High-dimensional instrumental variables regression and confidence sets

Eric Gautier

We present an instrumental variables method for inference in structural linear models with endogenous regressors in the high-dimensional setting where the number of possible regressors K can be much larger than the sample size n. Our new procedure, called STIV (Self Tuning Instrumental Variables) estimator, is realized as a solution of a conic optimization program. We allow for partial identification and very weak distributional assumptions including heteroscedasticity. We do not need prior knowledge of the variances of the errors. A key ingredient is sparsity, i.e., the vector of coefficients has many zeros. The main result of the paper is nested confidence sets, sometimes non-asymptotic, around the identified region under various level of sparsity. In the presence of very many instruments, a number exponential in the sample size, too many non-zero coefficients, or when there is an endogenous regressor which is only instrumented by weak instruments, our confidence sets have infinite volume (a variation on the STIV estimator is a robust method to estimate low dimensional models without sparsity and possibly many weak instruments). We obtain rates of estimation and show that, under appropriate assumptions, a thresholded STIV estimator correctly selects the non-zero coefficients with probability close to 1. In a non sparse setting, we obtain a sparse oracle inequality that shows how well the regression coefficients are estimated when the model can be well approximated by a sparse model. In our IV regression setting, the standard tools from the literature on sparsity, such as the restricted eigenvalue assumption are inapplicable. Therefore, for our analysis we develop a new approach based on data-driven sensitivity characteristics. We also obtain confidence sets, when the endogenous regressors have a sparse reduced form and we use a two-stage procedure, akin to two-stage least squares. There both stages involve high-dimensional regressions, and the second stage involves endogenous regressors. We finally study the properties of a two-stage procedure to deal with the detection of non-valid (endogenous) instruments. (joint work with Alexandre Tsybakov)