Cleaning correlation matrices, HCIZ integrals & Instantons

J.P Bouchaud
with: M. Potters, L. Laloux, R. Allez, J. Bun, S. Majumdar

Thank you Alain for having taught me so much!

Empirical Correlation Matrices

- 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: $N = 500 - 2000$; $T = 500 - 2500$ days (10 years)
  $\longrightarrow q := N/T = O(1)$

- In many application (e.g. portfolio optimisation) one needs to invert the correlation matrix – dangerous!

- How should one estimate/clean correlation matrices?

Rotational invariance hypothesis (RIH)

- In the absence of any cogent prior on the eigenvectors, one can assume that $\mathbf{C}$ is a member of a *Rotationally Invariant* Ensemble – "RIH"

- In finance: surely not true for the "market mode"

  $\vec{v}_1 \approx (1, 1, \ldots, 1)/\sqrt{N}$, with $\lambda_1 \approx N\rho$, but OK in the bulk

- "Cleaning" $\mathbf{E}$ within RIH: keep the eigenvectors, play with eigenvalues

  $\longrightarrow$ The simplest, classical scheme, shrinkage:

$$\mathbf{C} = (1-\alpha)\mathbf{E} + \alpha\mathbf{I} \rightarrow \widehat{\lambda}_C = (1-\alpha)\lambda_E + \alpha, \qquad \alpha \in [0, 1]$$

RMT: from $\rho_C(\lambda)$ to $\rho_E(\lambda)$

- Solution using different techniques (replicas, diagrams, free matrices) gives the resolvent $G_E(z) = N^{-1}\text{Tr}(\mathbf{E} - z\mathbf{I})$ as:

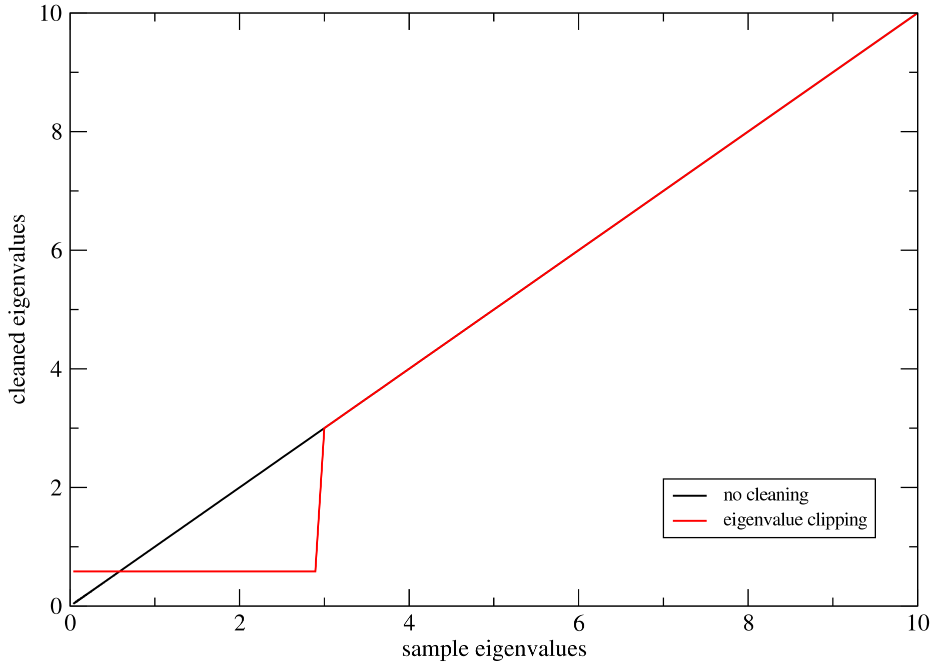
$$G_E(z) = \int d\lambda\, \rho_C(\lambda) \frac{1}{z - \lambda(1 - q + qzG_E(z))},$$

- Example 1: $\mathbf{C} = \mathbf{I}$ (null hypothesis) $\rightarrow$ Marcenko-Pastur [67]

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

- Suggests a second cleaning scheme (Eigenvalue clipping, [Laloux et al. 1997]): any eigenvalue beyond the Marcenko-Pastur edge can be trusted, the rest is noise.

Eigenvalue clipping



$\lambda < \lambda_+$ are replaced by a unique one, so as to preserve $\text{Tr}\mathbf{C} = N$.

RMT: from $\rho_C(\lambda)$ to $\rho_E(\lambda)$

- Solution using different techniques (replicas, diagrams, free matrices) gives the resolvent $G_E(z)$ as:

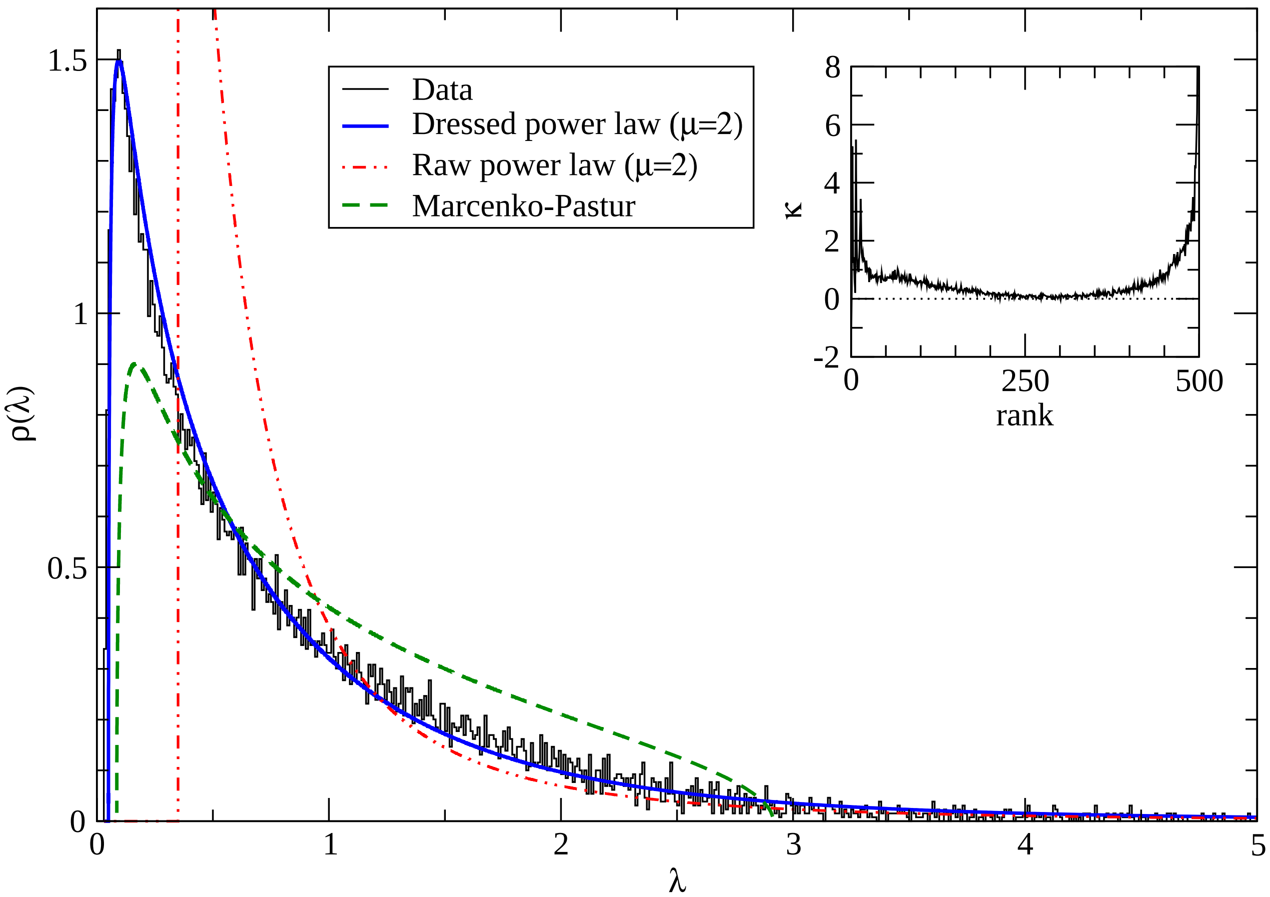
$$G_E(z) = \int d\lambda\, \rho_C(\lambda) \frac{1}{z - \lambda(1 - q + qzG_E(z))},$$

- Example 2: Power-law spectrum (motivated by data)

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

- Suggests a third cleaning scheme (Eigenvalue substitution, Potters et al. 2009, El Karoui 2010): λ_E is replaced by the theoretical λ_C with the same rank k

Empirical Correlation Matrix



MP and generalized MP fits of the spectrum

A RIH Bayesian approach

- All the above schemes lack a rigorous framework and are at best ad-hoc recipes

- A Bayesian framework: suppose $\mathbf{C}$ belongs to a RIE, with $\mathcal{P}(\mathbf{C})$ and assume Gaussian returns. Then one needs:

$$\langle \mathbf{C} \rangle|_{X_i^t} = \int \mathcal{D}\mathbf{C}\mathbf{C}\mathcal{P}(\mathbf{C}|\{X_i^t\})$$

with

$$\mathcal{P}(\mathbf{C}|\{X_i^t\}) = Z^{-1} \exp\left[-N \mathrm{Tr} V(\mathbf{C}, \{X_i^t\})\right];$$

where (Bayes):

$$V(\mathbf{C}, \{X_i^t\}) = \frac{1}{2q}\left[\log \mathbf{C} + \mathbf{E}\mathbf{C}^{-1}\right] + V_0(\mathbf{C}); \qquad V_0: \text{ prior}$$

A Bayesian approach: a fully soluble case

- $V_0(\mathbf{C}) = (1+b)\ln\mathbf{C} + b\mathbf{C}^{-1}$, $b > 0$: "Inverse Wishart"

- $\rho_C(\lambda) \propto \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda^2}$; $\lambda_\pm = (1 + b \pm \sqrt{(1+b)^2 - b^2/4})/b$

- In this case, the matrix integral can be done, leading exactly to the "Shrinkage" recipe, with $\alpha = f(b, q)$

- Note that b can be determined from the empirical spectrum of $\mathbf{E}$, using the generalized MP formula

The general case: HCIZ integrals

- A Coulomb gas approach: integrate over the orthogonal group $\mathbf{C} = \mathbf{O}\Lambda\mathbf{O}^\dagger$, where Λ is diagonal.

$$\int \mathcal{D}\mathbf{O} \exp\left[-\frac{N}{2q}\mathrm{Tr}\left[\log\Lambda + \mathbf{E}\mathbf{O}^\dagger\Lambda^{-1}\mathbf{O} + 2qV_0(\Lambda)\right]\right]$$

- Can one obtain a large N estimate of the HCIZ integral

$$F(\rho_A, \rho_B) = \lim_{N\to\infty} N^{-2} \ln \int \mathcal{D}\mathbf{O} \exp\left[\frac{N}{2q}\mathrm{Tr}\mathbf{A}\mathbf{O}^\dagger\mathbf{B}\mathbf{O}\right]$$

in terms of the spectrum of $\mathbf{A}$ and $\mathbf{B}$?

The general case: HCIZ integrals

- Can one obtain a large N estimate of the HCIZ integral

$$F(\rho_A, \rho_B) = \lim_{N \to \infty} N^{-2} \ln \int \mathcal{D}\mathbf{O} \exp\left[\frac{N}{2q} \mathrm{Tr} \mathbf{A}\mathbf{O}^\dagger \mathbf{B}\mathbf{O}\right]$$

  in terms of the spectrum of $\mathbf{A}$ and $\mathbf{B}$?

- When $\mathbf{A}$ (or $\mathbf{B}$) is of finite rank, such a formula exists in terms of the "R-transform" of B (with a different scaling in N) [Marinari, Parisi & Ritort, 1995].

- When the rank of $\mathbf{A}, \mathbf{B}$ are of order N, there is a formula due to Matytsin [94] (in the unitary case), later shown rigorously by Zeitouni & Guionnet, but its derivation is quite obscure...

An instanton approach to large N HCIZ

- Consider Dyson's Brownian motion matrices. The eigenvalues obey:

$$\mathrm{d}x_i = \sqrt{\frac{2}{\beta N}}\mathrm{d}W + \frac{1}{N}\mathrm{d}t \sum_{j \neq i} \frac{1}{x_i - x_j},$$

- Constrain $x_i(t=0) = \lambda_{Ai}$ and $x_i(t=1) = \lambda_{Bi}$. The probability of such a path is given by a large deviation/instanton formula, with:

$$\frac{d^2 x_i}{dt^2} = -\frac{2}{N^2} \sum_{\ell \neq i} \frac{1}{(x_i - x_\ell)^3}.$$

An instanton approach to large N HCIZ

- Constrain $x_i(t=0) = \lambda_{Ai}$ and $x_i(t=1) = \lambda_{Bi}$. The probability of such a path is given by a large deviation/instanton formula, with:

$$\frac{d^2 x_i}{dt^2} = -\frac{2}{N^2} \sum_{\ell \neq i} \frac{1}{(x_i - x_\ell)^3}.$$

- This can be interpreted as the motion of massive particles interacting through an *attractive* two-body potential $\phi(r) = -(Nr)^{-2}$. Using the virial formula, one gets in the hydrodynamic limit Matytsin's equations:

$$\partial_t \rho + \partial_x[\rho v] = 0, \qquad \partial_t v + v \partial_x v = \pi^2 \rho \partial_x \rho.$$

with $\rho(x, t=0) = \rho_A(x)$ and $\rho(x, t=1) = \rho_B(x)$

An instanton approach to large N HCIZ

- Finally, the "action" associated to these trajectories is:

$$S \approx \frac{1}{2}\int \mathrm{d}x\rho\left[v^2 + \frac{\pi^2}{3}\rho^2\right] - \frac{1}{2}\left[\int \mathrm{d}x\mathrm{d}y\rho_Z(x)\rho_Z(y)\ln|x-y|\right]_{Z=A}^{Z=B}$$

- Now, the link with HCIZ comes from noticing that the propagator of the Brownian motion in matrix space is:

$$\mathcal{P}(\mathbf{B}|\mathbf{A}) \propto \exp -[\frac{N}{2}\,\mathrm{Tr}(\mathbf{A}-\mathbf{B})^2] = \exp -\frac{N}{2}[\mathrm{Tr}\mathbf{A}^2 + \mathrm{Tr}\mathbf{B}^2 - 2\mathrm{Tr}\mathbf{AOBO}^\dagger]$$

Disregarding the eigenvectors of $\mathbf{B}$ (i.e. integrating over $\mathbf{O}$) leads to another expression for $P(\lambda_{Bi}|\lambda_{Aj})$ in terms of HCIZ that can be compared to the one using instantons

- The final result for $F(\rho_A, \rho_B)$ is exactly Matytsin's expression, up to small details (!)

An instanton approach to large N HCIZ

- An alternative path: use the Kawasaki-Dean equation describing the density of Dyson random walks:

$$\partial_t \rho(x,t) + \partial_x J(x,t) = 0$$

with:

$$J(x,t) = \frac{1}{N}\xi(x,t)\sqrt{\rho(x,t)} - \frac{1}{2N}\partial_x \rho(x,t) - \rho(x,t)\int \mathrm{d}y\, \partial_x V(x-y)\rho(y,t),$$

where $V(r) = -\ln r$ is the "true" two-body interaction potential ($\neq \phi(r)$!), $\xi(x,t)$ is a normalized Gaussian white noise.

An instanton approach to large N HCIZ

- One then writes the weights of histories of $\{\rho(x,t)\}$ using Martin-Siggia-Rose path integrals:

$$\mathcal{P}(\{\rho(x,t)\}) \propto \left\langle \int \mathcal{D}\psi \, e^{\left[\int_0^1 \mathrm{d}t \int \mathrm{d}x N^2 i\psi(x,t)(\partial_t \rho + \partial_x J)\right]} \right\rangle_\xi$$

- Performing the average over ξ:

$$\mathcal{S} = N^2 \int_0^1 \mathrm{d}t \int \mathrm{d}x \left[\psi \partial_t \rho + F(x,t)\rho\partial_x \psi - \frac{\psi}{2N}\partial_{xx}^2 \rho + \frac{1}{2}\rho(\partial_x \psi)^2\right]$$

with $F(x,t) = \int \mathrm{d}y \partial_x V(x-y)\rho(y,t)$.

An instanton approach to large N HCIZ

- Taking functional derivatives with respect to ρ and ψ then leads to:

$$\partial_t \rho = \partial_x(\rho F) + \partial_x(\rho \partial_x \psi) + \frac{1}{2N}\partial^2_{xx}\rho$$

and

$$\partial_t \psi - \frac{1}{2}(\partial_x \psi)^2 = F\partial_x \psi - \frac{1}{2N}\partial^2_{xx}\psi - \partial_x \int \mathrm{d}y\, V(x-y)\rho(y,t)\partial_y \psi(y,t)$$

- The Euler-Matystin equations are again recovered, after a little work, by setting $v(x,t) = -F(x,t) - \partial_x \psi(x,t)$.

Back to eigenvalue cleaning...

- Estimating HCIZ at large N is only the first step, but...

- ...one still needs to apply it to $\mathbf{B} = \mathbf{C}^{-1}$, $\mathbf{A} = \mathbf{E} = X^{\dagger}\mathbf{C}X$ and to compute also correlation functions such as

$$\langle O_{ij}^2 \rangle_{\mathbf{E} \to \mathbf{C}^{-1}}$$

  with the HCIZ weight – in progress

- As we were working on this we discovered the work of Ledoit-Péché that solves the problem exactly using tools from RMT...

The Ledoit-Péché “magic formula”

- The Ledoit-Péché [2011] formula is a non-linear shrinkage, given by:

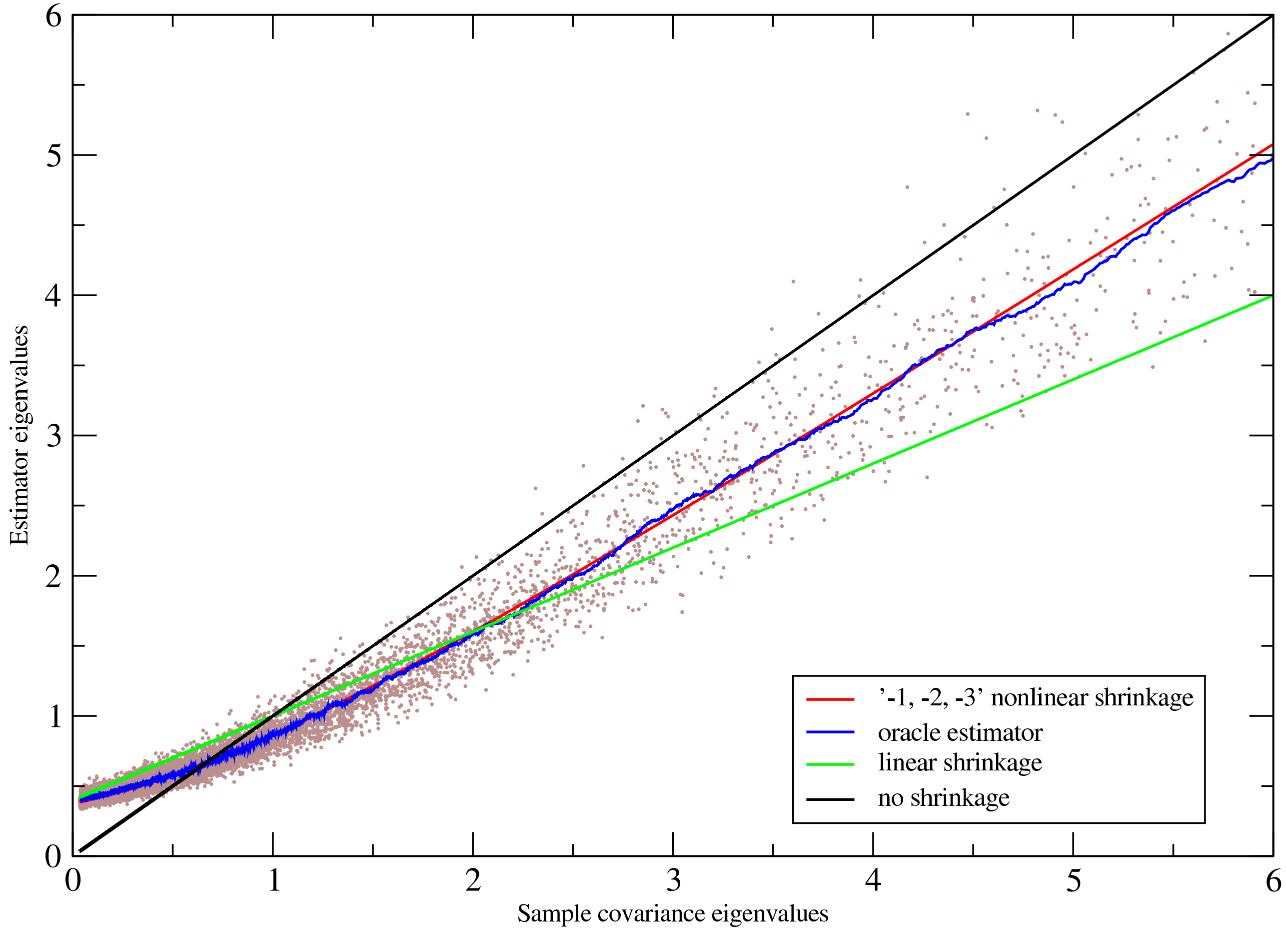
$$\widehat{\lambda}_C = \frac{\lambda_E}{|1 - q + q\lambda_E \lim_{\epsilon \to 0} G_E(\lambda_E - i\epsilon)|^2}.$$

- Note 1: Independent of $\mathbf{C}$: only G_E is needed (and is observable)!

- Note 2: When applied to the case where $\mathbf{C}$ is inverse Wishart, this gives again the linear shrinkage

- Note 3: Still to be done: reobtain these results using the HCIZ route (many interesting intermediate results to hope for!)

Eigenvalue cleaning: Ledoit-Péché



Fit of the empirical distribution with $V_0'(z) = a/z + b/z^2 + c/z^3$.

What about eigenvectors?

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

- What about eigenvectors? What natural null-hypothesis beyond RIH?

- Are eigen-values/eigen-directions *stable* in time? $\longrightarrow$ Romain Allez

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

Bibliography

- J.P. Bouchaud, M. Potters, *Financial Applications of Random Matrix Theory: a short review*, in "The Oxford Handbook of Random Matrix Theory" (2011)

- R. Allez and J.-P. Bouchaud, *Eigenvectors dynamics: general theory & some applications*, arXiv 1108.4258

- P.-A. Reigneron, R. Allez and J.-P. Bouchaud, *Principal regression analysis and the index leverage effect*, Physica A, Volume 390 (2011) 3026-3035.

- J. Bun, J.-P. Bouchaud, S. Majumdar, M. Potters,*An instanton approach to large N Harish-Chandra-Itzykson-Zuber integrals*, Physical Review Letters, 113, 070201 (2014)