High-dimensional random tensors and applications to data science

Master 2 internship offer

Keywords. data science, high dimensional statistics, random tensors

1 Context

Due to the significant technical progress in computer science, electronics and information and communication technology, researchers and engineers are nowadays commonly dealing with high-dimensional data, in various fields such as signal and image processing (high resolution images, large sensor arrays), telecommunications (massive multi-antennas systems), genomics (large DNA microrrays), neurosciences (EEG), etc.

In high-dimensional statistical problems, it is often necessary to use dimension reduction techniques to represent the data into a new space of smaller dimension, while preserving useful statistical information. Such techniques usually involve dealing with large random matrices, such as Principal Component Analysis (PCA) and its variants, which require the study of the eigenvalues and eigenvectors of large sample covariance matrix, for which several results have been published in the field of random matrix theory [8, 7, 1, 2, 9].

Nonetheless, in several applications, the use of matrices to structure the data is not necessarily relevant, and one may resort to higher dimensional objects such as tensors; this is the case e.g. when using higher-order statistics (cumulants), in multidimensional signal processing or in machine learning. Unlike the literature on large random matrices, the study of large random tensors is more recent, and a few studies have been published in the past five years, regarding PCA in random tensors [10], detection of noisy tensors [9], or the behaviour of the spectrum of certain unfolded random tensors [3, 5].

2 Objectives

The goal of this internship is twofold:

(1) explore the benefits of using tensors in high dimensional data science problems. The student will get familiar with the mathematical properties of tensors and the standard algorithms for tensor decomposition [6], as well as the use of `Python` libraries for tensors, such as `TensorLy`. A short survey of the main applications of random tensors to machine learning and signal processing problems will also be conducted.

(2) study the statistical behaviour of certain high dimensional random tensors arising in specific problems such as noisy tensor detection [9, 4], Independent Component Analysis [5], or tensor PCA [10]. In particular, the student will explore various extensions of random matrix theory results to the case of random tensors, and evaluate the influence of the high dimensionality on these statistical inference problems.

3 Information and contact

Information. The internship will last 5/6 months, starting near February/March 2021, and will take place in the Institut de Mathématique de Bordeaux, at University of Bordeaux, with the following supervisory team:

- Jérémie Bigot, Institut de Mathématiques de Bordeaux/Université de Bordeaux
- Pascal Vallet, Laboratoire de l'Intégration du Matériau au Système/Bordeaux INP

Depending on the progress made during the internship and available fundings, an opportunity to pursue a Ph.D may be offered to the candidate.

Profile. The candidate should be master 2 or last year engineering school student, with a solid background in the field of statistics/data science and/or signal and image processing, and having good programming skills in `Python`.

Application. Applicants must send via e-mail to

jeremie.bigot@math.u-bordeaux.fr, pascal.vallet@ims-bordeaux.fr

a CV as well as a transcript.

References

[1] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[2] F. Benaych-Georges and R.R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.*, 111(0):120–135, 2012.

[3] R. Boyer and P. Loubaton. Large deviation analysis of the cpd detection problem based on random tensor theory. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 658–662, 2017.

[4] A. Chevreuil and P. Loubaton. On the non-detectability of spiked large random tensors. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*, pages 443–447. IEEE, 2018.

[5] P. Gouédard and P. Loubaton. On the behaviour of the estimated fourth-order cumulants matrix of a high-dimensional gaussian white noise. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017.

[6] M. Janzamin, R. Ge, J. Kossaifi, and A. Anandkumar. Spectral learning on matrices and tensors. *Foundations and TrendsÂ® in Machine Learning*, 12(5-6):393–536, 2019.

[7] I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 04 2001.

[8] V.A. Marcenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967.

[9] A. Montanari, D. Reichman, and O. Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank one perturbations of gaussian tensors. *IEEE Trans. Inf. Theory*, 63(3):1572–1579, 2017.

[10] E. Richard and A. Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2897–2905, 2014.