Computationally tractable statistical estimation when there are more variables than observations

Emmanuel Candes, California Institute of Technology, USA
emmanuel@acm.caltech.edu

Abstract: In many important statistical applications, the number of variables or parameters is much larger than the number of observations. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the unknown number of pixels. Examples in functional MRI and tomography immediately come to mind. Other examples of high-dimensional data in genomics, signal processing and many other fields abound. In the context of multiple linear regression for instance, this setup raises the question of whether or not it is possible to estimate a vector of parameters of size p from a vector of observations of size n when n ¡¡ p, or whether it is possible to estimate the mean response reliably under the same circumstances.

This talk will survey very recent progress in this area showing that l1-methods such as the Dantzig selector and/or the lasso enjoy remarkable statistical properties. For instance, we will show that under reasonable sparsity assumptions, the Dantzig selector achieves an accuracy which nearly equals that one would achieve with an oracle that would supply perfect information about which coordinates of the unknown parameter vector are nonzero and which were above the noise level. This is connected with the important model selection problem since we will show that one can effectively tune l1-based methods as to automatically select the subset of covariates with nearly the best predictive power, by solving convenient optimization programs. We will discuss a few engineering applications where this could have a large pay-off.