Improved Estimators for Semi-supervised High-dimensional Regression Model

We study a linear high-dimensional regression model in a semi-supervised setting, where for many observations only the vector of covariates X is given with no responses Y. We consider a linear regression model but do not make any sparsity assumptions on the vector of coefficients, and aim at estimating $Var(Y \mid X)$. We propose an estimator, which is unbiased, consistent, and asymptotically normal. This estimator can be improved by adding zero-estimators arising from the unlabeled data. Adding zero-estimators does not affect the bias and potentially can reduce the variance.

We further illustrate our approach for other estimators, and present an algorithm that improves estimation for any given variance estimator. Our theoretical results are demonstrated in a simulation study.