Combining New Dimension Reduction Tools for High-Dimensional Regression

R. Parzer[1], L. Vana-Gür[1], and P. Filzmoser[1]

[1]TU Wien, Institute of Statistics and Mathematical Methods in Economics, Wiedner Hauptstr. 8, 1040 Wien, Austria

26.04.2023

We examine a regression technique in the challenging high-dimensional ($p \gg n$) setting with correlated predictors to explain and predict relevant quantities. Common approaches in these high-dimensional settings are to use some dimension reduction method first and then fit a predictive model in the reduced space (such as Partial Least Squares or Principal Component regression), or to use regularized regression (e.g. Elastic Net). However, many of those commonly used methods are not designed for the high-dimensional case and fail to identify the important variables or have poor prediction performance when there are thousands of variables and only hundreds of observations. Also, computation time becomes a notable issue for these dimensions.

A new, fast approach is to combine a probabilistic variable screening step with a random projection step to obtain a sparse set of reduced variables. These are used to fit a simple linear model for prediction. This procedure of screening, projection and prediction is repeated several times to explore different reductions. The obtained set of smaller predictive models is then aggregated by model averaging to get overall predictions.

We propose to use a measure for variable importance in the screening step, which considers the response as well as all other given variables, and a sparse, data-informed random projection, which tries to consider the effect direction when combining variables. We evaluate the marginal models to form model weights and combine them for a final prediction model. In extensive simulations and a real data application we guide through the elements of our method and compare prediction performance to competitors. Our method performs best relative to competitors in settings with a high number (close to n) of truly active variables.