Eigenvector stability:
Random Matrix Theory
and Financial Applications

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

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

$$G = \sum_i w_i g_i$$

- 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 $\sigma_i^2$ is the variance of asset $i$, and $C_{ij}$ is the correlation matrix.

J.-P. Bouchaud
Markowitz Optimization

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

- In matrix notation:
  \[ w_C = G \frac{C^{-1}g}{g^T C^{-1}g} \]
  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 constraints, e.g. $\sum_i |w_i| \leq A$ – an NP complete, “spin-glass” problem! (see [JPB,Galluccio,Potters])

J.-P. Bouchaud
Markowitz Optimization

• In QM notation:

\[ |w\rangle \propto \sum_{\alpha} \lambda_\alpha^{-1} \langle \psi_\alpha | g \rangle |\psi_\alpha\rangle = |g\rangle + \sum_{\alpha} (\lambda_\alpha^{-1} - 1) \langle \psi_\alpha | g \rangle |\psi_\alpha\rangle \]

• Compared to the naive allocation \(|w\rangle \propto |g\rangle\):
  - Eigenvectors with \(\lambda \gg 1\) are projected out
  - Eigenvectors with \(\lambda \ll 1\) are overallocated

• Very important for “stat. arb.” strategies

J.-P. Bouchaud
Empirical Correlation Matrix

- Empirical Equal-Time Correlation Matrix $E$

\[ E_{ij} = \frac{1}{T} \sum_t \frac{X^t_i X^t_j}{\sigma_i \sigma_j} \]

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

- "In-sample" risk
  \[ R_{\text{in}}^2 = w^{T}_E E w_E = \frac{1}{g^{T}E^{-1}g} \]

- True minimal risk
  \[ R_{\text{true}}^2 = w^{T}_C C w_C = \frac{1}{g^{T}C^{-1}g} \]

- "Out-of-sample" risk
  \[ R_{\text{out}}^2 = w^{T}_E C w_E = \frac{g^{T}E^{-1}CE^{-1}g}{(g^{T}E^{-1}g)^2} \]
Risk of Optimized Portfolios

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

$$R_{\text{in}}^2 \leq R_{\text{true}}^2 \leq R_{\text{out}}^2$$
In Sample vs. Out of Sample

Return vs. Risk

- Raw in-sample
- Cleaned in-sample
- Cleaned out-of-sample
- Raw out-of-sample

J.-P. Bouchaud
Possible Ensembles

- **Null hypothesis Wishart ensemble:**
  \[ \langle X_t^i X_{t'}^j \rangle = \sigma_i \sigma_j \delta_{ij} \delta_{tt'} \]
  
  **Constant volatilities** and **$X$ with a finite second moment**

- **General Wishart ensemble:**
  \[ \langle X_t^i X_{t'}^j \rangle = \sigma_i \sigma_j C_{ij} \delta_{tt'} \]
  
  **Constant volatilities** and **$X$ with a finite second moment**

- **Elliptic Ensemble**
  \[ \langle X_t^i X_{t'}^j \rangle = s \sigma_i \sigma_j C_{ij} \delta_{tt'} \]
  
  **Random common volatility**, with a certain $P(s)$
  
  (Ex: Student)

*J.-P. Bouchaud*
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 \( E_{ij}^t = (X_i^t X_j^t)/T \)

- **Free random matrix theory:** R-transform are additive →

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

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

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

*J.-P. Bouchaud*
Null hypothesis $C = I$

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

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

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

$$
E_{t+1} = (1 - \varepsilon)E_t + \varepsilon X^t X^t
$$

J.-P. Bouchaud
General $C$ Case

- The general case for $C$ 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 $C$

- Remark 2: One should work from $\rho_C \rightarrow G_E$ and postulate a parametric form for $\rho_C(\lambda)$, i.e.:
  \[
  \rho_C(\lambda) = \frac{\mu A}{(\lambda - \lambda_0)^{1+\mu}} \Theta(\lambda - \lambda_{\text{min}})
  \]

J.-P. Bouchaud
Empirical Correlation Matrix

- Data
- Dressed power law ($\mu=2$)
- Raw power law ($\mu=2$)
- Marcenko-Pastur
Eigenvalue cleaning

- Raw in-sample
- Cleaned in-sample
- Cleaned out-of-sample
- Raw out-of-sample

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...

J.-P. Bouchaud
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.

J.-P. Bouchaud
Eigenvectors exchange

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

• Example: $2 \times 2$ matrices

\[
H_{11} = a, \quad H_{22} = a + \epsilon, \quad H_{21} = H_{12} = c, \quad \lambda_{\pm} \approx_{\epsilon \to 0} a + \frac{\epsilon}{2} \pm \sqrt{c^2 + \frac{\epsilon^2}{4}}
\]

• 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 → labelling problem

J.-P. Bouchaud
Subspace stability

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

$$|\psi_{k+1}\rangle, |\psi_{k+2}\rangle, \cdots |\psi_k+P\rangle \rightarrow |\psi'_{k+1}\rangle, |\psi'_{k+2}\rangle, \cdots |\psi'_{k+P}\rangle$$

• 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

$$Q = \frac{1}{P} \ln |\det G|$$

is a measure of how well the first subspace can be approximated by the second

J.-P. Bouchaud
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])

$$Q = \int_0^1 ds \ln s \, \rho(s)$$

with:

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

- **Non equal time correlation matrices**

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

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

*J.-P. Bouchaud*
Intermezzo: Singular values

- **Singular values**: Square root of the non zero eigenvalues of $GG^T$ or $G^T G$, with associated eigenvectors $u^k_\alpha$ and $v^k_i \to 1 \geq s_1 > s_2 > ... s_{(M,N)-} \geq 0$

- **Interpretation**: $k = 1$: best linear combination of input variables with weights $v^1_i$, to optimally predict the linear combination of output variables with weights $u^1_\alpha$, 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

J.-P. Bouchaud
Benchmark: no cross-correlations

- **Null hypothesis:** No correlations between $X$s and $Y$s:
  \[ G_{\text{true}} \equiv 0 \]

- But arbitrary correlations *among* $X$s, $C_X$, and $Y$s, $C_Y$, are possible

- Consider exact **normalized principal components** for the sample variables $X$s and $Y$s:
  \[ \hat{X}_t^i = \frac{1}{\sqrt{\lambda_i}} \sum_j U_{ij} x_{tj}; \quad \hat{Y}_t^\alpha = ... \]

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

*J.-P. Bouchaud*
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: \( T \rightarrow N, N = M \rightarrow P \)

\[ J.-P. \text{ Bouchaud} \]
Sectorial Inflation vs. Economic indicators

\[ N = 50, \ M = 16, \ T = 265 \]
Back to eigenvectors: perturbation theory

- Consider a randomly perturbed matrix:

\[ H = H_0 + \epsilon H_1 \]

- Perturbation theory to second order in \( \epsilon \) yields:

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

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

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

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

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 [Allez, JPB]

J.-P. Bouchaud
Stability of eigenspaces: GOE

Blue: $Z = 0$, Red: $Z =$ theoretical value ($\hat{\epsilon} = 0.01$; fixed $a$, changing $b$)
The case of correlation matrices

- Consider the empirical correlation matrix:

\[ E = C + \eta \quad \eta = \frac{1}{T} \sum_{t=1}^{T} (X_t X_t^t - C) \]

- The noise \( \eta \) is correlated as:

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

- From which one derives:

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

(and a similar equation for eigenvalues)

J.-P. Bouchaud
Stability of eigenvalues: Correlations

Eigenevalues clearly change: well known correlation crises

J.-P. Bouchaud
Stability of eigenspaces: Correlations

8 meaningful eigenvectors

J.-P. Bouchaud
Stability of eigenspaces: Correlations

J.-P. Bouchaud
Stability of eigenspaces: Correlations

Empirical results

Numerical simulations

\[ P = 5 \]

\[ J.-P. \ Bouchaud \]
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 $\lambda_1 \gg \lambda_2$, and for EMA:
  – Ornstein-Uhlenbeck processes on the unit sphere around $\theta = 0$
  – Explicit solution for the distribution $P(\theta)$
  – $\det G = \cos(\theta - \theta')$

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