Eigenvector stability:
Random Matrix Theory
and Financial Applications

J.P Bouchaud
with: R. Allez, M. Potters

CAPITAL FUND MANAGEMENT

http://www.cfm.fr

Portfolio theory: Basics

- Portfolio weights w_i, Asset returns X_i^t

- If expected/predicted gains are g_i then the expected gain of the portfolio is

$$\mathcal{G} = \sum_i w_i g_i$$

- Let risk be defined as: variance of the portfolio returns (maybe not a good definition !)

$$R^2 = \sum_{ij} w_i \sigma_i C_{ij} \sigma_j w_j$$

where σ_i^2 is the variance of asset i, and

C_{ij} is the correlation matrix.

Markowitz Optimization

- Find the portfolio with maximum expected return for a given risk or equivalently, minimum risk for a given return ($\mathcal{G}$)

- In matrix notation:

$$\mathbf{w}_C = \mathcal{G}\frac{\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^T\mathbf{C}^{-1}\mathbf{g}}$$

where all gains are measured with respect to the risk-free rate and $\sigma_i = 1$ (absorbed in g_i).

- Note: in the presence of non-linear contraints, e.g.

$$\sum_i |w_i| \leq A$$

an NP complete, “spin-glass” problem! (see [JPB,Galluccio,Potters])

Markowitz Optimization

- More explicitely:

$$\mathbf{w} \propto \sum_\alpha \lambda_\alpha^{-1} (\mathbf{\Psi}_\alpha \cdot \mathbf{w}) \, \mathbf{\Psi}_\alpha = \mathbf{g} + \sum_\alpha (\lambda_\alpha^{-1} - 1) \, (\mathbf{\Psi}_\alpha \cdot \mathbf{w}) \, \mathbf{\Psi}_\alpha$$

- Compared to the naive allocation $\mathbf{w} \propto \mathbf{g}$:
  - Eigenvectors with $\lambda \gg 1$ are projected out
  - Eigenvectors with $\lambda \ll 1$ are overallocated

- Very important for "stat. arb." strategies

Empirical Correlation Matrix

- Empirical Equal-Time Correlation Matrix $\mathbf{E}$

$$E_{ij} = \frac{1}{T} \sum_t \frac{X_i^t X_j^t}{\sigma_i \sigma_j}$$

Order N^2 quantities estimated with NT datapoints.

When $T < N$, $\mathbf{E}$ is not even invertible.

Typically: $N = 500 - 1000$; $T = 500 - 2500$

Risk of Optimized Portfolios

- "In-sample" risk

$$R^2_{\text{in}} = \mathbf{w}_E^T \mathbf{E} \mathbf{w}_E = \frac{1}{\mathbf{g}^T \mathbf{E}^{-1} \mathbf{g}}$$

- True minimal risk

$$R^2_{\text{true}} = \mathbf{w}_C^T \mathbf{C} \mathbf{w}_C = \frac{1}{\mathbf{g}^T \mathbf{C}^{-1} \mathbf{g}}$$

- "Out-of-sample" risk

$$R^2_{\text{out}} = \mathbf{w}_E^T \mathbf{C} \mathbf{w}_E = \frac{\mathbf{g}^T \mathbf{E}^{-1} \mathbf{C} \mathbf{E}^{-1} \mathbf{g}}{(\mathbf{g}^T \mathbf{E}^{-1} \mathbf{g})^2}$$

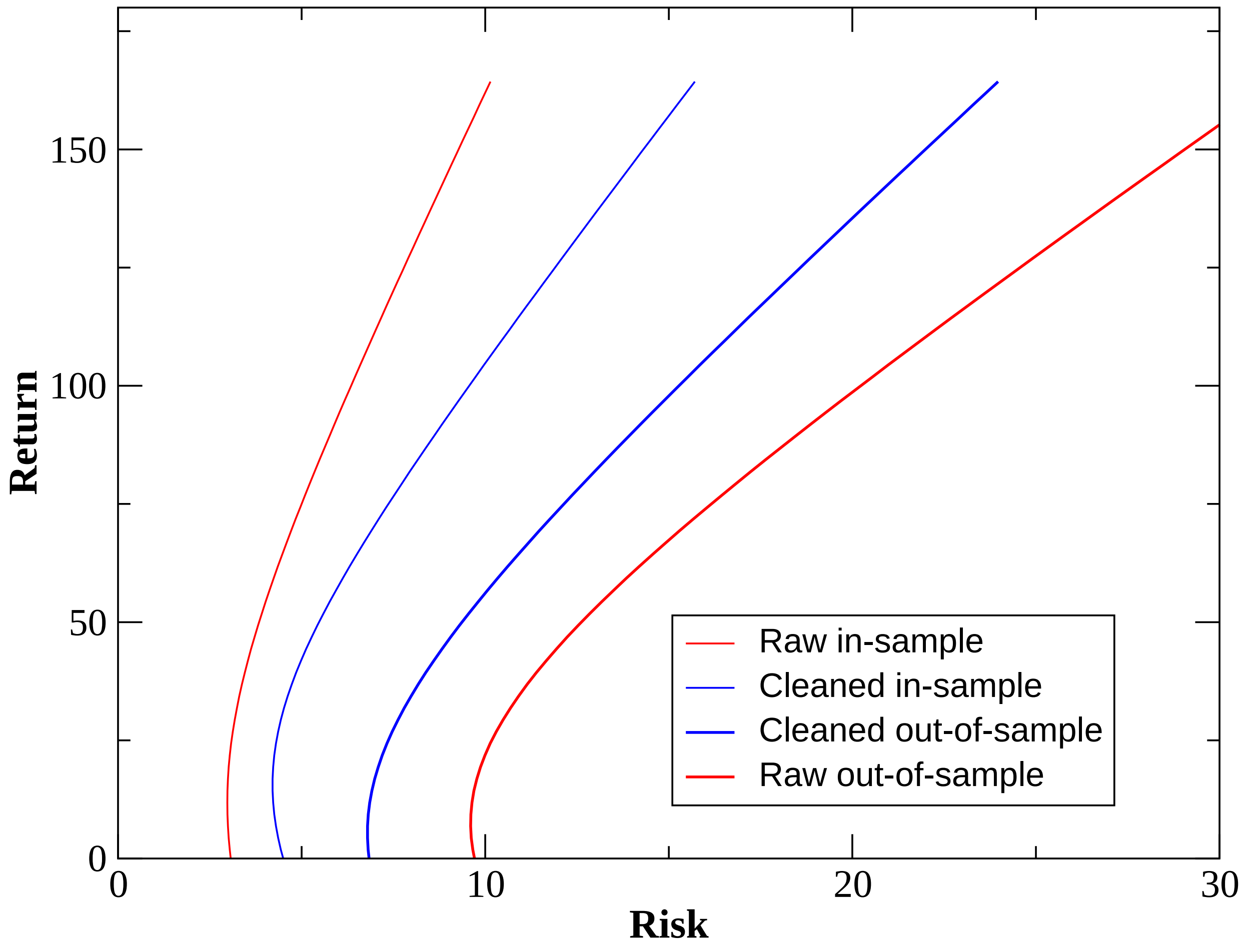
Risk of Optimized Portfolios

- Let $\mathbf{E}$ be a noisy, unbiased estimator of $\mathbf{C}$. Using convexity arguments, and for large matrices:

$$R^2_{\text{in}} \leq R^2_{\text{true}} \leq R^2_{\text{out}}$$

- If $\mathbf{C}$ has some time dependence, one expects an even worse underestimation

In Sample vs. Out of Sample



Possible Ensembles (stationary case)

- Null hypothesis Wishart ensemble:

$$\langle X_i^t X_j^{t'} \rangle = \sigma_i \sigma_j \delta_{ij} \delta_{tt'}$$

Constant volatilities and X with a finite second moment

- General Wishart ensemble:

$$\langle X_i^t X_j^{t'} \rangle = \sigma_i \sigma_j C_{ij} \delta_{tt'}$$

Constant volatilities and X with a finite second moment

- Elliptic Ensemble

$$\langle X_i^t X_j^{t'} \rangle = s\, \sigma_i \sigma_j C_{ij} \delta_{tt'}$$

Random common volatility, with a certain $P(s)$

(Ex: Student)

Null hypothesis $\mathbf{C} = \mathbf{I}$

- Goal: understand the eigenvalue density of empirical correlation matrices when $q = N/T = O(1)$

- E_{ij} is a sum of (rotationally invariant) matrices $E^t_{ij} = (X^t_i X^t_j)/T$

- Free random matrix theory: R-transform are additive $\rightarrow$

$$\rho_E(\lambda) = \frac{\sqrt{4\lambda q - (\lambda + q - 1)^2}}{2\pi\lambda q} \qquad \lambda \in [(1-\sqrt{q})^2, (1+\sqrt{q})^2]$$

  [Marcenko-Pastur] (1967) (and many rediscoveries)

- Any eigenvalue beyond the Marcenko-Pastur band can be deemed to contain some information (but see below)

Null hypothesis $\mathbf{C} = \mathbf{I}$

- Remark 1: $-G_E(0) = \langle \lambda^{-1} \rangle_E = (1-q)^{-1}$, allowing to compute the different risks:

$$\mathcal{R}_{\text{true}} = \frac{\mathcal{R}_{\text{in}}}{\sqrt{1-q}}; \qquad \mathcal{R}_{\text{out}} = \frac{\mathcal{R}_{\text{in}}}{1-q}$$

- Remark 2: One can extend the calculation to EMA estimators [Potters, Kondor, Pafka]:

$$\mathbf{E}_{t+1} = (1-\varepsilon)\mathbf{E}_t + \varepsilon X^t X^t$$

General $\mathbf{C}$ Case

- The general case for $\mathbf{C}$ cannot be directly written as a sum of "Blue" functions.

- Solution using different techniques (replicas, diagrams, S-transforms):

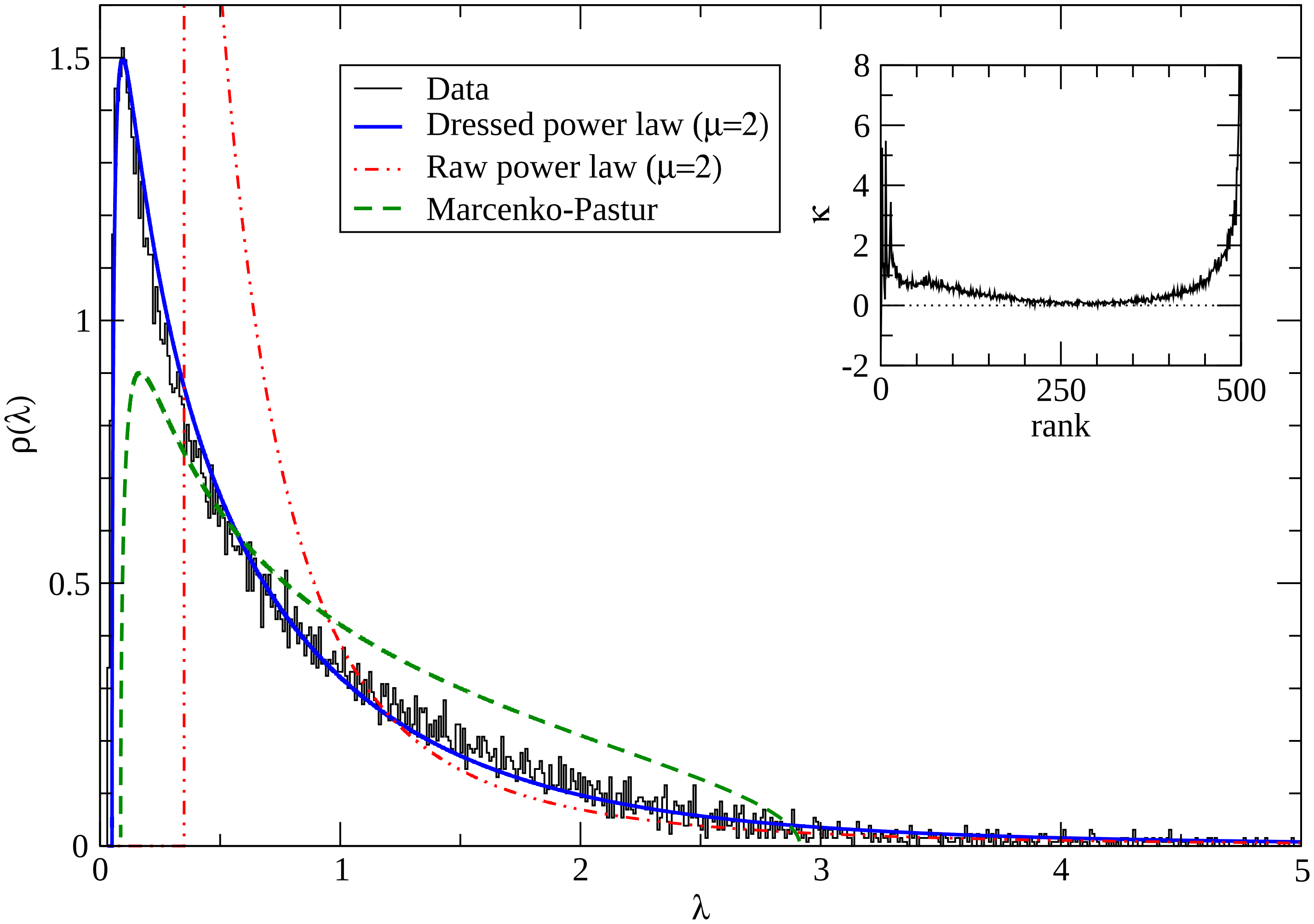
$$G_E(z) = \int d\lambda\, \rho_C(\lambda) \frac{1}{z - \lambda(1 - q + qzG_E(z))},$$

- Remark 1: $-G_E(0) = (1-q)^{-1}$ independently of $\mathbf{C}$

- Remark 2: One should work from $\rho_C \longrightarrow G_E$ and postulate a parametric form for $\rho_C(\lambda)$, i.e.:

$$\rho_C(\lambda) = \frac{\mu A}{(\lambda - \lambda_0)^{1+\mu}} \Theta(\lambda - \lambda_{\text{min}})$$

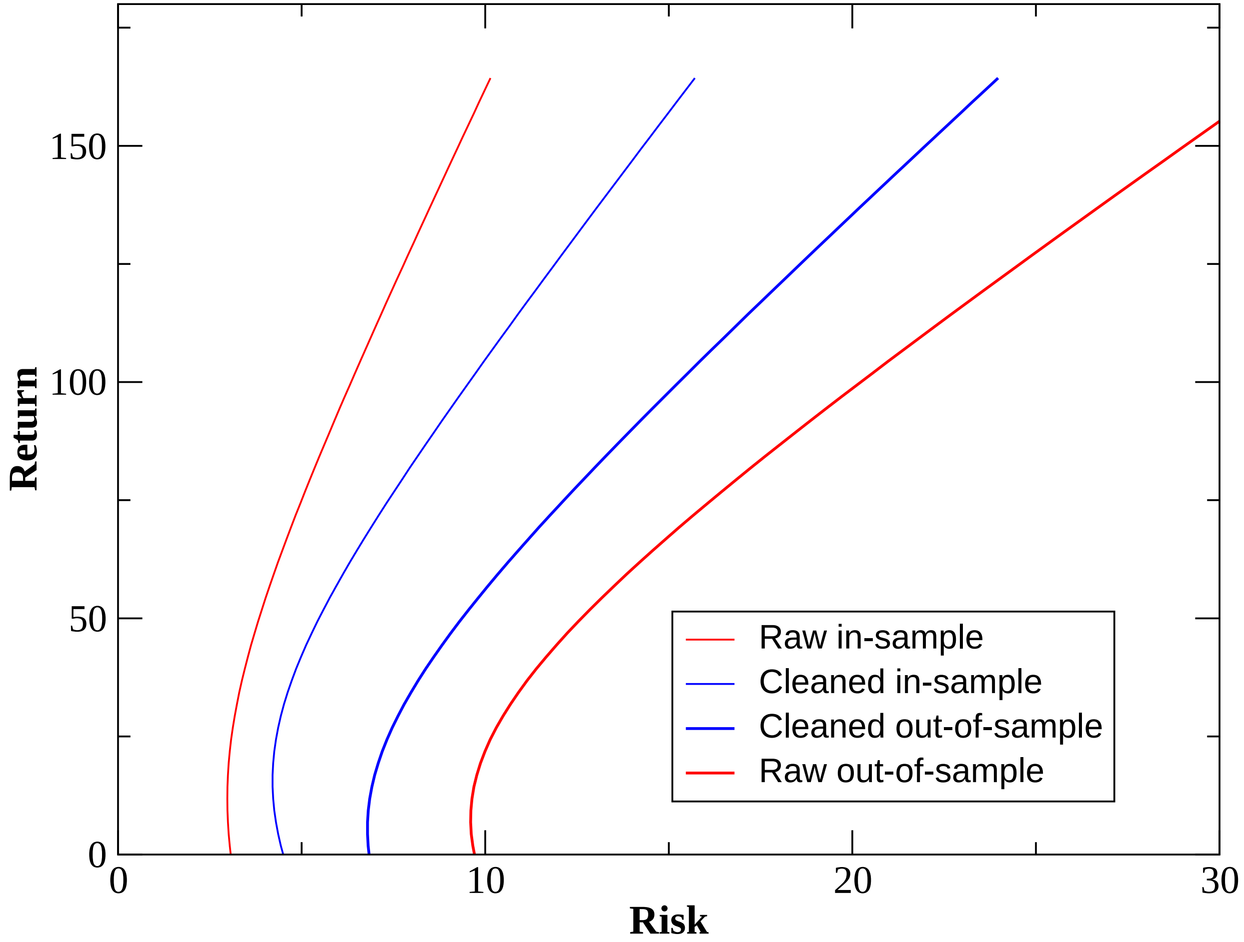
Empirical Correlation Matrix



CAPITAL FUND MANAGEMENT

J.-P. Bouchaud

Eigenvalue cleaning



CAPITAL FUND MANAGEMENT

J.-P. Bouchaud

What about eigenvectors?

- Up to now, most results using RMT focus on eigenvalues

- What about eigenvectors? What natural null-hypothesis?

- Are eigen-directions *stable* in time?

- Important source of risk for market/sector neutral portfolios: a sudden/gradual rotation of the top eigenvectors!

- ..a little movie...

What about eigenvectors?

- Correlation matrices need a certain time T to be measured

- Even if the "true" $\mathbf{C}$ is fixed, its empirical determination fluctuates:

$$\mathbf{E}_t = \mathbf{C} + \text{noise}$$

- What is the dynamics of the empirical eigenvectors induced by measurement noise?

- Can one detect a genuine evolution of these eigenvectors beyond noise effects?

What about eigenvectors?

- More generally, can one say something about the eigenvectors of randomly perturbed matrices:

$$\mathbf{H} = \mathbf{H}_0 + \epsilon \mathbf{H}_1$$

where $\mathbf{H}_0$ is deterministic or random (e.g. GOE) and $\mathbf{H}_1$ random.

Eigenvectors exchange

- An issue: upon pseudo-collisions of eigenvectors, eigenvalues exchange

- Example: 2×2 matrices

$$H_{11} = a, \qquad H_{22} = a + \epsilon, \qquad H_{21} = H_{12} = c, \longrightarrow$$

$$\lambda_{\pm} \approx_{\epsilon \to 0} a + \frac{\epsilon}{2} \pm \sqrt{c^2 + \frac{\epsilon^2}{4}}$$

- Let c vary: quasi-crossing for $c \to 0$, with an exchange of the top eigenvector: $(1, -1) \to (1, 1)$

- For large matrices, these exchanges are extremely numerous $\to$ labelling problem

Subspace stability

- An idea: follow the subspace spanned by P-eigenvectors:

$$|\psi_{k+1}\rangle, |\psi_{k+2}\rangle, \dots |\psi_{k+P}\rangle \longrightarrow |\psi'_{k+1}\rangle, |\psi'_{k+2}\rangle, \dots |\psi'_{k+P}\rangle$$

- Form the $P \times P$ matrix of scalar products:

$$G_{ij} = \langle \psi_{k+i} | \psi'_{k+j} \rangle$$

- The determinant of this matrix is insensitive to label permutations and is a measure of the overlap between the two P-dimensional subspaces

  – $Q = \frac{1}{P} \ln |\det \mathbf{G}|$ is a measure of how well the first subspace can be approximated by the second

Null hypothesis

- Note: if P is large, Q can be “accidentally” large
- One can compute Q exactly in the limit $P \to \infty$, $N \to \infty$, with fixed $p = P/N$:
- Final result:([Wachter] (1980); [Laloux,Miceli,Potters,JPB])

$$Q = \int_0^1 \mathrm{d}s \ln s \; \rho(s)$$

with:

$$\rho(s) = \frac{1}{p} \frac{\sqrt{s^2(4p(1-p) - s^2)^+}}{\pi s(1-s^2)}.$$

Intermezzo

- Non equal time correlation matrices

$$E_{ij}^{\tau} = \frac{1}{T} \sum_t \frac{X_i^t X_j^{t+\tau}}{\sigma_i \sigma_j}$$

$N \times N$ but not symmetrical: 'leader-lagger' relations

- General rectangular correlation matrices

$$G_{\alpha i} = \frac{1}{T} \sum_{t=1}^{T} Y_\alpha^t X_i^t$$

N 'input' factors X; M 'output' factors Y

 – Example: $Y_\alpha^t = X_j^{t+\tau}$, $N = M$

Intermezzo: Singular values

- Singular values: Square root of the non zero eigenvalues of GG^T or G^TG, with associated eigenvectors u_α^k and v_i^k $\rightarrow$ $1 \geq s_1 > s_2 > ...s_{(M,N)^-} \geq 0$

- Interpretation: $k = 1$: best linear combination of input variables with weights v_i^1, to optimally predict the linear combination of output variables with weights u_α^1, with a cross-correlation $= s_1$.

- s_1: measure of the predictive power of the set of Xs with respect to Ys

- Other singular values: orthogonal, less predictive, linear combinations

Benchmark: no cross-correlations

- Null hypothesis: No correlations between Xs and Ys:

$$G_{\text{true}} \equiv \mathbf{0}$$

- But arbitrary correlations *among* Xs, C_X, and Ys, C_Y, are possible

- Consider exact normalized principal components for the sample variables Xs and Ys:

$$\hat{X}_i^t = \frac{1}{\sqrt{\lambda_i}} \sum_j U_{ij} X_j^t; \qquad \hat{Y}_\alpha^t = \ldots$$

and define $\hat{G} = \hat{Y}\hat{X}^T$.

Benchmark: Random SVD

- Final result:([Wachter] (1980); [Laloux,Miceli,Potters,JPB])

$$\rho(s) = (m+n-1)^+\delta(s-1) + \frac{\sqrt{(s^2-\gamma_-)(\gamma_+-s^2)}}{\pi s(1-s^2)}$$

  with

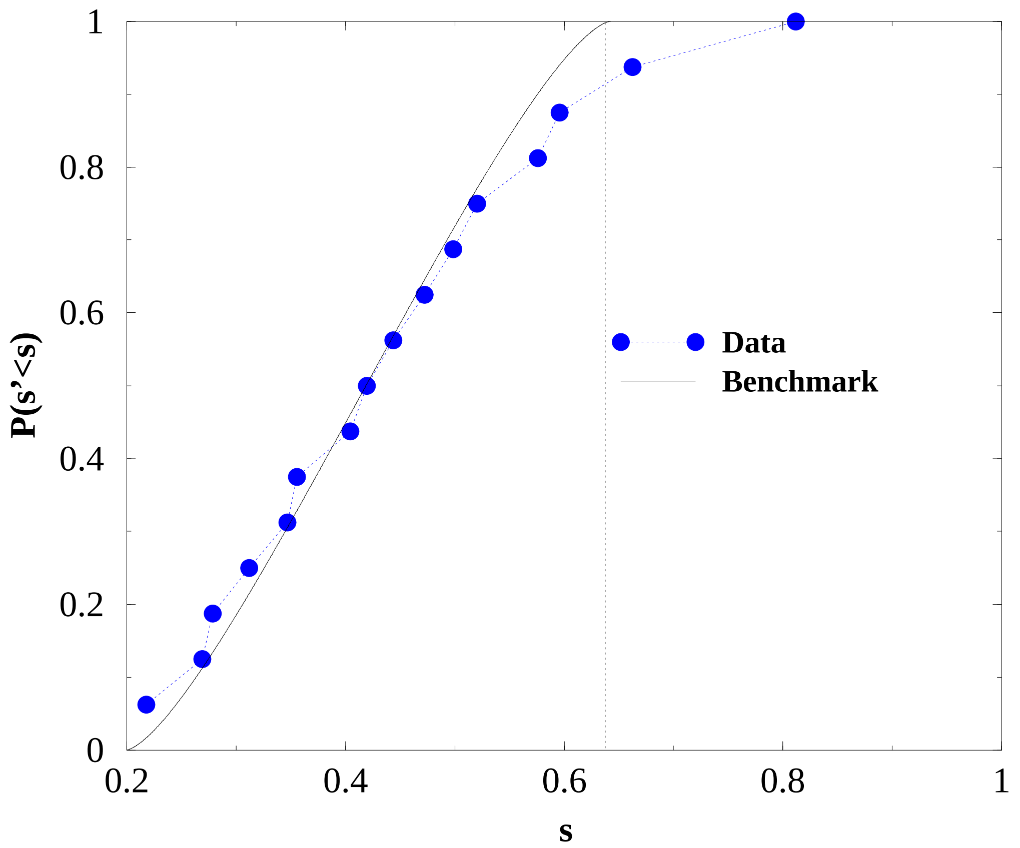
$$\gamma_\pm = n+m-2mn \pm 2\sqrt{mn(1-n)(1-m)}, \quad 0 \leq \gamma_\pm \leq 1$$

- Analogue of the Marcenko-Pastur result for rectangular correlation matrices

- Many applications; finance, econometrics ('large' models), genomics, etc.

- Same problem as subspace stability: $T \longrightarrow N$, $n = m \longrightarrow p$

Sectorial Inflation vs. Economic indicators



$N = 50, M = 16, T = 265$

Back to eigenvectors: perturbation theory

- Consider a randomly perturbed matrix:

$$\mathbf{H} = \mathbf{H}_0 + \epsilon \mathbf{H}_1$$

- Perturbation theory to second order in ϵ yields:

$$|\det(G)| = 1 - \frac{\epsilon^2}{2} \sum_{i \in \{k+1,\ldots,k+P\}} \sum_{j \notin \{k+1,\ldots,k+P\}} \left(\frac{\langle \psi_i | \mathbf{H}_1 | \psi_j \rangle}{\lambda_i - \lambda_j} \right)^2 .$$

The case of correlation matrices

- Consider the empirical correlation matrix:

$$\mathbf{E} = \mathbf{C} + \eta \qquad \eta = \frac{1}{T}\sum_{t=1}^{T}(X^t X^t - \mathbf{C})$$

- The noise η is correlated as:

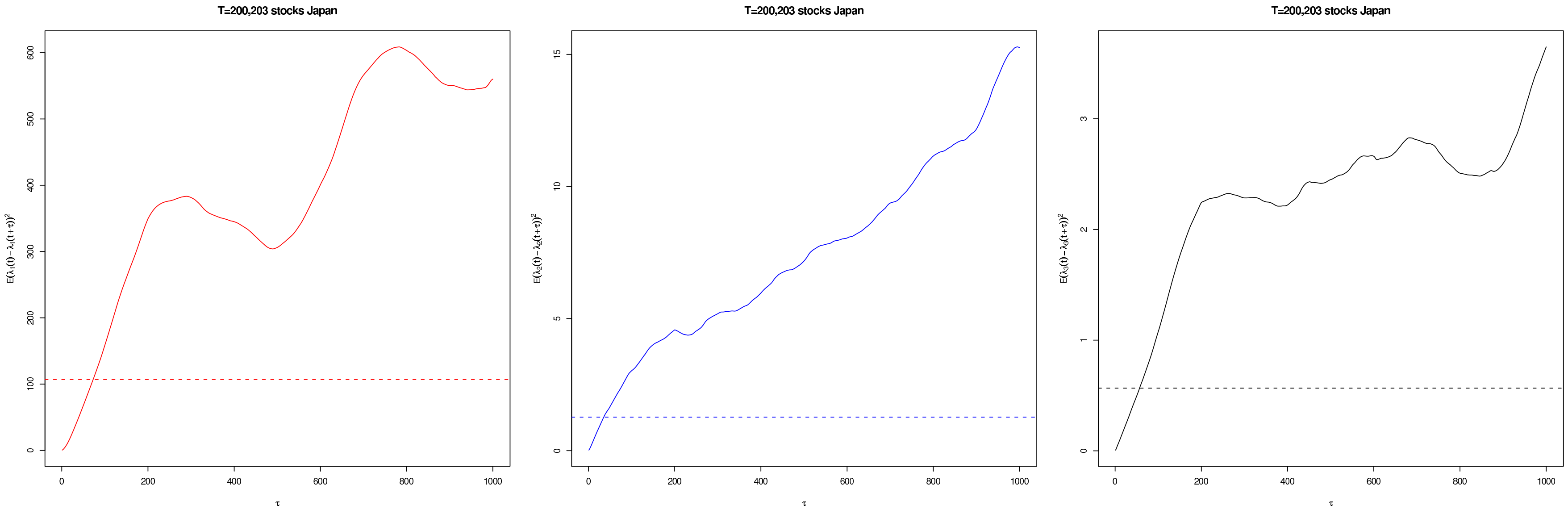
$$\left\langle \eta_{ij}\eta_{kl} \right\rangle = \frac{1}{T}(C_{ik}C_{jl} + C_{il}C_{jk})$$

- from which one derives:

$$\left\langle |\det(\mathbf{G})|^{1/P} \right\rangle \approx 1 - \frac{1}{2TP}\left[\sum_{i=1}^{P}\sum_{j=P+1}^{N}\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\right].$$

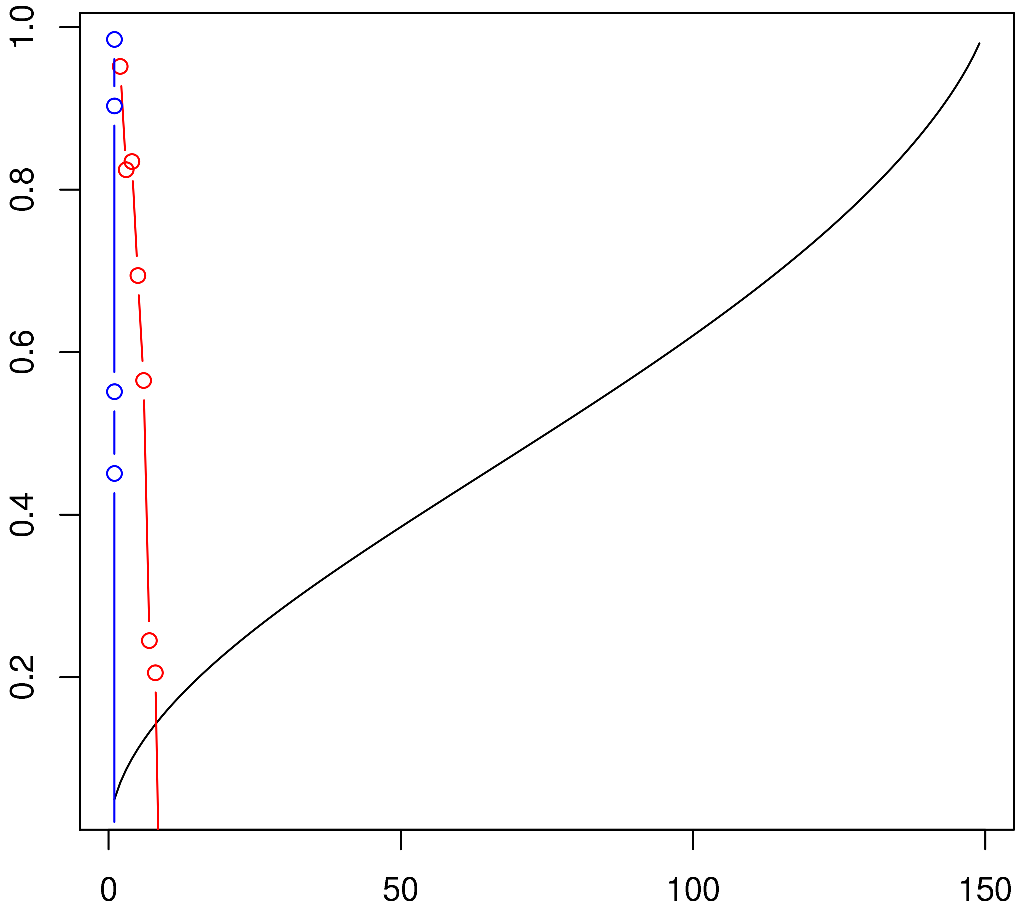
(and a similar equation for eigenvalues)

Stability of eigenvalues: Correlations



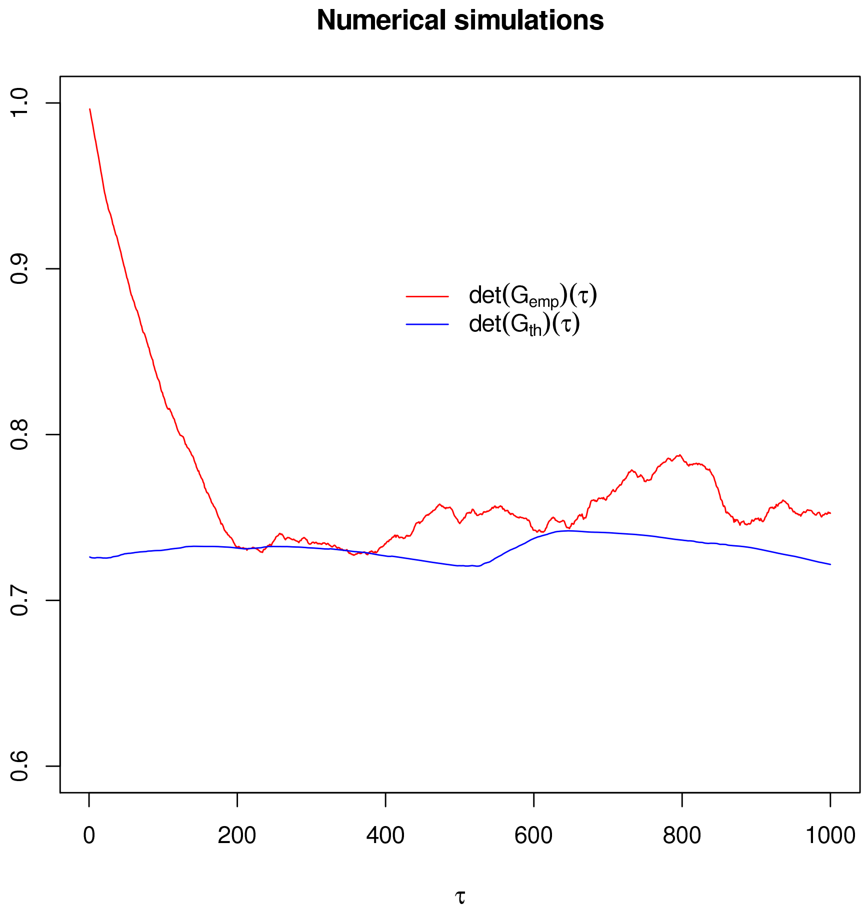
Eigenvalues clearly change: well known correlation crises

Stability of eigenspaces: Correlations



8 meaningful eigenvectors

Stability of eigenspaces: Correlations



$P = 5$

The case of correlation matrices

- Empirical results show a faster decorrelation $\rightarrow$ real dynamics of the eigenvectors

- The case of the top eigenvector, in the limit $\lambda_1 \gg \lambda_2$, and for EMA:

  – An Ornstein-Uhlenbeck process on the unit sphere around $\theta = 0$

  – Explicit solution for the full distribution $P(\theta)$ and time correlations

  – $\det G = \cos(\theta - \theta')$

- Full characterisation of the dynamics for arbitrary P? (Random rotation of a solid body)