Sparsity Oriented Importance Learning for High-dimensional Linear Regression

With now well-recognized non-negligible model selection uncertainty, data analysts should no longer be satisfied with the output of a single final model from a model selection process, regardless of its sophistication. To improve reliability and reproducibility in model choice, one constructive approach is to make good use of a sound variable importance measure. Although interesting importance measures are available and increasingly used in data analysis, little theoretical justification has been done. In this paper, we propose a new variable importance measure, sparsity oriented importance learning (SOIL), for high-dimensional regression from a sparse linear modeling perspective by taking into account the variable selection uncertainty via the use of a sensible model weighting. The SOIL method is theoretically shown to have the inclusion/exclusion property: When the model weights are properly around the true model, the SOIL importance can well separate the variables in the true model from the rest. In particular, even if the signal is weak, SOIL rarely gives variables not in the true model significantly higher important values than those in the true model. Extensive simulations in several illustrative settings and real data examples with guided simulations show desirable properties of the SOIL importance in contrast to other importance measures.