A robust method for ultra-high dimensional regression analysis

Y. Wang[1*] and S. Van Aelst[1]

[1] *Department of Mathematics, KU Leuven;*
yixin.wang@wis.kuleuven.be, stefan.vanaelst@wis.kuleuven.be
[*] *Presenting author*

Keywords. *Sure Screening; Factor Profiling; Robust Factor Analysis; Least Trimmed Squares; Robust Regression.*

To increase the estimation accuracy and reduce the computational cost in ultra-high dimensional regression analysis, **?** proposed Sure Independence Screening (SIS) which selects a subset of the variables before estimating the regression coefficients. Predictor variables are selected according to the magnitude of their marginal correlations with the response variable. **?** proved that SIS shares the Sure Screening Property that the selected model will cover all the important variables with an overwhelming probability under certain model assumptions. However, the performance of SIS deteriorates greatly with increasing dependence among the predictors. To solve this problem, **?** proposed Factor Profiled Sure Independence Screening (FPSIS) based on the assumption that the predictors' correlation structure can be represented by a low-dimensional factor model. After profiling out the correlation structure by projecting the data onto the orthogonal complement of the subspace spanned by the factors, the screening performance can be largely improved. Since FPSIS is sensitive to outliers, a robust procedure is proposed.

In this work, we first estimate the subspace robustly based on the least trimmed squares estimator. Other robust techniques, such as ROBPCA, can also be applied. To increase the statistical efficiency, we apply a reweighting step in which we select the observations that are only outlying according to their score distances. The profiled variables are obtained by projecting all variables onto the orthogonal complement of the subspace spanned by the factors estimated from the reduced data. Finally, a robust regression method is used on the profiled data to estimate the marginal contribution of each predictor to the response variable. The results given by both the classical and the robust procedures with different types of outliers are compared.

References

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, **70**, 849-911.

Wang, H. (2012). Factor Profiled Sure Independence Screening. *Biometrika*, **99**, 15-28.