Eigenvector stability:
Random Matrix Theory
and Financial Applications

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

CAPITAL FUND MANAGEMENT

ttp *c*

Portfolio theory: Basics

- Portfolio weights w_i, Asset returns X_i^t
-

Markowitz Optimization

- Find the portfolio with maximum expected return for a given

Markowitz Optimization

- In QM notation:

$$|w \quad \sum_{\alpha} \lambda_{\alpha}^{-1} \Psi_{\alpha}|g \; |\Psi_{\alpha} \;=\; |g \;+\; \sum_{\alpha} (\lambda_{\alpha}^{-1} - 1) \Psi_{\alpha}|g \; |\Psi_{\alpha}$$

- Compared to the naive allocation $|w \quad |g$:
  - Eigenvectors with λ 1 are projected out
  - Eigenvectors with λ 1 are overallocated

- Very important for "stat. arb." strategies

Empirical Correlation Matrix

- **Empirical Equal-Time Correlation Matrix 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.

If $T < N$ E is not even invertible.

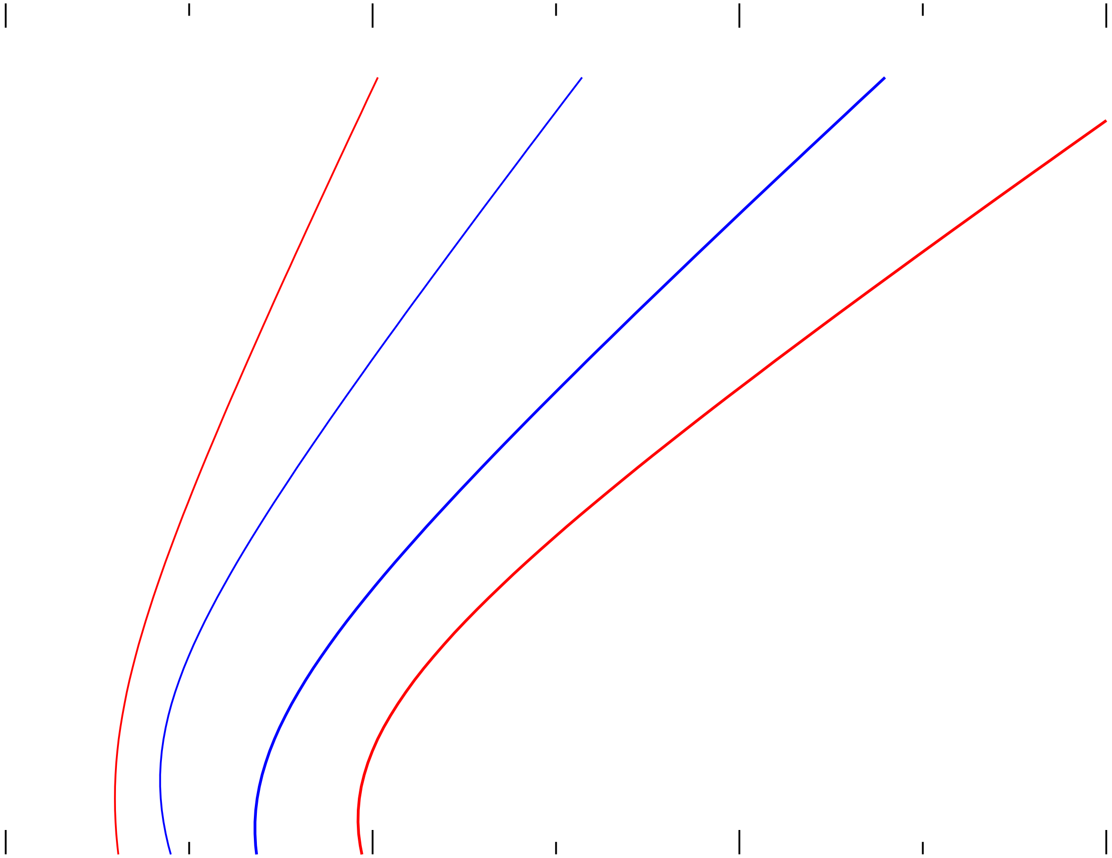
Typically: N = 500 – 1000; T = 500 – 2500

Risk of Optimized Portfolios

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

$$R_{in}^2 \leq R^2$$

In Sample vs. Out of Sample



Possible Ensembles

Null hypothesis C = 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

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

[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 $C = I$

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

$$R_{true} = \frac{R_{in}}{\sqrt{1-q}}; \qquad R_{out} = \frac{R_{in}}{1-q}$$

General C Case

- The general case for C cannot be directly written as a sum of “Blue” functions.

- Solution using di□erent 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 C

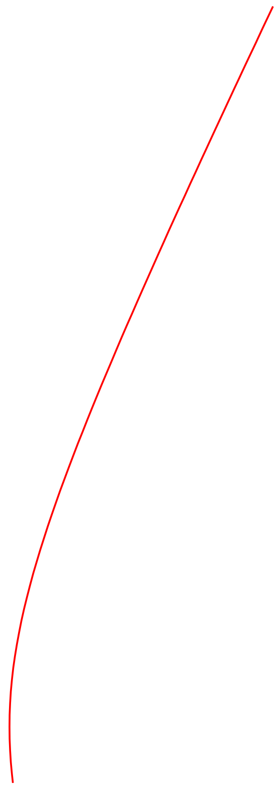
- Remark 2: One should work from ρ_C – G_E and postulate a parametric form for $\rho_C(\lambda)$, i.e.:

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

CAPITAL FUND MANAGEMENT

Empirical Correlation Matrix

Eigenvalue cleaning



What about eigenvectors?

- Up to now, most results using RMT focus on

What about eigenvectors?

- Correlation matrices need a certain time T to be measured

- Even if the “true” C is fixed, its empirical determination fluctuates:

$$E_t = C + \text{noise}$$

- 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:

$$H = H_0 + \epsilon H_1$$

where H_0 is deterministic or random (e.g. GOE) and H_1 random.

Eigenvectors exchange

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

- **Example:** 2 2 matrices

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

$$\lambda_{\pm} \quad \square \quad {}_0\, a + \frac{\square}{2} \pm \sqrt{c^2 + \frac{\square^2}{4}}$$

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

- For large matrices, these exchanges are extremely numerous

  labelling problem

CAPITAL FUND MANAGEMENT

Intermezzo

- **Non equal time correlation matrices**

$$E_{ij}^{\tau} = \quad 1$$

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
  $1 \geq s_1 > s_2 > \dots s_{(M,N)} \geq 0$

- **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:

$$G_{true} \equiv 0$$

- **But** arbitrary correlations *on* 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: $T - N, N = M - P$

Sectorial Inflation vs. Economic indicators

Back to eigenvectors: perturbation theory

- Consider a randomly perturbed matrix:

$$H = H_0 + \epsilon H_1$$

- Perturbation theory to second order in ϵ yields:

$$|\det(G)| = 1 - \frac{\epsilon^2}{2} \sum_{i \in \{k+1,\ldots,k+P\}} \sum_{j \in \{k+1,\ldots,k+P\}} \left(\frac{\langle \psi_i | H_1 | \psi_j \rangle}{\lambda_i - \lambda_j} \right)^2 .$$

The GOE case

- Take H_0 and H_1 to be GOE matrices, and consider the sub-space of eigenvectors in a finite interval $[a, b]$ of the Wigner spectrum $[-2, 2]$

- Let $\square = \hat{\square}/\sqrt{\ln N}$, then, when N , P :

$$Q \quad -\frac{\hat{\square}^2}{2}\frac{\rho(a)^2 + \rho(b)^2}{\int_a^b \rho(\lambda)d\lambda} + \frac{Z\hat{\square}^2}{\ln N}$$

  with:

$$\frac{P}{N} = \int_a^b \rho(\lambda)d\lambda.$$

  and Z a numerical constant that only depends on the two-point correlation function of eigenvalues [

Stability of eigenspaces: GOE

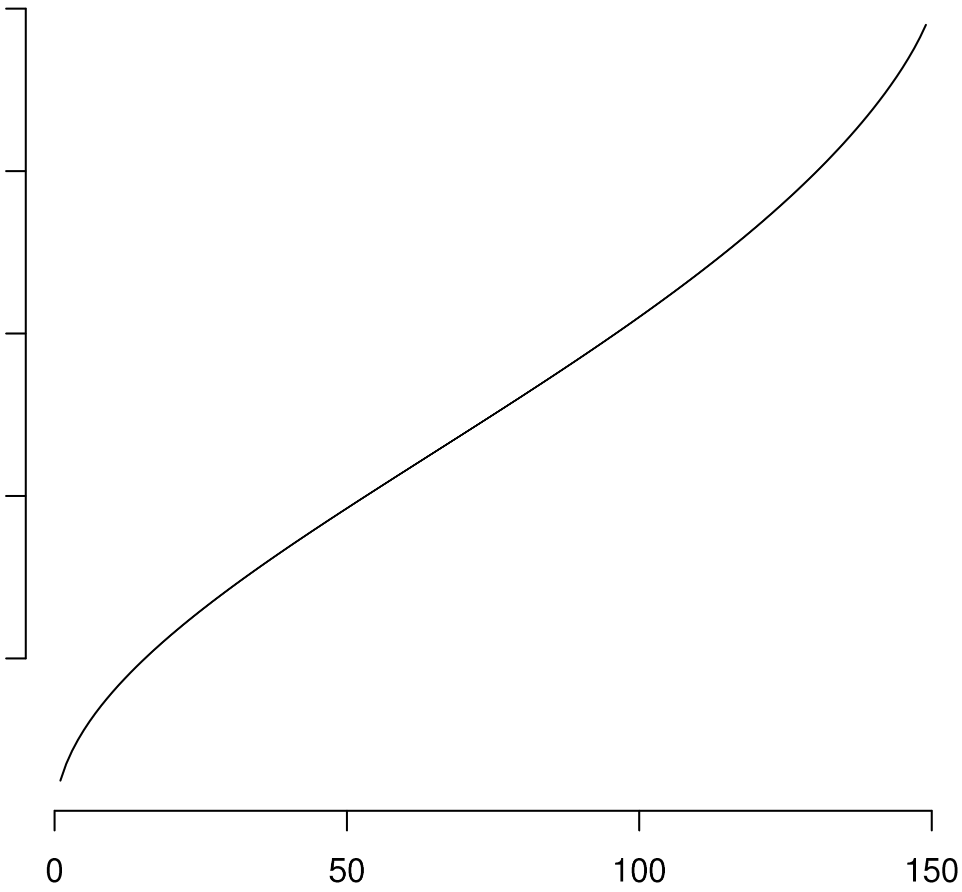
The case of correlation matrices

- Consider the empirical correlation matrix:

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

Stability of eigenvalues: Correlations

Stability of eigenspaces: Correlations



Stability of eigenspaces: Correlations

Stability of eigenspaces: Correlations

The case of correlation matrices

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

- The case of the top eigenvector, in the limit λ_1 λ_2, and for EMA: