ERCIM "Alain Bensoussan"
Fellowship Scientific Report

Fellow: Cristian Gatu
Visited Location : VTT Technical Research Centre of Finland, Espoo, Finland
Duration of Visit: 01/07/2007 – 30/06/2008

I - Scientific activity

The research during the twelve months of the ERCIM "Alain Bensoussan" fellowship was focused on the investigation of computational methods and exploration techniques able to solve combinatorial statistical model selection problems and on the application of these methods to real biomedical data. Specifically, the following contributions have been brought.

1. A directed graph approach which can be employed in statistical model selection has been proposed. The combinatorial problem of generating all possible subset regression models has been formalized by means of the regression graph. Specifically, the graph provides an optimal computational procedure for deriving the best subset regression models [Gatu:CSDA07].
2. A computationally efficient branch-and-bound algorithm for finding the subsets of the most statistically significant variables of a vector autoregressive (VAR) model was proposed. The candidate submodels were obtained by deleting columns from the coefficient matrices of the original VAR process [Gatu:JEDC08].
3. A fast algorithm for solving the non-negativity constrained model selection problem was proposed. The new algorithm estimates only unconstrained least squares, and in addition it verifies whether the non-negativity restrictions are satisfied. It has its foundation on an alternative approach to quadratic programming that derives the non-negative least squares by solving the normal equations for a number of unrestricted least squares subproblems [Gatu:TR08a].
4. Various algorithms to compute least trimmed squares regression for a range of coverage values were introduced. The algorithms are based on an adding row strategy and allow to efficiently compute and investigate a set of least trimmed squares estimators for distinct breakdown values [Hofmann:TR08].
5. An optimization strategy that aims to identify the best-subset regression models was proposed. The special case of fat-structure data, i.e. when the number of variables exceeds the number of available samples, was considered. The new approach consists of two stages that combine the heuristic and the exhaustive search aiming to reduce the computational time and to obtain quality solutions, respectively. The proposed strategy promised to be an effective statistical model selection method for fat-structure data, which may replace or complement the existing methods [Gatu:COMPSTAT08].

II- Publication(s) during your fellowship

[Gatu:CSDA07] C. Gatu, P. Yanev and E.J. Kontoghiorghes. A graph approach to generate all possible regression submodels. ***Computational Statistics & Data Analysis,*** **52 (2007), pp. 799-815.**

Abstract :
A regression graph to enumerate and evaluate all possible subset regression models is introduced. The graph is a generalization of a regression tree. All the spanning trees of the graph are minimum spanning trees and provide an optimal computational procedure for generating all possible submodels. Each minimum spanning tree has a different structure and characteristics. An adaptation of a branch-and-bound algorithm which computes the best-subset models using the regression graph framework is proposed. Experimental results and comparison with an existing method based on a regression tree are presented and discussed.

[Gatu :JEDC08] C. Gatu, E.J. Kontoghiorghes, M. Gilli and P. Winker. An efficient branch-and-bound strategy for Subset Vector Autoregressive model selection. ***Journal of Economic Dynamics & Control,*** **32 (2008), pp. 1949-1963.**

Abstract :
A computationally efficient branch-and-bound strategy for finding the subsets of the most statistically significant variables of a vector autoregressive (VAR) model from a given search subspace is proposed. Specifically, the candidate submodels are obtained by deleting columns from the coefficient matrices of the full-specified VAR process. The strategy is based on a regression tree and derives the best-subset VAR models without computing the whole tree. The branch-and-bound cutting test is based on monotone statistical selection criteria which are functions of the determinant of the estimated residual covariance matrix. Experimental results confirm the computational efficiency of the proposed algorithm.

[Gatu :TR08a] C. Gatu and E.J. Kontoghiorghes. Regression subset selection with non-negative coefficients. 2008, (To be submitted).

Abstract :
The problem of subset selection of the linear regression model where the regression coefficients are known to satisfy non-negativity constraints is considered. An algorithm that derives the constrained solution by solving a number of unrestricted least squares subproblems is introduced. The method is based on a regression tree structure that generates all possible submodels. The main computational tool is the QR factorization and its modification. The adaptation of a branch-and-bound device that prunes non-optimal subtrees while searching for the best submodels is also described. Experimental results that show the efficacy of the new method are presented and analyzed.

[Hofmann:TR08] M. Hofmann, C. Gatu and E.J. Kontoghiorghes. A least-trimmed-squares algorithm for a range of coverage values, 2008, (Revised version submitted).

Abstract :
A new adding row algorithm (ARA) to compute least trimmed squares regression for a range of coverage values is presented. The ARA employs a tree-based strategy. New nodes are generated by updating the QR decomposition after adding one observation to the linear model. A priori knowledge of the coverage parameter is not required. The ARA can be used to identify the degree of contamination of the data by efficiently computing Least-trimmed-squares estimators for a range of coverage values. A branch-and-bound algorithm (BBA) is designed based on the ARA. The BBA is an exhaustive algorithm that uses a cutting test to prune nonoptimal subtrees. It significantly improves upon the ARA in computational performance. The BBA is further enhanced by preordering the observations. A computationally efficient and numerically stable calculation of the bounds using Givens rotations is devised. The explicite updating of a triangular factor by an observation is avoided. This reduces the overall computational load of the Least-trimmed-squares algorithm by approximately half. The strategy

is illustrated by an example. Experimental results confirm the computational efficiency of the proposed algorithms.

[Gatu:COMPSTAT08] C. Gatu, M. Sysi-aho and M. Oresic. A regression subset-selection strategy for fat data. Proceedings of *International Conference on Computational Statistics*, Porto, Portugal, 24-29 August 2008, (Forthcoming).

Abstract :
A strategy is proposed for finding the most significant linear regression submodel for fat-structure data, that is when the number of variables n exceeds the number of available observations m. The method consists of two stages. First, a heuristic is employed to preselect a number of variables n_S such that $n_S \leq m$. The second stage performs an exhaustive search on the reduced list of variables. It employs a regression tree structure that generates all possible subset models. Non-optimal subtrees are pruned using a branch-and-bound device. Cross validation experiments on a real biomedical dataset are presented and analyzed.

[Gatu:TR08b] C. Gatu, K. Pietiläinen, M. Sysi-Aho, A. Rissanen, P. Gopalacharyulu, T.Seppänen-Laakso, I. Mattila, J. Naukkarinen, L. Peltonen, H.~Yki-Järvinen, J. Kaprio and M.~Oresic. Subset regression of metabolic profiles and pathways in obesity discordant twins, 2008 (Under preparation).

III -Attended Seminars, Workshops, and Conferences

Conference presentation:

1. First Workshop of the ERCIM Working Group on *Computing & Statistics* and 2nd International Workshop on *Computational and Financial Econometrics*, Neuchâtel, Switzerland, June 19 – 21, 2008. (Presentation: "**Regression subset selection with non-negative coefficients**").
2. The 14th International Conference on *Computing in Economics and Finance*, Paris, France, June 26 – 28, 2008. (Presentation: "**Subset Selection of the Linear Regression Model with Constraints**").
3. International Conference on *Computational Statistics* (CompStat'2008), Porto, Portugal, August 24 – 29, 2008. (Presentation: "**A Regression Subset-Selection Strategy for Fat-Structure Data**").

IV – Research Exchange Programme (12 month scheme)

1. February 13 – 27, 2008. Department of Public and Business Administration, University of Cyprus, Cyprus. Host: Prof. Erricos. J. Kontoghiorghes (chairman of the ERCIM working group on "Computing & Statistics").
2. March, 2 – 12, 2008. Computer Science Department, University of Neuchâtel, Switzerland. Host: Prof. Peter Kropf, head of the department.

During the two visits, besides the specific research activities, the aspects related to the organization of the conferences mentioned below have been also considered. The editing of a second special issue on "Statistical Algorithms and Software" of the *Computational Statistics & Data Analysis* journal has been discussed and initiated.

V – Other professional activities

Conference organization:

1. **Co-chair** of the 1st Workshop of the ERCIM Working Group on *Computing & Statistics*, Neuchâtel, Switzerland, June 19 – 21, 2008.
2. **Member of the Scientific Programme Committee** of 5th International Conference on *Computational Management Science*, Imperial College, London, UK, March 26 – 28, 2008.
3. **Member of the Local Organizing Committee** of
    - 1st Workshop of the ERCIM Working Group on *Computing & Statistics*, Neuchâtel, Switzerland, June 19 – 21, 2008.
    - 2nd International Workshop on *Computational and Financial Econometrics*, Neuchâtel, Switzerland, June 19 – 21, 2008.
    - 5th International Workshop on *Parallel Matrix Algorithms and* Applications, Neuchâtel, Switzerland, June 20 – 22, 2008.q

Editorial activities:

C. Gatu and B. D. McCulloughC, guest editors. 2nd Special Issue on *Statistical Algorithms and Software* of *Computational Statistics & Data Analysis*, 2008, (Forthcoming).