中央研究院統計科學研究所
學 術 演 講

講　題：Covariate screening in high dimensional data: applications to forecasting and text data

演講人：Dr. Adeline Lo

Department of Politics, Princeton University, USA.

時　間：2019年1月16日（星期三）下午14:00-15:30

地　點：中央研究院統計科學研究所6005會議室(環境變遷研究大樓A棟)

Abstract

High dimensional (HD) data, where the number of covariates and/or meaningful covariate interactions might exceed the number of observations, is increasing used in prediction in the social sciences. An important question for the researcher is how to select the most predictive covariates among all the available covariates. Common covariate selection approaches use ad hoc rules to remove noise covariates, or select covariates through the criterion of statistical significance or by using machine learning techniques. These can suffer from lack of objectivity, choosing some but not all predictive covariates, and failing reasonable standards of consistency that are expected to hold in most high-dimensional social science data. The literature is scarce in statistics that can be used to directly evaluate covariate predictivity. We address these issues by proposing a variable screening step prior to traditional statistical modeling, in which we screen covariates for their predictivity. We propose the influence (I) statistic to evaluate covariates in the screening stage, showing that the statistic is directly related to predictivity and can help screen out noisy covariates and discover meaningful covariate interactions. We illustrate how our screening approach can removing noisy phrases from U.S. Congressional speeches and rank important ones to measure partisanship. We also show improvements to out-of-sample forecasting in a state failure application. Our approach is applicable via an open-source software package.

中 央 研 究 院
統計科學研究所