A Letters Journal Exploring the Frontiers of Physics

December 2015

EPL, **112** (2015) 50001
doi: 10.1209/0295-5075/112/50001

www.epljournal.org

Estimating the principal components of correlation matrices from all their empirical eigenvectors

Rémi Monasson[1] and Dario Villamaina[1,2]

[1] *Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, associé au CNRS et à l'Université Pierre et Marie Curie - 24 rue Lhomond, 75005 Paris, France*
[2] *Institut de Physique Théorique Philippe Meyer - 24 rue Lhomond, 75005 Paris, France*

received 8 September 2015; accepted in final form 24 November 2015
published online 17 December 2015

PACS 02.50.Sk – Multivariate analysis
PACS 02.50.Tt – Inference methods
PACS 05.10.Ln – Monte Carlo methods

Abstract – We consider the problem of estimating the principal components of a population covariance matrix from a limited number of measurement data. Using a combination of random matrix and information-theoretic tools, we show that all the eigenmodes of the sample correlation matrices are informative, and not only the top ones. We show how this information can be exploited when *prior* information about the principal component, such as whether it is localized or not, is available by mapping the estimation problem onto the search for the ground state of a spin-glass–like effective Hamiltonian encoding the prior. Results are illustrated numerically on the spiked covariance model.

Introduction. – The availability of large-scale measurements of complex systems, such as in biology, finance, sociology,... calls for new methods to extract information from those data. Of crucial importance is the characterization of the correlation structure of the data, which reflects the underlying interaction network between the system components. A widely used technique is the principal component analysis (PCA), which retains only the components corresponding to the largest eigenvalues of the empirical correlation matrix computed from the data, considered as the most informative ones. PCA applications range from computer vision [1] to finance [2], to neuroscience [3] and many others. PCA can, however, be inefficient in some cases [4], in particular when the number T of available data is comparable to the number N of system components, a situation referred to as *high-dimensional* data analysis [5].

In this letter we focus on one aspect of this question, namely, how to estimate the main eigenmode(s) of the "true" system correlation matrix at large N/T ratio. We show that considering only the main components of the empirical correlation matrix, as PCA does, is generally not optimal, and that taking into account the eigenmodes associated to the low eigenvalues can greatly improve the quality of the predictions.

To fix notations let us consider a collection of N Gaussian random variables x_i ($i = 1,\dots,N$), with zero means and (population) covariances C_{ij}. Assume we have observed T-independent realizations of those variables, which define the ($N \times T$)-dimensional rectangular matrix X, *e.g.* $X_{3,2}$ corresponds to the second observation of variable x_3. The empirical covariance matrix, $\hat{C} \equiv \frac{1}{T} X \cdot X^\dagger$, also called sample covariance matrix, is an estimator of the population covariance matrix C. Let us call as $\hat{\boldsymbol{\xi}}_m$, $m = 1, 2, \dots, N$, the normalized eigenvectors of $\hat{C}$, corresponding to eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_m$. Similarly we call as $\boldsymbol{\xi}_m$ and λ_m the normalized eigenvectors and eigenvalues (ranked in decreasing order) of C. Suppose now that N is kept fixed and the number of observations T is increased. A perfect sampling condition corresponds to the limit of infinite measurements ($T \to \infty$), where the matrix estimator $\hat{C}$ approaches to the true covariance C. This scenario changes if the number of variables N increases at the same pace as the number of measures T, with a fixed ratio $r = N/T$, hereafter called sampling noise. In that case, the estimator $\hat{C}$ differs from the true covariance matrix since it is affected by finite sampling effects, a common situation in experiments. Perfect sampling is recovered in the limit case $r \to 0$.

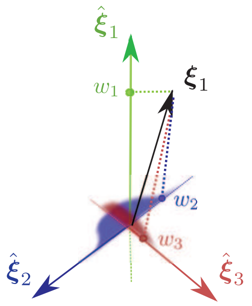
$\hat{\boldsymbol{\xi}}_1$
w_1
$\boldsymbol{\xi}_1$
w_2
w_3
$\hat{\boldsymbol{\xi}}_2$
$\hat{\boldsymbol{\xi}}_3$

Fig. 1: (Colour online) Eigenvectors $\hat{\boldsymbol{\xi}}_m$ of the sample correlation matrix and "true" component $\boldsymbol{\xi}_1$ in dimension $N = 3$. If the statistics is sufficient, *i.e.* r is low enough, the overlap $w_1 = \boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_1$ is finite in the large-N limit, while the overlaps $w_m = \boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_m$ (with $m \geq 2$) vanish as $O(1/\sqrt{N})$.

While the typical [6] properties of the distribution of the eigenvalues of $\hat{C}$ are well characterized (and theoretical results for rare events are available in the case of purely uncorrelated variables [7,8]), much less is known about its eigenvectors, see [9] or [10] and references therein. We want to estimate the top component $\boldsymbol{\xi}_1$. A simple estimator is provided by $\hat{\boldsymbol{\xi}}_1$, which is naturally expected to be exact when $r \to 0$ (in the absence of eigenvalue degeneracy), see fig. 1. For finite r, however, $\hat{\boldsymbol{\xi}}_1$ is generally not a perfect estimator, and we show below how the knowledge of the other empirical eigenvectors $\hat{\boldsymbol{\xi}}_m$, with $m \geq 2$, may considerably help to improve the estimate of $\boldsymbol{\xi}_1$.

Let us consider the scalar products $w_m \equiv \boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_m$ (fig. 1). Those overlaps are stochastic variables with zero mean, and variances $\langle w_m^2 \rangle$, where $\langle \cdot \rangle$ denotes the average over X. Each eigenvector $\hat{\boldsymbol{\xi}}_m$ taken individually is very weakly informative about $\boldsymbol{\xi}_1$ as the overlap vanishes as $N^{-1/2}$. On the contrary we show below that the mutual information between an extensive (of the order of N) number of eigenvectors $\hat{\boldsymbol{\xi}}_m (m \geq 2)$ and $\boldsymbol{\xi}_1$ remains finite when $N \to \infty$. We then present one application where the knowledge of the overlaps w_m helps us to improve our prediction of the top component $\boldsymbol{\xi}_1$ in the presence of prior information about this vector (here, it has "large" components).

The spiked covariance model. – We will illustrate our approach on the spiked covariance model, a popular model in random matrix theory, in which all the eigenvalues of C, but one, say, $\lambda_1 \equiv \gamma$, are equal to unity. Eigenvalue γ represents the "signal", with its associated eigenvector $\boldsymbol{\xi}_1$. We consider the case $\gamma > 1$ below, but similar results are found for $\gamma < 1$. Let $\hat{\rho}(\hat{\lambda}) = \frac{1}{N} \sum_m \langle \delta(\hat{\lambda} - \hat{\lambda}_m) \rangle$ be the average density of eigenvalues of $\hat{C}$ and $\rho(\lambda)$ the density of eigenvalues of C.

For "weak" signals, $\gamma < \gamma_c(r) \equiv 1 + \sqrt{r}$, $\hat{\rho}$ coincides with the spectrum of the covariance matrix of N independent variables, the so-called Marchenko-Pastur (MP) distribution [11], defined as

$$\hat{\rho}_{MP}(\hat{\lambda}) = \frac{\sqrt{(\hat{\lambda}_+ - \hat{\lambda})(\hat{\lambda} - \hat{\lambda}_-)}}{2\pi \hat{\lambda} r}, \tag{1}$$

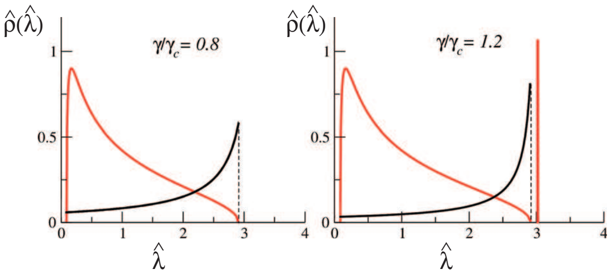

Fig. 2: (Colour online) The spiked covariance model below (left, $\gamma < \gamma_c(r)$) and above (right, $\gamma > \gamma_c(r)$). The eigenvalue spectrum given by the Marchenko-Pastur distribution in eq. (1) is shown in red; for $\gamma > \gamma_c(r)$, the signal eigenvalue $\hat{\gamma}(\gamma)$, see eq. (2), is represented by a vertical red line. The squared overlap function $W^2(\gamma, \hat{\lambda})$ in eq. (5) is shown in black over the interval $[\hat{\lambda}_-, \hat{\lambda}_+]$; the vertical dashed line locates the edge $\hat{\lambda}_+$. Note that W^2 is rescaled by a factor 0.1 to fit in the figure.

where $\hat{\lambda}_\pm(r) = (1 \pm \sqrt{r})^2$ are the edges of the distribution[1]. For "strong" signal, $\gamma > \gamma_c(r)$, the spectrum $\hat{\rho}$ is equal to the MP spectrum, and includes one eigenvalue, isolated from the MP bulk and centered in

$$\hat{\gamma}(\gamma) = \gamma + r \, \frac{\gamma}{\gamma - 1}. \tag{2}$$

The onset of a signal-related eigenvalue $\hat{\gamma}$ at the critical value of $\gamma = \gamma_c(r)$ was first reported in [12] and mathematically proven in [13]. Similar "retarded learning" transitions are encountered in models of neural networks [14] and in the Gaussian matrix ensemble [15]. The transition is pictorially represented in fig. 2.

We define $W^2(\lambda, \hat{\lambda})$ as the mean squared overlap, multiplied by N, between the eigenvectors of C and $\hat{C}$ associated to, respectively, the eigenvalues λ and $\hat{\lambda}$, namely

$$W^2(\lambda, \hat{\lambda}) = \sum_{\ell, m} \frac{\Big\langle (\boldsymbol{\xi}_\ell \cdot \hat{\boldsymbol{\xi}}_m)^2 \, \delta(\lambda - \lambda_\ell) \delta(\hat{\lambda} - \hat{\lambda}_m) \Big\rangle}{N \, \rho(\lambda) \, \hat{\rho}(\hat{\lambda})}. \tag{3}$$

The mean squared scalar products with the top component introduced above are given by

$$\langle w_m^2 \rangle = \frac{1}{N} W^2(\gamma, \hat{\lambda}_m). \tag{4}$$

We will therefore fix $\lambda = \gamma$ in the following.

W^2 can be computed using statistical physics approaches to random matrix theory [16,17]. For $\hat{\lambda} \in [\hat{\lambda}_-(r); \hat{\lambda}_+(r)]$ spanning the MP bulk spectrum one has [18]

$$W^2(\gamma, \hat{\lambda}) = \frac{\hat{\gamma}(\gamma) - \gamma}{\hat{\lambda} - \hat{\gamma}(\gamma)}. \tag{5}$$

W^2 in eq. (5) is an increasing function of $\hat{\lambda}$, which diverges for $\hat{\lambda} = \hat{\gamma}(\gamma)$. Note that this divergence is always located outside the MP spectrum (as in both panels of fig. 2), and

[1] We consider here the case $r < 1$. For $r > 1$ a δ-peak in $\hat{\lambda} = 0$ of mass $1 - \frac{1}{r}$ is present.

coincides with the MP edge $\hat{\lambda}_+(r)$ for the critical signal eigenvalue $\gamma = \gamma_c(r)$ only.

It is easy to obtain the expression for the overlap between the top eigenvectors of C and $\hat{C}$ by exploiting the standard relation for orthonormal bases, $\sum_m^N w_m^2 = 1$, or its continuous version in the infinite-N limit:

$$\int_{\hat{\lambda}_-}^{\hat{\lambda}_+} W^2(\gamma,\hat{\lambda})\hat{\rho}_{MP}(\hat{\lambda})\mathrm{d}\hat{\lambda} + \langle w_1^2\rangle = 1. \tag{6}$$

We deduce from eqs. (1), (5), (6) that the mean squared overlap between the top components of the population and sample covariance matrices is given by

$$\langle w_1^2\rangle = \begin{cases} 0, & \gamma \le \gamma_c(r), \\ \dfrac{\gamma^2 - \hat{\gamma}(\gamma)}{\hat{\gamma}(\gamma)\,(\gamma-1)}, & \gamma > \gamma_c(r), \end{cases} \tag{7}$$

and is non-zero in the strong-signal regime only. This result agrees with previous applications, *e.g.*, to signal detection [19], or to the inference of relevant modes in inverse problems [20].

Mutual information between the empirical eigenvectors $\hat{\boldsymbol{\xi}}_m$ and the top component $\boldsymbol{\xi}_1$. – As a consequence, close to the transition point (both from above and from below) the mean squared overlap $W^2(\gamma,\hat{\lambda})$ has a strong asymmetric shape (see black curves in fig. 2) showing how the sample eigenvectors close to the right edge are highly correlated to the top component. This correlation is informative and can be exploited to infer the principal component with appropriate algorithms, as we show in the next section.

To quantify this information about the component $\boldsymbol{\xi}_1$ contained in the eigenvectors $\hat{\boldsymbol{\xi}}_m$ we introduce the mutual information I between those variables. I is defined as the difference between the entropy of variable $\boldsymbol{\xi}_1$ and the entropy of $\boldsymbol{\xi}$ conditioned to $\{\hat{\boldsymbol{\xi}}_m\}$; it measures how much the knowledge of $\{\hat{\boldsymbol{\xi}}_m\}$ reduces the uncertainty on the estimate of $\boldsymbol{\xi}$. I is non-negative and vanishes if and only if $\boldsymbol{\xi}$ and $\{\hat{\boldsymbol{\xi}}_m\}$ are independent [21].

Our goal is to understand whether the knowledge of many (of the order of N) very small (of the order of $1/\sqrt{N}$) overlaps with the empirical eigenvectors is helpful to determine the top component, *i.e.* if I has a finite limit in the $N \to \infty$ limit. As the exact computation of I is hard, we assume that the correlations between the overlaps are negligible for large N. Within this assumption, we calculate the mutual information I between $\boldsymbol{\xi}_1$ and $\{\hat{\boldsymbol{\xi}}_m; 2 \le m \le f\,N\}$, with $f \le 1$ is the fraction of retained empirical eigenvectors. The details of the calculation, based on the use of the replica approach, are given in the appendix. Within the replica symmetric framework, we obtain

$$\frac{1}{N} I\big(\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\}\big) = -\frac{1}{2}\min_{q,\hat{q}}\Big[\big(q - \Omega(f)\big)\,\hat{q} + \log(1-q) + \int_{\Lambda(f)}^{\hat{\lambda}_+} \mathrm{d}\hat{\lambda}\ \hat{\rho}(\hat{\lambda})\ \log\big(1 + \hat{q}\,W^2(\gamma,\hat{\lambda})\big)\Big], \tag{8}$$

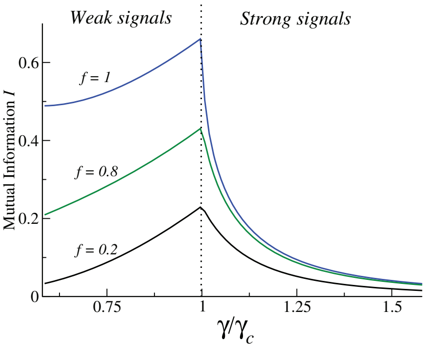

Fig. 3: (Colour online) Mutual information $I(\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\})$, eq. (8) for the spiked covariance model with $r = 0.5$, divided by the number N of variables. Qualitatively similar curves are obtained when r is varied. The vertical dotted line corresponds to the transition point, below which the top empirical eigenvector is completely uncorrelated with the true (population) one, as entailed by eq. (7). Remarkably, the mutual information, I in eq. (8), between an extensive number of empirical eigenvectors corresponding to lower eigenvalues and the true top components is finite positive even in this low sampling regime.

where $\Lambda(f)$ is such that $f = \int_{\Lambda(f)}^{\hat{\lambda}_+(r)} \mathrm{d}\hat{\lambda}\,\hat{\rho}(\hat{\lambda})$ and $\Omega(f) = \int_{\Lambda(f)}^{\hat{\lambda}_+(r)} \mathrm{d}\hat{\lambda}\,\hat{\rho}(\hat{\lambda}) W^2(\gamma,\hat{\lambda})$. The mutual information is plotted in fig. 3 as a function of γ, and for various values of f. It is strictly positive for all values of γ. $I(\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\})$ reaches a maximum at the transition point $\gamma_c(r)$, separating the weak- and strong-signal regimes. Moreover, it is increasing with the fraction f. Our calculation therefore provides clear evidence for the fact that sample eigenvectors in the bulk of the MP spectrum are informative about the principal component of the population covariance matrix. This result is remarkable especially in the case of weak signals, where the top empirical eigenvector is not informative at all.

Inference of the principal component with prior knowledge. – Based on the study of the overlaps w_m above we may express the principal component $\boldsymbol{\xi}_1$ as a weighted sum of the sample eigenvectors,

$$\boldsymbol{\xi}_1 = \sqrt{\langle w_1^2\rangle}\ \hat{\boldsymbol{\xi}}_1 + \sum_{m=2}^{N} \sigma_m\ \sqrt{\langle w_m^2\rangle}\ \hat{\boldsymbol{\xi}}_m, \tag{9}$$

where the σ_m's are independent Gaussian variables of zero means, and unit variances ($m \ge 2$). Equation (9) implicitly defines the likelihood of the component $\boldsymbol{\xi}_1$ given the sample eigenvectors. The rationale underlying eq. (9) is that it gives back the right statistics for the scalar products w_m. In particular, the expected value (over the σ_m's) of $\boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_m$ vanishes, while the average value of $(\boldsymbol{\xi}_1 \cdot \hat{\boldsymbol{\xi}}_m)^2$ coincides with $\langle w_m^2\rangle$.

In the absence of any prior information the average value of $\boldsymbol{\xi}_1$ is simply equal to $\sqrt{\langle w_1^2\rangle}\,\hat{\boldsymbol{\xi}}_1$, corresponding to the standard estimate widely used in the literature. Actually this estimate discards the information contained

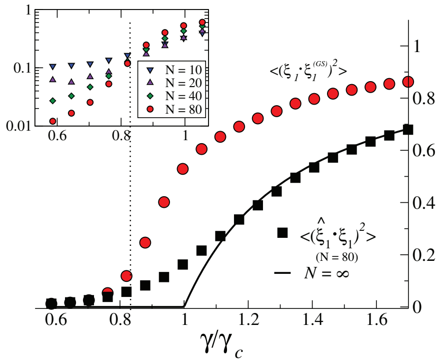

Fig. 4: (Colour online) Overlap of $\boldsymbol{\xi}_1 \equiv (1,0,\ldots,0)$ with the ground state $\boldsymbol{\xi}_1^{(GS)}$ of $-$IPR (red circles) and with the sample top component $\hat{\boldsymbol{\xi}}_1$ (black squares) as a function of γ for the spiked covariance model with $r = 0.5$ ($\gamma_c \simeq 1.7071$) and $N = 80$ variables. Inset: zoom of the region slightly below γ_c; overlap of $\boldsymbol{\xi}_1$ with the ground state of $-$IPR for different sizes N. All lines seem to intersect around $\gamma/\gamma_c \simeq 0.83$ in the poor sampling phase.

in the sample eigenvectors $\hat{\boldsymbol{\xi}}_m$, and vanishes in the weak-signal regime. Our purpose here is to improve over this simple estimate, by exploiting some prior information over the top component.

In many practical applications, indeed, prior knowledge over the principal components is available, such as the entries of those components are positive, sparse, bounded from above, etc.. . . . A physically sound prior knowledge we consider hereafter is the localization of principal components, found to be important for the identification of a site in contacts on the three-dimensional structure of proteins [22], or in the study of phonons in liquid crystals [23,24]. Drawing our inspiration from condensed-matter physics we consider the inverse participation ratio

$$\mathrm{IPR}(\boldsymbol{\xi}_1) = \sum_{i=1}^{N} [(\boldsymbol{\xi}_1)_i]^4, \tag{10}$$

and look for estimates of the principal component with large IPR. This prior favors vectors with large entries, *i.e.* non-vanishing in the large-N limit. More precisely the objective function to be maximized is the log-posterior distribution of $\boldsymbol{\xi}_1$, which sums the IPR in eq. (10) and the log-likelihood implicitly defined by eq. (9). To simplify this computational problem we consider a discrete version of eq. (9), where the σ_m's are constrained to take ± 1 values. As a result we have at our disposal a pool of 2^{N-1} candidate components for $\boldsymbol{\xi}_1$ with *equal* log-likelihoods. We can then simply look for the binary configuration $\{\sigma_m; m \geq 2\}$ in this pool, which maximizes the IPR.

To find the ground state of $-$IPR, denoted by $\boldsymbol{\xi}_1^{(GS)}$, we resort to simulated annealing with a Monte Carlo scheme[2].

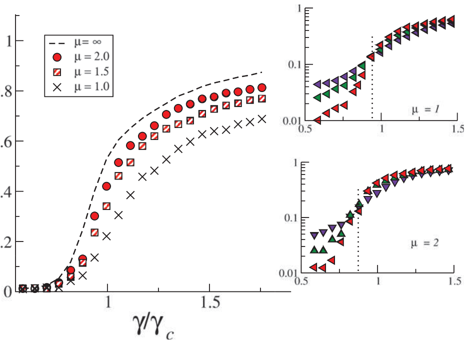

Fig. 5: (Colour online) Same simulation as in fig. 4 with a different top eigenvector $(\boldsymbol{\xi}_1)_i \propto \frac{1}{i^\mu}$, where the reconstruction is shown for different values of μ. Remarkably the second transition is present also in this case (with a threshold depending on μ), as shown in the insets for different values of N ($N = 20, 40, 80$).

The overlap between $\boldsymbol{\xi}_1^{(GS)}$ and $\boldsymbol{\xi}_1$ is shown for different sizes N in fig. 4. We find a better (higher) value than the one corresponding to the naive estimate based on $\hat{\boldsymbol{\xi}}_1$. The improvement is maximal for γ close to $\gamma_c(r)$, that is, in the critical region separating weak from strong signals. Remarkably, while the naive estimate breaks down for $\gamma < \gamma_c(r)$ regime in the large-N limit this does not seem to be the case for our procedure. This result is in agreement with the prediction given by the mutual information (eq. (8)) in fig. 3, which is positive even for low signals and reaches its maximum in correspondence of the transition. Therefore, the contribution of lower eigenvectors in the estimation is prominent in this region, as expected.

This scenario does not qualitatively depend on the choice of the eigenvector $\boldsymbol{\xi}_1$. In particular we have studied both the cases of a very sparse eigenvector (see fig. 4, where $\boldsymbol{\xi}_1 = (1,0,\ldots,0)$) and of a slow, power-law decay, $(\boldsymbol{\xi}_1)_i \propto \frac{1}{i^\mu}$ with $\mu \geq \frac{1}{2}$. As shown in fig. 5, our procedure results in a better prediction of the top eigenvector for all the values of μ we have tested. We stress again that the estimator in eq. (9) exploits the information contained in the lower eigenmodes. A random search around $\hat{\xi}_1$, for instance by considering an estimator such as $a\,\hat{\xi}_1 + \sqrt{1-a^2}\,\eta$, where $a = \sqrt{\langle w_1^2 \rangle}$ and η is a random vector orthogonal to $\hat{\xi}_1$, would not produce comparable results, especially in the weak-signal region.

Transition in prior knowledge-based inference. – As reported above, the insets of fig. 4 and fig. 5 suggest the existence of a transition point, well inside the weak-signal regime, above which $\boldsymbol{\xi}_1$ may be approximately inferred with the help of prior knowledge, even for large system sizes. This transition bears a strong analogy with transitions taking place in the Hopfield model *below the critical capacity*. Indeed, the IPR in eq. (10), once expressed in terms of the σ_m's, may be interpreted as minus the Hamiltonian of an effective spin system, with a mixture

[2] The inverse temperature β is slowly incremented by steps of $\delta\beta = 5$, from $\beta = 50$ to $\beta = 200$. For each temperature we attempt $20\,N$ Monte Carlo spin flips. At the end of this procedure, the configuration with lowest energy is retained. The average is taken over $\sim 10^3$ realizations of the measurement matrix X.

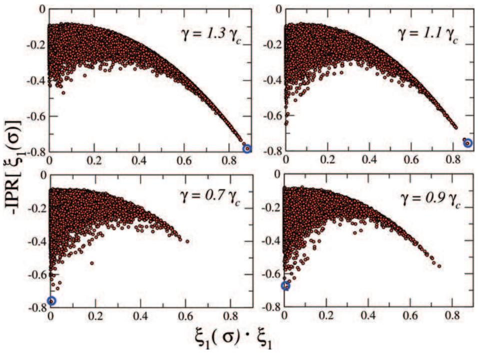

Fig. 6: (Colour online) Typical distributions of energies of the states $\boldsymbol{\sigma}$ *vs.* the overlaps with the true eigenvector, above (top panels) and below (bottom panels) the transition value. The ground state is shown by the blue circle. Same model as in fig. 4, with $N = 15$ and $r = 0.5$. For large N we expect the transition to take place at $\gamma/\gamma_c \simeq 0.83$.

of k-body interactions, where k ranges from 1 to 4. The interactions

$$J_{m_1,m_2,m_3,m_4} = \sum_{i=1}^{N} \prod_{\ell=1}^{4} \sqrt{\langle w_{m_\ell}^2 \rangle}\, \hat{\xi}_{i,m_\ell} \qquad (11)$$

are non-linear combinations of the eigenmodes components, and may have positive or negative signs. This spin-glass Hamiltonian is strongly reminiscent of the Hopfield model [25]. Each entry $(\boldsymbol{\xi}_1)_i$, $i = 1, \ldots, N$, of the principal component may be interpreted as the "magnetization" M_i of the spin configuration $\boldsymbol{\sigma}$ along the "pattern" i, whose m-th entry is $\sqrt{\langle w_m^2 \rangle}(\hat{\boldsymbol{\xi}}_m)_i$, or, equivalently,

$$M_i(\boldsymbol{\sigma}) = \sum_{m=1}^{N} \sigma_m \sqrt{\langle w_m^2 \rangle}(\hat{\boldsymbol{\xi}}_m)_i \qquad (12)$$

with $\sigma_1 \equiv 1$. Note that our Hamiltonian,

$$-\mathrm{IPR}\big(\boldsymbol{\xi}_1(\boldsymbol{\sigma})\big) = \sum_{i=1}^{N} M_i(\boldsymbol{\sigma})^4, \qquad (13)$$

is however quartic, and not quadratic in the magnetizations. Thanks to this analogy, the transition observed here can be interpreted in terms of the phase diagram of the Hopfield model [26]: At low temperatures and intermediate loads (comprised between $\simeq 0.05$ and the critical capacity $\simeq 0.14$) the patterns to be stored are local minima of the Hopfield Hamiltonian, and are uncorrelated with the ground state. This scenario holds in our case, too. We show in fig. 6 the distribution of the energies *vs.* the overlap, obtained through exhaustive searches of the configuration space for small sizes $N < 25$. Below the transition point, the ground-state vector $\boldsymbol{\xi}_1^{(GS)}$ is clearly not aligned along $\boldsymbol{\xi}_1$.

Conclusions. – As described by random matrix theory, the overlap between lower empirical eigenvectors and the top true one is very small, of the order of $N^{-1/2}$, and vanishes in the infinite-N limit. In spite of this, in this paper we have shown how an extensive number of sample eigenvectors with low eigenvalues is strongly informative about the population principal components, by presenting a calculation of mutual information for the spike covariance model.

Based on this result, we have introduced a general procedure to exploit that information in the presence of prior knowledge, by mapping the inference problem onto the search for the ground state of a spin-glass–like Hamiltonian encoding the prior. We have shown the efficiency of the approach when one knows *a priori* that top components have large entries, which considerably improves the standard inference and allows us to recover the component in the weak-signal region, where naive inference fails. This finding agrees with recent results on non-negative PCA [27].

It would be interesting to understand how efficient is our procedure for other priors, or in cases where the value of the overlap distribution is unknown and eigenvalue-cleaning techniques [18,28] must be used for its estimation. A limitation of our approach is the use of discrete (binary) variables σ_m in eq. (9). In a forthcoming publication we plan to study more refined algorithms by considering continuous variables instead, in order to test the generality of the (second) transition found in the weak-signal regime.

We stress that our approach could also be extended to infer more than one principal components. While the case of a finite number of separated eigenvalues (multiple-spiked covariance model) is straightforward, it would be interesting to consider $O(N)$-dimensional degenerate subspaces, as in [29].

* * *

We are very grateful to S. Cocco and G. Robichon for their participation in the initial stage of this work. We thank M. P. Brenner, M. Huntley, P. Vivo for discussions, and P. Vivo for a careful reading of the manuscript, too. This work was partly funded by the Agence Nationale de la Recherche Coevstat project (ANR-13-BS04-0012-01) and the CNRS-InphyNiTi project INFERNEURO.

Appendix: replica calculation of the mutual information. – Here, we present the derivation of the mutual information given in eq. (8). We assume that the components of $\hat{\boldsymbol{\xi}}_m$ are independent and identically Gaussianly distributed, with zero means and unit variances, and that the dot products w_m are also independent normal variables with zero means and variances W_m^2/N. The index m runs from 2 to $M \equiv f\,N$, where f is the fraction of eigenvectors retained. Equation (9) thus fully defines the joint distribution of the eigenvectors, $P[\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\}]$. We define $Z(n) = \int \mathrm{d}\boldsymbol{\xi}_1 \prod_m \mathrm{d}\hat{\boldsymbol{\xi}}_m \; P[\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\}]^n$. $Z(1)$ is obviously equal to unity by normalisation of P, and the mutual information I is simply related to the derivative of Z in $n=1$, see below. Along the lines of the replica method, we will first consider that n is an integer, and will next perform an analytic continuation to real-valued n on the outcome. After integrating out the eigenvector components and within the replica symmetric hypothesis, we obtain that $F_n \equiv \frac{1}{N}\log Z(n)$ is given, up to some irrelevant additive constant, by the saddle-point value of

$$\begin{aligned} F_n = &-\frac{n(n-1)}{2} q\, s - \frac{n}{2}\,\tilde{q}\,\tilde{s} \\ &- \frac{1}{2}\log\left[n\left(1 - \sum_m W_m^2 + \frac{\tilde{q}-q}{n}\right)^{n-1}\right] \\ &- \frac{n}{2N}\sum_m \log\left(1 + W_m^2(s-\tilde{s})\right) \\ &- \frac{1}{2N}\sum_m \log\left(1 - \frac{n\, s\, W_m^2}{1+W_m^2(s-\tilde{s})}\right) \qquad \text{(A.1)} \end{aligned}$$

over its four arguments $\tilde{q}, q, \tilde{s}, s$. The parameters $\tilde{q}, q$ are equal to, respectively, $\sum_m \overline{\langle w_m^2\rangle}, \sum_m \overline{\langle w_m\rangle^2}$, where the overbar denotes the averages over the eigenvectors and the angular brackets denote the averages over the Gaussian measure over the overlaps at fixed eigenvectors; $\tilde{s}, s$ are the conjugated Lagrange parameters.

A straightforward calculation shows that the mutual information I in eq. (8) is given by $\frac{1}{N} I\big(\boldsymbol{\xi}_1, \{\hat{\boldsymbol{\xi}}_m\}\big) = -\frac{\partial F_n}{\partial n}\Big|_{n=1}$. We are therefore left with the resolution of the saddle-point equations over $\tilde{q}, q, \tilde{s}$, and s. First, let us remark that F_1 depends on $\tilde{q}$ and $\tilde{s}$ only: $F_1 = -\frac{\tilde{s}\,\tilde{q}}{2} - \frac{1}{2N}\sum_m \log\left(1 - \tilde{s}\, W_m^2\right)$. The values of the order parameters at the saddle point are thus $\tilde{q} = \frac{1}{N}\sum_m W_m^2$, $\tilde{s}=0$. We next consider $F_{n=1+\epsilon} = F_1 - \epsilon I + O(\epsilon^2)$, where

$$\begin{aligned} I = &\frac{s\,q}{2} + \frac{1}{2N}\sum_m \log\left(1 + s\, W_m^2\right) \\ &- \frac{s}{2N}\sum_m W_m^2 + \frac{1}{2}\log(1-q). \qquad \text{(A.2)} \end{aligned}$$

To the lowest order in ϵ, the values of $\tilde{q}$ and $\tilde{s}$ are unchanged with respect to the case $n=1$. The saddle-point equations for q and s give $s = \frac{1}{1-q}, q = \frac{s}{N}\sum_m \frac{W_m^4}{1+s\,W_m^2}$. The corresponding expression for I is given in eq. (8).

REFERENCES

[1] De la Torre F. and Black M. J., *Proceedings of the Eighth IEEE International Conference on the Computer Vision (ICCV)*, Vol. **1** (IEEE) 2001, p. 362.

[2] Laloux L., Cizeau P., Bouchaud J. P. and Potters M., *Phys. Rev. Lett.*, **83** (1999) 1467.

[3] Cunningham J. P. and Byron M. Y., *Nat. Neurosci.*, **17** (2014) 1500.

[4] Yeung K. Y. and Ruzzo W. L., *Bioinformatics*, **17** (2001) 763.

[5] Donoho D. L., *AMS Math Challenges Lecture* (2000).

[6] Silverstein J. W., *J. Multivar. Anal.*, **55** (1995) 331.

[7] Dean D. S. and Majumdar S. N., *Phys. Rev. Lett.*, **97** (2006) 160201.

[8] Majumdar S. N. and Vivo P., *Phys. Rev. Lett.*, **108** (2012) 200601.

[9] Hoyle D. C. and Rattray M., *Phys. Rev. E*, **75** (2007) 016101.

[10] Allez R. and Bouchaud J. P., *Phys. Rev. E*, **86** (2012) 046202.

[11] Marchenko V. A. and Pastur L. A., *Mat. Sb.*, **114** (1967) 507.

[12] Hoyle D. C. and Rattray M., *Phys. Rev. E*, **69** (2004) 026124.

[13] Baik J., Ben Arous G. and Péché S., *Ann. Probab.*, **33** (2005) 1643.

[14] Watkin T. L. and Nadal J. P., *J. Phys. A: Math. Gen.*, **27** (1994) 1899.

[15] Edwards S. F. and Jones R. C., *J. Phys. A: Math. Gen.*, **9** (1976) 1595.

[16] Baik J. and Silverstein J. W., *J. Multivar. Anal.*, **97** (2006) 1382.

[17] Paul D., *Stat. Sin.*, **17** (2007) 1617.

[18] Ledoit O. and Péché S., *Probab. Theory Relat. Fields*, **151** (2011) 233.

[19] Nadler B., *Ann. Stat.*, **36** (2008) 2791.

[20] Cocco S., Monasson R. and Sessak V., *Phys. Rev. E*, **83** (2011) 051123.

[21] Cover T. M. and Thomas J. A., *Elements of Information Theory* (John Wiley & Sons) 2012.

[22] Cocco S., Monasson R. and Weigt M., *PLoS Comput. Biol.*, **9** (2013) e1003176.

[23] Chen K., Still T., Schoenholz S., Aptowicz K. B., Schindler M., Maggs A. C., Liu A. J. and Yodh A. G., *Phys. Rev. E*, **88** (2013) 022315.

[24] Maggs A. C. and Schindler M., *EPL*, **109** (2015) 48005.

[25] Hopfield J. J., *Proc. Natl. Acad. Sci. U.S.A.*, **79** (1982) 2554.

[26] Amit D. J., Gutfreund H. and Sompolinsky H., *Phys. Rev. A*, **32** (1985) 1007.

[27] Montanari A. and Richard E., arXiv preprint, arXiv:1406.4775 (2014).

[28] Bun J., Allez R., Bouchaud J. P. and Potters M., arXiv preprint, arXiv:1502.06736 (2015).

[29] Mestre X., *IEEE Trans. Inf. Theory*, **54** (2008) 5113.