Spectra of Empirical Auto-Covariance Matrices

Reimer Kühn, Peter Sollich
Disordered System Group, Department of Mathematics, King's College London

VII Brunel–Bielefeld Workshop
Random Matrix Theory and Applications in Theoretical Sciences

Bielefeld, Dec 15 – 17, 2011
Overview

- Empirical Auto-Covariance Matrices of Time Series
- Spectral Density and Resolvent (Edwards and Jones, 1976)
  - Annealed Average
  - Exploit Szegö’s theorem
  - Approximations
  - Scaling
- Results
- Summary
Empirical Auto-Covariance Matrices of Time Series

- Auto-covariance matrix of stationary stochastic process: Use $N \times M$ matrix $X = (x_{it})$ with entries $x_{it} = x_{i+t}$ to compute

$$C_{ij} = \frac{1}{M}(XX^T)_{ij} = \frac{1}{M} \sum_{t=0}^{M-1} x_{i+t}x_{j+t}.$$  

Expect finite sample fluctuation around mean

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

$\Rightarrow C$ is randomly perturbed Toeplitz matrix.

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

Known result as $\alpha \to 0$: Szegő (Wiener-Khinchin) Theorem

$$\rho_0(\lambda) = \int_0^{2\pi} \frac{dq}{2\pi} \delta(\lambda - \hat{C}(q))$$
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.

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

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

  $\Rightarrow$ Marcenko Pastur-Law

$$\rho_\alpha(\lambda) = \left(1 - \frac{1}{\alpha}\right) \delta(\lambda) + \frac{\sqrt{4\alpha - (\lambda - (1 + \alpha))^2}}{2\pi\alpha\lambda} \mathbb{I}_{\lambda \in [\lambda_- , \lambda_+]}$$
Principal Differences

- Rows of $X$ for the auto-covariance problem are sections of a single time series $(x_t)_{t \in \mathbb{Z}}$

$$X = \begin{pmatrix}
x_1 & x_2 & x_3 & \cdots & x_M \\
x_2 & x_3 & x_4 & \cdots & x_{1+M} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_N & x_{N+1} & x_{N+2} & \cdots & x_{N+M}
\end{pmatrix}$$

- Number of random variables in the problem is $O(N)$, rather than $O(N^2)$ 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, \ldots)

- Little is known about the auto-covariance problem
Spectral Density and Resolvent

- Spectral density of sample covariance matrix from resolvent

\[ \rho(\lambda) = \lim_{N \to \infty} \frac{1}{\pi N} \text{Im} \ Tr \left\langle \left[ \lambda_\varepsilon \mathbb{1} - C \right]^{-1} \right\rangle, \quad \lambda_\varepsilon = \lambda - i\varepsilon \]

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

\[
\begin{align*}
\rho(\lambda) &= \lim_{N \to \infty} \frac{1}{\pi N} \text{Im} \ \frac{\partial}{\partial \lambda} \ Tr \left\langle \ln \left[ \lambda_\varepsilon \mathbb{1} - C \right] \right\rangle \\
&= \lim_{N \to \infty} - \frac{2}{\pi N} \text{Im} \ \frac{\partial}{\partial \lambda} \left\langle \ln Z_N \right\rangle,
\end{align*}
\]

where \( Z_N \) is a Gaussian integral:

\[
Z_N = \int \prod_{k=1}^{N} \frac{du_k}{\sqrt{2\pi/i}} \exp \left\{ -\frac{i}{2} \sum_{k,\ell} u_i (\lambda_\varepsilon \delta_{k\ell} - C_{k\ell}) u_\ell \right\}
\]
Performing the Average

- Standard Approach – Replica Method

\[ \langle \ln Z_N \rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle Z^*_N \rangle \]

- For integer \( n \), \( Z^*_N \) 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 (\( n = 1 \)).

\[ \Rightarrow \text{Do annealed calculation from the start} \]

\[ \langle Z_N \rangle = \left\langle \int \prod_k \frac{du_k}{\sqrt{2\pi/i}} \exp \left\{ -\frac{i}{2} \lambda \varepsilon \sum_k u_k^2 + \frac{i}{2} \sum_{k\ell} C_{k\ell} u_k u_\ell \right\} \right\rangle, \]
• Insert definition of $C$

$$
\langle Z_N \rangle = \left\langle \int \prod_k \frac{du_k}{\sqrt{2\pi/i}} \exp \left\{ -\frac{i}{2} \lambda \sum_k u_k^2 + \frac{i}{2} \alpha \sum_{i=1}^M \left( \frac{1}{\sqrt{N}} \sum_k x_{k+i} u_k \right)^2 \right\} \rightangle
$$

• Disorder dependence of $Z_N$ only through the variables

$$
z_i = \frac{1}{\sqrt{N}} \sum_{k=1}^N x_{k+i} u_k .
$$

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

$$
\langle z_i \rangle = 0 , \quad \langle z_i z_j \rangle = \frac{1}{N} \sum_{k\ell} \langle x_{k+i} x_{\ell+j} \rangle u_k u_\ell \equiv Q_{i,j}
$$

and $Q$ given in terms of true process auto-covariance

$$
Q_{i,j} = \langle z_i z_j \rangle = \frac{1}{N} \sum_{k\ell} \bar{\mathcal{C}}(i - j + k - \ell) u_k u_\ell
$$
• \{z_i\} average is Gaussian

\[ \langle Z_N \rangle = \int \prod_k \frac{d u_{k\alpha}}{\sqrt{2\pi/i}} \exp \left\{ -\frac{i}{2} \lambda \varepsilon \sum_k u_k^2 - \frac{1}{2} \ln \det(1 - i\alpha Q) \right\} \]

• \( Q \) is a Toeplitz matrix. \( \Rightarrow \) evaluate \( \ln \det (1 - i\alpha Q) \) using

• **Szegö’s theorem**: Given an \( N \times N \) Toeplitz matrix \( A \) with elements \( A_{ik} = a(i - k) \), where \( a = a(n) \in \ell_1(\mathbb{Z}) \). Then the spectral density has a weak limit

\[ \rho_N(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \overset{w}{\rightarrow} \int_{-\pi}^{\pi} \frac{dq}{2\pi} \delta(\lambda - \hat{a}(q)) , \]

as \( N \to \infty \), where \( \hat{a}(q) \) is called the ‘symbol’, and is nothing but the Fourier transform of \( a \)

\[ \hat{a}(q) = \sum_{n=-\infty}^{\infty} a(n)e^{iqn} . \]
Fourier-transform of $\bar{C}$

$$\hat{C}(q) = \sum_{n=-\infty}^{\infty} \bar{C}(n)e^{iqn}, \quad q \in [0, 2\pi]$$

Inverse

$$\bar{C}(n) = \int_{-\pi}^{\pi} \frac{dq}{2\pi} \hat{C}(q)e^{-iqn}$$

Need to keep track of finite $N$/finite $M$ expressions, so use

$$\hat{C}(q_\mu) = \sum_{n=-(M-1)/2}^{(M-1)/2} \bar{C}(n)e^{i\mu n}, \quad \text{with} \quad q_\mu = \frac{2\pi}{M}\mu$$

with inverse

$$\bar{C}(n) = \frac{1}{M} \sum_{\mu=-(M-1)/2}^{(M-1)/2} \hat{C}(q_\mu)e^{-i\mu n}$$
• Gives

\[ Q_{ij} = \frac{1}{M} \sum_{\mu=-(M-1)/2}^{(M-1)/2} e^{-i\mu(i-j)} Q_\mu \]

where

\[ Q_\mu = \frac{1}{N} \sum_{k\ell} \hat{C}(q_\mu) e^{-i(q_k-q_\ell)} u_k u_\ell = \hat{C}(q_\mu) |\hat{u}(q_\mu)|^2 \equiv Q(q_\mu) \]

with

\[ \hat{u}(q_\mu) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} e^{i\mu k} \]

• Szegö

\[ \ln \det(\mathbb{I} - i\alpha Q) \sim \sum_{\mu=-M/2}^{M/2} \ln \left(1 - i\alpha Q_\mu\right) \]

• Extract \( \{u_k\} \) dependence (via the \( \{Q_\mu\} \)) from \( \ln \det \) evaluation by using \( \delta \)-functions and their Fourier representations.
• Entails that the $u_k$ integrals are Gaussian, giving

$$\langle Z_N \rangle = \int\prod_{\mu=0}^{(M-1)/2} \frac{d\hat{Q}_\mu dQ_\mu}{2\pi} \exp \left\{ - \sum_{\mu=0}^{(M-1)/2} i\hat{Q}_\mu Q_\mu \right. \right.$$

$$\left. - \sum_{\mu=0}^{(M-1)/2} \ln(1 - i\alpha Q_\mu) - \frac{1}{2} \ln \det(\lambda_\varepsilon \mathbb{I} - R) \right\}$$

with elements of $R$ given by

$$R_{k\ell} = \frac{2}{N} \sum_{\mu=0}^{(M-1)/2} \hat{Q}_\mu \hat{C}(q_\mu) \cos(q_\mu(k - \ell))$$.

• Next, evaluate $Q_\mu$ integrals using residues

$$\int \frac{dQ_\mu}{2\pi} \frac{\exp \left\{ - i\hat{Q}_\mu Q_\mu \right\}}{1 - i\alpha Q_\mu} = \left\{ \begin{array}{ll}
\frac{1}{\alpha} \exp \left\{ - \hat{Q}_\mu / \alpha \right\} ; & \hat{Q}_\mu > 0 , \\
0 ; & \text{else ,}
\end{array} \right.$$
After substitution $Q_\mu/\alpha \to Q_\mu$, gives

$$\langle Z_N \rangle = \left\langle \exp \left\{ -\frac{1}{2} \ln \det(\lambda_\varepsilon \mathbb{I} - R) \right\} \right\rangle_{\{\hat{Q}_\mu\}},$$

with

$$\langle \ldots \rangle_{\{\hat{Q}_\mu\}} = \int \prod_{\mu=0}^{(M-1)/2} \left\{ d\hat{Q}_\mu e^{-\hat{Q}_\mu} \right\}(\ldots),$$

and now

$$R_{k\ell} = \frac{2}{M} \sum_{\mu=0}^{(M-1)/2} \hat{Q}_\mu \hat{C}(q_\mu) \cos(q_\mu(k - \ell)).$$

The matrix $R$, too, is Toeplitz. Evaluate $\ln \det(\lambda_\varepsilon \mathbb{I} - R)$ using Szegö,

$$\ln \det(\lambda_\varepsilon \mathbb{I} - R) \sim \sum_{\nu = -N/2}^{N/2} \ln \left( \lambda_\varepsilon - R_\nu \right).$$
with

\[ R_\nu = \sum_{\mu=0}^{M/2} \hat{Q}(q_\mu) \hat{C}(q_\mu) S_{\nu\mu} \equiv R(p_\nu) , \quad p_\nu = \frac{2\pi}{N} \nu \]

and

\[ S_{\nu\mu} := \frac{1}{M} \sum_{\sigma = \pm 1} \frac{\sin(N(p_\nu - \sigma q_\mu)/2)}{\sin((p_\nu - \sigma q_\mu)/2)} \]

- Enforce \( R_\nu \) definitions via \( \delta \)-functions

\[
\langle Z_N \rangle = \left\langle \int \prod_{\nu=0}^{(N-1)/2} \left\{ \frac{d\hat{R}_\nu dR_\nu e^{-i\hat{R}_\nu R_\nu}}{2\pi} \frac{e^{i\hat{R}_\nu \sum_{\mu=0}^{(M-1)/2} \hat{Q}_\mu \hat{C}(q_\mu) S_{\nu\mu}}}{\lambda_\nu - R_\nu} \right\} \right\rangle_{\{\hat{Q}_\mu\}}
\]

- Do \( R_\nu \) integrals using residues
• Gives

\[
\langle Z_N \rangle = (i)^{N/2} \left( \prod_{\nu=0}^{(N-1)/2} F_{\nu} \right) \{Q_\mu\}
\]

with

\[
F_{\nu} = \int_0^\infty d\hat{R}_\nu \ e^{-i\hat{R}_\nu \left( \lambda_\varepsilon - \sum_\mu \hat{Q}_\mu \bar{C}(q_\mu) S_{\nu\mu} \right)}
\]

• \(S\)-kernel couples \(\hat{Q}_\mu\) for \(\mu \in I_\nu = \{\mu; |\mu - \nu/\alpha| \leq 1/\alpha\}\).

\[
S_{\nu\mu} \text{ at } \nu = 10 \text{ as a function of } \mu, \text{for } \alpha = 0.1
\]
• Approximations (using smoothness of $\hat{C}(q_\mu)$ on $q_\mu$ scale)

\[
(i) \quad \sum_\mu \hat{Q}_\mu \hat{C}(q_\mu) S_{\nu\mu} \simeq \frac{\alpha}{2} \hat{C}(p_\nu) \sum_{\mu \in I_\nu} Q_\mu
\]

\[
(ii) \quad \left\langle \prod_{\nu=0}^{(N-1)/2} F_\nu \right\rangle \{\hat{Q}_\mu\} \simeq \prod_{\nu=0}^{(N-1)/2} \left\langle F_\nu \right\rangle \{\hat{Q}_\mu\}
\]

• Allow closed form expression of $\langle Z_N \rangle$, and hence $\rho(\lambda)$

\[
\langle Z_N \rangle \simeq (i)^{N/2} \prod_{\nu=0}^{(N-1)/2} \left\{ \frac{2}{\alpha \hat{C}(p_\nu)} \int_0^\infty dy \frac{e^{-iy\lambda\epsilon}2/(\alpha \hat{C}(p_\nu))}{(1 - iy)^{2/\alpha}} \right\}
\]

• Gives

\[
\rho(\lambda) = \int_0^\pi dq \frac{1}{\pi \hat{C}(q)} \rho^{(0)}_\alpha \left( \frac{\lambda}{\hat{C}(q)} \right)
\]
Our approximations give

\[ \rho^{(0)}_\alpha(\lambda) = - \lim_{\varepsilon \to 0} \frac{1}{\pi} \text{Im} \frac{\partial}{\partial \lambda} \ln I_\alpha \left( \frac{2}{\alpha} \lambda \varepsilon \right) \]

with

\[ I_\alpha(x) = i (-x)^{-1+2/\alpha} e^{-x} \Gamma(1 - 2/\alpha, -x) , \quad \text{Im} x < 0 . \]

As \( \hat{C}(q) \equiv 1 \) for uncorrelated data, have to identify \( \rho^{(0)}_\alpha \) with the spectral density for auto-covariance matrices of sequences of i.i.d. (uncorrelated) data.

Thus have a remarkable scaling-relation

\[ \rho(\lambda) = \int_0^\pi \frac{dq}{\pi} \frac{1}{\hat{C}(q)} \rho^{(0)}_\alpha \left( \frac{\lambda}{\hat{C}(q)} \right) \]
Results

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

Simulation results (green); analytic approximation for $\rho^{(0)}_{\alpha}(\lambda)$ (red), Marčenko-Pastur law (blue-dashed).
• (Logarithmic) Spectral density for AR-1 process @ $\alpha = 0.1$

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

**Left:** $a_1 = 0$ (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).
(Logarithmic) Spectral density for AR-1 process @ $\alpha = 0.8$

Left: $a_1 = 0$ (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).
• (Logarithmic) Spectral density for AR-2 processes @ $\alpha = 0.1$

\[ x_n + a_1 x_{n-1} + a_2 x_{n-2} = \sigma \xi_n , \quad \sigma \text{ s.t. } \bar{C}(0) = 1. \]

**Left:** $a_1 = 0.5$ $a_2 = -3/16$, **Right:** $a_1 = 0.5$ $a_2 = 5/16$. Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).
(Logarithmic) Spectral density for AR-2 processes @ $\alpha = 0.8$

Left: $a_1 = 0.5$ $a_2 = -3/16$, Right: $a_1 = 0.5$ $a_2 = 5/16$. Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).
Summary

• Computed DOS of sample auto-covariance matrices using annealed calculation.

• Key ingredient: Szegö’s theorem for Toeplitz matrices

• Triangular window and decorrelation approximation ⇒ Closed form approximation.

• Use of Szegös theorem suggests a scaling form for DOS.
  – results suggest that scaling is exact
  – ideas for independent proof

• Results may become useful for time-series analysis

• Thanks! K. Anand, L. Dall’Asta, P. Vivo