

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
BERKELEY-STANFORD JOINT COLLOQUIUM

4:15 p.m., Tuesday, May 1st, 2007
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Noureddine El Karoui
Department of Statistics
UC Berkeley

Estimation of large dimensional covariance matrices

With data manipulation and storage increasingly easy, the size of the datasets statisticians are analyzing is becoming very large. The starting point of the analysis in this setting is often an $n \times p$ data matrix X , with n the number of observations and p the number of variables. A key feature of a number of “modern” datasets is that p and n are of the same order of magnitude, typically in the 100’s.

In this setting, it is known (from results in random matrix theory and theoretical statistics) that the sample covariance matrix is a very poor estimator of the population covariance. In particular, it can be shown that it estimates population eigenvalues - let alone eigenvectors - very poorly. It is therefore natural to ask how we can improve upon it.

In the first part of the talk, I will propose a method to (consistently) estimate the distribution of the eigenvalues of the population covariance from the eigenvalues of the sample covariance matrix. The technique requires minimal assumptions on the structure of the population covariance. The corresponding algorithm is fast and performs quite well in practice.

In the second part of the talk, I will turn to the narrower class of “sparse” covariance matrices. I will develop a notion of sparsity for matrices that is well suited for the study of spectral properties and reasonable from an application point of view. By exploiting this sparsity, we can come up with very good estimators that consistently estimate all eigenvalues and any well-separated eigenspaces.