Variable selection in high dimensional regression when variables are heavily correlated

Alois Kneip
University of Bonn

Joint work wth Pascal Sarda, Université Paul Sabatier, Toulouse

Abstract:
The talk considers linear regression problems where the number of predictor variables is possibly larger than the sample size. The basic motivation of the study is to combine the points of view of model selection and functional regression.

Model selection procedures like Lasso can be used if the predictor vector can be decomposed into a sum of two independent random components reflecting common factors and specific variabilities of the explanatory variables. However, the usual assumption of sparseness of coefficients is restrictive in this context. Common factors may possess a significant influence on the response variable which cannot be captured by the specific effects of a small number of individual variables. We therefore propose to include principal components as additional explanatory variables in an augmented regression model. We give finite sample inequalities for estimates of these components. It is then shown that model selection procedures can be used to estimate the parameters of the augmented model and we state theoretical properties of the estimators.

In a second part of the talk functional explanatory variables are considered. We study a generalization of the classical functional linear regression model. It is assumed that there exists an unknown number of "points of impact", i.e. a number of discrete observation times where the corresponding functional values possess significant influence on the response variable. Problems of identifiability as well as corresponding estimation procedures are discussed.