Spectra of Sample Covariance Matrices for Multiple Time Series

Reimer Kühn, Peter Sollich

Disordered System Group, Department of Mathematics, King's College London

VIIIth Brunel-Bielefeld Workshop on Random Matrices, Dec 14-15, 2012

Sample Covariance Matrices of Multiple Time Series

- Covariance matrix of stationary stochastic process $\boldsymbol{x}_t = (x_t^a)$, $t \in \mathbb{Z}$, $1 \le a \le p$:

$$C_{ij}^{ab} = \frac{1}{M} \sum_{t=1}^{M} x_{i+t}^a x_{j+t}^b = \frac{1}{M} (XX^T)_{ij}^{ab} \; .$$

  Here $X = (\boldsymbol{x}_{it})$ is $pN \times M$ matrix with entries $\boldsymbol{x}_{it} = \boldsymbol{x}_{i+t}$.
  Expect finite sample fluctuation around mean

$$C_{ij}^{ab} = \langle x_i^a x_j^b \rangle \pm \mathcal{O}(1/\sqrt{M}) = \bar{C}^{ab}(i-j) \pm \mathcal{O}(1/\sqrt{M})$$

  $\Rightarrow$ C is randomly perturbed block Toeplitz matrix.

- Spectrum of C as $N \to \infty$, $M \to \infty$ @ fixed p and $\alpha = N/M$?
  Known result as $\alpha \to 0$: Szegö's Theorem

$$\rho_0(\lambda) = \frac{1}{p} \sum_{s=1}^{p} \int_0^{2\pi} \frac{\mathrm{d}q}{2\pi} \, \delta(\lambda - \widehat{C}_s(q))$$

Compare with Wishart–Laguerre Ensemble

- Empirical covariances for N data, evaluated on the basis of M measurements for each variable. Use $N \times M$ matrices $X = (x_{it})$ with i.i.d. entries x_{it} to compute:

$$C_{ij} = \frac{1}{M}(XX^T)_{ij} = \frac{1}{M}\sum_{t=1}^{M} x_{it}x_{jt} \ .$$

  Expect finite sample fluctuation around mean.

$$C_{ij} = \langle x_i x_j \rangle \pm \mathcal{O}(1/\sqrt{M}) = \delta_{ij} \pm \mathcal{O}(1/\sqrt{M})$$

- Spectrum of C as $N \to \infty$, $M \to \infty$ @ fixed $\alpha = N/M$?
  $\Rightarrow$ Marčenko Pastur-Law

$$\rho_\alpha(\lambda) = \frac{1}{2\pi\alpha\lambda}\,\sqrt{4\alpha - (\lambda - (1+\alpha))^2}$$

Principal Differences

- Rows of X for the multi-time series covariance problem groups of shifted sections of a set of p time series $(\boldsymbol{x}_t)_{t\in\mathbb{Z}}$, $\boldsymbol{x}_t \in \mathbb{R}^p$.

$$X = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \dots & \boldsymbol{x}_M \\ \boldsymbol{x}_2 & \boldsymbol{x}_3 & \boldsymbol{x}_4 & \dots & \boldsymbol{x}_{1+M} \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{x}_N & \boldsymbol{x}_{N+1} & \boldsymbol{x}_{N+2} & \dots & \boldsymbol{x}_{N+M} \end{pmatrix}$$

- Number of random variables in the problem is $\mathcal{O}(N)$, rather than $\mathcal{O}(N^2)$ as in the Wishart Laguerre ensemble.

- Extensive body of knowledge about the Wishart-Laguerre ensemble and its variants (applications in multivariate statistics, signal-processing, finance, . . .)

- Very little is known. Existence proofs (Basak, Bose, Sen 2011). $p = 1$-case solved only recently. (RK, P Sollich, EPL 2012)

Spectral Density and Resolvent

- Spectral density of sample covariance matrix from resolvent

$$\rho(\lambda) = \lim_{N\to\infty} \frac{1}{\pi N p} \mathrm{Im}\ \mathrm{Tr} \left\langle \left[\lambda_\varepsilon 1\!\!\mathrm{I} - C\right]^{-1} \right\rangle , \qquad \lambda_\varepsilon = \lambda - \mathrm{i}\varepsilon$$

- Express as (S F Edwards & R C Jones, JPA, 1976)

$$\begin{aligned} \rho_\alpha(\lambda) &= \lim_{N\to\infty} \frac{1}{\pi N_p} \mathrm{Im} \frac{\partial}{\partial\lambda} \mathrm{Tr} \left\langle \ln\left[\lambda_\varepsilon 1\!\!\mathrm{I} - C\right] \right\rangle \\ &= \lim_{N\to\infty} -\frac{2}{\pi N_p} \mathrm{Im} \frac{\partial}{\partial\lambda} \left\langle \ln Z_{N_p} \right\rangle , \end{aligned}$$

where $N_p = Np$ and Z_{N_p} is a Gaussian integral:

$$Z_{N_p} = \int \prod_{k,a} \frac{\mathrm{d}u_{ka}}{\sqrt{2\pi/\mathrm{i}}} \exp\left\{ -\frac{\mathrm{i}}{2} \sum_{ka,\ell b} u_{ka}(\lambda_\varepsilon \delta_{ab}\delta_{k\ell} - C^{ab}_{k\ell}) u_{\ell b} \right\} .$$

Performing the Average

- Standard Approach – Replica Method

$$\left\langle \ln Z_{N_p} \right\rangle = \lim_{n\to 0} \frac{1}{n} \ln \left\langle Z_{N_p}^n \right\rangle$$

- For integer n, $Z_{N_p}^n$ is partition function of n identical copies of the system (n-th power of Gaussian integral)

- Experience: final result has structure of replica-symmetric high-temperature solution $\Leftrightarrow$ annealed calculation ($n = 1$). $\langle \ln Z_{N_p} \rangle \simeq \ln \langle Z_{N_p} \rangle$ $\Rightarrow$ Do annealed calculation from the start

$$\langle Z_{N_p} \rangle = \left\langle \int \prod_{k,a} \frac{\mathrm{d}u_{ka}}{\sqrt{2\pi/\mathrm{i}}} \; \exp\left\{ -\frac{\mathrm{i}}{2} \sum_{ka,\ell b} u_{ka} (\lambda_\varepsilon \delta_{ab} \delta_{k\ell} - C_{k\ell}^{ab}) u_{\ell b} \right\} \right\rangle$$

Performing the Average (contd.)

- Insert definition of C, and $\alpha_p = \alpha p$,

$$\langle Z_{N_p} \rangle = \left\langle \int \prod_{ka} \frac{\mathrm{d}u_{ka}}{\sqrt{2\pi/\mathrm{i}}} \; \exp\left\{ -\frac{\mathrm{i}}{2}\lambda_\varepsilon \sum_{ka} u_{ka}^2 + \frac{\mathrm{i}}{2}\alpha_p \sum_{i=1}^{M} z_i^2 \right\} \right\rangle$$

- with disorder dependence of Z_{N_p} only through the M variables

$$z_i = \frac{1}{\sqrt{N_p}} \sum_{ka} x_{k+i}^a u_{ka} \ , \qquad 1 \le i \le M \ .$$

- By CLT (for weakly dependent rv's) normally distributed for large M with

$$\langle z_i \rangle = 0 \ , \qquad \langle z_i z_j \rangle = \frac{1}{N_p} \sum_{ka,\ell b} \langle x_{k+i}^a x_{\ell+j}^b \rangle u_{ka} u_{\ell b} \equiv Q_{ij}$$

and Q given in terms of true process auto-covariance

$$Q_{ij} = \langle z_i z_j \rangle = \frac{1}{N_p} \sum_{ka,\ell b} \bar{C}^{ab}(i - j + k - \ell)\, u_{ka} u_{\ell b}$$

Exploiting Szegö's Theorem for Spectral Sums

- $\{z_i\}$ average is Gaussian; with $\alpha_p = \alpha p$:

$$\langle Z_{N_p}\rangle = \int \prod_{ka} \frac{\mathrm{d}u_{ka}}{\sqrt{2\pi/\mathrm{i}}} \exp\left\{ -\frac{\mathrm{i}}{2}\lambda_\varepsilon \sum_{ka} u_{ka}^2 - \frac{1}{2}\ln\det(\mathbb{1} - \mathrm{i}\alpha_p Q)\right\}$$

- Q is a Toeplitz matrix. $\Rightarrow$ evaluate $\ln\det\,(\mathbb{1} - \mathrm{i}\alpha_p Q)$ using **Szegö's theorem**:

$$\ln\det(\mathbb{1} - \mathrm{i}\alpha_p Q) \sim \sum_{\mu=-(M-1)/2}^{(M-1)/2} \ln\Big(1 - \mathrm{i}\alpha_p Q_\mu\Big)$$

where

$$Q_\mu = \frac{1}{N_p}\sum_{ka,\ell b} \hat{C}^{ab}(q_\mu)\mathrm{e}^{-\mathrm{i}q_\mu(k-\ell)}\, u_{ka}u_{\ell b} = \frac{1}{p}\sum_{ab} \hat{C}^{ab}(q_\mu)\hat{u}_a^*(q_\mu)u_b(q_\mu)$$

with $\hat{u}_a(q_\mu) = \frac{1}{\sqrt{N}}\sum_{k=1}^N \mathrm{e}^{\mathrm{i}q_\mu k}\, u_{ka}$ and $q_\mu = \frac{2\pi}{M}\mu$.

Closed Form Approximation & Scaling

- Get closed form expression of $\langle Z_{N_p}\rangle = \prod_{\nu\geq 0} I_\nu$ with

$$I_\nu = \frac{\mathrm{i}^p}{\prod_s \widehat{C}_s(p_\nu)} \int_0^\infty \prod_{s=1}^{p} \mathrm{d}x_s \; \frac{\mathrm{e}^{-\mathrm{i}\sum_s x_s \lambda_\varepsilon / \widehat{C}_s(p_\nu)}}{\left(1 - \mathrm{i}\frac{\alpha}{2}\sum_s x_s\right)^{2/\alpha}}$$

  where $\widehat{C}_s(p_\nu)$, $s = 1, \ldots, p$ are eigenvalues of $\widehat{\boldsymbol{C}}(p_\nu) = (\widehat{C}^{ab}(p_\nu))$

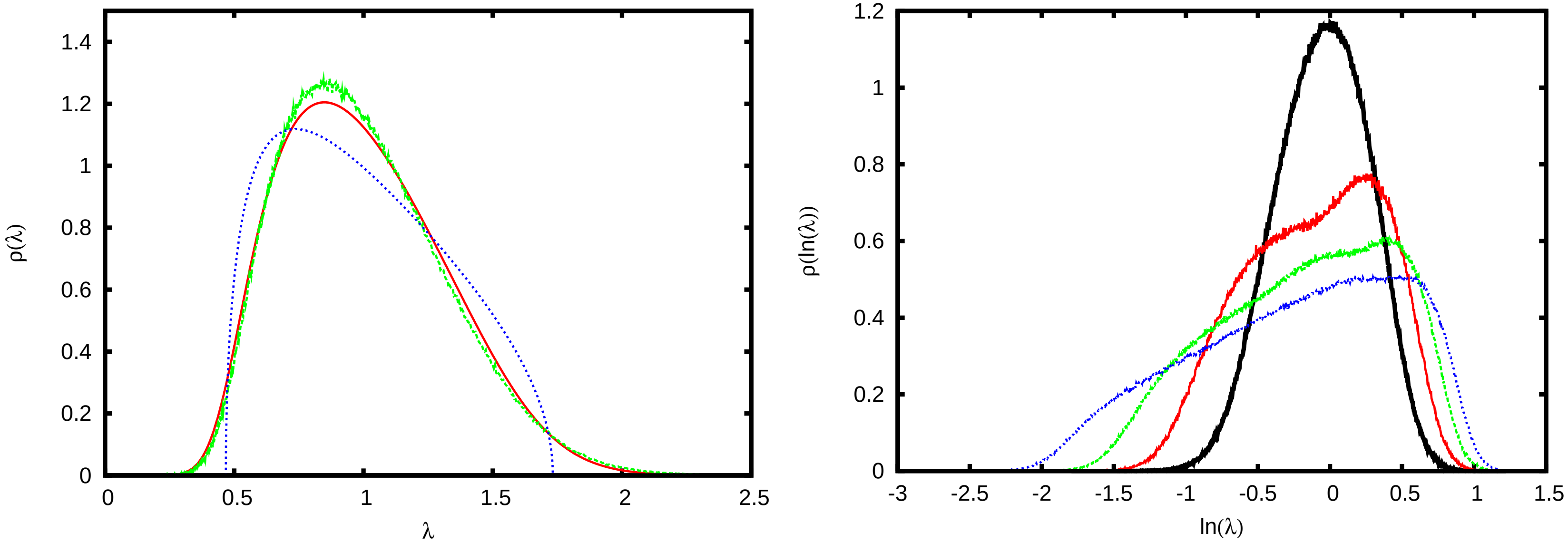
- Gives

$$\rho_\alpha(\lambda) = \frac{1}{p}\int_0^\pi \frac{\mathrm{d}q}{\pi} \; \sum_{s=1}^{p} \frac{1}{\widehat{C}_s(q)} \rho_s\left(\left\{\frac{\lambda}{\widehat{C}_s(q)}\right\}\right)$$

- For uncorrelated data $\widehat{C}_s(q) \equiv 1$, and ρ_s is independent of $s \Rightarrow$ identify ρ_s with the spectral density for covariance matrices of p time series of i.i.d. (uncorrelated) data.

Numerical Tests

- Spectral density for $x_n \sim \mathcal{N}(0,1)$ i.i.d. @ $\alpha = 0.1$



(Left) $p = 1$: simulation results (green); analytic approximation for $\rho_\alpha^{(0)}(\lambda)$ (red), Marčenko-Pastur law (blue-dashed) (. From RK, P Sollich, EPL 2012.)

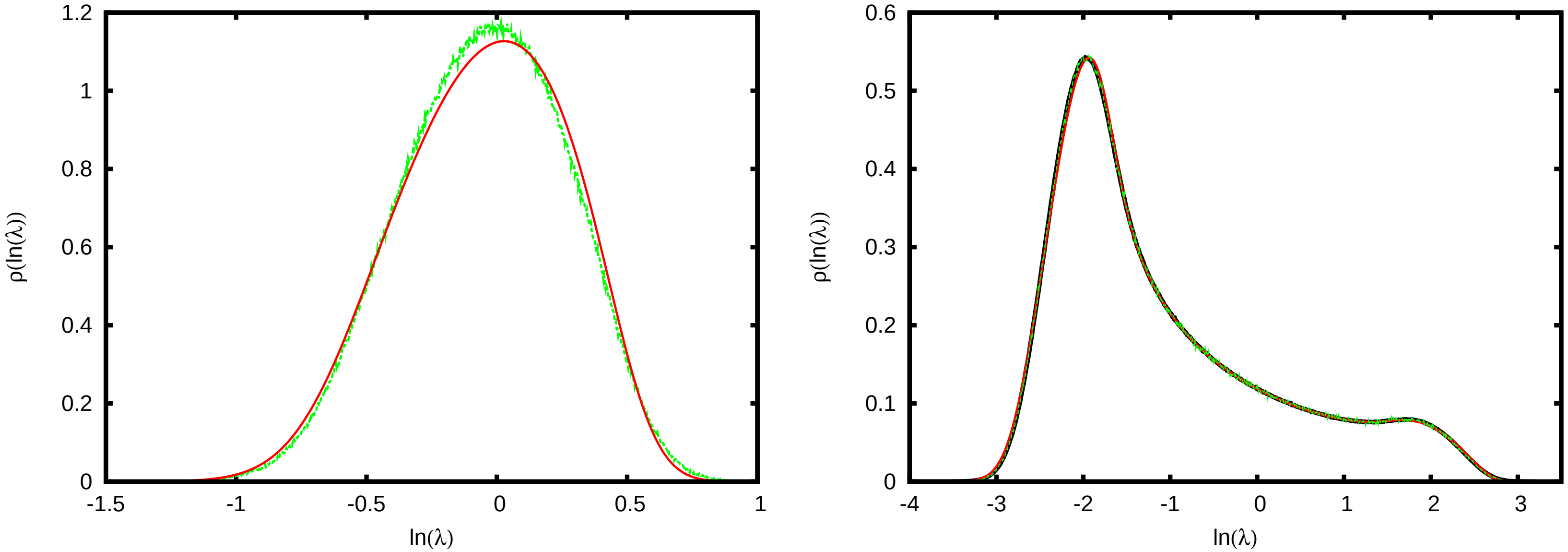
(Right) logarithmic spectral density; simulation results for $p = 1, \ldots, 4$. In all cases @

$\alpha = 0.1$

AR-1 Process @ $\alpha = 0.1$, $p = 1$

$$x_n = a_1\, x_{n-1} + \sqrt{1 - a_1^2}\, \xi_n$$

- (Logarithmic) Spectral density for AR-1 process @ $\alpha = 0.1$



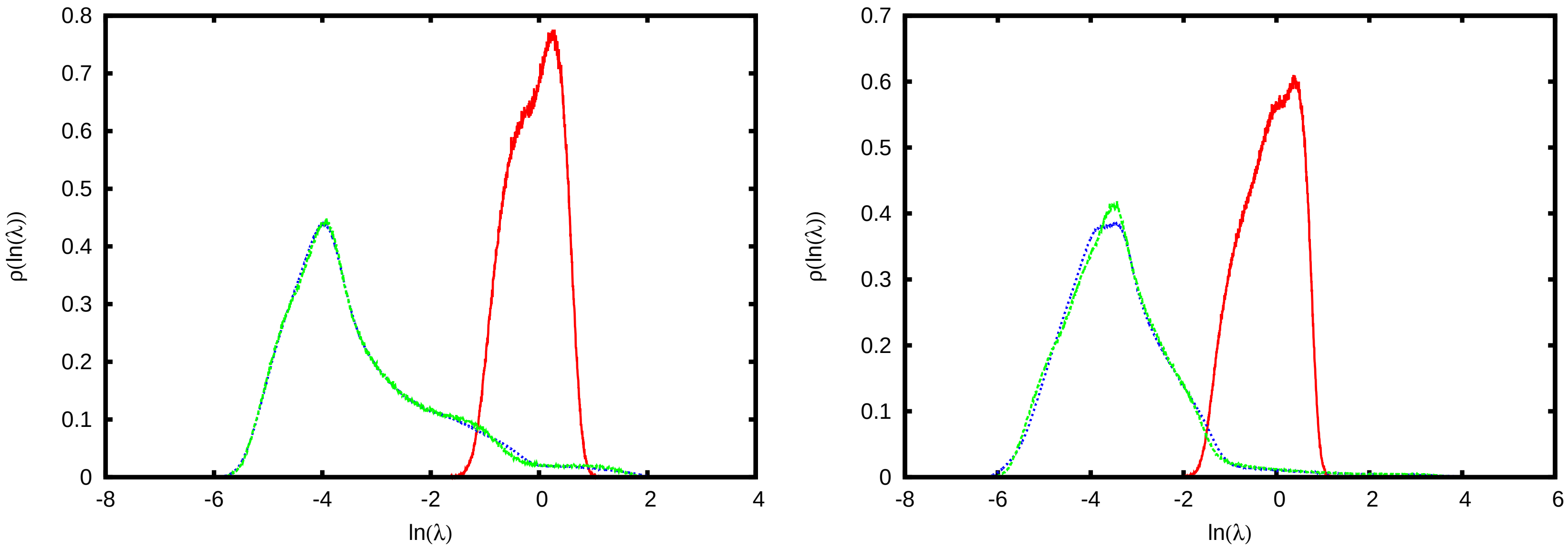
Left: i.i.d. data, simulation (green) and analytic result (red).

Right $a_1 = 0.8$. Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).

AR-1 Process @ $\alpha = 0.1$, $p = 2$ and $p = 3$

- (Logarithmic) Spectral density for AR-1 process @ $\alpha = 0.1$

$$\boldsymbol{x}_n = A\,\boldsymbol{x}_{n-1} + \sigma\,\boldsymbol{\xi}_n$$



(Left) Spectra for $p = 2$, uncorrelated data and $A = [0.8\ 0.1; 0.1\ 0.8]$, **(Right)** $p = 3$, uncorrelated data and $A = [0.8\ 0.2\ 0.1; 0.2\ 0.6\ 0.1; 0.1\ 0.1\ 0.5]$. In both cases $\sigma = 0.2$, and simulation results are compared with scaling using an approximate evaluation of scaling integrals.

Summary

- Computed DOS of sample covariance matrices for multiple time-series using annealed calculation.

- Key ingredient: Szegö's theorem for (block) Toeplitz matrices

- Rectangular window and decorrelation approximation $\Rightarrow$ Closed form approximation.

- Use of Szegös theorem suggests a scaling form for DOS.

  – scaling is requires knowledge of a function on $\mathbb{R}^p$! DOS for i.i.d. data is insufficient.

  – currently working on effective methods to evaluate scaling function for $p > 1$.

- Lots of possible applications.