Model selection for high-dimensional regressions

銀慶剛博士
中央研究院統計科學研究所

Abstract

A fundamental difficulty in model selection for high-dimensional regressions is that the number of observations is much less than the number of candidate variables. Recently, Buhlmann (2006, *Ann. Statist.*) showed that when $p_n = O(\exp(n^\theta))$ with $0 < \theta < 1$, where n is the number of observations and p_n is the number of candidate variables, the mean-squared prediction error of the L_2-Boosting predictor can converge to the variance of the random noise. He also proposed a corrected AIC criterion to determine the number of iterations in the corresponding boosting process. However, he did not give any theoretical justification for the proposed corrected AIC, and hence no consistency result for variable selection was reported. On the other hand, Zhao and Yu (2006, *J. Mach. Learning Res.*) provided the first consistency result for variable selection using LASSO in situations where $p_n = O(\exp(n^\theta))$ with $0 < \theta < 1$. However, to show consistency, they needed to impose the so called "irrepresentable" condition, which essentially says that the relevant and irrelevant regressors in the model are nearly uncorrelated. In fact, as shown in their paper, this irrepresentable condition is very restrictive and can hardly be met in general.

To obtain consistent variable selection in high-dimensional regressions without imposing the irrepresentable condition, we proposed L_2-Boosting + corrected BIC. Here, the corrected BIC is given by

$$\text{CBIC(k)} = \log \hat{\sigma}(k) + \frac{kC_n \log p_n}{n},$$

where C_n is a sequence of positive numbers tending to infinity with n at a certain rate, and $\hat{\sigma}(k)$ is the residual mean squared error after k boosting iterations. The major difference between CBIC and the traditional BIC is the appearance of $\log p_n$. This component, something related to the extreme value theory, is one of the most interesting findings of this study. Simulations show that using only 100 observations, our method can correctly identify 10 relevant regressors among 400 highly correlated regressors (390 of them are irrelevant) well above 90% of the time.