High-Dimensional Covariance Estimation with Applications
University of Rome -Tor Vergata, May 16-21, 2016

Mohsen Pourahmadi
Department of Statistics
Texas A& M University

Motivation: Nowadays it is common in econometrics, finance and other application areas to work with high-dimensional data where the number of variables p is as large as the sample size n, and both may grow to infinity. In such situations, the sample covariance matrix, though unbiased and the maximum likelihood estimator (MLE) under normality, is a notoriously bad estimator of its population counterpart due to **excessive spread and bias of the sample eigenvalues**. Consequently, the statistical properties of "plug-in" estimators are drastically different from the commonly used "textbook plug-in" results where p is fixed and n goes to infinity, with devastating real-life decisions. For example, it is known that a Markowitz portfolio selection strategy with "plug-in" mean-covariance *overestimates* the target return and *underestimates* the risk of the portfolio. Similarly, in principal component analysis (PCA) the larger sample eigenvalues *overestimate* the contributions of the larger PCs, and hence lead to retaining too few of the PCs.

The Course Plan: We present an overview of the recent trends, and the details for a few useful and viable alternatives to the sample covariance matrix for high-dimensional data. In general, the presentation will follow along two complementary perspectives: (1) generalized linear models (GLMs) or parsimony and use of covariates in low dimensions; (2) regularization or sparsity for high-dimensional data. An emerging and powerful trend in both perspectives is that of **reducing a covariance estimation problem to that of estimating a mean vector or a sequence of regression problems**.

The coverage in the first two days is based on and inspired by Stein's (1955, 1975) idea of shrinking the eigenvalues of the sample covariance matrix toward a central value, Bayesian estimators and the well-conditioned Ledoit-Wolf estimators. The role of the ratio p/n and its limit in capturing the excessive spread of the sample eigenvalues will be described succinctly and rigorously by studying the asymptotic behavior of their empirical distribution through a milestone result due to Marčenko and Pastur (1967). Similar results for the sample eigenvectors and their implications for the PCA will be discussed.

A review of LASSO regression, penalized likelihood covariance estimation, Gaussian graphical models and elementwise regularization of the sample covariance matrix such as thresholding, banding and tapering will be covered during the third day.

The last two days will review and cover generalized linear models (GLMs) for covariance matrices inspired by Anderson's (1973) linear covariance models. Some advantages and limitations of the regression-based Cholesky decomposition relative to the classical spectral decomposition and factor models will be presented. It turns out that only the Cholesky decomposition provides an unconstrained and statistically interpretable reparameterization, and guarantees the positive-definiteness of the estimated covariance at no additional computational cost.

The Objectives: Introduce participants to the concepts and some useful techniques pertinent to high-dimensional data and covariance modeling in the context of cross-sectional and panel data. Provide the opportunity for extensive classroom interactions, discussions, and possibly analyzing data using existing R packages; participants may analyze their own data in tandem with the topics covered in the course and share their findings in subsequent sessions during the short course.