On the behaviour of the singular values of empirical autocovariance matrices in the high-dimensional case

D. Tieplova

University Paris-Est Marne la Vallee
B. Verkin Institute for Low Temperature Physics and Engineering

14. 08. 2019

Joint work with P. Loubaton, L. Pastur

We consider a M – dimensional multivariate time series $(y_n)_{n\in\mathbb{Z}}$ generated as

$$y_n = u_n + v_n$$

Where v_n is centred Gaussian "noise" part such that $\mathbb{E}(v_n v_{n+k}^*) = R_N \delta_k$ and u_n is non observable "information" part which admits causal state space representations

$$\begin{aligned} x_{n+1} &= Ax_n + B\nu_n \\ u_n &= Cx_n + D\nu_n \end{aligned}$$

The problematic is to retrieve information on u_n based on $(y_n)_{n=1,\ldots,N}$, e. g. estimate A. The standard low-dimensional approaches use the fact that

- the dimension P of minimal state space representation coincides with the rank of the autocovariance matrix $R_{f|p}^{(L)}$ between $u_n^L = (u_n^T, \ldots, u_{n+L-1}^T)^T$ (the past) and $u_{n+L}^L = (u_{n+L}^T, \ldots, u_{n+2L-1}^T)^T$ (the future) , if $L \geq P$.

Therefore an important information can be obtained from the singular value decomposition of $R_{f|p}^{(L)}$

In practice true matrix $R_{f|p}^{(L)}$ is replaced by the empirical estimate $\hat{R}_{f|p}^{(L)} = \frac{Y_f^{(L)}(Y_p^{(L)})^*}{N}$ where matrices $Y_p^{(L)}$ and $Y_f^{(L)}$ are defined by

$$Y_p^{(L)} = \begin{pmatrix} y_1 & y_2 & \cdots & y_N \\ y_2 & y_3 & \cdots & y_{N+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_{N+L-1} \end{pmatrix}, \quad Y_f^{(L)} = \begin{pmatrix} y_{L+1} & y_{L+2} & \cdots & y_{N+L} \\ y_{L+2} & y_{L+3} & \cdots & y_{N+L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{2L} & y_{2L+1} & \cdots & y_{N+2L-1} \end{pmatrix}$$

If $N \to +\infty$ while M, L and P are fixed parameters, $\|\hat{R}_{f|p}^{(L)} - R_{f|p}^{(L)}\| \to 0$ which is not true in high dimensional regime, e. g. if $M \to +\infty$, $N \to +\infty$ such that $c_N = \frac{ML}{N} \to c_*$, $0 < c_* < +\infty$.

To be able to improve the classical schemes it is necessary to

- study the behaviour of the eigenvalues of $\hat{R}_{f|p}^{(L)}(\hat{R}_{f|p}^{(L)})^*$

and for this we begin with

- studying the model when the signal is absent, to understand the contribution of the eigenvalues due to the noise.

Previous works

- Z. Li, G. Pan, J. Yao, ”On singular value distribution of large-dimensional autocovariance matrices”, J. Multivariate Analysis, vol. 137, pp. 119-140, May 2015.
- Z. Li, Q. Wang, J. Yao, ”Identifying the number of factors from singular values of a large sample auto-covariance matrix”, to appear in Annals of Statistics, December 2016.

Here we are in the regime when $M, N \to +\infty$, $L = 1$, $R_N = I_M$.

Stieltjes transform

The Stieltjes transform f of a measure ν,

$$f(z) = \int \frac{\nu(d\lambda)}{\lambda - z}, \quad \Im z \neq 0.$$

There is the one-to-one correspondence between finite nonnegative measures and their Stieltjes transforms.

Normalized Counting Measure

Denote by $\{\hat{\lambda}_i\}_{i=1}^{ML}$ the eigenvalues of matrix $\hat{R}_{f|p}^{(L)}\hat{R}_{f|p}^{(L)*}$. Then the Normalized Counting Measure $\hat{\nu}_N$ is defined as

$$\hat{\nu}_N(\Delta) = \frac{\#\{\hat{\lambda}_i \in \Delta\}}{ML}$$

Deterministic equivalent

For each $z \in \mathbb{C}^+$, the equation

$$t_N(z) = \frac{1}{M}\mathrm{Tr} R_N \left(-zI_M - \frac{zc_N t(z)}{1 - z(c_N t_N(z))^2} R_N\right)^{-1}$$

has a unique solution that belongs to $\mathbb{C}^+$. Moreover, function $z \to t_N(z)$ is the Stieltjes transform of a measure carried by $\mathbb{R}^+$. We construct the function $s_N(z) = \frac{1}{M}\mathrm{Tr}\left(-zI_M - \frac{zc_N t_N(z)}{1 - z(c_N t_N(z))^2} R_N\right)^{-1}$ which is also the Stieltjes transform of some measure ν_N. Moreover, $\hat{\nu}_N - \nu_N \to 0$ weakly almost surely when $N \to +\infty$.

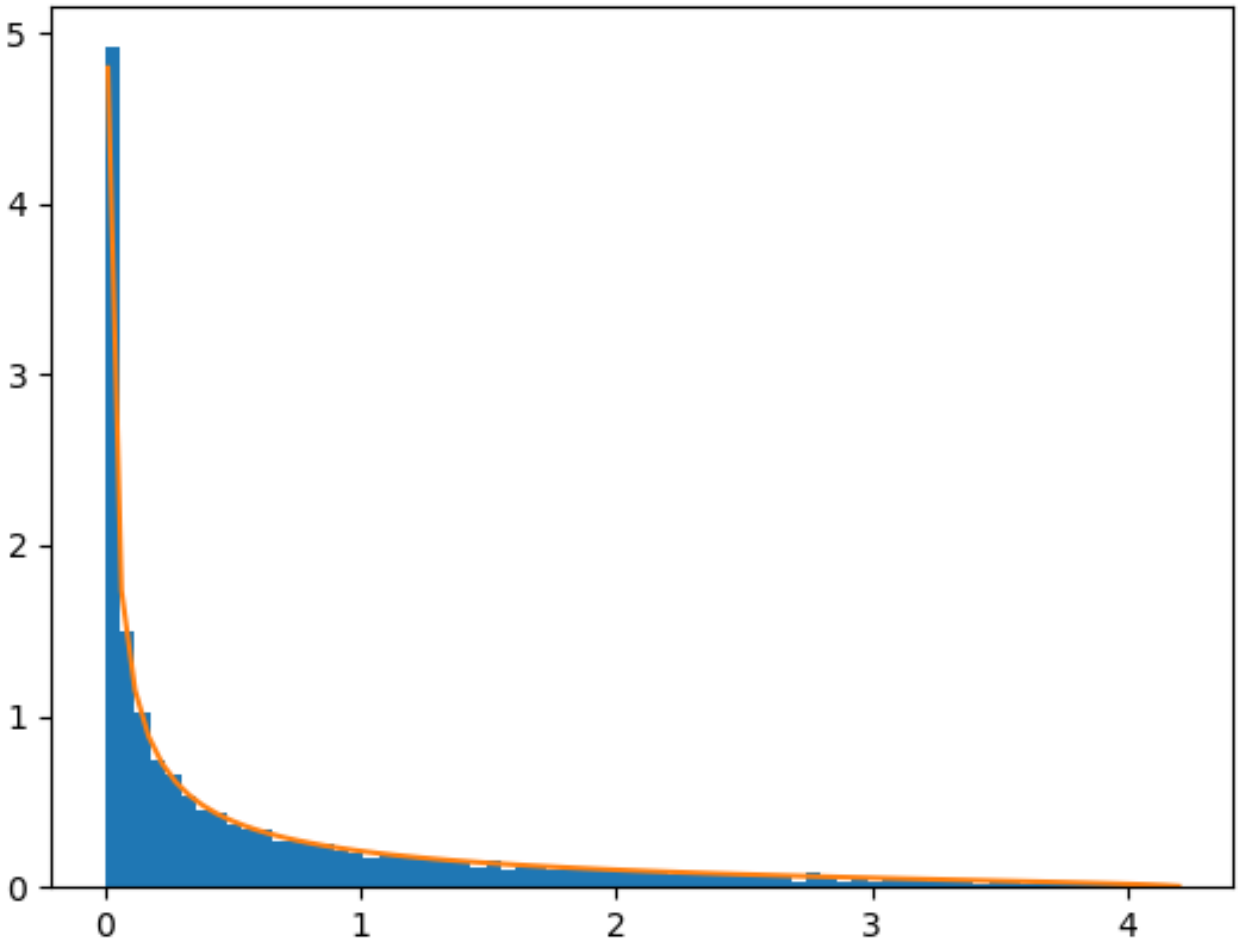

Figure: Histogram of the eigenvalues of $\hat{R}_{f|p}^{(L)}\hat{R}_{f|p}^{(L)*}$ and graph of the density of $\nu_N(x)$ for $M = 500, N = 1500, L = 2$

Location of the eigenvalues

We denote by $\mathcal{S}_N$ the support of measure ν_N, the following holds:

Theorem

Assume that there exists a positive quantity $\epsilon > 0$, two real values $e, f \in \mathbb{R}$ and an integer N_0 such that

$$(e - \epsilon,\ f + \epsilon) \cap \mathcal{S}_N = \varnothing \qquad \forall N \in \mathbb{N},\ N \geq N_0.$$

Then, almost surely, no eigenvalue of $\hat{R}_{f|p}^{(L)} \hat{R}_{f|p}^{(L)*}$ appears in $[e,\ f]$ for all N large enough.

When signal is present, that is $y_n = u_n + v_n$, it is important to characterise the support of deterministic equivalent more precisely.

In the presence of signal

We assume that dimension K of ν_n is also fixed, then the rank r of $\mathbb{E}(u_n^L u_n^{*L})$, is constant for N big enough and we can use finite rank perturbation methods. One can expect that the number of eigenvalues s, that can escape the support $\mathcal{S}_N$ is not greater than P (the rank of $N^{-1} U_f^{(L)} (U_p^{(L)})^*$). In reality we obtain that s can take values between 0 and $2r$ which can be much bigger than P.