Incorporating Auxiliary Information for Improved Prediction in High Dimensional Datasets: An Ensemble of Shrinkage Approaches

Philip Boonstra*, University of Michigan

Jeremy Taylor, University of Michigan

Bhramar Mukherjee, University of Michigan

Abstract

We consider predicting a continuous outcome Y using a length-p covariate X, where p is large, with a sample of moderate size (less than p). W, a noisy surrogate for X, is also available in this sample, and Y and W are measured for a large number of additional observations. We propose shrinkage approaches where the regression coefficients of Y on X are shrunk towards targets that use information derived from the larger, noisy dataset. We compare the proposed shrinkage approaches with ridge regression of Y on X, which does not use W. We present a unified theoretical framework to view all the estimators as targeted ridge estimators. Finally, we propose a hybrid estimator that is able to balance efficiency and robustness in a data-adaptive way. The methods are evaluated in terms of their mean squared prediction error via analytical and simulation studies. In several plausible scenarios, our proposed shrinkage estimators are able to beat the classical ridge estimator in terms of predictive performance.

Keywords: Cross-validation, Mean squared prediction error; Measurement error; Ridge regression; Generalized ridge.

* Presenting author