Inference of principal components of noisy correlation matrices with prior information

Rémi Monasson
Laboratory of Theoretical Physics, CNRS & Ecole Normale Supérieure, PSL Research, 24 rue Lhomond, Paris, France
Email: see http://www.physique.ens.fr/$\sim$monasson

***Abstract*—The problem of infering the top component of a noisy sample covariance matrix with prior information about the distribution of its entries is considered, in the framework of the spiked covariance model. Using the replica method of statistical physics the computation of the overlap between the top components of the sample and population covariance matrices is formulated as an explicit optimization problem for any kind of entry-wise prior information. The approach is illustrated on the case of top components including large entries, and the corresponding phase diagram is shown. The calculation predicts that the maximal sampling noise level at which the recovery of the top population component remains possible is higher than its counterpart in the spiked covariance model with no prior information.**

***Index Terms*—Random Matrix Theory, Spiked Covariance Model, Prior Information, Replica Method, Phase Transitions**

I. Introduction

In the era of big data inferring features of complex systems, characterized by many degrees of freedom, is of crucial importance. The high-dimensional setting, where the number of features to extract is not small compared to the number of available data, makes this task statistically or computationally hard. One case of practical interest is the inference of the largest component (eigenvector) of correlation matrices. Consider T independently drawn observations of N interacting Gaussian variables, *i.e.* such that the population covariance matrix C is not the identity matrix. If T is much larger than N the empirical covariance matrix $\hat{C}$ computed from the T observations converges to C, and recovering the top eigenvector is easy. The case where both N, T are large (sent to infinity at fixed ratio $r = N/T$) has received a lot of attention, both theoretically and practically [1]. From a theoretical point of view, it has been shown, in the case of a covariance matrix C with one (or few compared to N) eigenvalues larger than unity, say, γ, that recovery is possible if r is smaller than the critical value $r_c = (\gamma - 1)^2$ [2], [3]. For larger sampling noise ($r > r_c$), the top eigenvector of $\hat{C}$ is essentially orthogonal to the top component of C, and is therefore not informative. It is reasonable to expect that the situation will improve in the presence of additional, prior information about the structure of the top component to be recovered, and that recovery will be possible even when r is (not too much) larger than r_c. That this is indeed the case has been rigorously shown when all entries are nonnegative [4], and is supported by strong numerical evidence when the top component is known to have large entries (finite as $N \to \infty$) [5]. In the present work, using techniques from statistical physics we propose explicit conjectures about the critical noise level and its dependence on the signal eigenvalue (γ) and on prior knowledge. The framework is general and can be applied to any kind of entry-wise prior probability, *i.e.* factorized over the entries ξ_i of the top component $\boldsymbol{\xi}$. We show how rigorous results in the nonnegative case of [4] are recovered, and present new results for the large entry prior.

The motivation to consider the latter prior stems from computational biology, more precisely, from the study of co-evolution between amino acids in protein families. Sequences of proteins diverged from a common ancestor widely differ across many organisms, while the protein structure and function are often very well conserved. The constraints induced by structural and functional conservation manifest themselves as correlations between amino acids (the N variables, where N is the protein length) across the different organisms (the T observations). Recently, it was shown that the eigenmodes $\boldsymbol{\xi}$ of the amino-acid covariance matrix corresponding to *low* eigenvalues were informative about three-dimensional contacts on the protein structure [6]. These modes show large entries on the protein sites and amino-acid types in contact; as the other entries contain diffuse, non-structural signal [7], the components $\boldsymbol{\xi}$ cannot be thought of as being sparse. The presence of large entries in structurally-informative components was empirically assessed through the so-called inverse participation ratio, $\sum_i \xi_i^4$ (for normalized $\boldsymbol{\xi}$), a quantity that remains finite for components with (few) large entries and otherwise vanishes for $N \to \infty$. We hereafter use this quantity as a prior over the components to facilitate their recovery.

II. Probabilistic framework

A. Spiked covariance model

We consider the popular Spiked Covariance Model, in which the entries of N-dimensional vectors, $\mathbf{x} = (x_1, x_2, ..., x_N)$, are Gaussian random variables with zero means and population covariance matrix $\mathbf{C}$. All eigenvalues of $\mathbf{C}$ but one are equal to unity, while the remaining eigenvalue is $\gamma \neq 1$, with associated eigenvector $\mathbf{u}$. As usual we choose $\gamma > 1$ but our results could be transposed to the case $\gamma < 1$ with minor modifications. We draw T independent samples $\mathbf{x}^t, t = 1, 2, ..., T$, and define the sample covariance matrix $\hat{\mathbf{C}}$, with entries $\hat{C}_{ij} = \frac{1}{T}\sum_t x_i^t x_j^t$. The top eigenvector of $\hat{\mathbf{C}}$ is denoted by $\boldsymbol{\xi}$. In the double limit

$N, T \to \infty$ at fixed ratio $r = N/T$, there exists a phase transition at a critical value of the sampling noise $r_c = (\gamma-1)^2$ separating the *high-noise regime*, $r > r_c$, in which $\boldsymbol{\xi}$ and $\mathbf{u}$ have asymptotically zero squared dot product, and the *low-noise regime*, $r < r_c$, where the squared dot product between $\boldsymbol{\xi}$ and $\mathbf{u}$ is strictly positive with high probability [2], [3].

B. Likelihood of principal component $\boldsymbol{\xi}$

The sample covariance matrix $\hat{\mathbf{C}}$ obeys a Wishart distribution, determined by $\mathbf{C}$, N and T. Using Bayes formula we may write the likelihood (density of probability) for the normalized top component $\boldsymbol{\xi}$ of $\hat{\mathbf{C}}$ as follows

$$\rho(\boldsymbol{\xi}) \propto \exp\left(\frac{r\,\beta}{2}\sum_{i,j}\xi_i\,\hat{C}_{ij}\,\xi_j\right)\delta(\boldsymbol{\xi}^2-1)\ , \qquad (1)$$

up to a normalization coefficient. Parameter β above is equal to $\beta_{Bayes} = 1 - \frac{1}{\gamma}$. However, it is convenient to consider β as a free parameter. The $\beta \to \infty$ limit corresponds to Maximum Likelihood inference, while working at low values of β may be useful to ensure rapid mixing of Monte Carlo Markov Chain sampling of distribution ρ, especially in the presence of prior information, see below.

C. Prior information over $\boldsymbol{\xi}$

We now assume that prior information over the population eigenvector $\mathbf{u}$ is available under the form of a potential V acting on the entries of $\boldsymbol{\xi}$. The posterior distribution over the top component now reads

$$\rho(\boldsymbol{\xi}) \propto \exp\left(\frac{r\,\beta}{2}\sum_{i,j}\xi_i\,\hat{C}_{ij}\,\xi_j - \sum_i V(\xi_i)\right)\delta(\boldsymbol{\xi}^2-1)\ , \qquad (2)$$

up to a normalization coefficient. Three choices for the potential V are shown in Fig. 1. Motivated by previous works on protein sequence analysis, see Introduction, we will hereafter mostly concentrate on $V(\xi) = -V_0\ \xi^4$, with $V_0 \geq 0$ (Fig. 1(a)). This potential favors the presence of large entries in the top component, but does not rule out the existence of many entries with small magnitude (typically, of the order of $N^{-1/2}$). It is therefore different from sparsity-enforcing potentials, such as $V(\xi) \propto |\xi|$ in Fig. 1(b). Exact results for the location of the transition in the nonnegative case (Fig. 1(c)) were recently derived [4]. Our formalism finds back those results, and can be applied to any potential V as shown below.

III. Calculation of phase diagram

A. General replica calculation

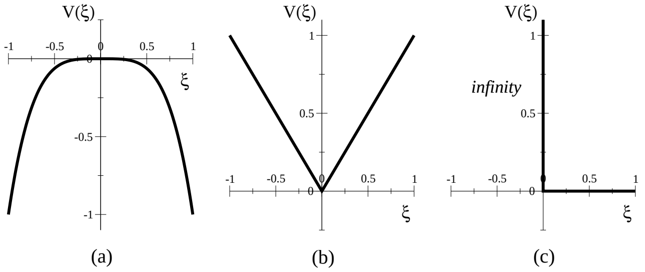

Fig. 1. Three potentials $V(\xi)$ corresponding to three prior information about the entries ξ of the principal component: (a) large entries are present, (b) L_1 penalty favoring zero entries, (c) all entries are nonnegative.

We assume that the logarithm, divided by N, of the normalization coefficient of ρ in Eq. (2),

$$Z(\hat{\mathbf{C}}) = \int_{\boldsymbol{\xi}^2=1} d\boldsymbol{\xi}\ \exp\left(\frac{r\,\beta}{2}\sum_{i,j}\xi_i\,\hat{C}_{ij}\,\xi_j - \sum_i V(\xi_i)\right) \qquad (3)$$

is concentrated around its expectation value $L \equiv \frac{1}{N}E_{\hat{\mathbf{C}}}[\log Z]$ in the $N, T \to \infty$ limit, and compute the latter with the help of the replica method, a non rigorous technique commonly used in the statistical physics of disordered systems [8], see [9], [10], [11] for recent applications to related high-dimensional inference problem. We obtain

$$\begin{aligned} L(r,\gamma,\beta) = \text{Extr}\bigg\{ & -\frac{1}{2r}\log\left(\frac{1}{\beta} - q_0 + q_1\right) - q_0\hat{q}_0 \\ & + \frac{q_1}{2r(\frac{1}{\beta} - q_0 + q_1)} + \big(\hat{q}_0 - \hat{q}_1\big)\big(1 + (\gamma-1)p^2\big) \qquad (4) \\ & + \hat{\mu} + q_1\hat{q}_1 + p\hat{p} - \frac{1}{2}\log\hat{\mu} + \frac{\hat{q}_1}{2\hat{\mu}} + \tilde{L}(\hat{\mu},\hat{p},\mathbf{u})\bigg\}\ , \end{aligned}$$

where the Extremum is computed over the order parameters $p^2 = \frac{1}{N}E_{\hat{\mathbf{C}}}[\langle(\sum_i u_i\xi_i)^2\rangle_{\hat{\mathbf{C}}}]$, $q_0 = \frac{1}{N}\sum_{i,j}C_{ij}E_{\hat{\mathbf{C}}}[\langle\xi_i\xi_j\rangle_{\hat{\mathbf{C}}}]$, $q_1 = \frac{1}{N}\sum_{i,j}C_{ij}E_{\hat{\mathbf{C}}}[\langle\xi_i\rangle_{\hat{\mathbf{C}}}\langle\xi_j\rangle_{\hat{\mathbf{C}}}]$, and the conjugated Lagrange multipliers $\hat{q}_0, \hat{q}_1, \hat{\mu}$. Here, $\langle\cdot\rangle_{\hat{\mathbf{C}}}$ denotes the expectation over the ρ distribution over $\boldsymbol{\xi}$ in Eq. (2). The term $\tilde{L}$ in Eq. (4) depends on the prior potential V and on the structure of the normalized population top component $\mathbf{u}$, more precisely, on how its entries scale with N. We now consider two cases of interest.

B. Case of nonnegative entries

We assume first that the components of $\mathbf{u}$ scales as $\frac{\tilde{u}_i}{\sqrt{N}}$, with $\tilde{u}_i$ finite, and denote by $\varphi(\tilde{u})$ the distribution of the $\tilde{u}_i$. We focus on the nonnegative entry prior, for which $V(\xi) = +\infty$ for $\xi < 0$ and 0 for $\xi \geq 0$. We obtain

$$\tilde{L} = \frac{\hat{p}^2}{2\hat{\mu}} + \frac{1}{2\hat{\mu}}\int d\tilde{u}\,\varphi(\tilde{u})\left[\log\text{erfc}\left(\frac{-\hat{p}\,\tilde{u} - z\sqrt{2\hat{q}_1}}{2\sqrt{\hat{\mu}}}\right)\right]_z\ , \qquad (5)$$

where erfc is the complementary error function, and $[F(z)]_z$ denotes the average of $F(z)$ over the Gaussian measure, $e^{-z^2/2}/\sqrt{2\pi}$. After some elementary algebra, we obtain the expression for the overlap

$$p = \frac{[\tilde{u}(x\,\tilde{u}+z)]_z^+}{[(x\,\tilde{u}+z)^2]_z^+}\ \text{with}\ x = \frac{\hat{p}}{\sqrt{2\hat{q}_1}} = \frac{(\gamma-1)\,p}{\sqrt{r(1+(\gamma-1)p^2)}} \qquad (6)$$

where $[F(z))]_z^+ = [\max(F(z),0)]_z$. These equations correspond to Eqs. (7), (8), (9), (21) and (23) of [4].

C. Case of large entries

We now assume that $\mathbf{u}$ has only K 'large' entries, $u_1, ..., u_K$, different from zero in the $N \to \infty$ limit (with finite K), and that the other entries decay fast enough with N, *e.g.* are of the order of $\frac{1}{\sqrt{N}}$. Then

$$\tilde{L} = -\sum_{i=1}^{K} \min_{-1 \leq \xi \leq 1} \left\{ \hat{\mu}\, \xi^2 + \hat{p}\, u_i\, \xi + V(\xi) \right\} , \qquad (7)$$

While the above formula is valid for generic V we restrict ourselves, in the remaining part of this article, to the potential $V(\xi) = -V_0\ \xi^4$. Furthermore, we assume that the K finite entries of $\mathbf{u}$ are all equal to $u = \frac{1}{\sqrt{K}}$; The calculation can be easily extended to $u < \frac{1}{\sqrt{K}}$, or to the case of nonhomogeneous entries u_i. In addition we assume that (A1) all ξ_is take identical values in Eq. (4) (homogeneous regime); (A2) $\boldsymbol{\xi}$ has no large (finite when $N \to \infty$) entry ξ_i on sites such that $u_i = 0$. The validity of these assumptions will be discussed in the next Section.

After some elementary algebra the extremization conditions reduce to the following set of $K+1$ coupled equations over $\hat{\mu}$ and the first K entries of $\boldsymbol{\xi}$:

$$\begin{aligned} &\xi_i - \frac{2\,V_0}{\hat{\mu}}\,\xi_i^3 = \frac{(\gamma-1)p}{r\sqrt{K}\left(\frac{2\hat{\mu}}{\beta}-1\right)} \qquad (i=1,...,K)\ , \\ &\sum_{i=1}^{K} \xi_i^2 = 1 - \frac{1}{2\hat{\mu}} - \frac{1+(\gamma-1)p^2 - \frac{1}{2\hat{\mu}}}{r\left(\frac{2\hat{\mu}}{\beta}-1\right)^2} \end{aligned} \qquad (8)$$

where

$$p = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i\ . \qquad (9)$$

Note that the K variables ξ_is obey the same cubic equation and, hence, can take at most three different values as i varies.

The equations above admit the solution $p = \xi_i = 0, \hat{\mu} = \frac{1}{2}$, corresponding to the 'unaligned' phase. In addition, in some well-defined regions of the four-dimensional control parameter space (r, γ, β, V_0) solutions with $p \neq 0$ may be found. We give in Section IV below results for three cases: (A) no prior ($V_0 = 0$); (B) weak exploitation of many data with prior information (small β, r for finite V_0); (C) maximum *a posteriori* inference (finite r, and infinite β and V_0 at fixed ratio V_0/β).

IV. Results for 'large entry' prior

A. Warm-up: no prior

We start with the case $V_0 = 0$. Extremization conditions over the parameters in Eq. (4) give the value of the squared overlap p^2 between $\mathbf{u}$ and $\boldsymbol{\xi}$ for any β. We find: $p^2 = 0$ for $r > r_c = (1-\gamma)^2$ whatever the value of β, and, for $r < r_c$,

$$p^2 = \left(1 - \frac{\beta(r)}{\beta}\right)\left(\frac{1-\frac{r}{r_c^2}}{1+\frac{r}{r_c}}\right) \text{ if } \beta > \beta(r) \equiv \frac{r}{r+\gamma-1}. \qquad (10)$$

Those expressions perfectly agree, in the $\beta \to \infty$ limit, with known results for the spiked covariance model [2], [3]. In addition our formalism gives access to the value of p^2 for finite β. Note that, for $r < r_c$, inference of the direction of $\mathbf{u}$ is possible, *i.e.* $p^2 > 0$, even for $\beta \leq \beta_{Bayes}$ (but larger than $\beta(r)$). At the critical noise, $\beta(r_c) = \beta_{Bayes}$.

B. Inference at low β with prior information

The above results imply that, in the absence of prior information ($V_0 = 0$), inference of the top component direction is possible at low $\beta \to 0$ provided the sampling noise r is smaller than $\beta/(1-\gamma)$. In other words, when both β and r tend to zero with a fixed ratio $\tilde{\beta} = \beta/r$, the aligned and not-aligned phases correspond, respectively, to $\tilde{\beta} > \tilde{\beta}_c(0) = \frac{1}{\gamma-1}$, and $\tilde{\beta} < \tilde{\beta}_c(0)$.

We expect the critical ratio $\tilde{\beta}_c(V_0)$ to be a drecreasing function of the prior strength V_0, as stronger prior should facilitate the recovery of the large-entry top component $\mathbf{u}$. Resolution of Eq. (8) gives the phase diagram shown in Fig. 2. Several regimes can be identified, depending on V_0:

- For $0 < V_0 \leq \frac{K}{4}$, the critical ratio $\tilde{\beta}_c(V_0)$ remains unchanged, see Fig. 2, and equal to $\frac{1}{\gamma-1}$. As $\tilde{\beta}$ crosses this critical value the overlap p continuously increases from 0 to a positive value, see Fig. 3.
- For $\frac{K}{4} < V_0 \leq K$, the aligned phase ($p \neq 0$) exist for $\tilde{\beta} > \tilde{\beta}_{c,1} = \frac{4\sqrt{V_0/K}\left(1-\sqrt{V_0/K}\right)}{\gamma-1}$, see dashed line in Fig. 2. As $\tilde{\beta}$ crosses $\tilde{\beta}_{c,1}$ the squared overlap p^2 discontinuously jumps from 0 to $1 - \frac{1}{2\sqrt{V_0}} > 0$, see Fig. 3.
- For $\frac{K}{4} < V_0 \leq V_0^+ \times K$, the aligned phase ($p \neq 0$) is thermodynamically stable, meaning that the value L_+ of L in Eq. (4) attached to this phase is larger than the one of the nonligned phase ($p = 0$), $L_0 = \frac{1}{2}(1+\tilde{\beta})$, for $\tilde{\beta} > \tilde{\beta}_{c,2}$, see full line in Fig. 2. The value of $\tilde{\beta}_{c,2}$ and of the overlap p_2 at the phase transition are the roots of the two coupled equations

  $$\begin{aligned} 0 &= \frac{p_2^2}{1-p_2^2} + \frac{1}{2}\log\left(1-p_2^2\right) - \frac{V_0}{K}\,p_2^4\ , \\ \tilde{\beta}_{c,2} &= \frac{1}{1-p_2^2} - 4\frac{V_0}{K}\,p_2^2\ , \end{aligned} \qquad (11)$$

  where the first equation actually implements the condition $L_+ = L_0$. The phase transition is illustrated in the middle panel of Fig. 3 for a specific value of V_0. The value of $V_0^+ \simeq 1.227703...$ is defined from

  $$V_0^+ = \min\left\{ V_0 \text{ s.t.} \max_{\xi \neq 0} \left(\log(1-\xi^2) + 2V_0\,\xi^4\right) > 0 \right\} \qquad (12)$$

- For $V_0 \geq V_0^+ \times K$ the prior strength is so strong that the inferred component $\boldsymbol{\xi}$ has few large entries whatever the value of β. For $\beta > 0$ it is aligned ($p \neq 0$) with $\mathbf{u}$ (Fig. 3), while for $\beta < 0$, it is not, see Fig. 2.

Assumption (A1), see Section III-C, is trivially valid for $K = 1$, but is not necessarily correct for $K \geq 2$ and strong prior strength, for which we expect that $\boldsymbol{\xi}$ will condensate and one component, say, ξ_1, will be larger than the other components, say, ξ_i, with $i = 2, ..., K$ (nonhomogeneous regime). The transition line between these two regimes is

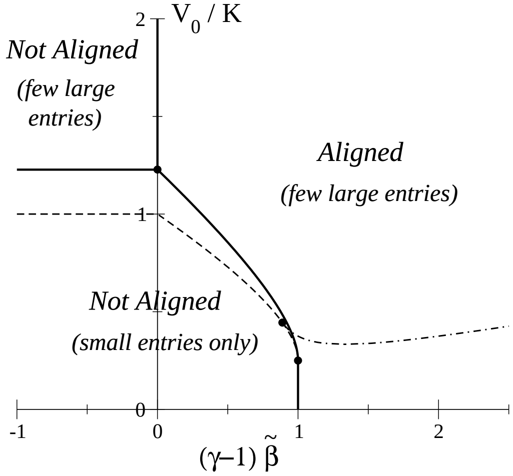

Fig. 2. Phase diagram of top component recovery in the $(\tilde{\beta}, V_0)$ plane; axis are rescaled by $(\gamma-1)$ and $\frac{1}{K}$ factors, where K is the number of nonzero components in $\mathbf{u} = \frac{1}{\sqrt{K}}(1,1,...,1,0,0,...,0)$. Dots locate the points $(1, \frac{1}{4})$, $(\frac{8}{9}, \frac{4}{9})$ and $(0, 1.228...)$, see text. Transitions between phases are shown by full lines. The dashed lines show the limit of existence of the Aligned phase, while the dot-dashed lines separate the regions with (above) and without (below line) homogeneity breaking among the K large entries of $\boldsymbol{\xi}$.

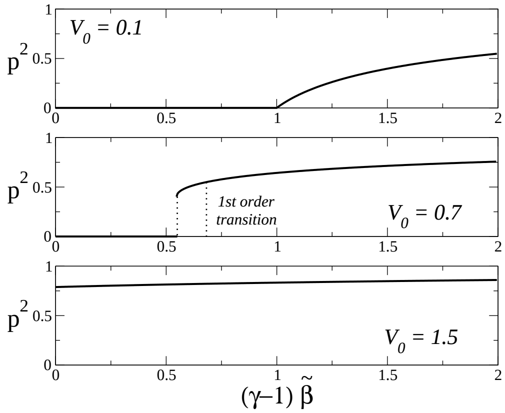

Fig. 3. Average squared overlap between top components of the population and sample covariance matrices, $p^2 \equiv E_{\hat{\mathbf{C}}}[(\mathbf{u}\cdot\boldsymbol{\xi})^2]$, vs. control parameter $(\gamma-1)\tilde{\beta}$ for three prior strengths and $K=1$.

identified upon imposing that the cubic equation over ξ_i in Eq. (4) admits a two-fold degenerate solution ξ, that is, $6V_0\,\xi^2 = \hat{\mu}$. The transition line is plotted in the phase diagram of Fig. 2 (dot-dashed line), and ends up in the point of coordinate $(\frac{4}{9}, \frac{8}{9})$ lying on the existence line (dashed line).

C. Inference at high β with strong prior information

We now focus on MAP inference at finite sampling noise r, whereas β and V_0 are both sent to infinity, with a fixed ratio $S = r\,V_0/(\beta)$. Parameter S, hereafter referred to as the slope, controls the relative magnitude of the $\hat{C}$-dependent and prior terms in ρ, see Eq. (2), while the multiplicative factor r is introduced in the definition of S to compensate for the explicit dependence upon r in ρ. For simplicity we present results for $K=1$ only, the extension to larger K being rather straightforward. Equations (8) admit the solution $p=0, \hat{\mu}=\frac{1}{2}$, and another solution, with $\hat{\mu}\to\infty$ as $\beta, V_0\to\infty$, with ratios $\beta/\hat{\mu}$, $V_0/\hat{\mu}$ having finite limits. After some simple algebra we obtain the following expresson for the slope as a function of the squared overlap for the latter solution:

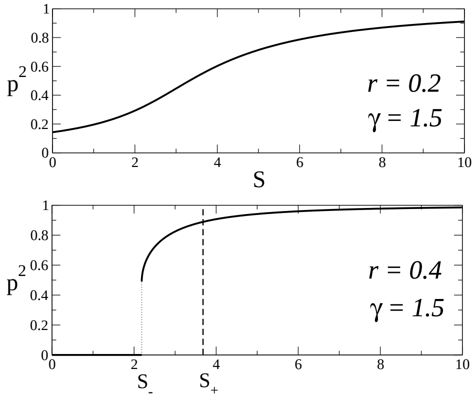

Fig. 4. Average squared overlap between top components of the population and sample covariance matrices, p^2, vs. slope S for sampling noises $r=0.2$ (top) and 0.4 (bottom). Note the presence of the discontinuous transition at $S_- \simeq 2.19$ in the latter case. The randomly condensed solution appears for $S > S_+ \simeq 3.68$. Here, $r_c = (\gamma-1)^2 = 0.25$.

$$S(p^2) = \frac{\big(r-(\gamma-1)y\big)\big(1+y\big)}{4\,p^2\,y} \;\text{ with } y = \sqrt{\frac{r(1-p^2)}{1+(\gamma-1)p^2}}. \tag{13}$$

We show in Fig. 4 the representative curve of p^2 vs. the slope S, for r below and above the critical noise level, $r_c = (\gamma-1)^2$, in the absence of prior. For $r < r_c$ the squared overlap is an increasing function of S, starting from a non zero value for $S=0$: the population eigenvector direction can be estimated without prior at low sampling noise [3], but the overlap is increased in the presence of prior. For $r > r_c$ a discontinuous jump is observed from $p=0$ to $p>0$ at a critical value of the slope,

$$S_- = \min_{p^2>0} S(p^2) \qquad (r > r_c)\ , \tag{14}$$

while the overlap further increases as S exceeds S_-. Remarkably, even for large sampling noise values, the presence of a sufficiently strong prior allows us to infer $\mathbf{u}$. The value of S_- as a function of the noise level r is shown in Fig. 5; for large noise levels we have $S_- = \frac{r}{4} + \frac{3}{4^{4/3}}\,\gamma^{1/3}\,r^{2/3} + O(r^{1/3})$.

This aligned phase competes with a nonaligned, but condensed phase, in which assumption (A2), see Section III-C, is violated. In other words, for S and r sufficiently large, $\boldsymbol{\xi}$ has few large entries ($\xi_j^2 > 0$ in the $N\to\infty$ limit), but not along the directions i corresponding to the large components of $\mathbf{u}$; hence, $p^2=0$. To describe this new phase we set u_i to 0 in the expression for $\tilde{L}$ in Eq. (7). The corresponding optimization equations can be solved in the $\beta, V_0\to\infty$ limit, with the result that the nonaligned, condensed regime exists for S larger than

$$S_+ = \min_{0<b<\frac{\sqrt{r}}{1+\sqrt{r}}} \left[\frac{r^2(1-b)^2}{4b\big(r(1-b)^2-b^2\big)}\right]\ . \tag{15}$$

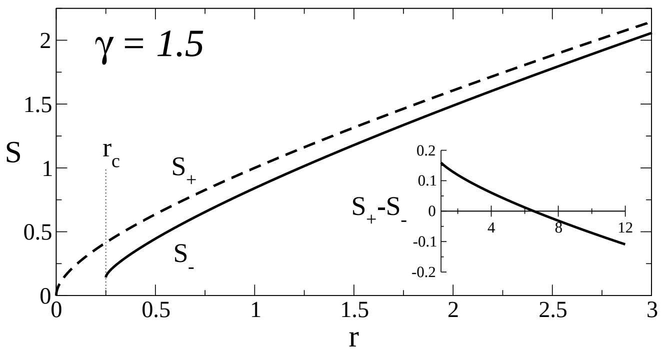

Fig. 5. Behaviours of S_- (full line) and S_+ (dashed line) as functions of the sampling noise r. Insert: $S_+ - S_-$ vs. r. The difference vanishes in $r_d \simeq 6.54$. Here, $r_c = 0.25$, *i.e.* $\gamma = 1.5$.

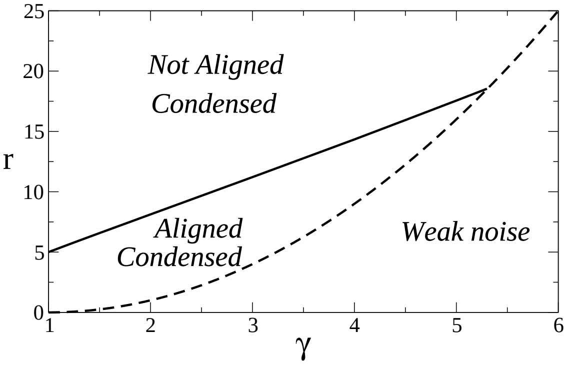

Fig. 6. Replica-symmetric phase diagram of top component inference with large-entry prior in the (γ, r) plane. The dashed line $r_c = (\gamma - 1)^2$ divides the plane into the weak noise region (below line), where inference is possible withour prior, and the strong noise region (above line). The full line shows the value of r_d, at which $S_- = S_+$, as a function of γ. In between the dashed and full lines, inference of the top component is possible in the presence of a prior with appropriate strength. The two lines merge for $\gamma \simeq 5.3$, $r \simeq 18.5$. Here, $K = 1$.

The value of S_+ as a function of the noise level r is shown in Fig. 5. For small r (slightly above r_c) we observe that S_+ is larger than S_-, as intuitively expected: it is favorable to condense $\boldsymbol{\xi}$ along the direction of $\mathbf{u}$ rather than any other direction. It can be checked that the value of L in Eq. (4) is higher for this phase than for the aligned condensed phase. Hence, as soon as S exceeds S_+ the overlap p vanishes.

For large noise levels, however, we have $S_+ = \frac{r}{4} + \frac{3}{4^{4/3}} r^{2/3} + O(r^{1/3})$, which is asymptotically smaller than S_-, see insert of Fig. 5. Indeed, the threshold slopes S_- and S_+ cross at a well-defined value of the noise, r_d, which depends on the top eigenvalue γ. We show in Fig. 6 the behaviour of r_d vs. γ, and compare it to the critical noise r_c in the absence of prior. We observe the presence of a region in the (γ, r) plane, above the critical line r_c, where the direction of $\mathbf{u}$ can be inferred thanks to the large-entry prior. Our replica symmetric theory predicts that the benefit of the prior does not extend to very large values of the signal eigenvalue γ and of the noise r, see Fig. 6.

V. Conclusion

The non rigorous calculations done in this paper suggest that inference of the top component of the population covariance matrix is possible in the presence of prior information, even above the critical noise level of the spiked covariance model, in agreement with rigorous results for the nonnegative case [4] and numerical investigations for the large-entry case [5]. Many interesting questions have not been investigated in the present paper: how hard is the recovery problem from a computational point of view? Are there 'dynamical' phase transitions separating subregions in the aligned phase, where the top component can be recovered in polynomial time or not? If so how do these line compare to the 'static' critical lines derived in this paper? In addition it would be interesting to investigate the validity of the replica-symmetric hypothesis used to derive the results above [8]. Though replica symmetry is generally expected to be correct for convex optimization problems what happens in nonconvex situations is not clear. For instance, inference of the top component with the nonnegative prior gives rise to a nonconvex optimization problem [4], but all rigorous results are exactly found back within our replica symmetric approach, see Section III.B. From this perspective it would be useful to investigate whether the present results are robust against replica symmetry breaking.

Acknowledgment

I am grateful to M.R. McKay for the invitation to the Asilomar 2016 conference, and to S. Cocco and D. Villamaina for useful discussions. This work has benefited from the financial support of the CNRS Inphyniti Inferneuro project.

References

[1] I.M. Johnstone, *High dimensional statistical inference and random matrices*, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 307-333, 2006.

[2] D.C. Hoyle and M. Rattray, *Principal-component-analysis eigenvalue spectra from data with symmetry breaking structure*, Phys. Rev. E 69, 026124, 2004.

[3] J. Baik, G. Ben Arous G. and S. Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Probab. 33, 1643-1697, 2005.

[4] A. Montanari and E. Richard, *Non-negative Principal Component Analysis: Message Passing Algorithms and Sharp Asymptotics*, IEEE Transactions on Information Theory 62, 1458-1484, 2016.

[5] D. Villamaina and R. Monasson, *Estimating the principal components of correlation matrices from all their empirical eigenvectors*, Europhys. Lett. 112, 50001, 2015.

[6] S. Cocco, R. Monasson and M. Weigt, *From principal component to direct coupling analysis of coevolution in proteins: Low–eigenvalue modes are needed for structure prediction*, PLoS Comp. Bio. 9, e:1003176, 2013.

[7] H. Jacquin, A. Gilson, E. Shakhnovich, S. Cocco, R. Monasson, *Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models*, PLoS Comp. Bio. 12: e1004889, 2016.

[8] A. Engel and C.V. den Broeck, *Statistical Mechanics of Learning*, Cambridge University Press, Cambridge, England, 2001.

[9] M. Advani and S. Ganguli, *Statistical mechanics of optimal convex inference in high dimensions*, Phys. Rev. X 6, 031034, 2016.

[10] Y. Nakanishi-Ohno, T. Obuchi, M. Okada, Y. Kabashima, *Sparse approximation based on a random overcomplete basis*, J. Stat. Mech. P063302, 1-30, 2016

[11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, L. Zdeborova, *Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices*, J. Stat. Mech. P08009, 1-57, 2012.