Sujet de Thèse, PHD Thesis Proposal

Large random matrix methods for high-dimensional time series analysis.

Proposed by Philippe Loubaton, Laboratoire d'Informatique Gaspard Monge, Université Paris-Est Marne la Vallée,

email : philippe.loubaton@u-pem.fr, http://www-syscom.univ-mlv.fr/~loubaton.

This PhD is supported by the ANR Project HIDITSA, HIgh DImensional Time Series Analysis

http://www-syscom.univ-mlv.fr/~loubaton/hiditsa.html

The proposed work is at the interface of statistical signal processing, large random matrices and statistics of multivariate time series. The successful applicant should have obtained a master degree in probability and / or statistics, but candidates with a good background in probability and statistics having obtained either an electrical and/or a computer engineering master degree are also encouraged to apply.

In order to apply, send as soon as possible to philippe.loubaton@u-pem.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 References (people who would agree to send a recommendation letter).

1 General context of the proposed work.

Large random matrices have been proved to be of fundamental importance in mathematics (high dimensional probability, operator algebras, combinatorics, number theory,...) and in physics (nuclear physics, quantum fields theory, quantum chaos,..) for a long time. The use of large random matrices is more recent in statistical signal processing and time series analysis. The corresponding tools turn out to be useful when the observation is a large dimension (say M) multivariate time series $(\mathbf{y}_n)_{n=1,...,N}$ and the sample size N is not much larger than M, a situation that becomes very comon due to the spectacular development of data acquisition devices and sensor networks. 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 $(\mathbf{y}_n\mathbf{y}_n^*)_{n=1,...,N}$ is known to perform poorly if N is not significantly larger than M. As a result, the conventional statistical inference schemes that are based on functionals of the empirical covariance matrix may perform poorly. In order to mitigate this conceptual difficulty, the most popular approaches were based on the design of inference schemes using some possible degree of sparsity of the underlying parameters. However, sparsity is a property that does not necessarily hold. The use of large random matrix theory is an appealing alternative because, under some assumptions on the observations $(\mathbf{y}_n)_{n=1,...,N}$, it is possible to precise the behaviour of certain functionals of the empirical covariance matrix when M and N are both large, and to use the corresponding results in order to design new improved performance inference schemes (see e.g. [4], [14], [17], [18]).

While these papers produced a number of valuable results, a considerable work remains to be done to exploit the potential of large random matrix technics in the context of statistics of high-dimensional Gaussian time series. In particular, a number of classical inference schemes are based on functionals of non parametric estimates of

the spectral density of time series $(\mathbf{y}_n)_{n\in\mathbb{Z}}$ (see e.g. [6]) or of non parametric estimates of the covariance matrix of the augmented ML-dimensional vector $\mathbf{y}_n^{(L)}$ defined by

$$\mathbf{y}_n^{(L)} = [(\mathbf{y}_{1,n},\ldots,\mathbf{y}_{1,n+L-1}),\ldots,(\mathbf{y}_{M,n},\ldots,\mathbf{y}_{M,n+L-1})]^T \tag{1}$$

where L is a relevant parameter, and where $(\mathbf{y}_{1,n})_{n\in\mathbb{Z}}, (\mathbf{y}_{2,n})_{n\in\mathbb{Z}},\ldots,(\mathbf{y}_{M,n})_{n\in\mathbb{Z}}$ are the M components of the observation $(\mathbf{y}_n)_{n\in\mathbb{Z}}$ ((see e.g. [10]). The goal of the present PhD thesis is to evaluate the behaviour of certain functionals of the above empirical estimates using large random matrix methods in asymptotic regimes where both M and N converge towards $+\infty$, and to take benefit of the results in order to revisit the problem of testing that the M components of the observation are mutually uncorrelated signals in the case where both M and N are large. Applications to the detection of a "useful" signal generated as the output of an unknown K inputs / M–outputs linear system driven by K non observable time series, and corrupted by a spatially uncorrelated additive Gaussian noise, will also be addressed.

The proposed research topics are thus at the interface between large random matrices and the statistics of multivariate time series. This interface, which has not yet been fully exploited in the technical literature, has a high academic and applicative potentials in the context of high-dimensional statistical inference problems.

2 More details on the proposed work.

We denote by $\mathbf{S}(\nu)$ the spectral density of time series $(\mathbf{y}_n)_{n\in\mathbb{Z}}$ defined for each frequency ν as the positive $M\times M$ matrix

$$\mathbf{S}(\nu) = \sum_{l\in\mathbb{Z}} \mathbf{R}_l\, e^{-2i\pi l\nu}$$

where for each integer l, $\mathbf{R}_l = \mathbb{E}(\mathbf{y}_{n+l}\mathbf{y}_n^*)$ represents the autocovariance matrix of $\mathbf{y}$ at lag l. A typical non parametric estimate of $\mathbf{S}(\nu)$ is the so-called frequency smoothed periodogram defined as the matrix $\hat{\mathbf{S}}(\nu)$ given by

$$\hat{\mathbf{S}}(\nu) = \frac{1}{2B}\sum_{b=-B}^{B} \boldsymbol{\xi}_{\mathbf{y}}(\nu - b/N)\,\boldsymbol{\xi}_{\mathbf{y}}(\nu - b/N)^* \tag{2}$$

where B is an integer less than N called the smoothing span, and where $\boldsymbol{\xi}_{\mathbf{y}}(\nu) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{y}_n e^{-2i\pi(n-1)\nu}$ is a normalized version of the discrete Fourier transform of M–dimensional vectors sequence $(\mathbf{y}_n)_{n=1,\ldots,N}$. We mention that positive factors could also weight the terms of the above sum, but we prefer to omit them for the sake of simplicity.

We denote by $\mathbf{R}^{(L)}$ the covariance matrix of augmented vector $\mathbf{y}_n^{(L)}$ defined by (1). This covariance matrix may be estimated by the empirical estimate $\hat{\mathbf{R}}^{(L)}$ defined by

$$\hat{\mathbf{R}}^{(L)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n^{(L)}\mathbf{y}_n^{*(L)} \tag{3}$$

where we ignore the end effects in the above sum.

2.1 Consistency of estimates $\hat{\mathbf{S}}(\nu)$ and $\hat{\mathbf{R}}^{(L)}$.

A preliminary problem is to find conditions on M, N, L, B and on the properties of the time series under which the above estimates are consistent in the spectral norm sense when M and N both converge towards $+\infty$, i.e.$\sup_{\nu\in[0,1]}\|\hat{\mathbf{S}}(\nu)-\mathbf{S}(\nu)\|\to 0$ and $\|\hat{\mathbf{R}}^{(L)}-\mathbf{R}^{(L)}\|\to 0$, where $\|\mathbf{A}\|$ represents the spectral norm of a matrix $\mathbf{A}$. While these important topics were addressed extensively in the past when M is fixed (see e.g. [2], [3], [6] in the context of scalar or low-dimensional multivariate linear processes, and [20], and [19] in the case of certain non linear models), the above mentioned problems do not seem to have been addressed when both

M and N converge towards $+\infty$ (except perhaps in the very preliminary analysis [5]) which will need the use of different tools. Intuitively, consistency of the estimates will probably hold only if $B \to +\infty$, $M/B \to 0$, $B/N \to 0$ (for $\hat{\mathbf{S}}(\nu)$) and $ML/N \to 0$ (for $\hat{\mathbf{R}}^{(L)}$).

2.2 Behaviour of functionals of the estimates when consistency is lost.

In practice, for large finite values of M, L, N, B, the above ratios may not be small enough in order to make reliable the performance predicted in the regime where consistency holds. It is therefore relevant to consider limit regimes where ML/N (for $\hat{\mathbf{R}}^{(L)}$) and M/B (for $\hat{\mathbf{S}}(\nu)$) converge towards non zero constants, and to study the behaviour of relevant functionals of matrices $\hat{\mathbf{S}}(\nu)$ and $\hat{\mathbf{R}}^{(L)}$. The main goal of this work is to address this more complicated scenario. For this, specific, but important, observation models will be considered. In particular, it will be assumed that the observation is a linear dynamic factor model, i.e.

$$\mathbf{y}_n = \mathbf{u}_n + \mathbf{v}_n \tag{4}$$

where $(\mathbf{v}_n)_{n\in\mathbb{Z}}$ represents an additive spatially uncorrelated Gaussian noise (i.e. the spectral density of $\mathbf{v}$ is a diagonal matrix), and where $(\mathbf{u}_n)_{n\in\mathbb{Z}}$ is a "useful" signal conveying some informations that should be extracted from the N available observations $(\mathbf{y}_n)_{n=1,\dots,N}$. In the context of the present work, $\mathbf{u}$ is assumed to be generated as

$$\mathbf{u}_n = \sum_{l=0}^{+\infty} \mathbf{H}_l \mathbf{s}_{n-l} \tag{5}$$

where $(\mathbf{H}_l)_{l\geq 0}$ are unknown deterministic matrices, and where $(\mathbf{s}_n)_{n\in\mathbb{Z}}$ is a K–dimensional time series whose components are mutually uncorrelated independent identically distributed Gaussian sequences that are sometimes called the factors. Here, the number of factors K will be assumed much lower than M, and will be considered as a fixed parameter when M, N, L, B converge towards $+\infty$. The spectral density of $\mathbf{y}$ is thus at each frequency ν the sum of a rank $K << M$ matrix with a diagonal matrix (i.e. the spectral density of the noise).

We will first study the behaviour of the empirical eigenvalue distribution of matrix $\hat{\mathbf{R}}^{(L)}$ when N, M, L converge towards $+\infty$ in such a way that ML/N converges towards a constant. Hopefully, this distribution will have a deterministic behaviour, which, intuitively means that the histogram of the eigenvalues of any realization of $\hat{\mathbf{R}}^{(L)}$ tend to concentrate around the graph of a certain probability distribution which will be characterized. This study will allow to evaluate the asymptotic behaviour of linear statistics of the eigenvalues $(\hat{\lambda}_k)_{k=1,\dots,ML}$ of $\hat{\mathbf{R}}^{(L)}$, i.e. terms defined by

$$\frac{1}{ML}\sum_{k=1}^{ML} \phi(\hat{\lambda}_k) \tag{6}$$

for some smooth function ϕ. The case where $\phi(\lambda) = \log\lambda$, i.e.

$$\frac{1}{ML}\sum_{k=1}^{ML} \phi(\hat{\lambda}_k) = \frac{1}{ML}\log\det(\hat{\mathbf{R}}^{(L)})$$

will be studied thoroughly. For this, we will use large random matrix theory tools that were developped in the context of more standard random matrix models (see e.g. [1] or [15]). It will also be important to study the behaviour of the largest eigenvalues of $\hat{\mathbf{R}}^{(L)}$ in order to evaluate the influence of useful signal $\mathbf{u}$ on these eigenvalues. Motivated by the problem of testing that the M components of $\mathbf{y}$ are uncorrelated signals (see below), we will also consider the same questions, but when matrix $\hat{\mathbf{R}}^{(L)}$ is replaced by a multiplicative deformed versions $\mathbf{C}^{1/2}\hat{\mathbf{R}}^{(L)}\mathbf{C}^{1/2}$, where $\mathbf{C}$ is a positive definite deterministic $ML \times ML$ matrix. The recent paper [13] should be a good starting point to address these issues.

Using again large random matrix technics, we will also study the asymptotic behaviour of the empirical eigenvalue distribution and the largest eigenvalues of estimate $\hat{\mathbf{S}}(\nu)$ for each frequency ν. For this, it might be useful to remark that for each ν, matrix $\hat{\mathbf{S}}(\nu)$ can be interpreted as the sample covariance matrix of $M-$dimensional vectors $(\boldsymbol{\xi}_{\mathbf{y}}(\nu - b/N))_{b=-B/2,\ldots,B/2}$, which should "be close" from nearly uncorrelated vectors $\left(\mathbf{H}(e^{2i\pi\nu})\boldsymbol{\xi}_{\mathbf{s}}(\nu - b/N) + \boldsymbol{\xi}_{\mathbf{v}}(\nu - b/N)\right)_{b=-B/2,\ldots,B/2}$ where $\mathbf{H}(e^{2i\pi\nu})$ is defined by $\mathbf{H}(e^{2i\pi\nu}) = \sum_{l=0}^{+\infty} \mathbf{H}_l\, e^{-2i\pi l\nu}$. As above, generalizations to multiplicative deformed versions of $\hat{\mathbf{S}}(\nu)$ will be considered.

2.3 Applications to testing that the M components of $\mathbf{y}$ are uncorrelated.

The components of $\mathbf{y}$ are mutually uncorrelated (or equivalently useful signal $\mathbf{u}$ is absent, i.e. $\mathbf{y}$ is reduced to the Gaussian noise $\mathbf{v}$) if and only if matrix $\mathbf{R}^{(L)}$ is block diagonal for each L, the (m,m) block being equal to the covariance matrix $\mathbf{R}_m^{(L)}$ of vector $(\mathbf{y}_{m,n}, \mathbf{y}_{m,n+1}, \ldots, \mathbf{y}_{m,n+L-1})^T$. An alternative frequency-domain characterization is that the spectral density $\mathbf{S}(\nu)$ is a diagonal matrix whose entries are the spectral densities of the various components of $\mathbf{y}$. Therefore, a "time-domain" statistics should test if matrix $\hat{\mathbf{R}}^{(L)}$ for L large enough is close to be block diagonal, and a frequency-domain statistics should check whether matrix $\hat{\mathbf{S}}(\nu)$ is close to be diagonal for each ν. We refer the reader to [11],[8] and [9] for existing works presenting time-domain and frequency-domain approaches when the observation is a low dimensional time series (i.e. M remains fixed).

Motivated by [16], [12] and [14] in which different, but related, problems are studied, we propose to study time-domain statistics that are based on the eigenvalues of matrix $\hat{\mathbf{R}}_{cor}^{(L)}$ defined by

$$\hat{\mathbf{R}}_{cor}^{(L)} = \left(\mathrm{Diag}((\hat{\mathbf{R}}_m^{(L)})_{m=1,\ldots,M})\right)^{-1/2}\ \hat{\mathbf{R}}^{(L)}\ \left(\mathrm{Diag}((\hat{\mathbf{R}}_m^{(L)})_{m=1,\ldots,M})\right)^{-1/2} \tag{7}$$

when M, N and L converge towards $+\infty$ in such a way that ML/N converges towards a non zero constant and $\hat{\mathbf{R}}_m^{(L)}$ represents the sample estimate of $\mathbf{R}_m^{(L)}$. Here, $\mathrm{Diag}((\hat{\mathbf{R}}_m^{(L)})_{m=1,\ldots,M})$ is the block diagonal matrix whose diagonal blocks are $L \times L$ matrices $(\hat{\mathbf{R}}_m^{(L)})_{m=1,\ldots,M}$. The index cor indicates that the above matrix is a kind of autocorrelation matrix. L should converge towards $+\infty$ to test the values of the autocovariance matrices of $\mathbf{y}$ at any lag. The general idea is to recognize that if the various matrix estimates in the right hand side of (7) were replaced by their true values, then the left hand side of (7) would be equal to the identity matrix. In this case, all its eigenvalues would coincide with 1, and its empirical eigenvalue distribution would reduce to the Dirac distribution at point 1. It is thus relevant to study the behaviour of the empirical eigenvalue distribution of matrix $\hat{\mathbf{R}}_{cor}^{(L)}$ when the M components of $\mathbf{y}$ are uncorrelated, i.e. when the useful signal $\mathbf{u}$ is absent in the expression (5), and when $\mathbf{u}$ is present. The study of the largest eigenvalues of $\hat{\mathbf{R}}_{cor}^{(L)}$ might also be relevant. When M, N, L converge towards $+\infty$ and that ML/N converges towards a constant, ratio L/N converges towards 0. Therefore, each empirical estimate $\hat{\mathbf{R}}_m^{(L)}$ should converge in the spectral norm sense towards its true value, and the eigenvalues of $\hat{\mathbf{R}}_{cor}^{(L)}$ should have the same first order behaviour than the eigenvalues of the multiplicative deformed version $\mathbf{C}^{1/2}\hat{\mathbf{R}}^{(L)}\mathbf{C}^{1/2}$ where matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \left(\mathrm{Diag}((\mathbf{R}_m^{(L)})_{m=1,\ldots,M})\right)^{-1}$$

Therefore, the problems presented in paragraph 2.1 appear intimately connected to our testing problem. However, more work will have to be done to study potential limit distributions of linear statistics of the eigenvalues of $\hat{\mathbf{R}}_{cor}^{(L)}$, or of the largest eigenvalues of $\hat{\mathbf{R}}_{cor}^{(L)}$.

These time-domain statistics have frequency domain analogs obtained by replacing matrix $\hat{\mathbf{R}}_{cor}^{(L)}$ by an estimator $\hat{\mathbf{S}}_{co}(\nu)$ of the spectral coherency matrix defined for each frequency ν by

$$\hat{\mathbf{S}}^{(co)}(\nu) = \left(\mathrm{Diag}((\hat{S}_m(\nu))_{m=1,\ldots,M})\right)^{-1/2}\ \hat{\mathbf{S}}(\nu)\ \left(\mathrm{Diag}((\hat{S}_m(\nu))_{m=1,\ldots,M})\right)^{-1/2} \tag{8}$$

where for each m, $\hat{S}_m(\nu)$ represents the empirical estimate of the spectral density of component m of $\mathbf{y}$. As above, if the M components of $\mathbf{y}$ were uncorrelated, the left hand side of (8) would reduce to the identity

matrix if the spectral density estimates in the right hand side of (8) were replaced by their true values. This is a motivation to build statistics depending on the empirical eigenvalue distribution and on the largest eigenvalues of $\hat{\mathbf{S}}^{(co)}(\nu)$. As ratio B/N is assumed to converge towards 0, each estimate $\hat{S}_m(\nu)$ should converge towards its true value, and the eigenvalues of $\hat{\mathbf{S}}^{(co)}(\nu)$ should behave as the eigenvalues of the multiplicative deformed version

$$(\mathrm{Diag}((S_m(\nu))_{m=1,\ldots,M}))^{-1/2}\ \hat{\mathbf{S}}(\nu)\ (\mathrm{Diag}((S_m(\nu))_{m=1,\ldots,M}))^{-1/2}$$

of $\hat{\mathbf{S}}(\nu)$. The frequency domain problems presented in paragraph 2.1 will thus allow to study the above frequency-domain statistics.

The technics that will be developed will finally be used in order to revisit in the high-dimensional case the sensor network detection problem considered in [12].

References.

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

[2] R. Bentkus, "Cumulants of estimates of the spectrum of a stationary time series", Lithuanian Math. J., vol. 16, no. 4, pp. 501-518, 1976.

[3] R. Bentkus, R. Rudzkis, "On the distribution of some statistical estimates of spectral density", Theory of Prob. Appl., vol. 27, no. 4, pp.795-814, 1982.

[4] P. Bianchi, M. Debbah, M. Maïda and J. Najim. "Performance of statistical tests for single source detection using random matrix theory" IEEE Trans. Inf. Theory, Vol. 57 (4), april 2011 , 2400–2419.

[5] H. Bohm, R. Von Sachs, "Shrinkage estimation in the frequency domain of multivariate time series", J. of Multivariate Analysis, vol. 100, pp. 913-935, 2009.

[6] D. Brillinger, "Time Series Analysis: Datas and Theory", Classics in Applied Mathematics, SIAM, 2001.

[7] 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.

[8] P. Duchesne, R. Roy, "Robust tests for independence of two time series", Statistica Sinica 13(2003), pp. 827-852.

[9] M. Eichler, " Testing nonparametric and semiparametric hypotheses in vector stationary processes", *Journal of Multivariate Analysis*, 99 (2008), pp. 968-1009.

[10] E.J. Hannan, M. Deistler, "The statistical theory of linear systems", John Wiley & Sons, 1988.

[11] Y. Hong, "Testing for independence between two covariance stationary time series", Biometrika, vol. 83, no. 3 (Sept. 1996), pp. 615-625.

[12] N. Klausner, M. Azimi-Sadjadi, L. Scharf, "Detection of spatially correlated time series from a network of sensor arrays", *IEEE Transactions of Signal Processing*, vol. 62, no. 6, pp. 1396-1407, March 15, 2014.

[13] P. Loubaton, X. Mestre, "Spectral convergence of large block-Hankel Gaussian random matrices", Colombo F., Sabadini I., Struppa D., Vajiac M. (eds) Advances in Complex Analysis and Operator Theory. Trends in Mathematics. Birkhäuser, Cham, 2017, also available on Arxiv, arXiv:1704.06651

[14] X. Mestre, P. Vallet, "Correlation tests and linear spectral statistics of the sample correlation matrix", *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4585 - 4618, July 2017.

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

[16] D. Ramirez, J. Via, I. Santamaria, L. Scharf, "Detection of spatially correlated time series", *IEEE Transactions of Signal Processing*, vol. 58, no. 10, pp. 5006-5015, October 2010.

[17] P. Vallet, X. Mestre, P. Loubaton, "Performance analysis of an improved MUSIC DoA estimator", IEEE Trans. on Signal Processing, vol. 63, no. 23, pp. 6407-6422, December 1 2015

[18] J. Vinogradova, R. Couillet and W. Hachem, "Statistical Inference in Large Antenna Arrays under Unknown Noise Pattern",, IEEE Transactions on Signal Processing, 61 (22), 2013, pages 5633-5645.

[19] W.B. Wu, P. Zaffaroni, "Uniform convergence of multivariate spectral density estimates", 2015, Preprint, ArXiv:1505.03659.

[20] H. Xiao, W.B. Wu, "Covariance matrix estimation for stationary time series", Annals of Statistics, vol. 40, pp. 466-493, 2012.