PHD PROJECT: RANDOM MATRIX STRUCTURES AND HIGH-DIMENSIONAL EXTREMES

JOHANNES HEINY

1. Introduction

Measuring and estimating the dependence between two random variables are fundamental problems in statistics. Starting with the early works of Pearson, Kendall, Hoeffding and Blum, several measures of dependence or association have been introduced and analyzed by numerous authors. An outstanding role is played by Pearson's correlation coefficient, a measure of the linear dependency of two random variables, about which most students learn early on in their studies. Motivated by its importance for statistical inference and estimation, many works are devoted to its stochastic properties in different frameworks. For example, in time series analysis, the notion of correlation plays a vital role in multivariate statistical analysis for parameter estimation, goodness-of-fit tests, change-point detection, etc.

Consider a p-dimensional population $\mathbf{x} \in \mathbb{R}^p$. For a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the population, we construct the matrix $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Since the entries of $\mathbf{X}_n$ are random variables, $\mathbf{X}_n$ is called a random matrix. Similarly the *sample covariance matrix* $\mathbf{S}_n = n^{-1}\mathbf{X}_n\mathbf{X}_n^\top$ and the *sample correlation matrix* $\mathbf{R}_n = \{\operatorname{diag}(\mathbf{S}_n)\}^{-1/2}\,\mathbf{S}_n\{\operatorname{diag}(\mathbf{S}_n)\}^{-1/2}$, where $\operatorname{diag}(\mathbf{S}_n)$ denotes the diagonal matrix with the same diagonal elements as $\mathbf{S}_n$, are random matrices. Note that $\mathbf{R}_n$ is the empirical version of Pearson correlation for multivariate data. In the field of random matrix theory, one is concerned with the spectral properties, that is, eigenvalues and eigenvectors, of random matrices such as $\mathbf{S}_n$ and $\mathbf{R}_n$ in a setting of growing dimension p.

It is worth mentioning that traditional multivariate analysis relies on the assumption that the dimension p remains fixed and thus is negligible compared to the sample size n. For this reason, results from traditional multivariate analysis are typically not applicable in other regimes. Spurred by these problems, new analysis tools for high-dimensional data were developed in recent years. Often it is assumed that dimension-to-sample-size ratio p/n tends to a positive constant as $n, p \to \infty$. This PhD project contributes to this line of research by providing asymptotic theory for random matrices that occur in statistical practice and also in physics, where eigenvalues of random matrices describe the energies of particles.

2. Purpose and aims

The focus of the PhD project is on high-dimensional probability theory and random matrix theory with possible applications in statistics. In particular, the research aims at describing the dependence structure of large data sets which is often studied via sample covariances, correlations, Spearman's ρ or Kendall's τ. Estimating and accurately assessing dependence has become the cornerstone of statistical inference and prediction in high dimension, where the sample size is at most of the order of the data dimension. In this effort, asymptotic probabilistic results from random matrix theory will be employed to obtain statistically meaningful results facilitating effective inference and prediction within the curse of dimensionality domain. Among a variety of important, difficult, and empirically meaningful topics in random matrix theory, the phase transitions and universality phenomena of large random matrices take a particularly prominent place. The underlying methodological flavor

Date: March 10, 2024.

of the research aims at combining elegant and sophisticated mathematical theories with a variety of fields of applications.

This project requires a **good knowledge of probability theory and statistics**, as well as self-motivation, enthusiasm and the willingness to carry out significant research within a lively area of modern mathematics. The supervisor and the PhD student will choose the PhD topic together to ensure that personal preferences and strengths are accounted for, with the student taking as much initiative as possible.

Aims of the PhD project could be to:

- explain the connections between the extremal entries and the spectral properties of large random matrices,
- develop novel statistical tests for high-dimensional data by combining methods from random matrix theory with other fields.

3. Significance

The dramatic increase and improvement of computing power and data collection devices have triggered the necessity to study and interpret the sometimes overwhelming amounts of data in an efficient and tractable way. Huge data sets arise naturally in wireless communication, finance, natural sciences and genetic engineering. The suggested research is tailored (but not limited) to popular dependence measures, such as correlations, covariances and other non-parametric measures of association for high-dimensional data. The proposed projects have the potential to considerably deepen our understanding of the connection between extremal and spectral properties of large random matrices. Ideally the planned research will provide new approaches and insights to high-dimensional data analysis.

More information about my research can be found at
https://www.su.se/english/profiles/johe3032-1.640812

Department of Mathematics, Stockholm University, Albano hus 1, 10691 Stockholm, Sweden
Email address: johannes.heiny@math.su.se