Variable selection in statistical modelling via a numerical linear algebra approach

Marilena Mitrouli

Department of Mathematics, National and Kapodistrian University of Athens,
Panepistemiopolis 15784, Athens, Greece. mmitroul@math.uoa.gr

Abstract

Variable selection requires the minimization of $||X\beta - y||_2$ with respect to β, where $X \in R^{n\times p}$ is the design matrix, $\beta \in R^p$ is a vector of predictors and $y \in R^n$ is the response of the model. The identification of predictors is important for statistical modelling and numerical analysis can bring to statistics community advanced linear algebra techniques for handling this issue. In this work, for a given model $y = X\beta + \epsilon$, where $\epsilon \in R^n$ is the vector of random errors, we study the following problems:

(P1) *Regularization and condition estimation*
It is crucial to decide whether the given model needs regularization or not for the derivation of the vector β. The notion of the effective condition number is introduced, which provides a measure for the stability of β due to a perturbation in y.

(P2) *Fast GCV estimates for correlated matrices*
When regularization is applied and therefore the minimization of $\{||X\beta - y||_2 + \lambda||\beta||_2\}$ is needed, the specification of appropriate values of the tuning parameter λ is an important issue. When the design matrix has correlated columns, its eigenvalue structure leads to a fast estimate for the generalized cross validation (GCV) function which can provide a good value for the parameter λ.

(P3) *Numerical methods for high dimensional data*
When the design matrix has much more columns than rows we deal with high dimensional data. In such cases, it is needed the appropriate computation of β in a way preserving the sparsity and the stability of the solution.

References

1. Buccini, A., De la Cruz Cabrera, O., Koukouvinos, C., Mitrouli, M., Reichel, L.: Variable selection in saturated and supersaturated designs via $l_p - l_q$ minimization.Communications in Statistics- Simulation and Computation, to appear. DOI:10.1080/03610918.2021.1961151
2. C. Koukouvinos, C., Jbilou, K., Mitrouli, M., Turek, O.: An eigenvalue approach for estimating the generalized cross validation function for correlated matrices, Electronic Journal of Linear Algebra 35 (2019) 482–496.
3. Winkler, J. R., Mitrouli, M., Koukouvinos, C.: The application of regularisation to variable selection in statistical modelling. J. Comp. Appl. Math 404:113884 (2022).