data. Besides, in many applications, the dimension of the data vectors (or parameters) may be larger than the sample size, i.e. there are many variables rather than individual observations. All these situations require the development of innovative statistical approaches for analysis. In depth, dealing with large amount of data (whatever its nature and complexity) requires of mathematical tools relying on matrix algebra, numerical analysis, statistical inference, theoretical and practical consistency of the estimators, efficient model parameters estimation, computational feasibility etc. In high-dimensional Statistics a crucial problem is to select among a set of covariates those that influence a response variable. This is also known as sparse regression, given the fact that in this type of problems the solution is sparse (i.e. only a reduce number of coefficients are distinct from zero).


Proposed project: Sparse array regression and variable selection in high-dimensional problems
The aim of the Ph.D. project is the development of new extensions, methods and algorithms of the so-called Generalized Linear Array Methods (GLAM, proposed by Currie et al. 2006 and Eilers et al. 2006) into high-
dimensional complex data problems. The Ph.D. student will study the array methodology in the context of L1-norm regularization methods such as Lasso or a combination of L1 and L2-norms where variable selection and shrinkage is desired. Preliminary results of efficient methods in Lee et al. (2013) and Rodríguez-Álvarez et. al (2015) will be considered as starting approaches. It is also expected the development of computationally-efficient estimation procedures, as well as the implementation of the algorithms in R packages for public use. The obtained results will be applied to multidisciplinary fields such as spatial epidemiology, or genetics, but also in other fields where large amount of data are available and variable selection and sparse regression will benefit the field.

The thesis project will be developed in the group Applied Statistics (AS) http://www.bcamath.org/en/research/lines/AS

**Keywords:** High-dimensional statistics, Computational Statistics, array methods, sparse regression.

**References:**


**Job Position description:**

*Include all the relevant information about the position role, responsibilities and skills required within the project/group*

The Ph.D. student will develop the aforementioned extensions of the Array methodology for High-dimensional statistics problems with special focus on sparse regression and variable selection in large scale problems.

**Requirements:**

- Master degree (preferable in Statistics, Applied Mathematics, Engineering or Computer Science). The candidate must have his/her Master Degree completed before the incorporation.
- Applicants must have an excellent academic record.

Skills:
- Good communication and interpersonal skills.
- Good programming skills in R, Python or Matlab.
- Ability to effectively communicate and present research ideas to researchers with different background (e.g., mathematicians, engineers, biologists, and geneticist).
- Ability to clearly present and publish research outcomes in spoken (talks) and written (papers) form.
- Good command of spoken and written English.

Group Leader:
Dae-Jin Lee, PhD (Mathematics, U. Carlos III, 2010) dlee@bcamath.org