High-dimensional instrumental variables regression and confidence sets

Eric Gautier (CREST ENSAE)
Alexandre B. Tsybakov (Université Paris 6)

Resumen We propose an instrumental variables method for estimation in linear models with endogenous regressors in the high-dimensional setting where the sample size n can be smaller than the number of possible regressors K, and $L \geq K$ instruments. We allow for heteroscedasticity and we do not need a prior knowledge of variances of the errors. We suggest a new procedure called the STIV (Self Tuning Instrumental Variables) estimator, which is realized as a solution of a conic optimization program. The main results of the paper are upper bounds on the estimation error of the vector of coefficients in l_p-norms for $1 \leq p \leq \infty$ that hold with probability close to 1, as well as the corresponding confidence intervals. All results are non-asymptotic. These bounds are meaningful under the assumption that the true structural model is sparse, i.e., the vector of coefficients has few non-zero coordinates (less than the sample size n) or many coefficients are too small to matter. In our IV regression setting, the standard tools from the literature on sparsity, such as ther restricted eigenvalue assumption are inapplicable. Therefore, for our analysis we develop a new approach based on data-driven sensitivity characteristics. We show that, under appropriate assumptions, a thresholded STIV estimator correctly selects the non-zero coefficients with probability close to 1. The price to pay for not knowing which coefficients are non-zero and which instruments to use is of the order $\sqrt{\log(L)}$ in the rate of convergence. We extend the procedure to deal with high-dimensional problems where some instruments can be non-valid. We obtain confidence intervals for non-validity indicators and we suggest a procedure, which correctly detects the non-valid instruments with probability close to 1.