Spectra of Sample Auto-Covariance Matrices Derived from Time Series

Reimer Kühn and Peter Sollich

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

Dresden, Sept 28, 2012

Details in Europhys. Lett. 99 20008 (2012), available from http://www.mth.kcl.ac.uk/∼kuehn

Outline

Outline

Sample Auto-Covariance Matrices of Time Series

- Auto-covariance matrix of stationary stochastic process $(x_t)_{t\in\mathbb{Z}}$:

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

  Here $X = (x_{it})$ is $N \times M$ matrix with entries $x_{it} = x_{i+t}$.
  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 $\alpha \to 0$: Szegö's Theorem

$$\rho_0(\lambda) = \int_0^{2\pi} \frac{\mathrm{d}q}{2\pi}\,\delta(\lambda - \hat{C}(q))$$

Outline

Comparison with Wishart-Laguerre Ensemble

- Empirical covariances for N data, evaluated on the basis of M measurements for each variable. Express in terms of $N \times M$ matrices $X = (x_{it})$ as

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

  Expect finite sample fluctuation around mean. For i.i.d. entries x_{it}

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

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

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

Principal Differences

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

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

- 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, ...)

Principal Differences (contd.)

- Comparatively little is known about the auto-covariance problem
  - Existence of limiting spectral densityfor auto-covariance matrices of moving average processes with i.i.d. driving (Basak et. al 2011)
  - Universality of results: independence of statistics of i.i.d. driving (numerical, Sen 2010)
  - Existence of limiting spectral density for random Toeplitz matrices with i.i.d. entries (Bryc 2007)

Outline

Spectral Density and Resolvent

- Spectral density of sample covariance matrix from resolvent

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

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

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

where Z_N is a Gaussian integral:

$$Z_N = \int \prod_{k=1}^{N} \frac{\mathrm{d}u_k}{\sqrt{2\pi/\mathrm{i}}}\ \exp\left\{ -\frac{\mathrm{i}}{2} \sum_{k,\ell} u_k (\lambda_\varepsilon \delta_{k\ell} - C_{k\ell}) u_\ell \right\}$$

Outline

Performing the Average

- Standard Approach – Replica Method

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

- For integer n, Z_N^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 ($n = 1$). $\langle \ln Z_N \rangle \simeq \ln \langle Z_N \rangle \Rightarrow$ Do annealed calculation from the start

$$\langle Z_N \rangle = \left\langle \int \prod_k \frac{\mathrm{d}u_k}{\sqrt{2\pi/\mathrm{i}}} \exp\left\{ -\frac{\mathrm{i}}{2}\lambda_\varepsilon \sum_k u_k^2 + \frac{\mathrm{i}}{2} \sum_{k\ell} C_{k\ell} u_k u_\ell \right\} \right\rangle ,$$

Performing the Average (contd.)

- Insert definition of C

$$\langle Z_N \rangle = \left\langle \int \prod_k \frac{\mathrm{d}u_k}{\sqrt{2\pi/\mathrm{i}}} \exp\left\{ -\frac{\mathrm{i}}{2}\lambda_\varepsilon \sum_k u_k^2 + \frac{\mathrm{i}}{2}\alpha \sum_{i=1}^{M} z_i^2 \right\} \right\rangle$$

- with disorder dependence of Z_N only through the M variables

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

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

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

and Q given in terms of true process auto-covariance

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

Exploiting Szegö's Theorem for Spectral Sums

- $\{z_i\}$ average is Gaussian

$$\langle Z_N \rangle = \int \prod_k \frac{\mathrm{d}u_{ka}}{\sqrt{2\pi/\mathrm{i}}} \exp\left\{ -\frac{\mathrm{i}}{2}\lambda_\varepsilon \sum_k u_k^2 - \frac{1}{2}\ln\det(\mathbb{1} - \mathrm{i}\alpha Q) \right\}$$

- Q is a Toeplitz matrix. $\Rightarrow$ evaluate $\ln\det(\mathbb{1} - \mathrm{i}\alpha Q)$ 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

$$\rho_N(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \delta(\lambda - \lambda_i) \xrightarrow{w} \int_{-\pi}^{\pi} \frac{\mathrm{d}q}{2\pi} \delta(\lambda - \hat{a}(q)) \ ,$$

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

$$\hat{a}(q) = \sum_{n=-\infty}^{\infty} a(n) \mathrm{e}^{\mathrm{i}qn} \ .$$

- Szegö (keeping track of finite-M finite-N expressions)

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

where

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

with

$$\hat{u}(q_\mu) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N} \mathrm{e}^{\mathrm{i}q_\mu k}\, u_k \ , \quad q_\mu = \frac{2\pi}{M}\mu$$

- Enforce Q_μ definitions using δ-functions. $\Rightarrow$ Get Gaussian u_k-integrals.

$$\begin{aligned}\langle Z_N\rangle \quad = \quad & \int \prod_{\mu=0}^{(M-1)/2} \frac{\mathrm{d}\hat{Q}_\mu \mathrm{d}Q_\mu}{2\pi}\, \exp\Big\{ - \sum_{\mu=0}^{(M-1)/2} \mathrm{i}\hat{Q}_\mu Q_\mu \\ & - \sum_{\mu=0}^{(M-1)/2} \ln(1 - \mathrm{i}\alpha Q_\mu) - \frac{1}{2}\ln\det(\lambda_\varepsilon \mathbb{1} - R)\Big\}\end{aligned}$$

- Do Q_μ integrals using residues; gives R matrix elements

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

with exponentially distributed $\hat{Q}_\mu$.

- R is Toeplitz matrix. Evaluate $\ln\det(\lambda_\varepsilon \mathbb{1} - R)$ using Szegö,

$$\ln\det(\lambda_\varepsilon \mathbb{1} - R) \sim \sum_{\nu=-(N-1)/2}^{(N-1)/2} \ln\left(\lambda_\varepsilon - R_\nu\right)$$

with

$$R_\nu = \sum_{\mu=0}^{(M-1)/2} \hat{Q}_\mu \hat{C}(q_\mu) S_{\nu\mu} \equiv R(p_\nu) \ , \qquad p_\nu = \frac{2\pi}{N}\nu$$

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_ν definitions via δ-functions

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

- Do R_ν integrals using residues. Gives

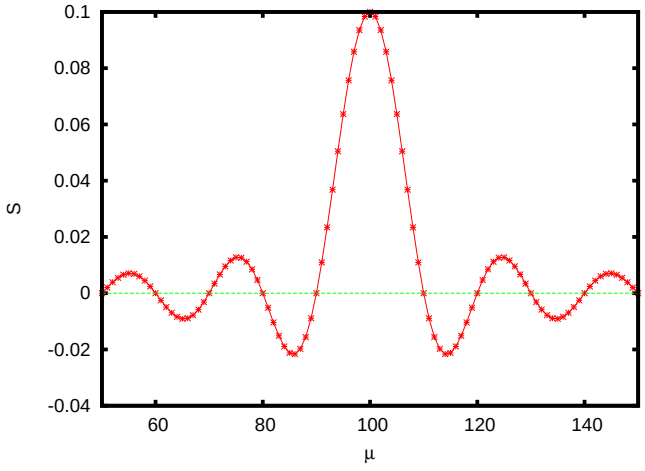
$$\langle Z_N \rangle = \left\langle \prod_{\nu=0}^{(N-1)/2} F_\nu \right\rangle_{\{Q_\mu\}}$$

with

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

Decoupling & Decorrelation Approximations

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



$S_{\nu\mu}$ at $\nu = 10$ as a function of μ,for $\alpha = 0.1$

- Approximations (using smoothness of $\hat{C}(q_\mu)$ on q_μ scale)

$$\textbf{(i)} \qquad \sum_\mu \hat{Q}_\mu \hat{C}(q_\mu) S_{\nu\mu} \simeq \frac{\alpha}{2} \hat{C}(p_\nu) \sum_{\mu \in I_\nu} Q_\mu$$

$$\textbf{(ii)} \qquad \left\langle \prod_{\nu=0}^{(N-1)/2} F_\nu \right\rangle_{\{\hat{Q}_\mu\}} \simeq \prod_{\nu=0}^{(N-1)/2} \left\langle F_\nu \right\rangle_{\{\hat{Q}_\mu\}}$$

Closed Form Approximation & Scaling

- Allow closed form expression of $\langle Z_N \rangle$, and hence $\rho_\alpha(\lambda)$

$$\langle Z_N \rangle = \prod_{\nu=0}^{(N-1)/2} \left\{ \frac{2\,\mathrm{i}}{\alpha \hat{C}(p_\nu)} \int_0^\infty \mathrm{d}y \, \frac{\mathrm{e}^{-\mathrm{i}y\lambda_\varepsilon 2/(\alpha \hat{C}(p_\nu))}}{\left(1-\mathrm{i}y\right)^{2/\alpha}} \right\}$$

- Gives

$$\rho_\alpha(\lambda) = \int_0^\pi \frac{\mathrm{d}q}{\pi} \, \frac{1}{\hat{C}(q)} \, \rho_\alpha^{(0)} \left(\frac{\lambda}{\hat{C}(q)} \right)$$

- As $\hat{C}(q) \equiv 1$ for uncorrelated data, we have to identify $\rho_\alpha^{(0)}$ with the spectral density for auto-covariance matrices of i.i.d. (uncorrelated) data.
- Our approximations give

$$\rho_\alpha^{(0)}(\lambda) = -\lim_{\varepsilon \to 0} \frac{1}{\pi} \mathrm{Im} \frac{\partial}{\partial \lambda} \ln I_\alpha \left(\frac{2}{\alpha} \lambda_\varepsilon \right)$$

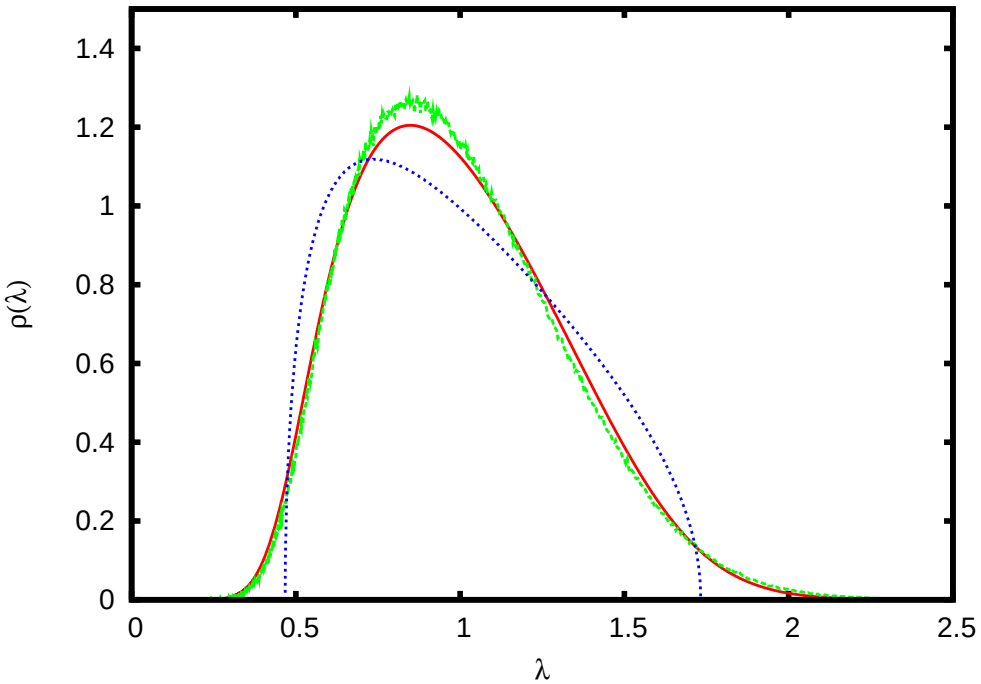
with

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

Outline

The Scaling Function — Spectrum for i.i.d. Data

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

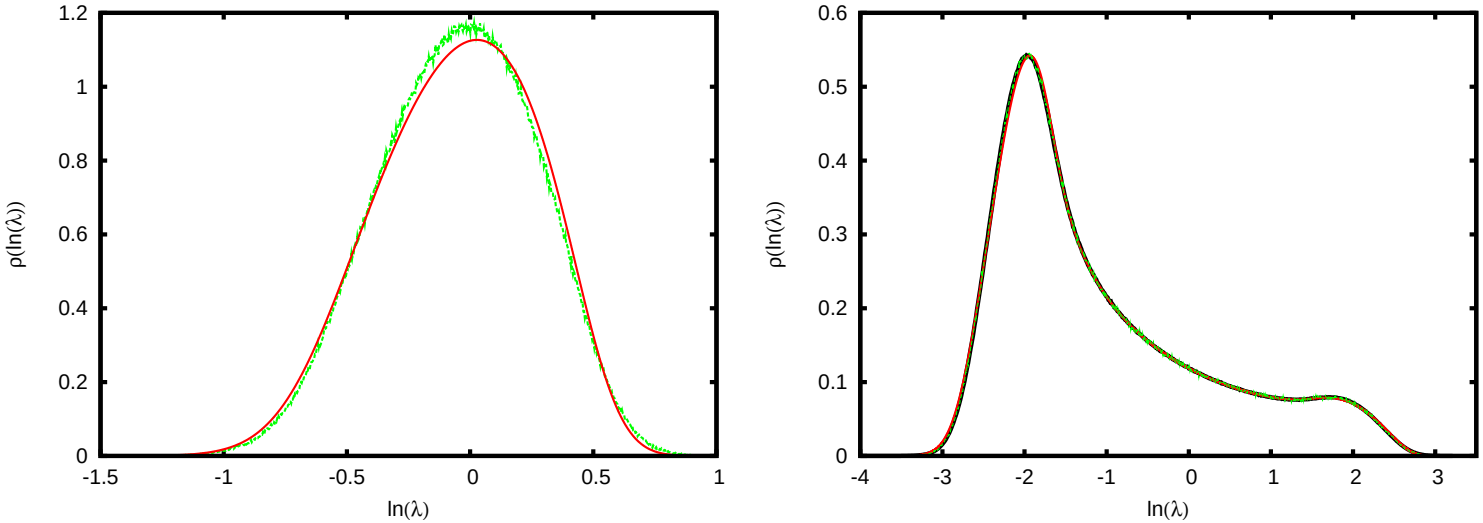


Simulation results (green); analytic approximation for $\rho_\alpha^{(0)}(\lambda)$ (red), Marčenko-Pastur law (blue-dashed).

AR-1 Process @ $\alpha = 0.1$

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

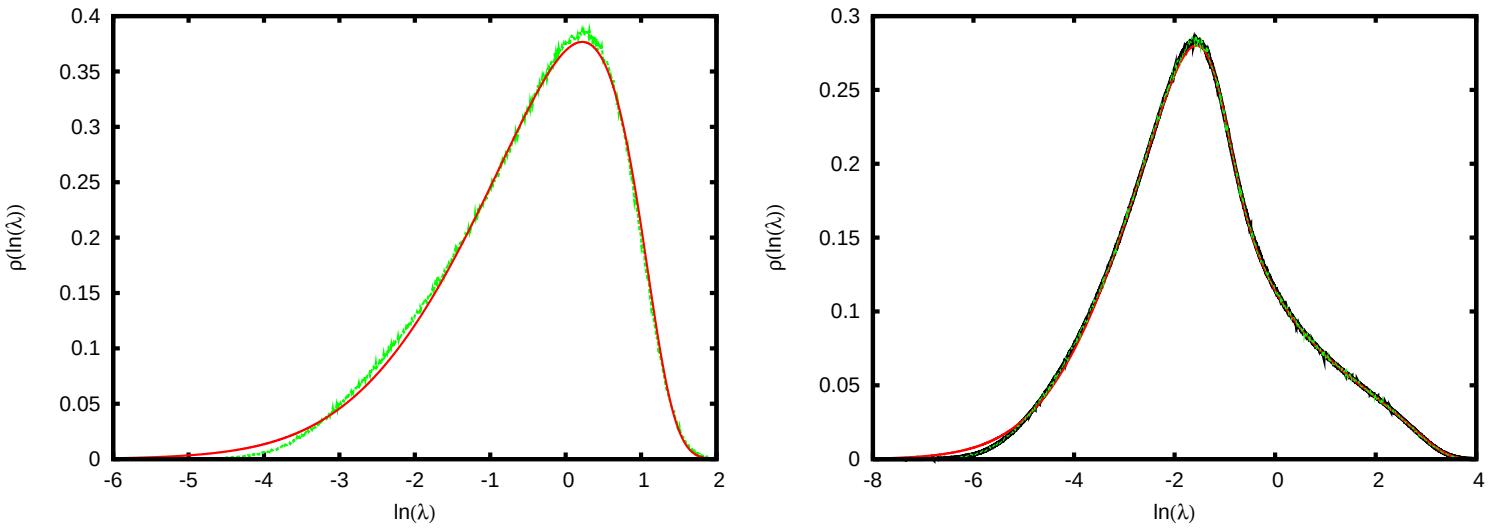
$$x_n = a_1 x_{n-1} + \sqrt{1 - a_1^2}\,\xi_n$$



Left: 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 @ $\alpha = 0.8$

- (Logarithmic) Spectral density for AR-1 process @ $\alpha = 0.8$



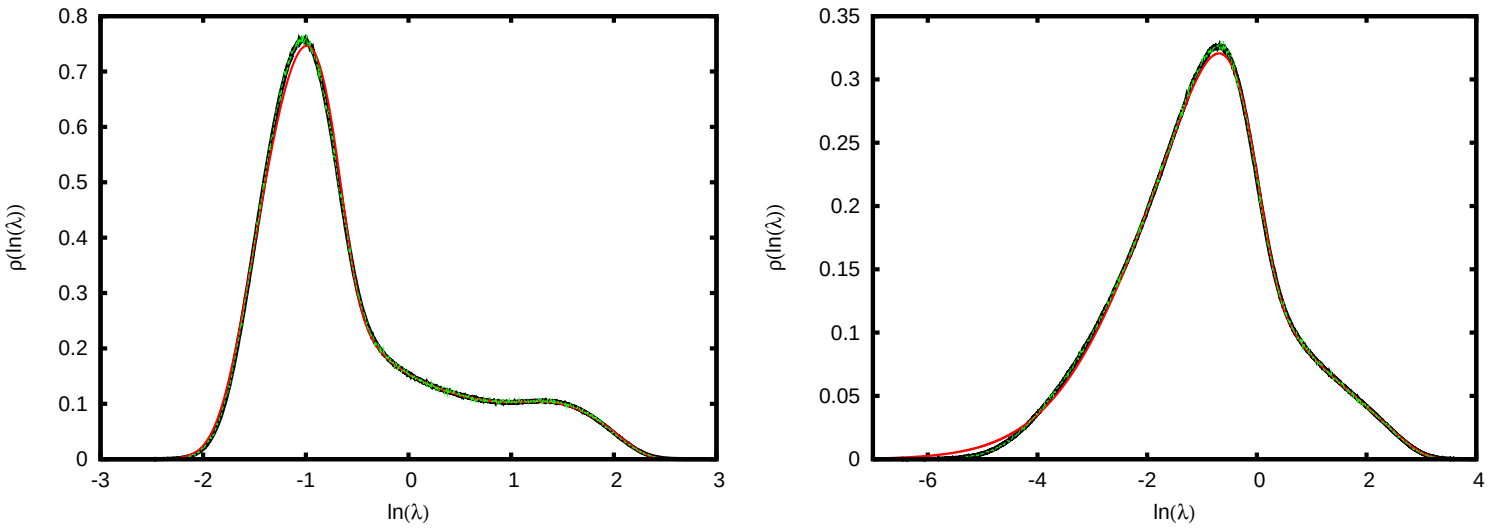
Left: i.i.d. data, simulation (green) and analytic result (red).
Right $a_1 = 0.8$. Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).

AR-2 Process (Two Real Eigenvalues)

- (Logarithmic) Spectral density for AR-2 process

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} = \sigma \xi_n$$

$$a_1 = 0.5,\ a_2 = -3/16\ , \qquad \sigma \text{ such that } \bar{C}(0) = 1\ .$$



Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).

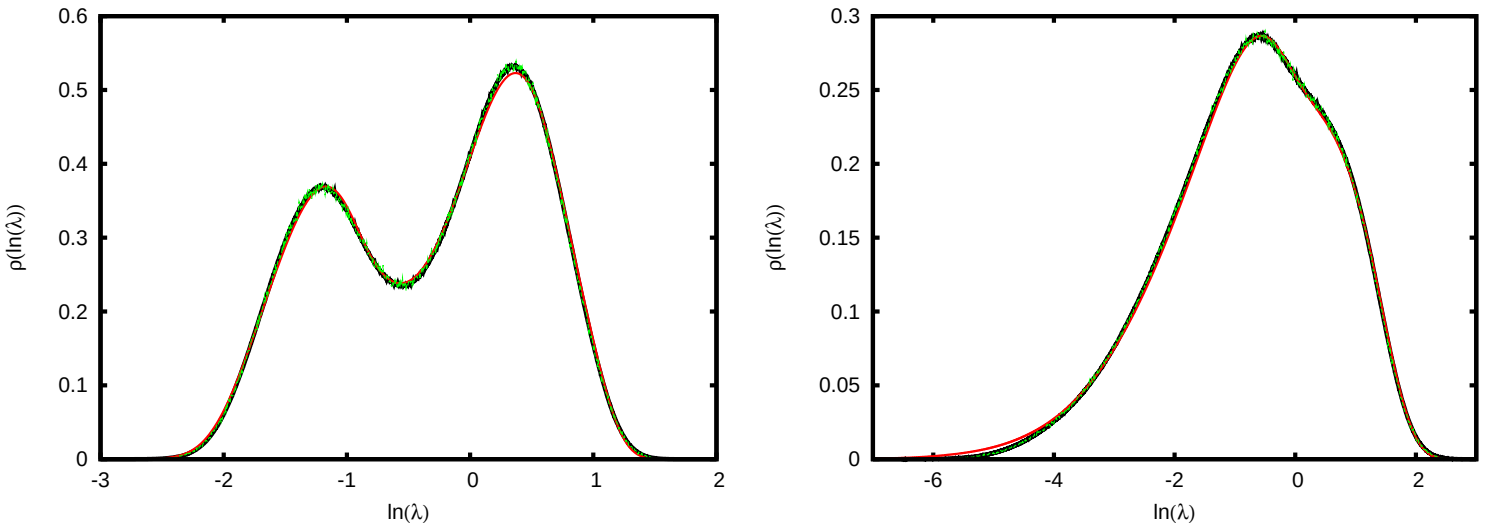
Left: $\alpha = 0.1$, **Right**: $\alpha = 0.8$.

AR-2 Process (Complex Conjugate Eigenvalues)

- (Logarithmic) Spectral density for AR-2 process

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} = \sigma \xi_n$$

$$a_1 = 0.5,\ a_2 = 5/16\ , \qquad \sigma \text{ such that } \bar{C}(0) = 1\ .$$



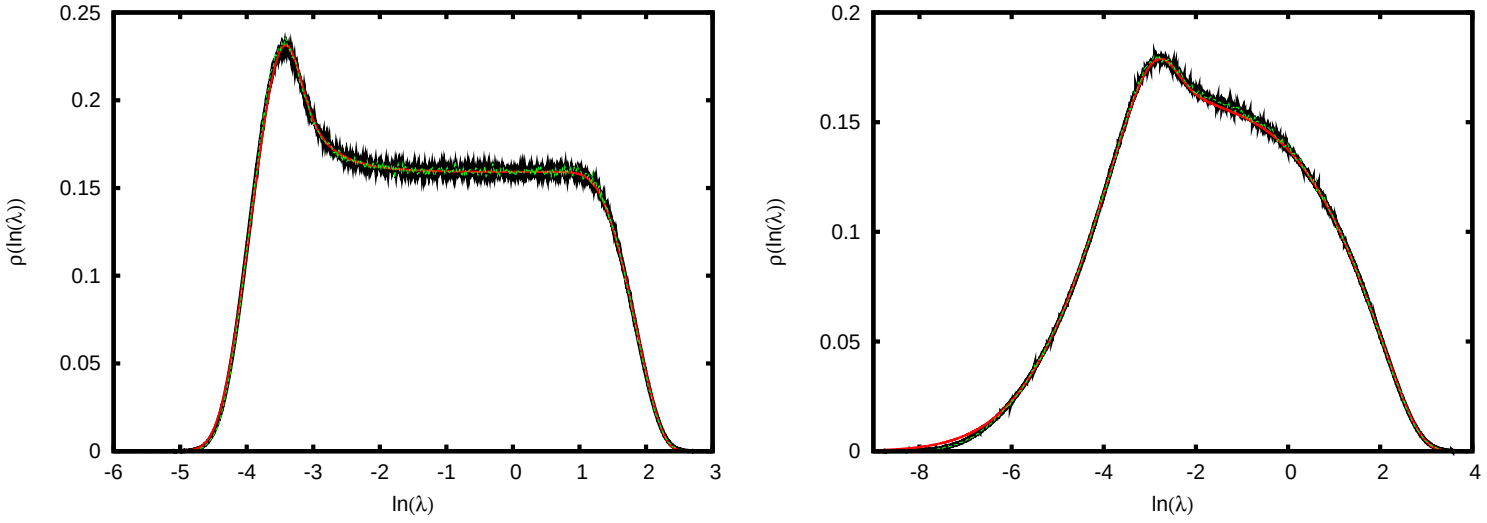
Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).

Left: $\alpha = 0.1$, **Right**: $\alpha = 0.8$.

A Process with Long Range Auto-Correlation

- A process with power-law decay of auto-correlation

$$\bar{C}(n) = \frac{1}{1+(n/2)^2} \ ,$$

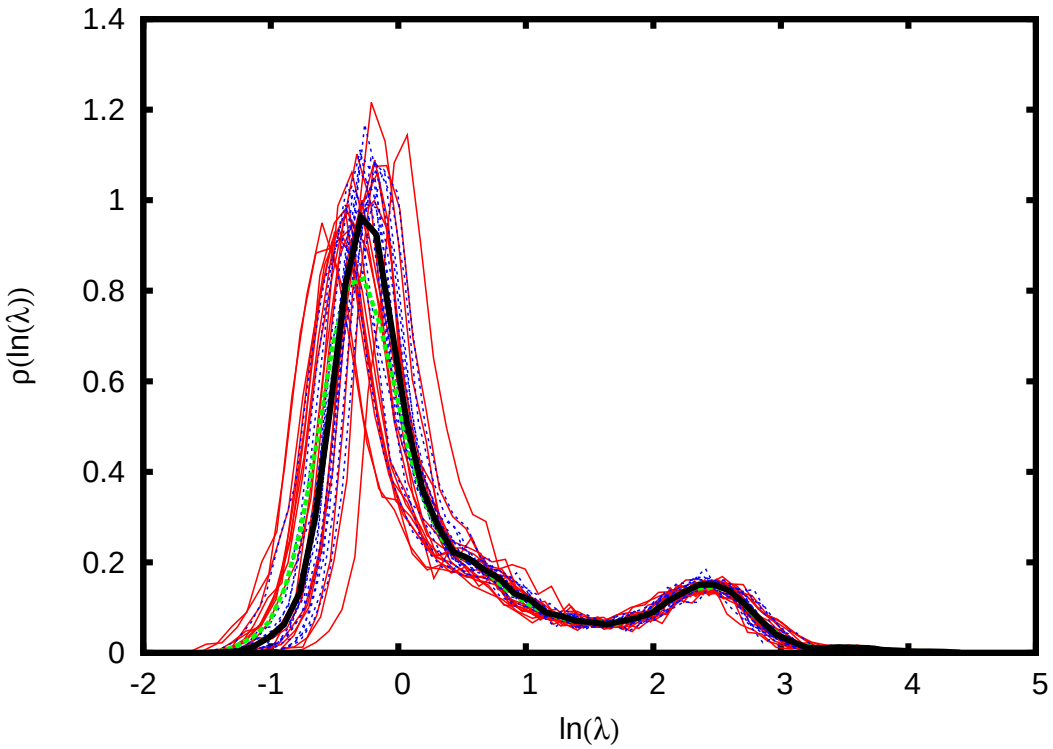


Comparing scaling based on the empirical scaling function (black) with that based on the analytic result (red) and simulations (green).

Left: $\alpha = 0.1$, **Right**: $\alpha = 0.8$.

An application (Work in Progress)

- DNA-Methylation Levels for different cancers



Spectral density of auto-covariance matrices of DNA mathylation levels with $N = 100$ at $\alpha = N/M = 0.025$. Thin red lines are for individual patients from the P class, dashed blue are for individual patients from the N class. The thicker green and black lines correspond to average spectra of the P and N classes, respectively.

Outline

Summary

- Computed DOS of sample auto-covariance matrices using annealed calculation.
- Key ingredient: Szegö's theorem for Toeplitz matrices
- Rectangular window and decorrelation approximation $\Rightarrow$ Closed form approximation.
- Use of Szegös theorem suggests a scaling form for DOS.
  - results suggest that scaling is exact
  - ideas for independent proof
- Applications: time-series analysis, signal processing, information theory, finance ...
- **Thanks!** K. Anand, L. Dall'Asta, P. Vivo