SIMULTANEOUS DIMENSION REDUCTION AND VARIABLE SELECTION IN MODELING HIGH DIMENSIONAL DATA

A dissertation presented by

JOSEPH RYAN G. LANSANGAN

to the

School of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy (Ph.D.) in Statistics

School of Statistics
University of the Philippines
Diliman, Quezon City

March 2014

ABSTRACT

High dimensional input in regression is usually associated with multicollinearity and with other estimation problems. As a solution, a constrained optimization method to address high-dimensional data issues is developed. The method simultaneously considers dimension reduction and variable selection while keeping the predictive ability of the model at a high level. The method uses an alternating and iterative solution to the optimization problem, and via soft thresholding, yields fitted models with sparse regression coefficients. Simulated data sets are used to assess the method for both p>>n and n>p cases. Results show that the method outperforms the SPCR, LASSO and EN procedures in terms of predictive ability and optimal selection of inputs (independent variables). Results also indicate that the method yields reduced models which have smaller prediction errors than the estimated full models from the PCR or the PCovR. That is, the method identifies a smaller set of inputs that captures the dimensionality (hidden factors) of the inputs, and at the same time gives the most predictive model for the response (dependent variable).

Keywords: high dimensionality, regression modeling, dimension reduction, variable selection, hidden factors, sparsity, soft thresholding, SPCA

Keywords: spatio-temporal mixed model; small area estimation; backfitting iteration; bootstrap

---