東京大学 大学院ゼミナールNo.2

「Hypothesis Testing and Random Matrix Theory」

2019年10月09日(水) 11:00~ 於 農学部1号館1階100番教室

生物測定学研究室　修士課程1年　Dang Tung Thanh

Abstract

One of the most common ways to analyze a collection of massive data is to extract top eigenvectors of a sample covariance matrix that represent directions of largest variance, often referred to as principal component analysis (PCA). The study of principal component analysis for sample covariance matrices is fundamental in multivariate analysis of scientific fields. For instance, genome-wide association studies construct a correlation matrix of expression levels, whereby PCA is able to identify collections of genes that work together. More broadly, it underlies much of exploratory data analysis, dimensionality reduction and visualization.

Classical random matrix theory provides a suite of tools to characterize the behavior of the eigenvalues of various random matrix models in high-dimensional settings. A direction initiated by Johnstone (2001)[1] has brought this powerful theory closer to statistical questions by introducing spiked models that are of the form “signal + noise”. Instead of applying the Marčenko–Pastur equation to cover the empirical distribution of most eigenvalues, Johnstone focused to consider the behavior of the largest eigenvalue because in distinguishing a “signal subspace” of higher variance from many noise variables, one expects the largest eigenvalue of a null (or white) sample covariance matrix to play a basic role. By applying the Tracy–Widom law was found by Tracy and Widom (1996), the mean growth of the largest eigenvalue is described specifically to detect the signal.

Alexander Soshnikov (2002)[2] extended Johnstone’s results in two directions. First of all, they prove that the joint distribution of the first, second, third, etc. eigenvalues of a Wishart matrix converges (after a proper rescaling) to the Tracy–Widom distribution. If a real (complex) sample covariance matrix satisfy the standard conditions of p-dimensional random matrices, then the joint distribution of the first, second, third, etc. largest eigenvalues converge to the Tracy–Widom law.

Jinho Baik et al. (2005)[3] compute the limiting distributions of the largest eigenvalue of a complex Gaussian sample covariance matrix when both the number of samples and the number of variables in each sample become large. When all but finitely many, say r, eigenvalues of the covariance matrix are the same, the dependence of the limiting distribution of the largest eigenvalue of the sample covariance matrix on those distinguished r eigenvalues of the covariance matrix is completely characterized in terms of an infinite sequence of new distribution functions that generalize the Tracy–Widom distributions of the random matrix theory. Especially a phase transition phenomenon is observed. Their results also apply to a last passage percolation model and a queueing model.

To continue the impressive successes of the combination between principal component analysis (PCA) and Random Matrix Theory (RMT) in the theoretical statistics, Kritchman and Nadler (2009)[4] approached this framework to study detection of the number of signals embedded in noise in signal and array processing. They presented a detailed statistical analysis of this problem, including an analysis of

the signal strength required for detection with high probability, and the form of the optimal detection test under certain conditions where such a test exists.

Bianchi et al. (2011)[5] extended Kritchman's results by proposing the Generalized Maximum Likelihood Test is studied and yields the analysis of the ratio between the maximum eigenvalue of the sampled covariance matrix and its normalized trace. Using recent results from random matrix theory, a practical way to evaluate the threshold and the p-value of the test is provided in the asymptotic regime where the number K of sensors and the number N of observations per sensor are large but have the same order of magnitude.

Patterson et al. (2006)[6] used results from Random Matrix Theory (RMT) to develop formal significance tests for population differentiation. They also uncovered a general ''phase change'' phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory they used, and had an important implication for the ability to discover structure in genetic data: for a fixed but large dataset size, divergence between two populations (as measured, for example, by a statistic like F_{ST}) below a threshold is essentially undetectable, but a little above threshold, detection will be easy. Based on their results, the dataset size needed to detect structure can be predicted.

Zhou et al. (2018)[7] followed and extended Patterson's results which were based on a standard Tracy–Widom limiting distribution for the largest eigenvalue, derived under white-noise assumptions. In order to overcome the drawbacks of Tracy–Widom laws, they explored several methods to identify appropriate null eigenvalue thresholds, while remaining sensitive to eigenvalues corresponding to population stratification. They introduced a novel block permutation approach, designed to produce an appropriate null eigenvalue distribution by eliminating long-range genomic correlation while preserving local correlation. They also proposed a fast approach based on eigenvalue distribution modeling, using a simple fit criterion and the general Marčenko–Pastur equation under a simple discrete eigenvalue model. Block permutation and the model-based approach work well for pure simulations and for data resampled from the 1000 Genomes project.

References

1. Johnstone, Iain M. "On the distribution of the largest eigenvalue in principal components analysis." The Annals of statistics 29.2 (2001): 295-327.
2. Soshnikov, Alexander. "A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices." Journal of Statistical Physics 108.5-6 (2002): 1033-1056.
3. Baik, Jinho, Gérard Ben Arous, and Sandrine Péché. "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices." The Annals of Probability 33.5 (2005): 1643-1697.
4. Kritchman, Shira, and Boaz Nadler. "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory." IEEE Transactions on Signal Processing 57.10 (2009): 3930-3941.
5. Bianchi, Pascal, et al. "Performance of statistical tests for single-source detection using random matrix theory." IEEE Transactions on Information theory 57.4 (2011): 2400-2419.
6. Patterson, Nick, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." PLoS genetics 2.12 (2006): e190.
7. Zhou, Yi-Hui, J. S. Marron, and Fred A. Wright. "Eigenvalue significance testing for genetic association." Biometrics 74.2 (2018): 439-447.