Applications of large random matrices to high dimensional statistical signal processing.

PHD subject proposed by Philippe Loubaton
Université Paris-Est Marne la Vallée and Institut Universitaire de France
Laboratoire d'Informatique Gaspard Monge, UMR 8049
email loubaton@univ-mlv.fr, URL : http://www-syscom.univ-mlv.fr/˜ loubaton/index.html

1 General context of the proposed work.

Large random matrices have been proved to be of fundamental importance in mathematics (high dimensional probability and statistics, operator algebras, combinatorics, number theory,...) and in physics (nuclear physics, quantum fields theory, quantum chaos,..) for a long time. The introduction of large random matrix theory in electrical engineering is more recent. It was introduced at the end of the nineties in the context of digital communications in order to analyse the performance of large CDMA and multi-antennas systems. The use of large random matrices is even more recent in statistical signal processing. The corresponding tools turn out to be useful when the observation is a large dimension (say M) multivariate time series $(\mathbf{y}_n)_{n=1,\ldots,N}$, and the sample size N is of the same order of magnitude as M. This context poses a number of new difficult statistical problems that are intensively studied by the high-dimensional statistics community. The most significant example is related with the fundamental problem of estimating the covariance matrix of the observation because the standard empirical covariance matrix defined as the empirical mean of the $\mathbf{y}_n \mathbf{y}_n^*$ is known to perform poorly if N is not significantly larger than M.

In the context of this project, the dimension of the observation corresponds to the number of elements of a large sensor network, and the components of vector $\mathbf{y}_n$ represent the signal received at time n on the various sensors. It turns out that a number of fundamental statistical signal processing schemes such as source detection, source localisation, estimation of linear prediction filters...fail in the case where M and N are large and of the same order of magnitude, a context modelled by the asymptotic regime M and N both converge to infinity in such a way that the ratio $\frac{M}{N}$ converges to a non zero constant. The purpose of proposed work is to develop mathematical tools leading to new high performance statistical signal processing algorithms in the above regime.

The PHD is funded by the projet DIONISOS from the "Agence Nationale de la Recherche" (URL http://www-syscom.univ-mlv.fr/˜ loubaton/dionisos.html), and the PHD student will beneficiate of the expertise of the various members of the project. As the proposed work is at the interface of signal processing and probability theory, the successful applicant either obtained an electrical and/or computer engineering master degree with a good background in probability and statistics, either obtained a master degree in probability and statistics and is interested by applications to signal processing and/or digital communications.

In order to apply, send as soon as possible to loubaton@univ-mlv.fr

- a motivation email
- a full CV especially showing the expertise in the fields, either in the form of courses or projects, the grades and ranking in the courses of interest
- 2 or 3 References (people who would agree to send a recommendation letter).

2 Presentation of the proposed work.

2.1 Brief state of the art.

The existing works concerning applications of large random matrices to statistical signal processing have been essentially motivated by the so-called narrow-band antenna array model in which the observation is a M−dimensional time series

$(\mathbf{y}_n)_{n=1,\ldots,N}$ defined as the superposition of a noise term with a low rank $K < M$ component that can be interpreted as the contributions of K narrowband non observable signals $(s_k)_{k=1,\ldots,K}$. More precisely, $\mathbf{y}_n$ can be written for $n = 1, \ldots, N$ as

$$\mathbf{y}_n = \sum_{k=1}^{K} s_{k,n}\mathbf{a}_k + \mathbf{v}_n = \mathbf{A}\mathbf{s}_n + \mathbf{v}_n \tag{2.1}$$

where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_K)$ is an unknown deterministic matrix and where $\mathbf{v}_n$ represents an additive temporally and spatially white complex Gaussian noise, *i.e.*, $\mathbb{E}(\mathbf{v}_m\mathbf{v}_n^*) = \sigma^2\delta(m-n)\mathbf{I}$ and $\mathbb{E}(\mathbf{v}_m\mathbf{v}_n^T) = 0$ for each m, n. A number of classical problems related to model (2.1) reduce to infer properties of the low rank component $\mathbf{A}\mathbf{s}_n$ from the observations. We mention in particular:

- Detection of narrow-band signals, i.e. testing $K = 0$ against $K \geq 1$, if $K \geq 1$, estimate K
- Estimate the subspace generated by the columns of $\mathbf{A}$
- When vectors $(\mathbf{a}_k)_{k=1,\ldots,K}$ are known up to finite number of parameters, estimate these parameters
- Estimate eigenvalues and/or eigenvectors of matrix $\mathbf{A}\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{s}_n\mathbf{s}_n^*\right)\mathbf{A}^*$

These questions were studied extensively in the past when the sample size N is much larger than the dimension M of the observation. When M and N are however of the same order of magnitude, a number of technics proposed in the context $M << N$ fail, and new approaches had to developed. Some recent works proposed to use large random matrices technics to address the above problems, see e.g. references [1] to [11]. In these works, the N observations $(\mathbf{y}_n)_{n=1,\ldots,N}$ are stacked into the $M \times N$ matrix $\mathbf{Y}$ defined by

$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N),$$

which depending on the assumptions on the source signals, correspond to random matrix models well documented in the literature. More precisely, if the source signals are i.i.d. sequences, the columns of matrix $\mathbf{Y}$ are independent and identically distributed (see e.g. the recent books [12], [14]), while if the source signals are modelled as deterministic sequences, $\mathbf{Y}$ correspond to the so-called information plus noise model (see [13]) and represents a Gaussian non zero mean random matrix with independent entries.

2.2 The context of wideband sources.

The main objective of the proposed thesis is to address more complicated source signal models. As expressed by Eq. (2.1), the contribution of a narrow band source to the observation lies in the 1–dimensional subspace generated by its associated vector $\mathbf{a}$. In a number of practical situations, this model appears irrelevant, and a more accurate model to represent the contribution of a source signal $s = (s_n)_{n\in\mathbb{Z}}$ to the observation is the output of an unknown a 1-input / M–outputs linear system driven by s, i.e.

$$\sum_{p=0}^{P-1}\mathbf{a}_p s_{n-p} = [\mathbf{a}(z)]s_n$$

where $\mathbf{a}(z) = \sum_{p=0}^{P-1}\mathbf{a}_p z^{-p}$ represents the M–variate transfer function of the above mentioned linear system, and where parameter P may take quite moderate values or may converge to $+\infty$ depending on the context. We refer to this model to as the wideband source model.

The goal of the PHD is to study detection and estimation problems when the observation $\mathbf{y}_n$ is a noisy version of the contribution of a wideband source (or of multiple wideband sources), i.e.

$$\mathbf{y}_n = [\mathbf{a}(z)]s_n + \mathbf{v}_n$$

where the additive noise $\mathbf{v}$ is as above. The problems addressed in the proposed PHD concern:

- the detection of the wideband source (presence of the source against absence of the source),
- the estimation of the coefficients $(\mathbf{a}_p)_{p=0,\ldots,P-1}$,
- the linear estimation of the sequence s_n from the observation,

both in the unsupervised case (sequence $(s_n)_{n=1,\ldots,N}$ is unobservable) and in the supervised case (sequence $(s_n)_{n=1,\ldots,N}$ is observable).

In the context of the above problems, for a number of reasons, the matrix on which are based most of the classical signal processing detection/estimation algorithms is not matrix $\mathbf{Y}$ defined above, but a block-Hankel matrix $\mathbf{Y}_L$ defined by

$$\mathbf{Y}_L = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \cdots & \mathbf{y}_{N-L+1} \\ \mathbf{y}_2 & \mathbf{y}_3 & \ddots & \ddots & \mathbf{y}_{N-L+2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{y}_L & \mathbf{y}_{L+1} & \cdots & \cdots & \mathbf{y}_N \end{pmatrix}$$

where parameter L should in principle be greater than P. In the context of this work, it will be assumed that N and ML are large and of the same order of magnitude, a context modelled by the asymptotic regime

$$N \to +\infty, ML \to +\infty \text{ in such a way that } \frac{ML}{N} \to d \text{ where } d > 0 \tag{2.2}$$

We note that the asymptotic regime (2.2) is able to capture situations in which M and N are of the same order of magnitude and $L << M$ as well as the case where M and L are much smaller than N but in such a way that ML and N are comparable. In order to address the above detection and estimation problems, it will be necessary, to study, among other questions, the location of the eigenvalues and the behaviour of the greatest eigenvalues of matrix $\frac{\mathbf{Y}_L \mathbf{Y}_L^*}{N}$ both in the absence and the presence of the signal. The difficulty of these questions is due to the particular structure of matrix $\mathbf{Y}_L$ which induces dependencies between its entries. We note that in the absence of signal, partial results show that the behaviour of the eigenvalues of $\frac{\mathbf{Y}_L \mathbf{Y}_L^*}{N}$ depends critically of L: if L remains finite (i.e. does not scale with M, N), the above eigenvalues behave as if matrix $\mathbf{Y}_L$ was i.i.d. in which case they lie in the neighborhood of the interval $[\sigma^2(1-\sqrt{d})^2, \sigma^2(1+\sqrt{d})^2]$, while if L is of the same order of magnitude than N, then the greatest eigenvalue tend to $+\infty$ (see e.g. [15]). The intermediate cases are still to be studied, as well as the context in which the signal is present. For this, in a first stage, the approach developed by Haagerup ([16]) and Pastur ([17]) and used e.g. in [18] and [19], will be adapted to the context of matrix $\mathbf{Y}_L$. While it is reasonable to expect that these technics can be successfull if $\frac{L}{N} \to 0$, the case where L and N are of the same order of magnitude will be probably harder.

Références

[1] F. Benaych-Georges, R.R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices, *Adv. in Math.*, vol. 227, no. 1 (2011), 494–521.

[2] F. Benaych-Georges, R.R. Nadakuditi,"The singular values and vectors of low rank perturbations of large rectangular random matrices", submitted, arXiv:1103.2221.

[3] P. Bianchi, M. Debbah, M. Maïda and J. Najim, "Performance of Statistical Tests for Single Source Detection using Random Matrix Theory", *IEEE Inf. Theory*, Vol. 57 (4), april 2011, 2400–2419.

[4] Walid Hachem, Philippe Loubaton, Jamal Najim, Xavier Mestre, Pascal Vallet, "Large information plus noise random matrix models and consistent subspace estimation in large sensor networks ", Random Matrices, Theory and Applications (RMTA), vol. 1, no. 2 (2012).

[5] Walid Hachem, Philippe Loubaton, Jamal Najim, Xavier Mestre, Pascal Vallet, "A subspace estimator for fixed rank perturbations of large random matrices", Journal of Multivariate Analysis, 114 (2013), pp. 427-447.

[6] S. Kritchman, B. Nadler, "Non-parametric Detection of the Number of Signals, Hypothesis Testing and Random Matrix Theory, *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3930–3941, 2009.

[7] Mestre, X. and Lagunas, M.A., "Modified subspace algorithms for DoA estimation with large arrays", *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598-614, February 2008.

[8] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates", IEEE Trans. Inform. Theory, 54 (11), pp. 5113-5129, 2008.

[9] Nadakuditi, R.R. and Silverstein, J.W., "Fundamental Limit of Sample Generalized Eigenvalue Based Detection of Signals in Noise Using Relatively Few Signal-Bearing and Noise-Only Samples", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 3, pp. 468 -480, June 2010.

[10] R.R. Nadakuditi, A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples", *IEEE Transactions of Signal Processing*, vol. 56, no. 7, pp. 2625-2637, July 2008.

[11] Pascal Vallet, Philippe Loubaton, Xavier Mestre, "Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case", *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1043-1068, February 2012.

[12] Z. Bai, J.W. Silverstein, "Spectral analysis of large dimensional random matrices", Springer Series in Statistics, 2nd ed., 2010.

[13] B. Dozier, J. Silverstein, "On the Empirical Distribution of Eigenvalues of Large Dimensional Information-Plus-Noise Type Matrices", *Journal of Multivariate Analysis*, 98(4) (2007), pp. 678-694.

[14] L.A. Pastur, M. Shcherbina, "Eigenvalue Distribution of Large Random Matrices", Mathematical Surveys and Monographs, Providence: American Mathematical Society, 2011.

[15] A. Basak, A. Bose, S. Sen, "Limiting spectral distribution of sample autocovariance matrices", to appear in *Bernouilli*, can be downloaded on Arxiv, arXiv:1108.3147v1.

[16] U. Haagerup, S. Thorbjornsen, "A new application of random matrices: $\mathrm{Ext}(C^*_{red}(F_2))$ is not a group", *Annals of Mathematics*, vol. 162, no. 2, 2005.

[17] L.A. Pastur, "A simple approach for the study of the global regime of large random matrices", Ukrainian Math. J., vol. 57, no. 6, pp. 936-966, June 2005.

[18] M. Capitaine, C. Donati-Martin, D. Féral, "The largest eigenvalue of finite rank deformation of large Wigner matrices: convergence and non-universality of the fluctuations", *Annals of Probability*, vol. 37, no. 1, pp. 1-47, 2009.

[19] Philippe Loubaton, Pascal Vallet, "Almost sure localization of the eigenvalues in a gaussian information plus noise model. Applications to the spiked models", *Electronic J. on Probability*, October 2011, pp. 1934-1959.