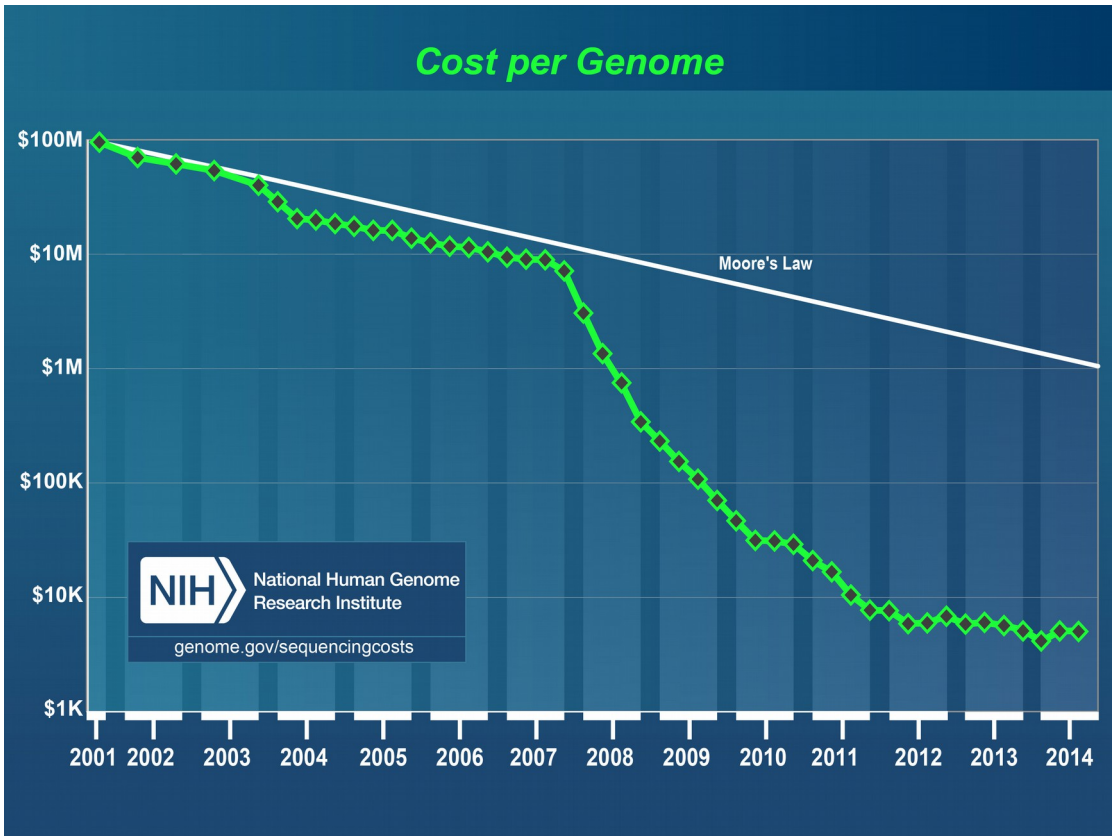
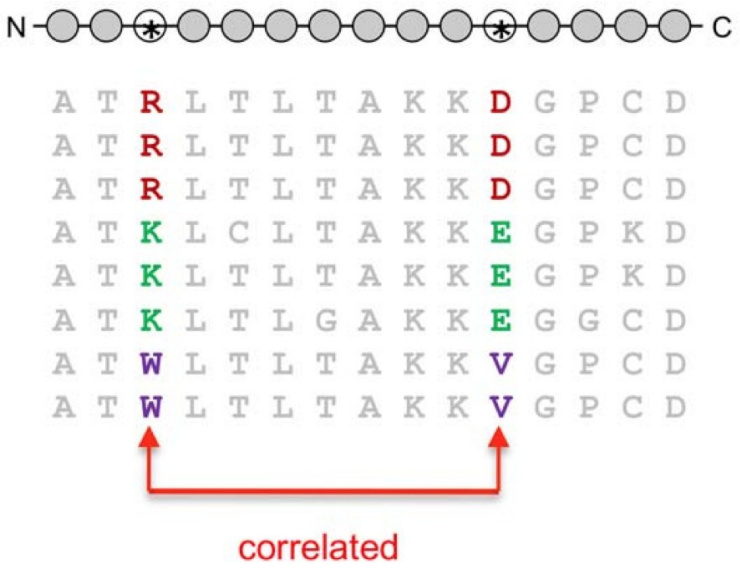
Dario Villamaina

École normale supérieure & Institut Philippe Meyer

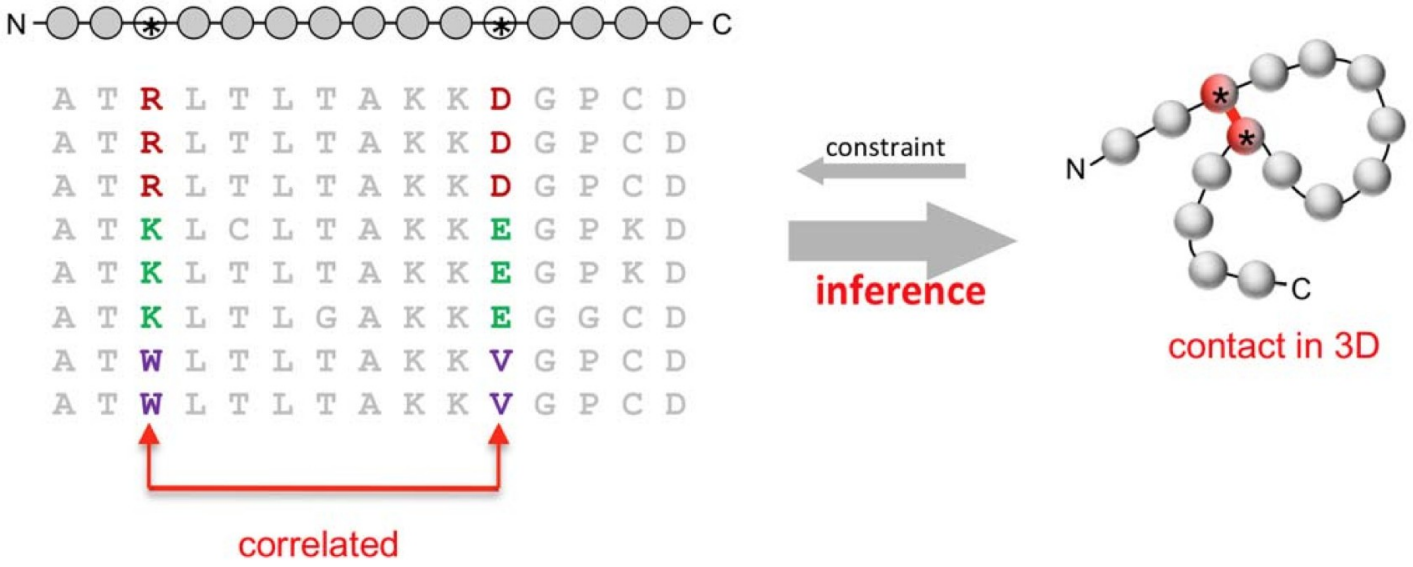
Random matrices and correlation data analysis



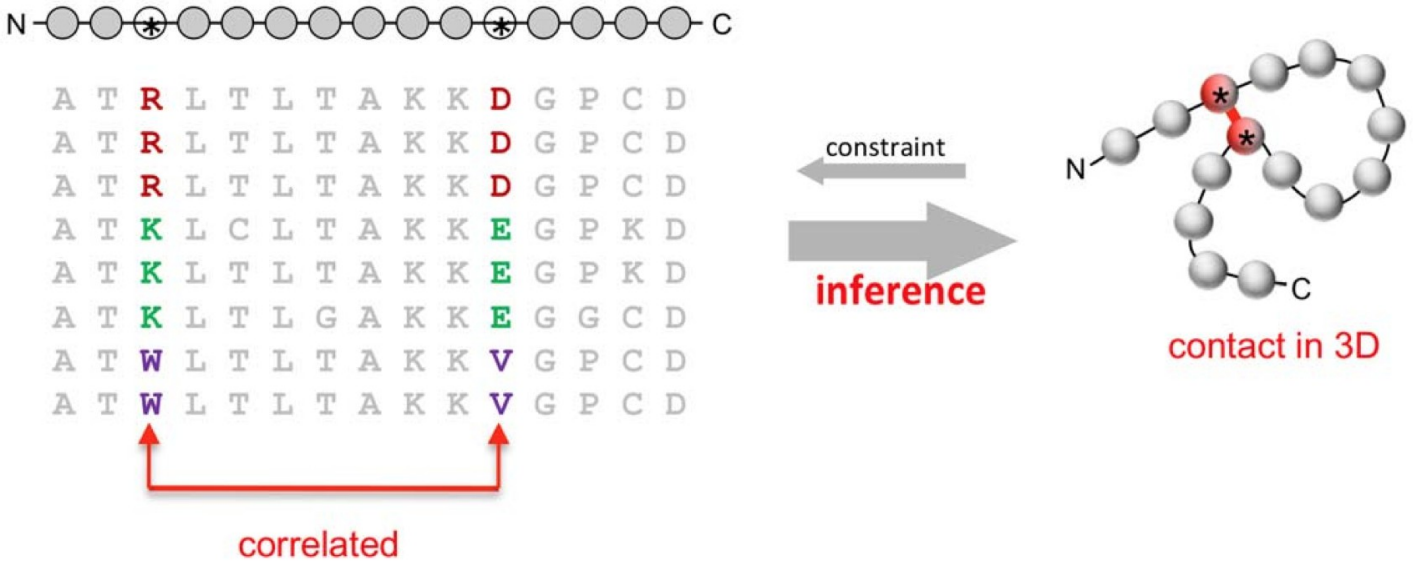
Motivation: Analysis of sequences



Motivation: Analysis of sequences



Motivation: Analysis of sequences



Data Sequences

Empirical Correlation matrix

$\hat{C}$

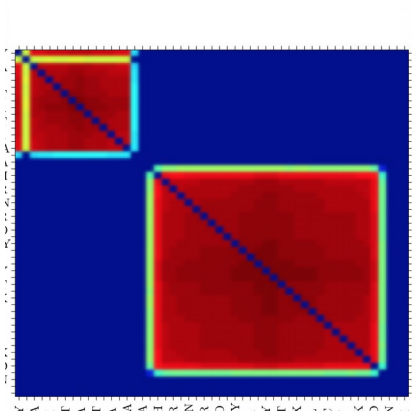
3D Contacts prediction

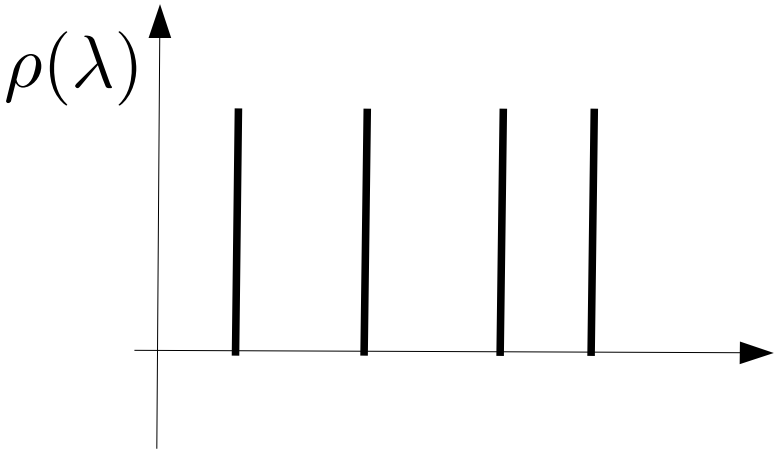
Sites in contact are associated with high localized eigenvectors of the correlation matrix

Population Covariance

Infinite statistics

C (One matrix)

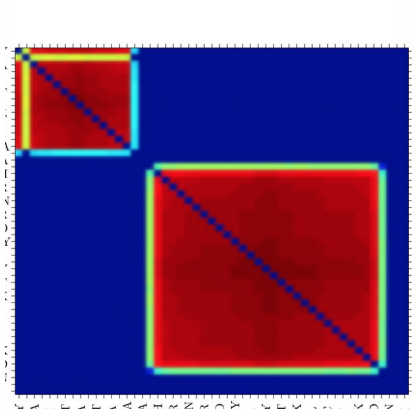




Population Covariance *vs* Empirical Covariance

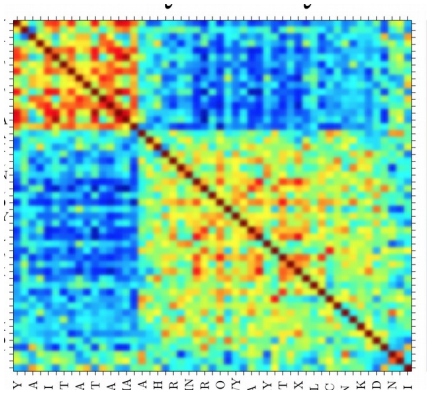
Infinite statistics

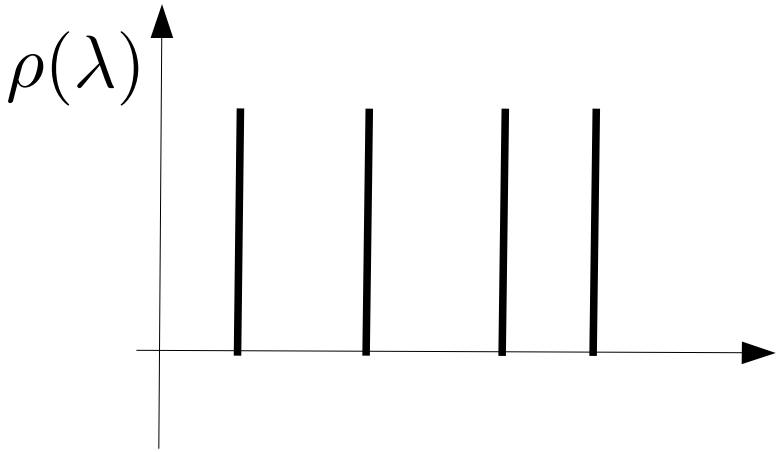
C (One matrix)

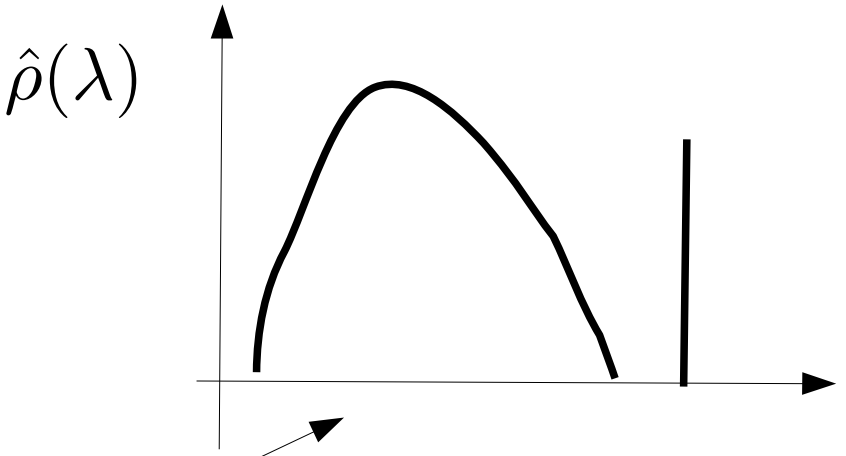


Finite statistics

$\hat{C}$ (Ensembles of matrices)







Part of the signal is hidden
by finite sampling effects!

Random Matrix Model

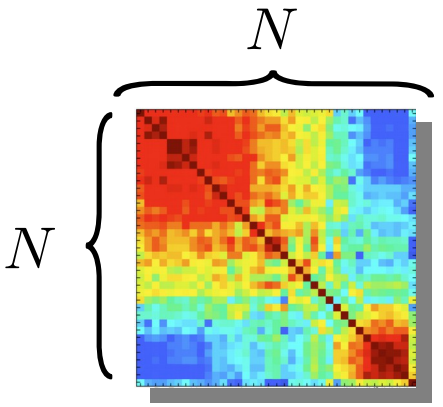
Number of measures (M)

$$\hat{X} = \overbrace{\begin{pmatrix} -0.2526 & -0.8387 & 0.9620 & -0.2771 & 0.2107 \\ 0.8294 & -0.2004 & -1.0053 & 0.3873 & -0.4419 \\ 0.5982 & 1.0760 & 0.5182 & 0.2671 & 0.4277 \\ -0.0285 & -1.0486 & -0.7054 & 1.1440 & -1.0764 \\ -1.1416 & -0.5782 & -0.1396 & -0.9523 & 0.2344 \end{pmatrix}}^{\text{Number of variables } (N)}$$

Gaussian sample with correlation matrix C

$$\hat{C} \propto \hat{X} \times \hat{X}^T$$

$$\hat{C}_{ij} = \frac{1}{M} \sum_t x_{i,t} x_{j,t}$$



Efficiency parameter $r = \frac{N}{M}$

Perfect sampling limit $\hat{C} \underset{r \to 0}{\to} C$

Toy model : Finite rank correlation matrix

$$C \propto \begin{pmatrix} \gamma & & 0 \\ & 1 & \\ 0 & & \ddots \\ & & & 1 \end{pmatrix} \qquad \boldsymbol{\xi}_1 \propto \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

continuous limit $\Longrightarrow$ $\rho_C(\lambda) = (1-\alpha)\delta(\lambda - 1) + \alpha\delta(\lambda - \gamma)$

Toy model : Finite rank correlation matrix

$$C \propto \begin{pmatrix} \gamma & & 0 \\ & 1 & \\ 0 & & \ddots \\ & & & 1 \end{pmatrix} \qquad \xi_1 \propto \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

continuous limit $\Rightarrow$ $\rho_C(\lambda) = (1-\alpha)\delta(\lambda - 1) + \alpha\delta(\lambda - \gamma)$

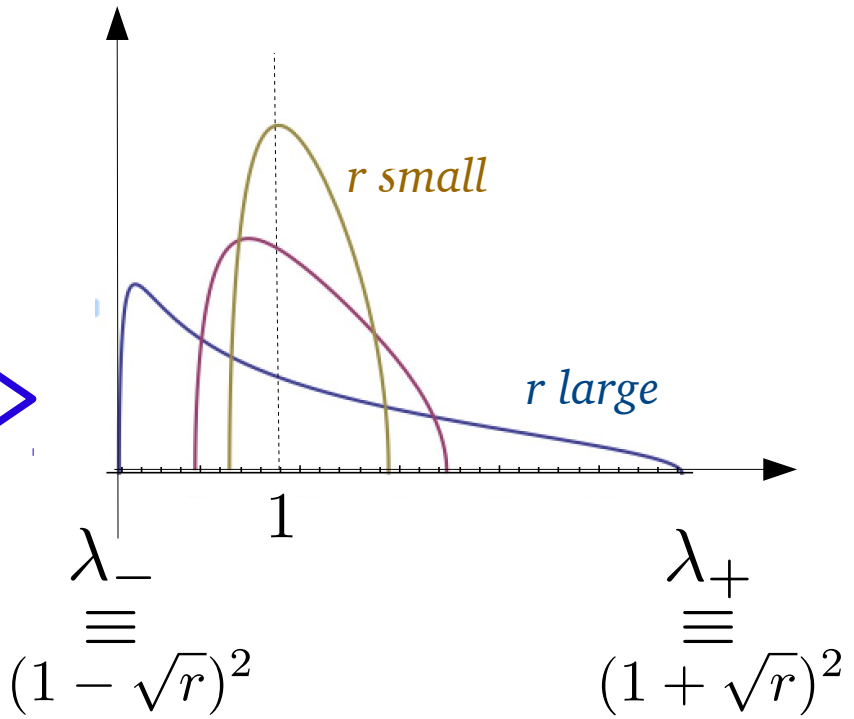
Case $\gamma = 1 \quad \alpha = 0$ (Null model)

$$\hat{C} \equiv XX^t$$

Wishart ensemble

$$\rho_{\hat{C}}(\lambda) = \frac{\sqrt{(\lambda - \lambda_-)(\lambda_+ - \lambda)}}{2\lambda\pi r}$$

Marcenko-Pastur distribution

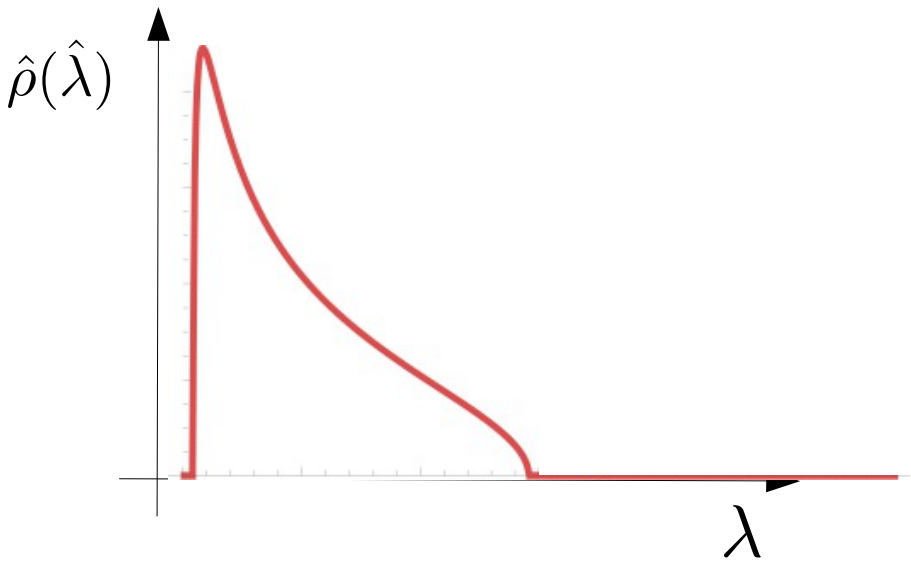


The transition in the eigenvalue spectrum

Case of γ finite and $\alpha \to 0$

Weak signal $\gamma < \gamma_c \equiv 1 + \sqrt{r}$

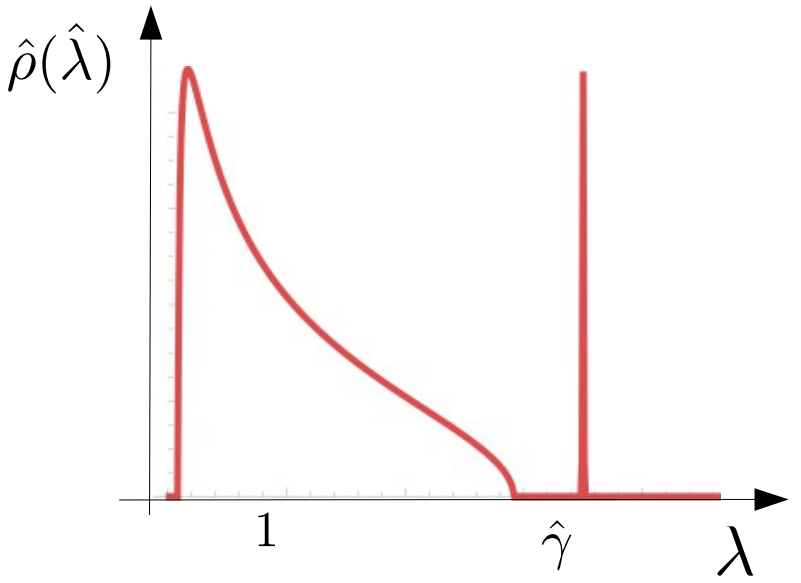
$$\hat{\rho}(\hat{\lambda}) \simeq \hat{\rho}_{bulk}(\hat{\lambda})$$



Strong signal $\gamma > \gamma_c$

$$\hat{\rho}(\hat{\lambda}) \simeq (1-\alpha)\hat{\rho}_{bulk}(\hat{\lambda}) + \alpha\,\delta(\hat{\lambda} - \hat{\gamma})$$

delta peak



Noise undressing of empirical correlation matrices

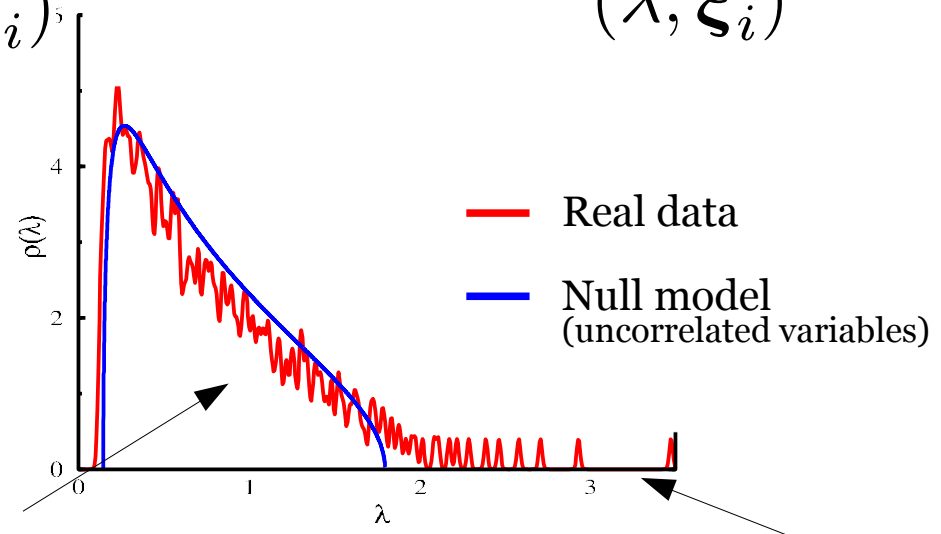
L Laloux, P Cizeau, JP Bouchaud, M Potters - Physical Review Letters, (1999)

(Measured)

$$\hat{C} \quad (\hat{\lambda}, \hat{\boldsymbol{\xi}}_i)$$

(True)

$$C \quad (\lambda, \boldsymbol{\xi}_i)$$



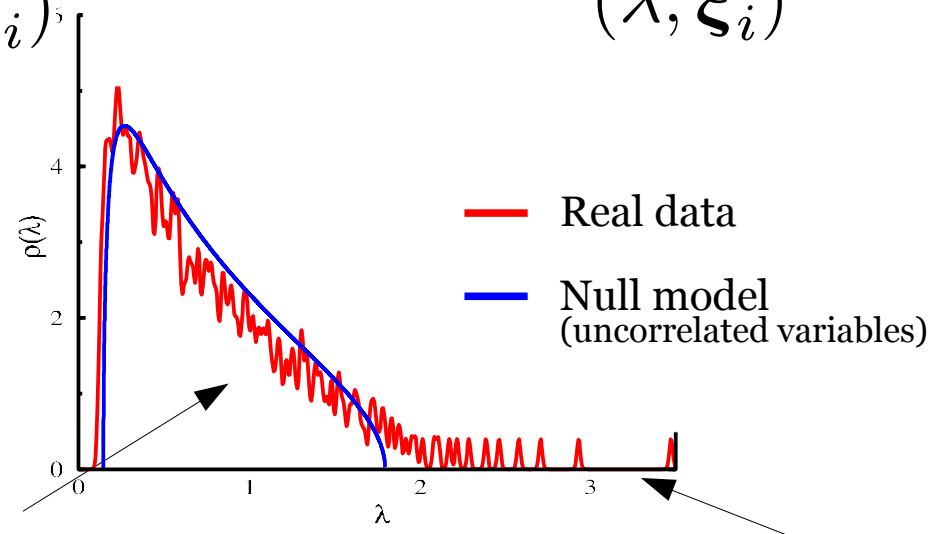
(Only noise ?)
Bulk eigenvalues

Signal

Noise undressing of empirical correlation matrices

L Laloux, P Cizeau, JP Bouchaud, M Potters - Physical Review Letters, (1999)

(Measured) $\hat{C}$ $(\hat{\lambda}, \hat{\boldsymbol{\xi}}_i)$ → (True) C $(\lambda, \boldsymbol{\xi}_i)$



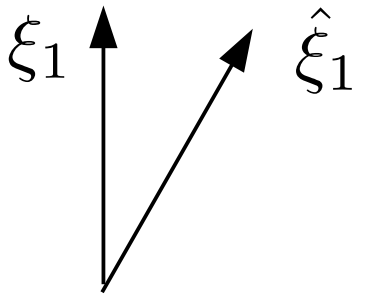
(Only noise ?)
Bulk eigenvalues

Signal

What about eigenvectors ?

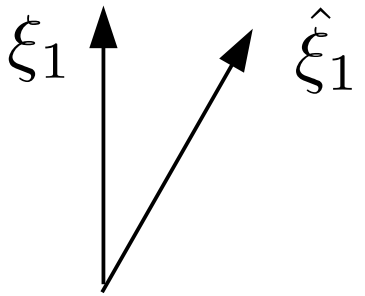
Estimation of true eigenvectors from the empirical ones

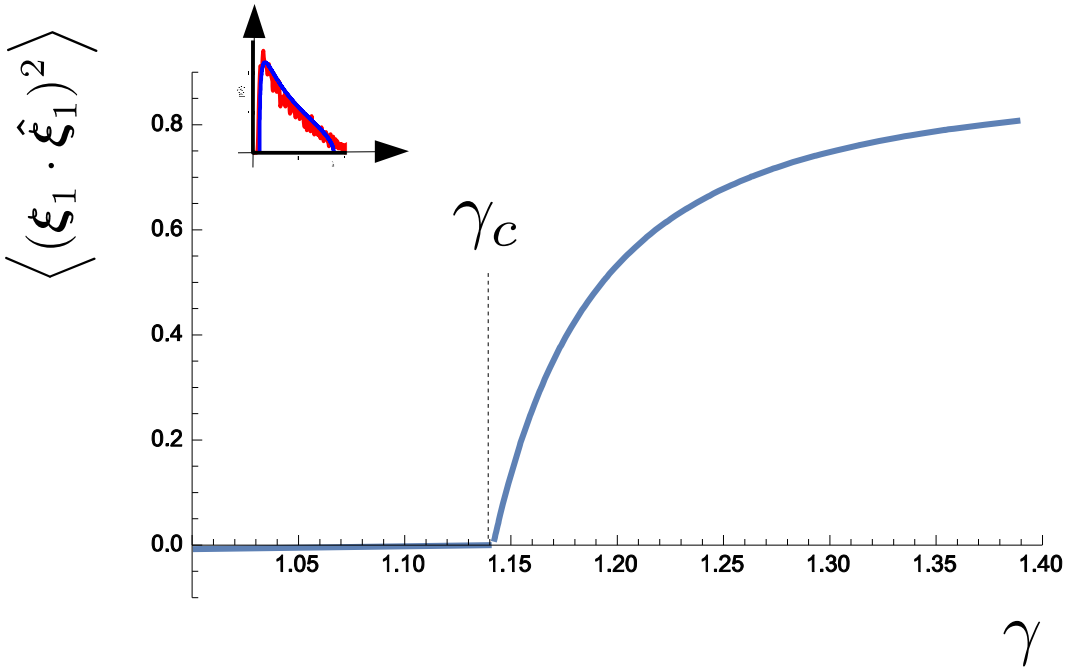
Scalar product between the largest true and empirical eigenvectors



Estimation of true eigenvectors from the empirical ones

Scalar product between the largest true and empirical eigenvectors

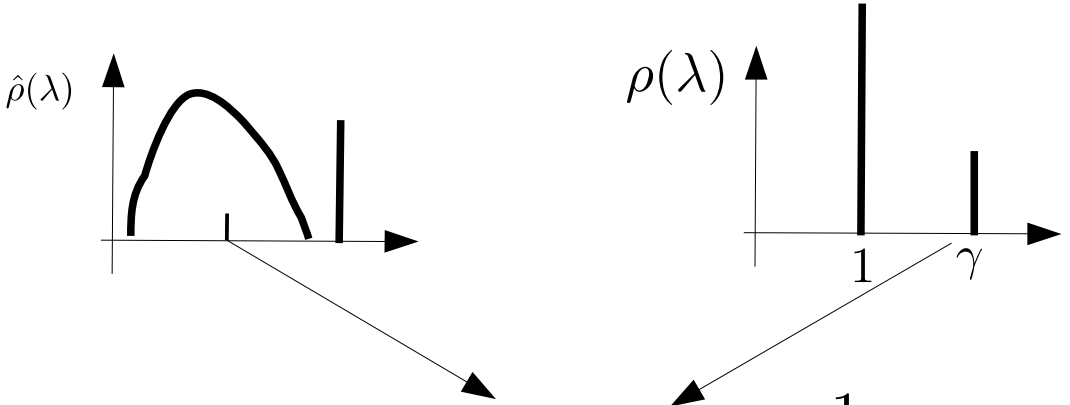




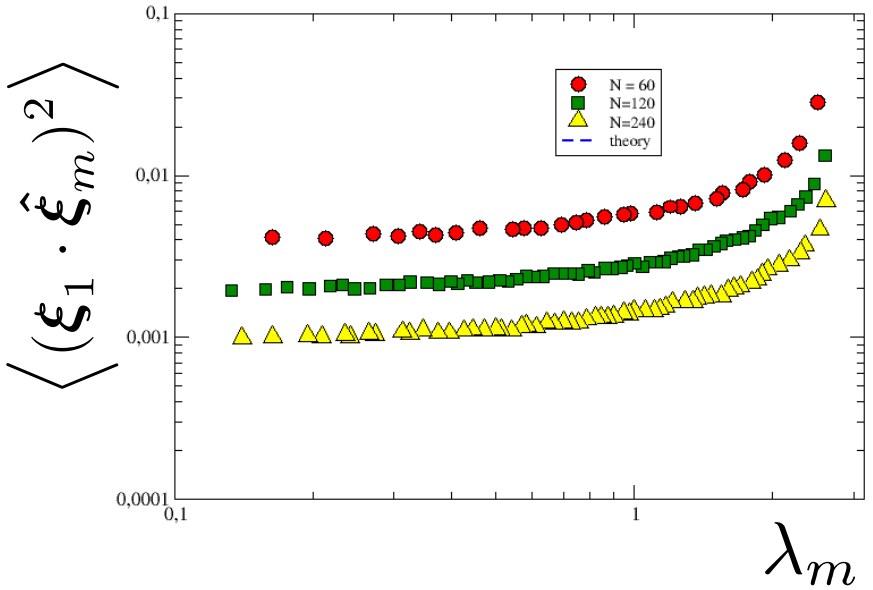
$$\left\langle (\boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_1)^2 \right\rangle \equiv w_1^2 = \begin{cases} 0 & \gamma < 1 + \sqrt{r} \\ \dfrac{(\gamma - 1)^2 - r}{(\gamma - 1)(\gamma + r - 1)} & \gamma > 1 + \sqrt{r} \end{cases}$$

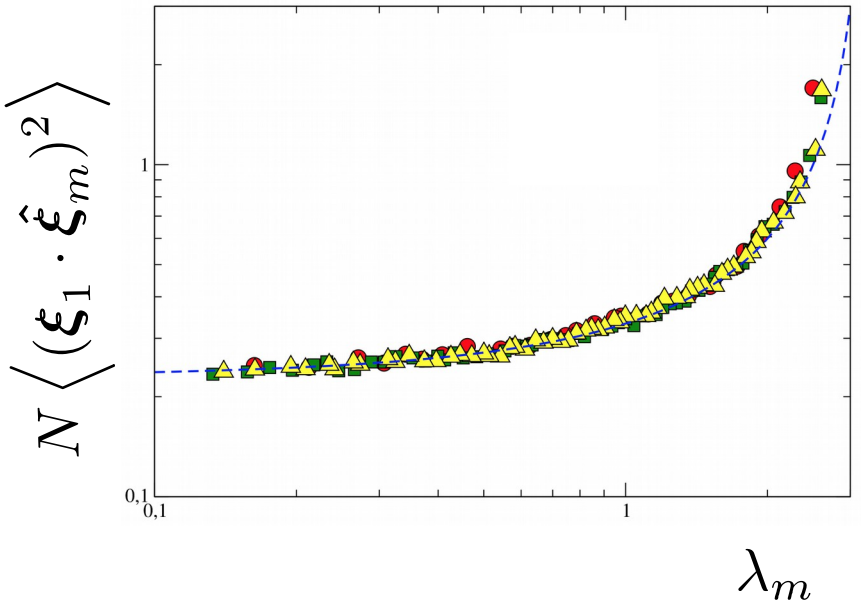
Estimation of true eigenvectors from the empirical ones

Scalar product between the largest true and bulk eigenvectors

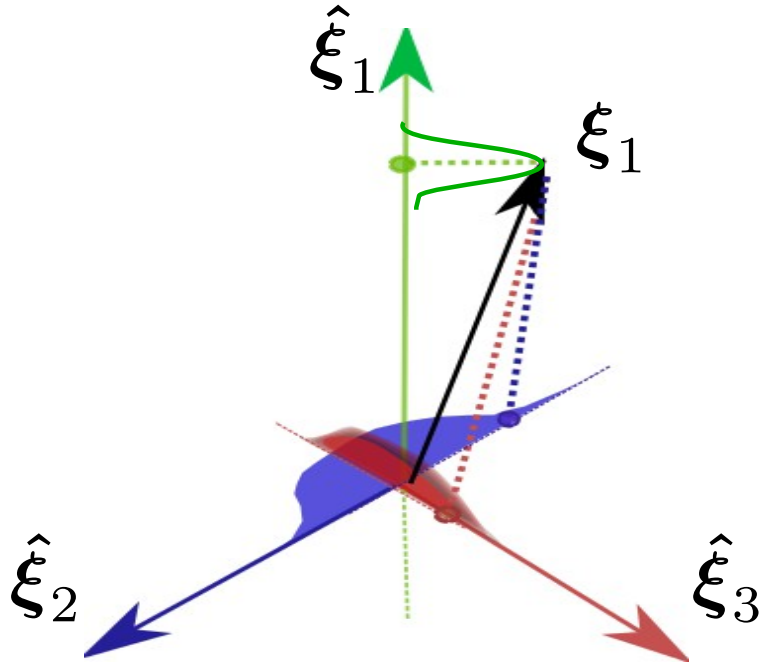


$$(\hat{\boldsymbol{\xi}}_m \cdot \boldsymbol{\xi}_1) \sim \frac{1}{\sqrt{N}}$$





Summary on empirical eigenvectors



$$\left\langle |\boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_1|^2 \right\rangle = \mathcal{O}(1) \qquad \left\langle |\boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_{m>1}|^2 \right\rangle = \mathcal{O}(N^{-1})$$

Two questions:

Is it possible to calculate analytically the scalar products ?

Is there information stored in bulk empirical eigenvectors ?

Part 1: How to calculate the scalar products?

$$g(\lambda,\hat{\lambda}) = \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\left[\left(\lambda - C\right)^{-1} \cdot \left\langle\left(\hat{\lambda} - \hat{C}\right)^{-1}\right\rangle\right]$$

Projection on the eigenvectors

$$\begin{cases} \left(\lambda - C\right)^{-1}_{ij} = \sum_m (\lambda - \lambda_m)^{-1} \xi_{m,i}\, \xi_{m,j} \\ \left(\hat{\lambda} - \hat{C}\right)^{-1}_{ij} = \sum_m (\hat{\lambda} - \hat{\lambda}_m)^{-1} \hat{\xi}_{m,i}\, \hat{\xi}_{m,j} \end{cases}$$

Continuous limit of the spectrum

$$\hat{\rho}(\hat{\mu}) = \frac{1}{N}\sum_m \left\langle \delta(\hat{\mu} - \hat{\lambda}_m)\right\rangle$$

Part 1: How to calculate the scalar products?

$$g(\lambda,\hat{\lambda}) = \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\left[(\lambda - C)^{-1}\cdot\langle(\hat{\lambda}-\hat{C})^{-1}\rangle\right]$$

Projection on the eigenvectors

$$\begin{cases} (\lambda - C)^{-1}_{ij} = \sum_m (\lambda-\lambda_m)^{-1}\xi_{m,i}\,\xi_{m,j} \\ (\hat{\lambda} - \hat{C})^{-1}_{ij} = \sum_m (\hat{\lambda}-\hat{\lambda}_m)^{-1}\hat{\xi}_{m,i}\,\hat{\xi}_{m,j} \end{cases}$$

Continuous limit of the spectrum

$$\hat{\rho}(\hat{\mu}) = \frac{1}{N}\sum_m \left\langle \delta(\hat{\mu}-\hat{\lambda}_m)\right\rangle$$

$$g(\lambda,\hat{\lambda}) = \int d\mu\, d\hat{\mu}\, \frac{\rho(\mu)}{\lambda-\mu}\,\frac{\hat{\rho}(\hat{\mu})}{\hat{\lambda}-\hat{\mu}}\, W^2(\mu,\hat{\mu})$$

$$N\left\langle (\hat{\xi}_{\hat{\mu}}\cdot\xi_\mu)^2\right\rangle$$

$$\mathrm{Im}\frac{1}{\lambda-\mu-i\epsilon} = \pi\delta(\lambda-\mu)$$

$$W^2(\lambda,\hat{\lambda}) = \frac{g(\lambda+i\epsilon,\hat{\lambda}+i\epsilon) - g(\lambda+i\epsilon,\hat{\lambda}-i\epsilon)}{2\pi\,\rho(\lambda)\,\hat{\rho}(\hat{\lambda})}$$

An extension of the Edwards-Jones technique

SF Edwards and RC Jones Journal of Physics A: Mathematical and General, 9(10), 1595 (1976)

$$\left\langle \mathrm{Tr}((\lambda - C)^{-1}(\hat{\lambda} - \hat{C})^{-1}) \right\rangle = \left\langle \int \mathcal{D}Y\mathcal{D}Z(Y \cdot Z)^2 \exp(-\frac{1}{2}(Y(\lambda - C)Y^{\dagger} + Z(\hat{\lambda} - \hat{C})Z^{\dagger}) \right\rangle$$

$$\langle ... \rangle \propto \int \prod_{i,t} dx_{i,t} \exp\left(-\frac{1}{2} x_{i,t} C_{ij}^{-1} x_{j,t}\right) \qquad \hat{C}_{ij} = \frac{1}{M} \sum_{t} x_{i,t} x_{j,t}$$

An extension of the Edwards-Jones technique

SF Edwards and RC Jones Journal of Physics A: Mathematical and General, 9(10), 1595 (1976)

$$\left\langle \mathrm{Tr}((\lambda - C)^{-1}(\hat{\lambda} - \hat{C})^{-1}) \right\rangle = \left\langle \int \mathcal{D}Y\mathcal{D}Z (Y \cdot Z)^2 \exp(-\frac{1}{2}(Y(\lambda - C)Y^{\dagger} + Z(\hat{\lambda} - \hat{C})Z^{\dagger}) \right\rangle$$

$$\langle ... \rangle \propto \int \prod_{i,t} dx_{i,t} \exp\left(-\frac{1}{2} x_{i,t} C_{ij}^{-1} x_{j,t}\right) \qquad \hat{C}_{ij} = \frac{1}{M} \sum_{t} x_{i,t} x_{j,t}$$

Gaussian integrals

Replica-based calculation

Saddle point equation

$$N \to \infty \quad M \to \infty \quad r = \frac{N}{M}$$

$$g(\lambda, \hat{\lambda}) = \frac{2}{N} \int d\phi \; d\hat{\phi} \frac{\partial}{\partial \eta}\bigg|_{\eta=0} e^{-N S_{\eta}(\phi, \hat{\phi})}$$

$$S_{\eta}(\phi, \hat{\phi}) = \frac{1}{2}\left(\int d\mu \rho(\mu) \ln((\lambda - \mu)(\hat{\lambda} - \mu\,\hat{\phi}) - \eta) + \frac{1}{r} \ln(1 - r\phi) + \phi\,\hat{\phi} \right)$$

Equations for the densities

O. Ledoit and S. Péché, Probability Theory and Related Fields 151.1-2 (2011).
JW Silverstein Journal of Multivariate Analysis, 55(2), (1995)

$$g(\lambda, \hat{\lambda}) = \int d\mu \frac{\rho(\mu)}{(\lambda - \mu)(\hat{\lambda} - \mu\,\hat{\phi}^*)}$$

Saddle point equation

$$\frac{\hat{\phi}^* - 1}{r\hat{\phi}^*} = \int d\mu \frac{\rho(\mu)}{(\hat{\lambda} - \hat{\mu}\hat{\phi}^*)}$$

$$\lambda \gg 1 \quad \Longrightarrow \quad \hat{\rho}(\hat{\lambda}) = \frac{1}{\pi}\mathrm{Im} \int d\mu \frac{\rho(\mu)}{(\hat{\lambda} - \mu\,\hat{\phi}^*)}$$

Equations for the densities

O. Ledoit and S. Péché, Probability Theory and Related Fields 151.1-2 (2011).
JW Silverstein Journal of Multivariate Analysis, 55(2), (1995)

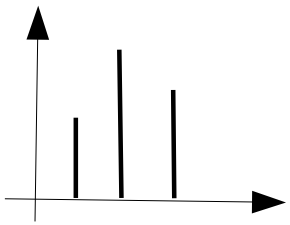
$$g(\lambda,\hat{\lambda}) = \int d\mu \frac{\rho(\mu)}{(\lambda-\mu)(\hat{\lambda}-\mu\,\hat{\phi}^*)}$$

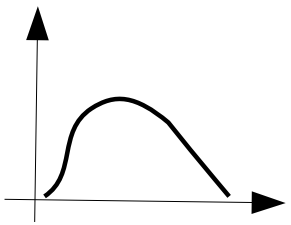
Saddle point equation

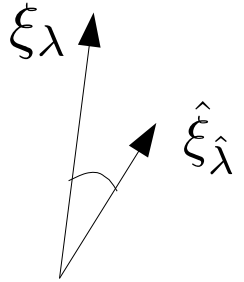
$$\frac{\hat{\phi}^*-1}{r\hat{\phi}^*} = \int d\mu \frac{\rho(\mu)}{(\hat{\lambda}-\hat{\mu}\hat{\phi}^*)}$$

$$\lambda \gg 1 \quad \Longrightarrow \quad \hat{\rho}(\hat{\lambda}) = \frac{1}{\pi}\mathrm{Im}\int d\mu \frac{\rho(\mu)}{(\hat{\lambda}-\mu\,\hat{\phi}^*)}$$

Input $\rho(\lambda)$ ⟹ **Noise dressing** $\hat{\rho}(\hat{\lambda})$ ⟹ **Scalar products** $W^2(\lambda,\hat{\lambda})$





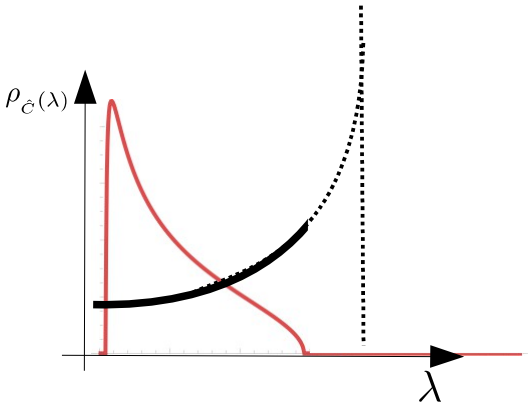


The transition from eigenvectors point of view

$$\rho(\lambda) = (1-\alpha)\delta(\lambda-1) + \alpha\delta(\lambda-\gamma)$$

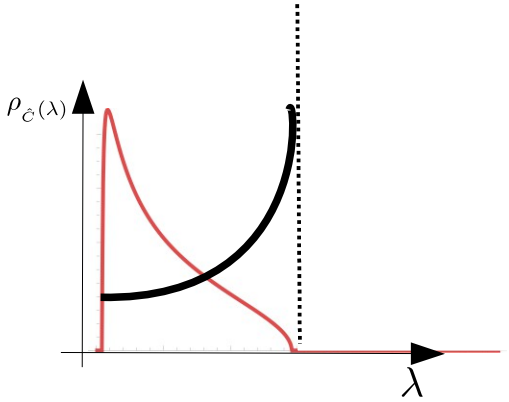
$$W^2(\gamma,\hat{\lambda}) = \frac{\gamma r}{\hat{\lambda} + \gamma(\gamma - \hat{\lambda} + r - 1)} \qquad \text{(with a divergence in } \hat{\lambda} = \gamma + \frac{r\gamma}{\gamma - 1}\text{)}$$

$\gamma < \gamma_c$

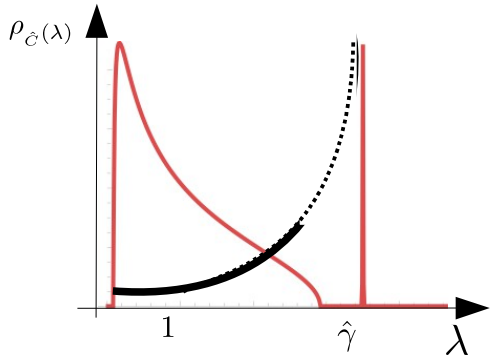


$$\int_{\lambda_-}^{\lambda_+} W^2(\gamma,\hat{\lambda})\rho(\hat{\lambda})d\hat{\lambda} = 1$$

$\gamma \equiv \gamma_c$



$\gamma > \gamma_c$



$$\int_{\lambda_-}^{\lambda_+} W^2(\gamma,\hat{\lambda})\rho(\hat{\lambda})d\hat{\lambda} = 1 - W_1^2$$

Estimation of the principal component

Inverse problem : How to estimate $\boldsymbol{\xi}_1$?

$$\boldsymbol{\xi}_1^{(est)} = w_1 \hat{\boldsymbol{\xi}}_1 + \sum_m^N w_m \hat{\boldsymbol{\xi}}_m$$

Estimation of the principal component

Inverse problem : How to estimate $\boldsymbol{\xi}_1$?

$$\boldsymbol{\xi}_1^{(est)} = w_1 \hat{\boldsymbol{\xi}}_1 + \sum_m^N w_m \hat{\boldsymbol{\xi}}_m$$

The *ws* are Gaussian variables (fairly verified for large *N*)

$$P\left(\boldsymbol{\xi}_1^{(est)} | \{\hat{\xi}\}\right)$$

"standard approach"

$$\left\langle \boldsymbol{\xi}_1^{(est)} \right\rangle = \langle w_1 \rangle \, \hat{\boldsymbol{\xi}}_1$$

Main limitation: reconstruction impossible for weak signals

$$\gamma < \gamma_c \longrightarrow \langle w_1 \rangle = 0$$

One question left:

Is it possible to calculate analytically the scalar products ?

YES !!

Is there information stored in bulk empirical eigenvectors ?

The mutual information

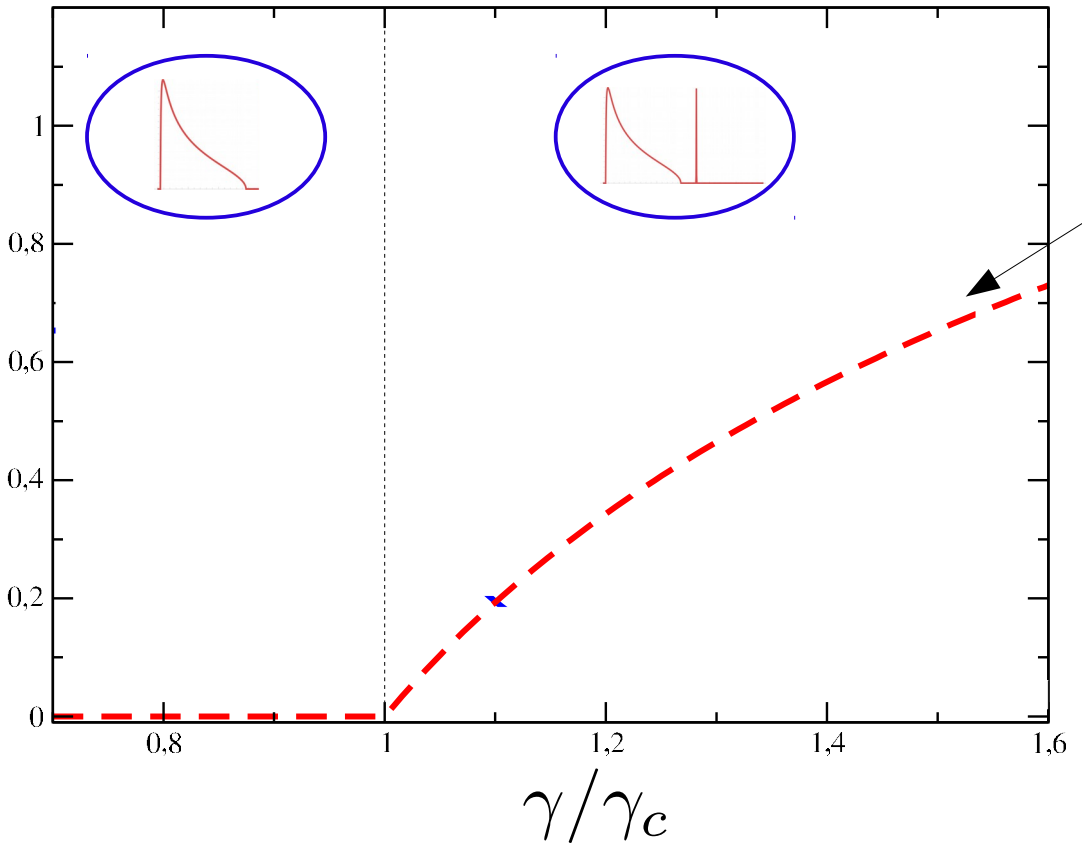
$$I(x;y) = \left\langle \log\left(\frac{P(x,y)}{P(x)P(y)}\right)\right\rangle$$

It measures how much knowing X reduces uncertainty about Y

The mutual information

$$I(x;y) = \left\langle \log\left(\frac{P(x,y)}{P(x)P(y)}\right)\right\rangle \quad \Longrightarrow \quad I\left(\boldsymbol{\xi}_1;\hat{\boldsymbol{\xi}}_1\right) = -\frac{1}{2}\mathrm{Log}\left(1-\left\langle\left(\hat{\boldsymbol{\xi}}_1\cdot\boldsymbol{\xi}_1\right)^2\right\rangle\right)$$

It measures how much knowing X reduces uncertainty about Y



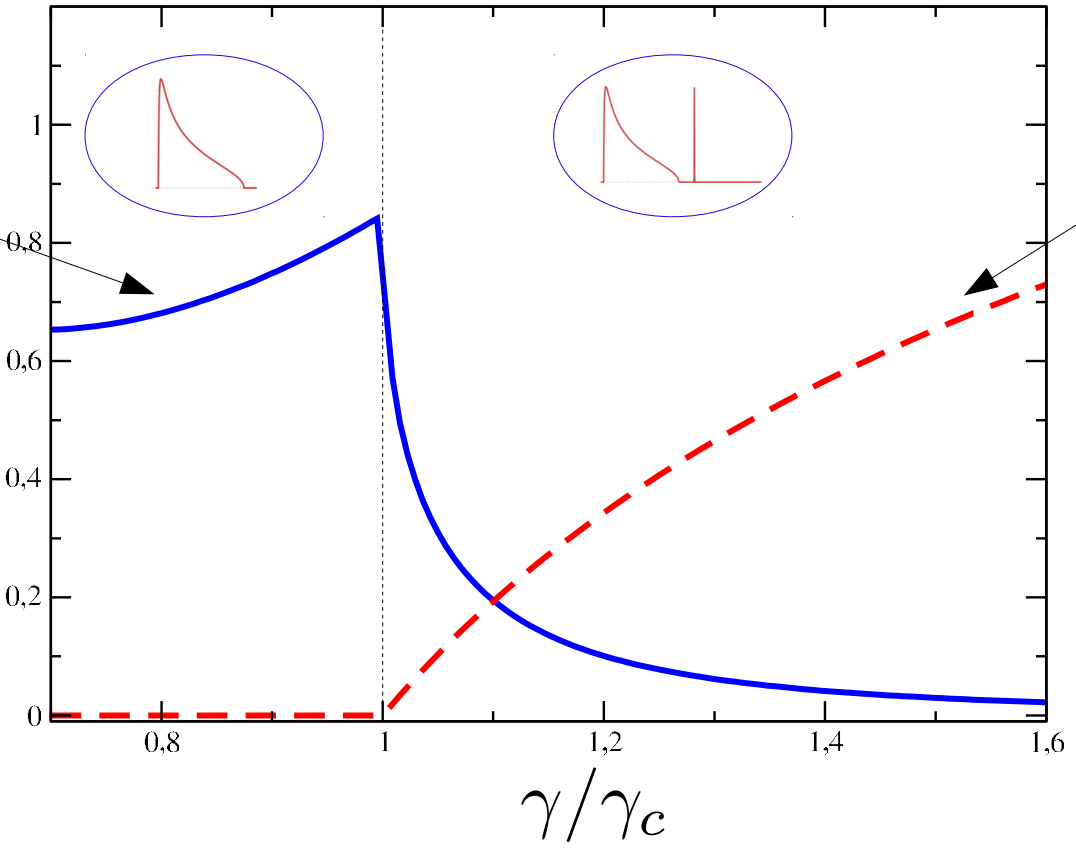
The mutual information

$$I(x;y) = \left\langle \log\left(\frac{P(x,y)}{P(x)P(y)}\right)\right\rangle \quad \Longrightarrow \quad I\left(\boldsymbol{\xi}_1;\hat{\boldsymbol{\xi}}_1\right) = -\frac{1}{2}\mathrm{Log}\left(1-\left\langle\left(\hat{\boldsymbol{\xi}}_1\cdot\boldsymbol{\xi}_1\right)^2\right\rangle\right)$$

It measures how much knowing X reduces uncertainty about Y

$I\left(\boldsymbol{\xi}_1;\{\hat{\boldsymbol{\xi}}_{m\neq 1}\}\right)$

Information related to bulk eigenvectors



$I\left(\boldsymbol{\xi}_1;\hat{\boldsymbol{\xi}}_1\right)$

Information related to the first eigenvector

Three questions:

Is it possible to calculate analytically the scalar products ?

YES !!

Is there information stored in bulk empirical eigenvectors ?

YES !!

Is it possible to exploit this information
to improve the noise undressing procedure ?

An application: reconstruction of localized eigenvectors

discrete approximation : $w_m \to \sqrt{\langle w_m^2 \rangle}\, \sigma_m \equiv \overline{w}_m \sigma_m$

(± 1)

$$\boldsymbol{\xi}_1^{(est)}(\boldsymbol{\sigma}) = \overline{w}_1 \hat{\boldsymbol{\xi}}_1 + \sum_m^N \overline{w}_m \hat{\boldsymbol{\xi}}_m \sigma_m$$

An application: reconstruction of localized eigenvectors

discrete approximation : $w_m \to \sqrt{\langle w_m^2 \rangle}\, \sigma_m \equiv \overline{w}_m \sigma_m$

(± 1)

$$\xi_1^{(est)}(\boldsymbol{\sigma}) = \overline{w}_1 \hat{\boldsymbol{\xi}}_1 + \sum_m^N \overline{w}_m \hat{\boldsymbol{\xi}}_m \sigma_m$$

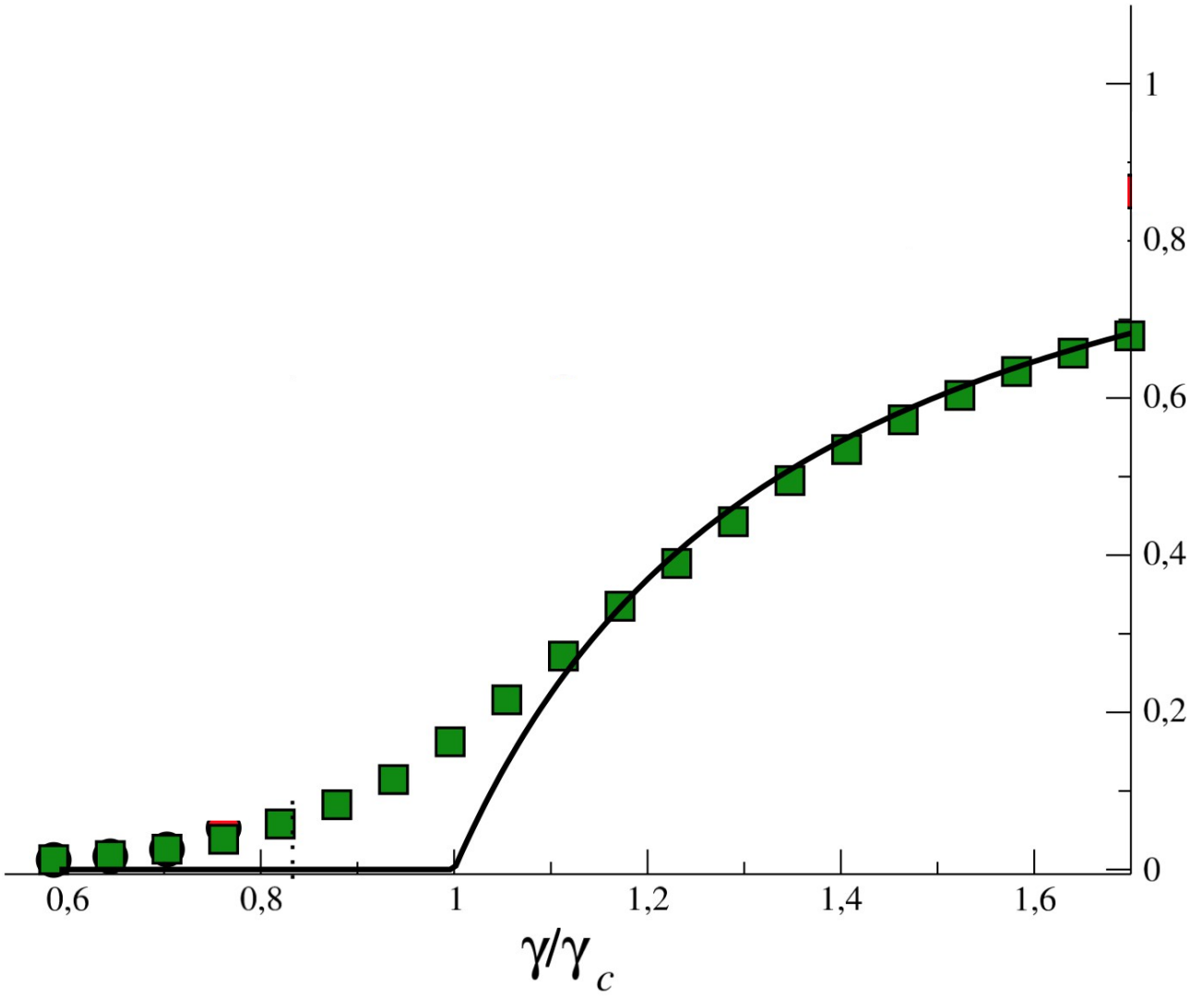
The estimation problem is mapped onto an Ising model

$$E(\boldsymbol{\sigma}) = -\sum_i \left(\xi_{1,i}^{(est)}(\boldsymbol{\sigma}) \right)^4$$

Estimate: finding $\overline{\boldsymbol{\sigma}}$ for which $E(\overline{\boldsymbol{\sigma}})$ attains its minimal value

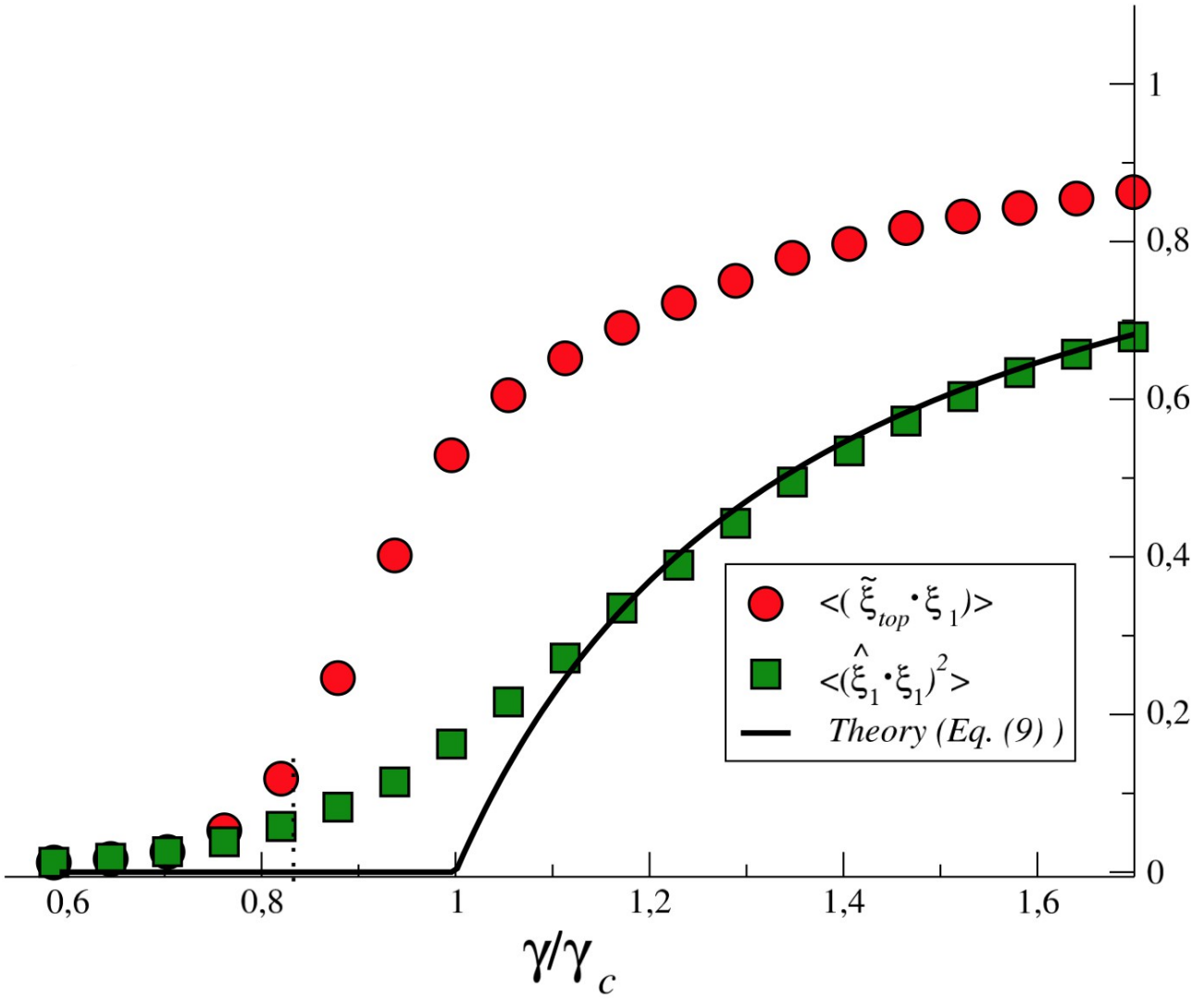
An application: reconstruction of localized eigenvectors

Montecarlo results (simulated annealing)



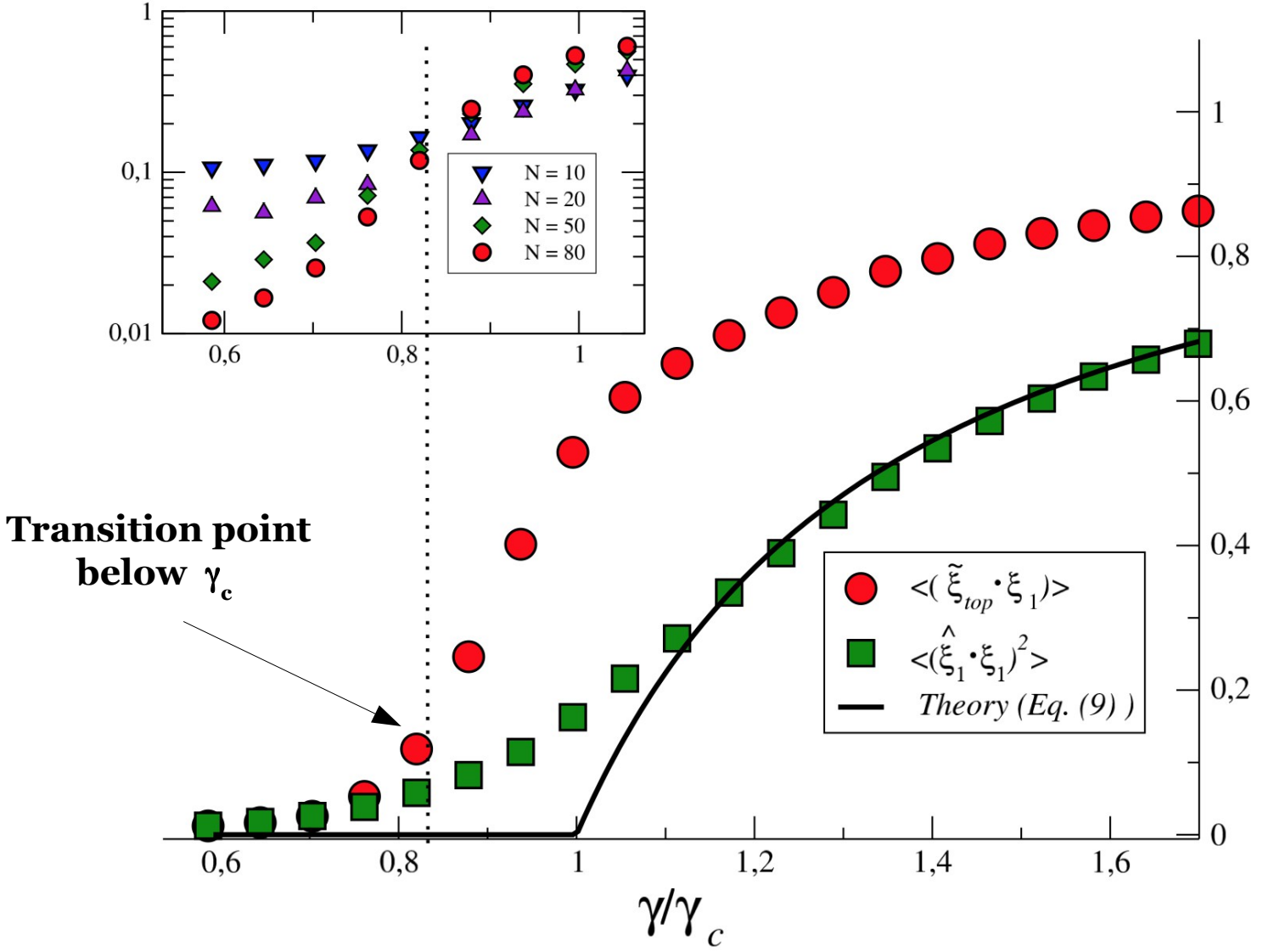
An application: reconstruction of localized eigenvectors

Montecarlo results (simulated annealing)



An application: reconstruction of localized eigenvectors

Montecarlo results (simulated annealing)



Conclusions and Perspectives

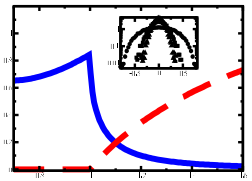
R. Monasson and D. Villamaina (arXiv:1503.00287)

General formalism for the calculation of the scalar products between the empirical eigenvectors and the true ones of a correlation matrix

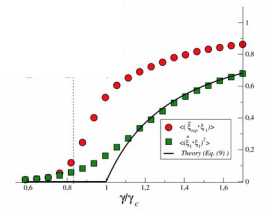
$$\mathrm{Tr}\left[(\lambda - C)^{-1} \cdot \langle(\hat{\lambda} - \hat{C})^{-1}\rangle\right]$$

Large deviations ?

Importance of minor components for the reconstruction of eigenvectors



Analogy between spin models and estimations of localized eigenvectors



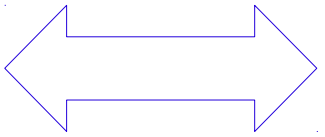
Smarter algorithms ?

Perspectives: analogy with the Hopfield model

Hopfield

$$m^\mu = \sum_j \xi_j^\mu \sigma_j$$

$$H_{opf}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_\mu (m^\mu)^2$$



Eigenvector estimation

$$m^i = \sum_m \overline{w}_m \hat{\xi}_{m,i} \sigma_m$$

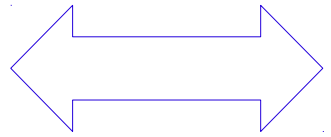
$$H(\boldsymbol{\sigma}) = -\sum_i (m^i)^4$$

Perspectives: analogy with the Hopfield model

Hopfield

$$m^\mu = \sum_j \xi_j^\mu \sigma_j$$

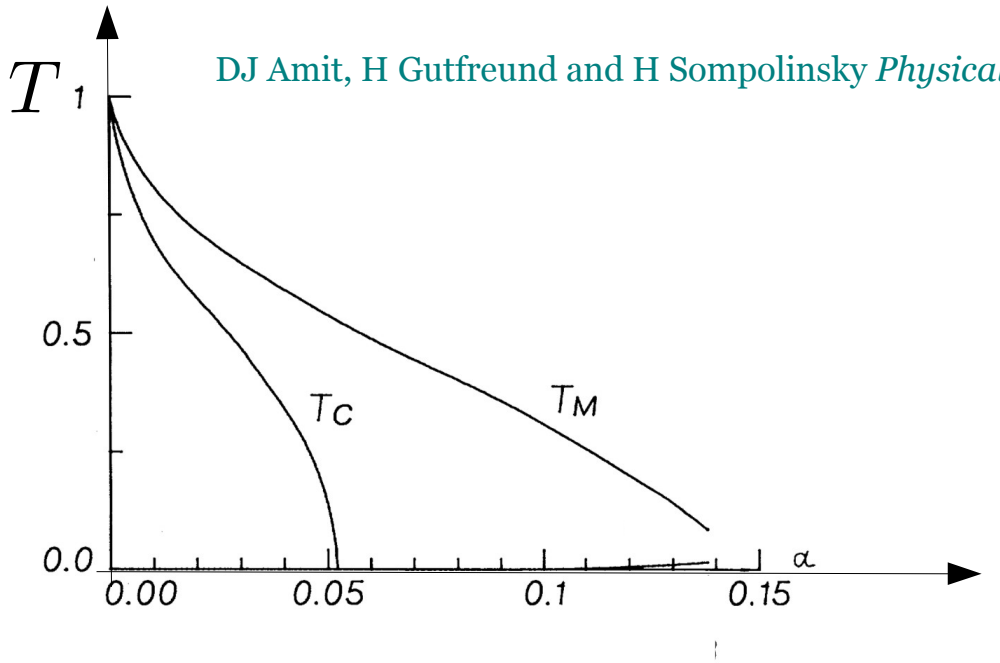
$$H_{opf}(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_\mu (m^\mu)^2$$



Eigenvector estimation

$$m^i = \sum_m \overline{w}_m \hat{\xi}_{m,i} \sigma_m$$

$$H(\boldsymbol{\sigma}) = -\sum_i (m^i)^4$$



DJ Amit, H Gutfreund and H Sompolinsky *Physical Review Letters* 55, 1530 (1985)

$$\alpha = \frac{\text{\# of patterns}}{\text{\# of spins}}$$

Perspectives: analogy with the Hopfield model

Hopfield

$$m^{\mu} = \sum_j \xi_j^{\mu} \sigma_j$$

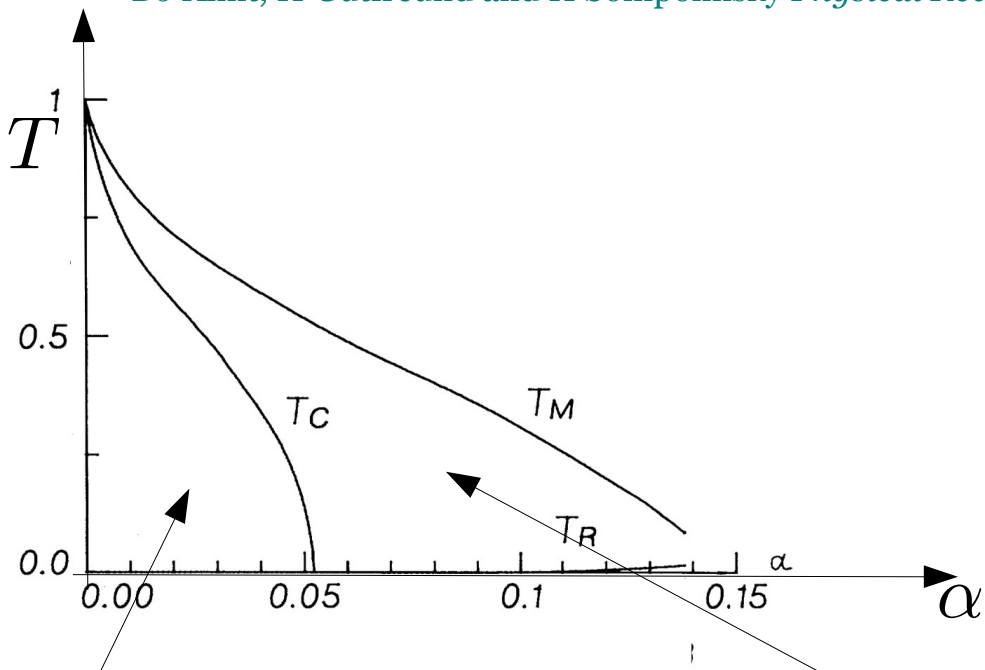
$$H_{opf}(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{\mu} (m^{\mu})^2$$

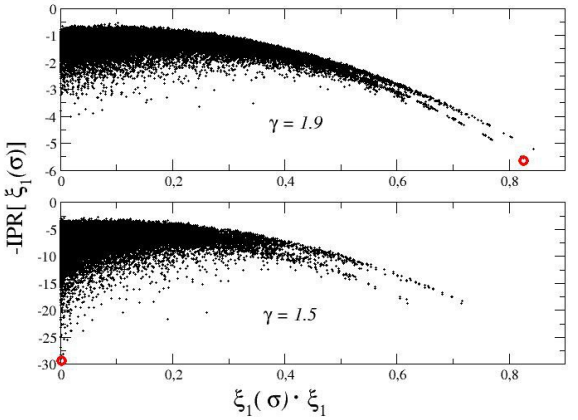
Eigenvector estimation

$$m^{i} = \sum_m \overline{w}_m \hat{\xi}_{m,i} \sigma_m$$

$$H(\boldsymbol{\sigma}) = -\sum_i (m^i)^4$$

DJ Amit, H Gutfreund and H Sompolinsky *Physical Review Letters* 55, 1530 (1985)





$$\alpha = \frac{\text{\# of patterns}}{\text{\# of spins}}$$

High γ

(small quenched disorder)

Small γ

(high quenched disorder)

References related to this talk

The transition in eigenvalue spectrum

SF Edwards and RC Jones, Journal of Physics A: Mathematical and General 9, 1595 (1976)

J Baik, G Ben Arous and S Péché, Annals of Probability, 1643 (2005)

Data analysis in finance and bioinformatics

L Laloux, P Cizeau, JP Bouchaud, M Potters, Physical Review Letters 83, 1467 (1999)
S. Cocco, R. Monasson, M. Weigt, PLoS computational biology 9, e1003176 (2013)

Eigenvalues and eigenvectors in empirical correlation matrices

O. Ledoit and S. Péché, Probability Theory and Related Fields 151, 233 (2011)

JW Silverstein, Journal of Multivariate Analysis 55, 331 (1995)

R. Monasson and D. Villamaina
Estimating the principal component of correlation matrices from all their eigenmodes
arXiv:1503.00287