Refitted Cross-validation in Ultrahigh Dimensional Regression

Jianqing Fan, Princeton University
Shaojun Guo, Instituted of Applied Mathematics, Academia Sinica
Ning Hao, Princeton University
jqfan@princeton.edu

Abstract

Variance estimation is a fundamental problem in statistical modeling. In ultrahigh dimensional linear regressions where the dimensionality is much larger than sample size, traditional variance estimation techniques are not applicable. Recent advances on variable selection in ultrahigh dimensional linear regressions make this problem more accessible. One of the major problems in ultrahigh dimensional regression is the high spurious correlation between the unobserved realized noise and some of the predictors. As a result, the realized noises are actually predicted when extra irrelevant variables are selected, leading to serious underestimate of the noise level. In this paper, we propose a two-stage refitted procedure via a data splitting technique, called refitted cross-validation (RCV), to attenuate the influence of irrelevant variables with high spurious correlations. Our asymptotic results show that the resulting procedure performs as well as the oracle estimator, which knows in advance the mean regression function. The simulation studies lend further support to our theoretical claims. The naive two-stage estimator which fits the selected variables in the first stage and the plug-in one stage estimators using LASSO and SCAD are also studied and compared. Their performances can be improved by the proposed RCV method. The methods are applied to assess the forecasting errors of home price indices in the core based statistical areas in the US.