Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure

D. C. Hoyle* and M. Rattray†

Department of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom

Principal component analysis (PCA) is a ubiquitous method of multivariate statistics that focuses on the eigenvalues λ and eigenvectors of the sample covariance matrix of a data set. We consider p, N-dimensional data vectors $\boldsymbol{\xi}$ drawn from a distribution with covariance matrix $\mathbf{C}$. We use the replica method to evaluate the expected eigenvalue distribution $\rho(\lambda)$ as $N \to \infty$ with $p = \alpha N$ for some fixed α. In contrast to existing studies we consider the case where $\mathbf{C}$ contains a number of symmetry-breaking directions, so that the sample data set contains some definite structure. Explicitly we set $\mathbf{C} = \sigma^2 \mathbf{I} + \sigma^2 \Sigma_{m=1}^{S} A_m \mathbf{B}_m \mathbf{B}_m^T$, with $A_m > 0 \ \forall \ m$. We find that the bulk of the eigenvalues are distributed as for the case when the elements of $\boldsymbol{\xi}$ are independent and identically distributed. With increasing α a series of phase transitions are observed, at $\alpha = A_m^{-2}$, $m = 1, 2, \ldots, S$, each time a single δ function, $\delta(\lambda - \lambda_u(A_m))$, separates from the upper edge of the bulk distribution, where $\lambda_u(A) = \sigma^2 [1 + A][1 + (\alpha A)^{-1}]$. We confirm the results of the replica analysis by studying the Stieltjes transform of $\rho(\lambda)$. This suggests that the results obtained from the replica analysis are universal, irrespective of the distribution from which $\boldsymbol{\xi}$ is drawn, provided the fourth moment of each element of $\boldsymbol{\xi}$ exists.

I. INTRODUCTION

The techniques of statistical physics have been applied, with some success, to the study of many different statistical learning methods [1]. Amongst such methods is that of principal component analysis (PCA) [2] where one aims to discover correlations between the different components of a data set, and thereby provide a means for reducing the complexity of the representation of the data. Given a data set consisting of p, N-dimensional mean centered vectors $\boldsymbol{\xi}_\mu$, $\mu = 1, \ldots, p$, the leading principal components are taken as the eigenvectors of the sample covariance matrix, $\hat{\mathbf{C}} = p^{-1} \Sigma_\mu \boldsymbol{\xi}_\mu \boldsymbol{\xi}_\mu^T$, that have the largest eigenvalues—i.e., those directions in the data space along which there is the greatest variation. If we use P principal components to represent (with loss) the original data set, the P leading eigenvectors of $\hat{\mathbf{C}}$ are in fact the maximum likelihood choice for reconstructing the data set. More recently PCA has been recast as a latent-variable model [3], allowing its extension to a mixture of PCA models [4].

Given a data set the question then remains: how many principal components of $\hat{\mathbf{C}}$ should one retain to model the true underlying covariance matrix $\mathbf{C}$? Typical methods of selecting principal components focus on comparing the eigenvalues λ of $\hat{\mathbf{C}}$ to the expected distribution of eigenvalues $\rho(\lambda)$ when the components of $\boldsymbol{\xi}_\mu$ are independent and identically distributed (i.i.d.) [5]. However it would be instructive to determine the expected distribution $\rho(\lambda)$ when the components of $\boldsymbol{\xi}_\mu$ are not i.i.d., but when the sample vectors $\boldsymbol{\xi}_\mu$ are drawn from a distribution $P(\boldsymbol{\xi})$ containing a finite number of symmetry-breaking directions.

Early work by Anderson [6] considered the asymptotic distribution of λ in the limit $p \to \infty$ with N finite, i.e., an increasing number of sample vectors. For some real applications one may have $p \ll N$ and so the limit considered by Anderson [6] will be somewhat unrealistic. Study of PCA for more general values of $\alpha = p/N$ has been done by applying results from standard matrix ensembles in random matrix theory [7]. These results have been extended to more general sample covariance matrices [8], although it is still only for the case where the elements of $\boldsymbol{\xi}$ are i.i.d. Therefore in this paper we aim to explicitly obtain the behavior of $\rho(\lambda)$ for general values of α and as $N \to \infty$, when the true covariance matrix $\mathbf{C}$ does contain a finite number of symmetry-breaking directions.

II. MODEL

The sample data vectors $\{\boldsymbol{\xi}_\mu\}_{\mu=1}^{p}$ are considered to contain both a signal and a noise component, i.e.,

$$\boldsymbol{\xi}_\mu = \boldsymbol{\zeta}_\mu + \mathbf{e}_\mu, \tag{1}$$

where the elements of the noise vector $\mathbf{e}_\mu$ are i.i.d. with mean zero and variance σ^2. Initially we restrict ourselves to the case where $\boldsymbol{\zeta}_\mu$ is given by a small number S of Gaussian distributed latent variables corresponding to orthogonal signal directions $\mathbf{B}_m$, $m = 1, \ldots, S$. Thus, $\boldsymbol{\zeta}_\mu = \Sigma_{m=1}^{S} z_m \mathbf{B}_m$, with $z_m \sim N(0, \sigma^2 A_m)$. Similarly we initially consider that $\mathbf{e}_\mu$ is drawn from a Gaussian distribution $N(\mathbf{0}, \sigma^2 \mathbf{I})$. In this case $\boldsymbol{\xi}_\mu$ has the Gaussian distribution

$$P(\boldsymbol{\xi}) = (2\pi)^{-N/2} (\det \mathbf{C})^{-1/2} \exp[-\tfrac{1}{2} \boldsymbol{\xi}^T \mathbf{C}^{-1} \boldsymbol{\xi}], \tag{2}$$

where the covariance matrix $\mathbf{C}$ is isotropic with variance σ^2 except for a small number of orthogonal symmetry-breaking directions, i.e.,

---

*Electronic address: david.c.hoyle@man.ac.uk; www.cs.man.ac.uk/~dchoyle

†Electronic address: magnus@cs.man.ac.uk www.cs.man.ac.uk/~magnus

$$\mathbf{C}=\sigma^2\mathbf{I}+\sigma^2\sum_{m=1}^{S}A_m\mathbf{B}_m\mathbf{B}_m^T, \quad \mathbf{B}_m^T\mathbf{B}_{m'}=\delta_{mm'}, \quad (3)$$

where $A_m>0\,\forall\,m$. Later we will relax the Gaussian assumption for the distribution of pattern vectors $\boldsymbol{\xi}_\mu$. The isotropic case corresponds to $A_m\equiv 0\,\forall\,m$, and for simplicity we will not consider the degenerate case, i.e., we restrict our analysis to $A_1>A_2>\cdots>A_S>0$. We denote the eigenvalues of $\mathbf{C}$ by $\{d_i\}_{i=1}^N$, so that for the covariance matrix given above $\{d_i\}_{i=1}^N=[\sigma^2(1+A_1),\ldots,\sigma^2(1+A_S),\sigma^2,\ldots,\sigma^2]$. The sample covariance matrix is defined by $\hat{\mathbf{C}}=p^{-1}\Sigma_\mu\boldsymbol{\xi}_\mu\boldsymbol{\xi}_\mu^T$. When the elements of $\boldsymbol{\xi}$ are i.i.d. Gaussian, $\hat{\mathbf{C}}$ is simply (up to a factor p^{-1}) the Wishart matrix $\mathbf{XX}^T$ [9] formed from the data matrix $\mathbf{X}=(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_p)$ (or strictly speaking an anti-Wishart matrix [10] when $p<N$). We are interested in the observed distribution of eigenvalues of $\hat{\mathbf{C}}$ when N is large but finite. The behavior at large N is often approximated by that found in the limit $N\to\infty$, with $p=\alpha N$ for some fixed α. For $\alpha<1$, $\hat{\mathbf{C}}$ is singular with a $(N-p)$-fold degenerate zero eigenvalue. However we are interested in the behavior of the largest eigenvalues of $\hat{\mathbf{C}}$ and how their average values depend upon $\{A_m\}_{m=1}^S$ and α.

The elements of $\hat{\mathbf{C}}$ are random variables and as such $\hat{\mathbf{C}}$ represents a random matrix. The study of the eigenvalue spectra of random matrices has a long history [11], with Bai [12] and Forrester *et al.* [13] providing recent comprehensive reviews of the different aspects of the literature. Often this study is restricted to ensembles of matrices that are invariant under certain classes of transformation, such as the Gaussian orthogonal ensemble (GOE). For Wishart matrices the distribution of the largest eigenvalue has been shown to be identical to that for the GOE [7]. Due to the presence of the symmetry-breaking directions in $\mathbf{C}$ we would not automatically expect the behavior of the largest eigenvalue in the ensemble of sample covariance matrices to be equivalent to that for the GOE, or any of the other common ensembles studied.

For the isotropic case the observed distribution of eigenvalues converges to, as $N\to\infty$ [14,15],

$$\rho(\lambda)=(1-\alpha)\Theta(1-\alpha)\delta(\lambda)+\frac{\alpha}{2\pi\lambda\sigma^2}$$
$$\times\sqrt{\max(0,(\lambda-\lambda_{min})(\lambda_{max}-\lambda))}, \quad (4)$$

where $\lambda_{max,min}=\sigma^2\alpha^{-1}(1\pm\sqrt{\alpha})^2$. Marčenko and Pastur [14] studied the case of a general covariance matrix and found that the limiting distribution satisfied

$$\rho(\lambda)=\lim_{\epsilon\to 0^+}\pi^{-1}\mathrm{Im}\,\alpha m_\rho(\lambda+i\epsilon), \quad (5)$$

where $m_\rho(z)$ is the Stieltjes transform of $\alpha^{-1}\rho(\lambda)$ and is determined by

$$z(m_\rho)=-\frac{1}{m_\rho}+\alpha^{-1}\int\frac{dH(t)}{t^{-1}+m_\rho}, \quad (6)$$

where the measure $H(t)$ is defined such that $N^{-1}\Sigma_i d_i^k$ converges to $\int t^k dH(t)\,\forall\,k$. An equivalent result is also derived by Wachter [16]. From this we can see that if our covariance matrix $\mathbf{C}$ has only a few symmetry-breaking directions, then in the limit $N\to\infty$ the variance along these directions will have zero measure as defined by H. Thus in the limit $N\to\infty$ the distribution of eigenvalues is the same as for the isotropic case. However it is the behavior, at finite but large N, of the eigenvalues corresponding to the few symmetry-breaking directions present in $\mathbf{C}$ that we are interested in. Understanding the distribution of the eigenvalues of $\hat{\mathbf{C}}$ may allow us to construct more robust hypothesis tests to select the number of signal components contained within a data set. At present a number of hypothesis tests use the distribution (4) as the basis for a null hypothesis test on the eigenvalues of $\hat{\mathbf{C}}$. For example, Johnstone [7] provides an inequality for the likelihood of an individual eigenvalue from $\hat{\mathbf{C}}$ in terms of the likelihood of the leading eigenvalue when $\mathbf{C}$ is isotropic. The leading eigenvalue from the GOE has a Tracy-Widom (TW) distribution [17] and Johnstone [7] has shown that when $\mathbf{C}$ is isotropic, the leading eigenvalue of $\hat{\mathbf{C}}$ also has a Tracy-Widom distribution. Therefore Johnstone's inequality provides a means of constructing a conservative hypothesis test on the observed eigenvalues of $\hat{\mathbf{C}}$. Whether the top eigenvalue of $\hat{\mathbf{C}}$ follows a TW distribution when $\mathbf{C}$ contains symmetry-breaking directions is still an open question.

The behavior of PCA when one symmetry-breaking direction $\mathbf{B}$ is present [with variance $\sigma^2(1+A)$] has been studied, using the replica method, in the context of unsupervised learning [18,19], where one considers the overlap $R=\mathbf{J}\cdot\mathbf{B}$ between $\mathbf{B}$ and the leading principal component $\mathbf{J}$ of $\hat{\mathbf{C}}$. One observes the phenomenon of retarded learning, whereby R^2 goes through a critical phase transition from $R^2=0$ for $\alpha<A^{-2}$ to $R^2>0$ for $\alpha>A^{-2}$. Formulating the calculation of the eigenvalue spectrum of $\hat{\mathbf{C}}$ in terms of its resolvent leads to a partition function very similar to that used in the study of the unsupervised learning performance of PCA, and therefore suggests using replicas to derive the eigenvalue spectrum of $\hat{\mathbf{C}}$. The trace of the resolvent, $\mathbf{G}(\lambda)$, of the matrix $\hat{\mathbf{C}}$ is defined as

$$\mathrm{tr}\,\mathbf{G}(\lambda)=\sum_{i=1}^{N}\frac{1}{\lambda-\lambda_i}, \quad (7)$$

from which the density of eigenvalues $\rho(\lambda)$ can be calculated as

$$\rho(\lambda)=\lim_{\epsilon\to 0^+}(N\pi)^{-1}\mathrm{Im}\,\mathrm{tr}\mathbf{G}(\lambda-i\epsilon). \quad (8)$$

The first application of replicas to evaluate the expected eigenvalue distribution of a random matrix was due to Edwards and Jones [20]. This approach has been extended in many directions (e.g., sparse random matrices [21] and asymmetric random matrices [22]). More recently Sengupta and Mitra [23] calculated the resolvent (as $N\to\infty$) using replicas and found that the resolvent obeys the relation

$$z = \frac{p}{\text{tr } \mathbf{G}(z)} - \sum_{i=1}^{N} \frac{1}{\text{tr } \mathbf{G}(z) - p d_i^{-1}}. \quad (9)$$

It is a trivial matter to confirm that $-p^{-1}\text{tr } \mathbf{G}(z)$ corresponds to the Stieltjes transform $m_\rho(z)$ and that Eq. (9) is identical to Eq. (6). The work of Sengupta and Mitra [23] demonstrates the utility of using replicas to study the resolvent of $\hat{\mathbf{C}}$. However, as already mentioned, for finite S and on taking the limit $N \to \infty$, Eq. (9) predicts a distribution $\rho(\lambda)$ identical to the isotropic case (4). Potentially one could solve Eq. (6) or (9) at finite N and then take the limit, although the validity of this approach is uncertain given that the relationships (6) and (9) have been obtained by taking the limit $N \to \infty$. For $S > 1$ solving for tr $\mathbf{G}(z)$ involves solution of a third- (or higher-) order polynomial. Thus determining the distribution $\rho(\lambda)$ explicitly may be more difficult.

In this paper we calculate the resolvent of $\hat{\mathbf{C}}$ using the replica method, but examine the explicit behavior of the saddle point equations with variation of the parameters $\{A_m\}_{m=1}^{S}$. Thus we are able to derive explicit results for the expectation values of the S highest eigenvalues, rather than the more general relationship for the resolvent obtained by Sengupta and Mitra [23]. For the case where $\mathbf{C}$ contains one symmetry-breaking direction we observe a phase transition in the eigenvalue spectrum of $\hat{\mathbf{C}}$ at $\alpha = A^{-2}$, thus coinciding not unsurprisingly with the transition observed in the order parameter R^2 analyzed in unsupervised learning studies. Below $\alpha_c = A^{-2}$, the calculated spectrum is identical to that obtained from the isotropic case. Above α_c the bulk of the spectrum is still identical to that for the isotropic case, but with a single eigenvalue (the largest) clearly separated from the bulk. Transitions in the eigenvalue spectra of random matrices have also been well documented in other branches of physics (see, for example, Ref. [24] for a recent review). The calculation of the resolvent using replicas can easily be generalized to the situation where $\mathbf{C}$ contains $S > 1$ (orthogonal) symmetry-breaking directions. In this case we observe a series of phase transitions at $\alpha = A_m^{-2}, m = 1, 2, \ldots, S$, each time a single eigenvalue separates from the upper edge of the bulk of the spectrum, which is still the same as for the isotropic case.

III. THEORY

The density of eigenvalues $\rho(\lambda)$ of the $N \times N$ matrix $\hat{\mathbf{C}}$ can be expressed in terms of its resolvent $\mathbf{G}(\lambda)$,

$$\rho(\lambda) = \lim_{\epsilon \to 0^+} \frac{1}{N\pi} \text{Im tr } \mathbf{G}(\lambda - i\epsilon). \quad (10)$$

The trace of the resolvent $\mathbf{G}(\lambda)$ can be represented as

$$\text{tr } \mathbf{G}(\lambda) = \frac{\partial}{\partial\lambda} \ln \det(\lambda \mathbf{I} - \hat{\mathbf{C}}) = -\frac{\partial}{\partial\lambda} \ln Z(\lambda). \quad (11)$$

Using the standard representation of the determinant of a matrix,

$$[\det \mathbf{A}]^{-1/2} = (2\pi)^{-N/2} \int \exp\left[-\frac{1}{2}\boldsymbol{\phi}^T \mathbf{A} \boldsymbol{\phi}\right] d\boldsymbol{\phi}, \quad (12)$$

we have

$$\ln Z(\lambda) = 2\ln \int \exp\left[-\frac{\lambda}{2}\|\boldsymbol{\phi}\|^2 + \frac{1}{2p}\sum_\mu (\boldsymbol{\phi} \cdot \boldsymbol{\xi}_\mu)^2\right] d\boldsymbol{\phi}$$
$$- N \ln 2\pi. \quad (13)$$

The belief is that the eigenvalue spectrum is self-averaging so that the calculation of $\rho(\lambda)$ for a specific realization of the sample covariance matrix $\hat{\mathbf{C}}$ can be replaced by an ensemble average. The ensemble average of $\ln Z(\lambda)$ can be performed using the replica method.

A. One symmetry-breaking direction

We initially restrict ourselves to the case where

$$\mathbf{C} = \sigma^2 \mathbf{I} + \sigma^2 A \mathbf{B} \mathbf{B}^T. \quad (14)$$

Performing a standard calculation involving replicas [18,19,25] (see the Appendix) the partition function $\langle \ln Z(\lambda) \rangle_\xi$ is approximated (assuming replica symmetry), for large N, $p = \alpha N$, by locating the extrema of the mean-field free energy,

$$-F(q_0, x, R) = \ln x + \frac{q_0 - R^2}{x} - \alpha \ln(1 - \alpha^{-1}\sigma^2 x)$$
$$+ \frac{\alpha\sigma^2 x}{\sigma^2 x - \alpha} - \frac{\alpha\sigma^2(q_0 + AR^2)}{\sigma^2 x - \alpha} - \lambda q_0, \quad (15)$$

where $q_0 = N^{-1}\|\boldsymbol{\phi}_\nu\|^2 \forall \nu$ and $x = q_0 - q_1$, with $q_1 = N^{-1}\boldsymbol{\phi}_\nu \cdot \boldsymbol{\phi}_{\nu'} \forall \nu, \nu' \neq \nu$ being the overlap between different replica fields $\boldsymbol{\phi}_\nu, \boldsymbol{\phi}_{\nu'}$. Similarly $R = N^{-1/2}\mathbf{B} \cdot \boldsymbol{\phi}_\nu \forall \nu$ is the overlap between the replica fields $\boldsymbol{\phi}_\nu$ and the symmetry-breaking direction $\mathbf{B}$.

Saddle point equations are

$$-\frac{\partial F}{\partial R} = -\frac{2R}{x} - \frac{2\alpha\sigma^2 AR}{\sigma^2 x - \alpha} = 0, \quad (16)$$

$$-\frac{\partial F}{\partial x} = \frac{1}{x} - \frac{q_0 - R^2}{x^2} - \frac{\alpha x \sigma^4}{(\sigma^2 x - \alpha)^2} + \frac{\alpha\sigma^4(q_0 + AR^2)}{(\sigma^2 x - \alpha)^2} = 0, \quad (17)$$

$$-\frac{\partial F}{\partial q_0} = -\frac{\partial F}{\partial x} + \frac{1}{x} - \frac{\alpha\sigma^2}{\sigma^2 x - \alpha} - \lambda = 0. \quad (18)$$

Equation (18) is quadratic in x and has the solutions

$$x = \frac{1}{2\lambda\sigma^2}[(\sigma^2 - \alpha\sigma^2 + \lambda\alpha) \pm \sqrt{(\sigma^2 - \alpha\sigma^2 + \lambda\alpha)^2 - 4\alpha\lambda\sigma^2}], \quad (19)$$

while Eq. (16) has a solution at $R=0 \forall \alpha$, and a solution with $|R|>0$ iff $x=\alpha\sigma^{-2}/(1+\alpha A)$. If $R=0$, then Eq. (17) yields $q_0=x$. The trace tr $\mathbf{G}(\lambda)$ is then determined by the value of q_0 at the saddle point and for $R=0$ we have $\mathrm{Im}q_0 = \mathrm{Im}x$. The contribution that the saddle point at $R=0$ makes to the density of eigenvalues $\rho(\lambda)$ comes from two sources—an imaginary part of x arising from the square root in Eq. (19) and the singularity in Eq. (19) at $\lambda=0$. For $\lambda < \lambda_{min}=\sigma^2\alpha^{-1}(1-\sqrt{\alpha})^2$, the solution branch of x with the positive sign before the root in Eq. (19) has the lower value of $\mathrm{Re}\,F$ for $\alpha<1$, while for $\alpha>1$ the branch with the negative sign before the root in Eq. (19) has the lower value of $\mathrm{Re}\,F$. For $\lambda > \lambda_{max}=\sigma^2\alpha^{-1}(1+\sqrt{\alpha})^2$ the branch with the negative sign before the root in Eq. (19) has the lower value of $\mathrm{Re}\,F$. For $\lambda_{min}\leq\lambda\leq\lambda_{max}$ the values of F for the two branches of x are complex conjugates of each other. Taking $\mathrm{Im}\,\lambda=-\epsilon, 0<\epsilon\ll 1$, this symmetry in $\mathrm{Re}\,F$ at the two solutions of x is broken and the solution with a negative sign in Eq. (19) has the lower value of $\mathrm{Re}\,F$. The square root in Eq. (19) can easily be rewritten to give a bulk contribution to $\rho(\lambda)$, namely,

$$\frac{\alpha}{2\pi\lambda\sigma^2}\sqrt{(\lambda-\lambda_{min})(\lambda_{max}-\lambda)}, \quad \lambda_{min}\leq\lambda\leq\lambda_{max}. \quad (20)$$

For $|\lambda|\ll 1$ we can expand the two roots in Eq. (19) as

$$\frac{1}{2\lambda\sigma^2}[2\sigma^2(1-\alpha)+O(\lambda^2)], \quad \frac{1}{2\lambda\sigma^2}[2\lambda\alpha+O(\lambda^2)], \quad (21)$$

and on using the representation $(y-i\epsilon)^{-1}=\mathrm{PP}y^{-1}+i\pi\delta(y)$ as $\epsilon\to 0^+$, the solution with the positive sign before the root in Eq. (19) makes a contribution $\Theta(1-\alpha)(1-\alpha)\delta(\lambda)$ to the density $\rho(\lambda)$.

When $|R|>0$ we require $x=\alpha\sigma^{-2}/(1+\alpha A)$ at the saddle point and Eq. (17) has solution

$$R^2\left(1+\frac{1}{\alpha A}\right)+\left(1-\frac{1}{\alpha A^2}\right)(x-q_0)=0. \quad (22)$$

If $|R|>0$ we would expect the overlap, q_1, between different replica vectors to be positive as the replicas $\boldsymbol{\phi}_\nu$ become increasingly aligned with the symmetry-breaking direction $\mathbf{B}$. With $x-q_0=-q_1$ we can see that a positive solution for R^2 is only obtainable if $\alpha>A^{-2}$. The transition point $\alpha=A^{-2}$ coincides with the retarded learning phase transition observed in unsupervised learning studies of PCA [18]. The relation $x=\alpha\sigma^{-2}/(1+\alpha A)$ and Eq. (18) specify two conditions on x and therefore can both be satisfied only at distinct values of λ, namely,

$$\lambda_u=\sigma^2(1+A)\left(1+\frac{1}{\alpha A}\right). \quad (23)$$

Interestingly, λ_u given by Eq. (23) agrees with the result for the largest PCA eigenvalue obtained in our previous analysis of the learning problem [26]. For $\lambda=\lambda_u$, Eqs. (16) and (18)

specify identical conditions on x and therefore the saddle point equations can determine values for only two of the three order parameters. There is in fact a continuous line of saddle points corresponding to $q_0\in[x,\infty)$, $x=\alpha\sigma^{-2}/(1+\alpha A)$ with R^2 given by Eq. (22). Note that the line of saddle points extends to $R=0$ and in effect we have only one distinct solution branch to the saddle point equations. Along this line of saddle points the free energy F is constant. Correspondingly the appropriate Hessian $\mathbf{H}(\lambda)$ evaluated along this line of saddle points will have a zero eigenvalue. The replica partition function can in principle be calculated at $\lambda=\lambda_u$ by diagonalizing the Hessian on the line of saddle points. Fluctuations orthogonal to the line of saddle points can be integrated out, leaving a final integration over q_0. In fact to evaluate the density $\rho(\lambda)$ at $\lambda=\lambda_u$ we need only to evaluate the replica partition function at $\lambda_u-i\epsilon$ for $\epsilon\to 0^+$. Since for $\lambda\neq\lambda_u$ only an isolated saddle point at $R=0$ exists, to obtain $O(N^{-1})$ corrections to the bulk density (20) we need only to calculate the Hessian at $R=0$. Evaluating the Hessian at the saddle point with $R=0$ is performed in the Appendix, and as $n\to 0$ gives

$$\ln\det\mathbf{H}(\lambda)=\tfrac{1}{2}n\ln\{\tfrac{1}{4}[\alpha^{-1}(\lambda x-1)^2-1]\}$$
$$+n\ln[1+A-A\lambda x]+O(n^2). \quad (24)$$

Only the second term in Eq. (24) depends upon A and can therefore make a contribution to $\rho(\lambda)$ with a weight that changes across the transition point. With $\mathrm{Im}\,\lambda=-\epsilon$ and x given by the appropriate solution branch in Eq. (19) this gives a contribution

$$\frac{1}{N}\Theta(\alpha-\alpha_c)\delta(\lambda-\lambda_u)-\frac{1}{2N\pi}\Theta(\lambda-\lambda_{min})\Theta(\lambda_{max}-\lambda)$$
$$\times\frac{[2\sigma^2(1+A)-\lambda-\sqrt{\lambda_{max}\lambda_{min}}]}{(\lambda-\lambda_u)\sqrt{(\lambda-\lambda_{min})(\lambda_{max}-\lambda)}}. \quad (25)$$

It is easily verified that this contribution has zero integral, $\forall\,\alpha$, over the entire range of λ. The first term in Eq. (24) also gives a $O(N^{-1})$ correction to the bulk eigenvalue density (20) of the form

$$N^{-1}\left[\frac{1}{4}\delta(\lambda-\lambda_{min})+\frac{1}{4}\delta(\lambda-\lambda_{max})\right.$$
$$\left.-\frac{\Theta(\lambda-\lambda_{min})\Theta(\lambda_{max}-\lambda)}{2\pi\sqrt{(\lambda-\lambda_{min})(\lambda_{max}-\lambda)}}\right]. \quad (26)$$

We find that this has zero integral over the interval $[\lambda_{min},\lambda_{max}]$. The correction in Eq. (26) is essentially of the same form as that derived by Sollich (not within the context of a replica calculation) for the isotropic case [27,28], and similar to the zero integral $O(N^{-1})$ correction to the Wigner semicircle law for the GOE obtained by Dhesi and Jones [29]. The correction in Eq. (26) is divergent at the edges of the bulk density, i.e., at $\lambda=\lambda_{min},\lambda_{max}$. For λ close to the spectral edges at $\lambda_{max},\lambda_{min}$ the stationary points represented by x as given by Eq. (19) are close together. In this region

and at finite values of N critical-like fluctuations lead to poor convergence of the naive perturbative expansion of $\rho(\lambda)$ in powers of N^{-1} [29]. Close to the critical points $\lambda_{min}, \lambda_{max}$ the calculation can be rescaled to yield an expression for $\rho(\lambda)$ that is convergent as $\lambda \to \lambda_{max}, \lambda_{min}$, at large finite values of N [29,30]. Adapting the work of Dhesi and Jones [29] we obtain, away from the transition point $A_c = \alpha^{-1/2}$,

$$\rho(\lambda = \lambda_{max} + \tilde{\delta} N^{-2/3}) \simeq \frac{1}{\pi} N^{-1/3} \left(c \frac{\sqrt{3}}{2} - c^2 \frac{\sqrt{3}}{6} \tilde{\delta} \right), \quad \tilde{\delta} \to 0,$$
(27)

where $c = \alpha \sigma^{-2} (1 + \sqrt{\alpha})^{-4/3}$ and $|A - \alpha^{-1/2}| N^{1/3} \gg 1$.

In the main we are interested in the effect the signal component, specified by A, has upon the observed spectrum, rather than the precise form of the edge of the bulk. Therefore in this study for large finite N the only $O(N^{-1})$ correction to the leading-order result (20) we retain is that due to the isolated contribution at $\lambda = \lambda_u$, and we approximate the observed distribution of eigenvalues of $\hat{\mathbf{C}}$ as

$$\rho(\lambda) = (1 - \alpha) \Theta(1 - \alpha) \delta(\lambda) + \frac{1}{N} \delta(\lambda - \lambda_u) \Theta(\alpha - A^{-2})$$

$$+ [1 - N^{-1} \Theta(\alpha - A^{-2})] \frac{\alpha}{2 \pi \lambda \sigma^2}$$

$$\times \sqrt{\max[0, (\lambda - \lambda_{min})(\lambda_{max} - \lambda)]}.$$
(28)

Although the various terms in Eq. (28) are of different orders, with respect to N, the above approximation for $\rho(\lambda)$ captures the salient features that we are interested in studying. In particular, we regard the transition point at $\alpha = \alpha_c$ as being a phase transition in the expected eigenvalue distribution $\rho(\lambda)$. The contribution $\Theta(\alpha - \alpha_c) \delta(\lambda - \lambda_u)$ has been derived entirely from the $R = 0$ saddle point and no change in the branch of solution to the saddle point equations occurs. Despite this the logarithm of the replica partition function is singular, for finite N, at $\lambda = \lambda_u$ and the transition point coincides with the phase transition observed in the learning of the symmetry-breaking direction $\mathbf{B}$.

It is interesting to observe that all the terms in the approximation (28) to the expected eigenvalue density $\rho(\lambda)$ have been derived by expanding about the saddle point at $R = 0$. At this saddle point $q_1 = 0$, i.e., different replica fields ϕ_ν are uncorrelated. This suggests that the ensemble average of the resolvent may be performed as an annealed average. The annealed average approximates $\langle \ln Z(\lambda) \rangle_\xi$ by $\ln \langle Z(\lambda) \rangle_\xi$, leading to a mean-field free energy $\frac{1}{2} N F_{an}$, with F_{an} given by

$$-F_{an}(q_{an}, R_{an}) = \ln(q_{an} - R_{an}^2)$$

$$- \alpha \ln[1 - \alpha^{-1} \sigma^2$$

$$\times (q_{an} - A R_{an}^2)] - \lambda q_{an}.$$
(29)

Here q_{an} and R_{an} represent the length and overlap with $\mathbf{B}$ of a stochastic vector, and the subscript an is used to denote that these are quantities determined within the context of an an-

nealed average. The free energy in Eq. (29) has a minimum at $R_{an} = 0$. The resulting expression for the leading-order contribution to the eigenvalue density is the same as Eq. (20), and we find an $O(N^{-1})$ correction identical to Eq. (25). However an $O(N^{-1})$ correction twice that of Eq. (26) is also obtained, because the replica-replica overlap $q_{\nu\nu'}$ has non-negligible fluctuations at the saddle point. Thus the annealed average yields an expression for $\rho(\lambda)$, which is very similar but not identical to that obtained from the quenched average, differing in the precise form of the $O(N^{-1})$ corrections. Although surprising, it has previously been observed that the annealed average leads to the correct leading-order result for isotropic data [31]. The use of an annealed average is also implicit within the derivation by Edwards and Jones [20] of the Wigner semicircle law, since terms involving replica-replica correlations are ultimately dropped from their calculation. Within the context of PCA it is somewhat surprising that the annealed average yields a similar expression for $\rho(\lambda)$ as the quenched average, in particular the contribution $\Theta(\alpha - \alpha_c) \delta(\lambda - \lambda_u)$, since within the context of learning for $\alpha > \alpha_c$ one has $|R = \mathbf{J} \cdot \mathbf{B}| > 0$ and replica-replica correlations are nonzero [18,26]. Thus to analyze the performance of learning the actual symmetry-breaking direction, evaluating the quenched average is essential.

B. Simulation results—one symmetry-breaking direction

Simulations were carried out in order to test the validity of our analytical results. We have chosen $N = 2000$ and set $\sigma^2 = 1$ and $\alpha = 0.5$. Figure 1(a) shows the top 20 eigenvalues from a sample covariance matrix generated according to Eq. (14) with $A^2 = 10$. The highest eigenvalue is clearly separated from the general trend followed by the others. The inset shows the empirical distribution of nonzero eigenvalues (except the highest), with the solid line being the theoretical distribution of nonzero eigenvalues obtained from Eq. (4). One can see that the distribution of the bulk of the eigenvalues follows the expected behavior. Figure 1(b) shows the top 20 eigenvalues from a sample covariance matrix generated according to Eq. (14) with $A^2 = 1.5$. The highest eigenvalue follows the same general trend as the others with the bulk of the nonzero eigenvalues being distributed according to Eq. (4)—see inset.

We can study the behavior, with α, of $\Delta = \lambda_1 - \lambda_2$, the gap between the top eigenvalue λ_1 and the next highest eigenvalue λ_2. We have evaluated $\langle \Delta \rangle$ by averaging over 1000 simulations, setting $A^2 = 10$ and $\sigma^2 = 1$. Plotted in Fig. 2(a), is $\ln \langle \Delta \rangle$ against $\ln \alpha$ for a number of different values of N. One can clearly see a change in the behavior of $\langle \Delta \rangle$ as α increases, with $\langle \Delta \rangle$ passing through a minimum at α_{min}. For these finite values of N this minimum does not coincide with the actual predicted transition point of $\alpha_c = 0.1$. This will be due, in part, to the fact that at the transition point α_c the separation between the top eigenvalue λ_1 and the bulk of the spectrum will be small. Indeed, α will need to be increased above α_c before λ_1 exceeds the typical eigenvalue spacing in the upper tail of the bulk. However for increasing N one can see from Fig. 2(a) that α_{min} moves closer to the value of α_c, with $\alpha_{min} \simeq 0.135$ when $N = 1000$. For $\alpha < \alpha_{min}$ the behavior
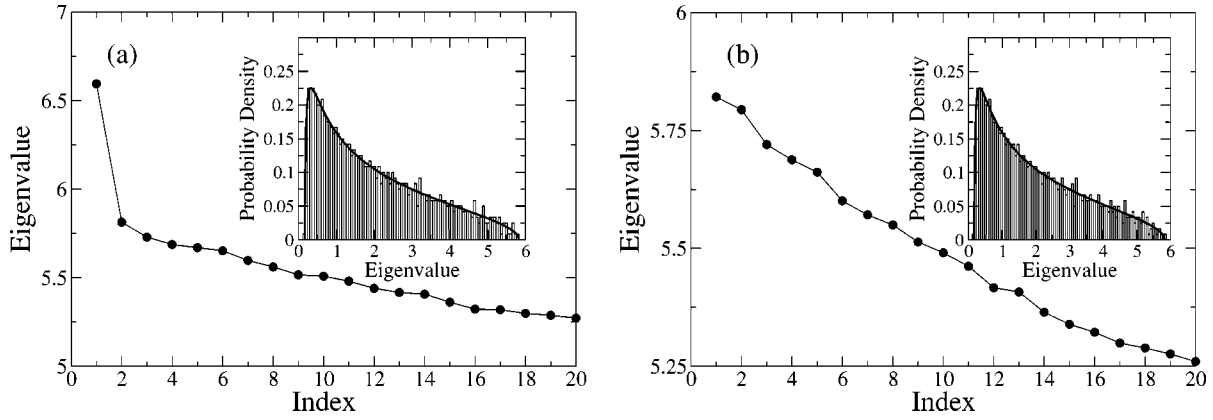
FIG. 1. (a) Plot of eigenvalue λ against rank, for $\sigma^2 = 1$, $N = 2000$, $\alpha = 0.5$, and $A^2 = 10$, so that $\alpha > \alpha_c$. Eigenvalues are calculated from a single instance of the sample covariance matrix $\hat{\mathbf{C}}$. The inset shows the distribution of all nonzero eigenvalues except the largest. The solid line corresponds to the distribution of nonzero eigenvalues, obtained from Eq. (4), when $\mathbf{C}$ is isotropic. (b) As for (a) except that $A^2 = 1.5$ so that $\alpha < \alpha_c$. The inset shows distribution of all nonzero eigenvalues and again the solid line corresponds to the case when $\mathbf{C}$ is isotropic.

appears to be approximately $\langle \Delta \rangle \sim \alpha^{-1}$. Soshnikov [8] has shown that the joint distribution of the eigenvalues of a Wishart matrix converges (as $N \rightarrow \infty$), after a location and scale transformation of each eigenvalue, to that for the GOE. The scale transformation consists of $\lambda \rightarrow \lambda / \sigma_{N,p}$, where

$$\sigma_{N,p} = (N^{1/2} + p^{1/2})(N^{-1/2} + p^{-1/2})^{1/3}. \quad (30)$$

Since for $\alpha < \alpha_c$ the distribution $\rho(\lambda)$ is identical to that when $\mathbf{C}$ is isotropic, we shall assume that the joint distribution is as for the GOE. Thus for large N and below the transition we can expect

$$\langle \Delta \rangle = \text{const} \times \alpha^{-1} (1 + \alpha^{1/2})(1 + \alpha^{-1/2})^{1/3}, \quad \alpha < \alpha_c. \quad (31)$$

Above the transition point we approximate $\langle \Delta \rangle$ as the difference between λ_u, given by Eq. (23), and λ_{max}, the upper limit of the distribution of the bulk of the remaining eigenvalues, i.e.,

$$\langle \Delta \rangle = \sigma^2 (1 + A)\left(1 + \frac{1}{\alpha A}\right) - \sigma^2 (1 + \alpha^{-1/2})^2, \quad \alpha \geq \alpha_c. \quad (32)$$

Figure 2(b) shows the simulation results for $\langle \Delta \rangle$ for $N = 1000$ replotted from Fig. 2(a). Also shown in Fig. 2(b) is the theoretical result from Eqs. (31) and (32). The constant in Eq. (31) has been fitted so that theory and simulation agree for the smallest value of α shown. Below the transition point of $\alpha_c = 0.1$ the agreement between theory and simulation is good, apart from the differences due to finite N around α_c. Above the transition point there are more obvious finite size discrepancies between the simulation results and the theory given by Eq. (32). The theoretical result (32) is constructed as the difference between $\lambda_u(A)$ and λ_{max}. For α just above α_c this will be a poor approximation, but as the top eigenvalue becomes more and more separated from the bulk spectrum, with increasing α, this approximation will become more accurate. The convergence of the simulation result with the theoretical result (32) as α increases can clearly be seen in Fig. 2(b).
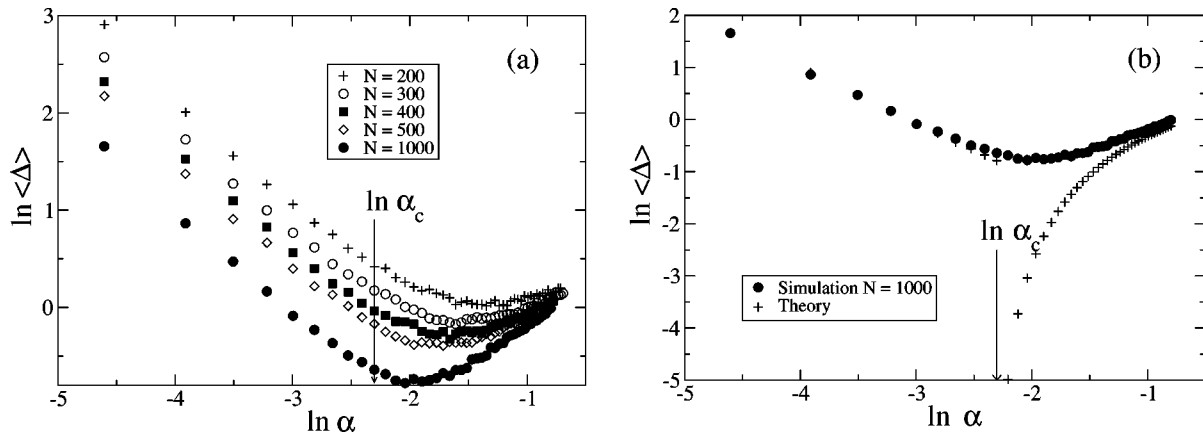

FIG. 2. (a) Plot of $\ln\langle \Delta \rangle$ obtained from simulation averages against $\ln \alpha$ for $\sigma^2 = 1$, $A^2 = 10$, and various values of N. Standard errors associated with each simulation point are typically less than the size of the plotted symbols. (b) Plot of simulation values of $\ln\langle \Delta \rangle$ against $\ln \alpha$ for $\sigma^2 = 1$, $A^2 = 10$, $N = 1000$ (solid circles ●), compared to theoretical values given by Eqs. (31) and (32) (crosses +).
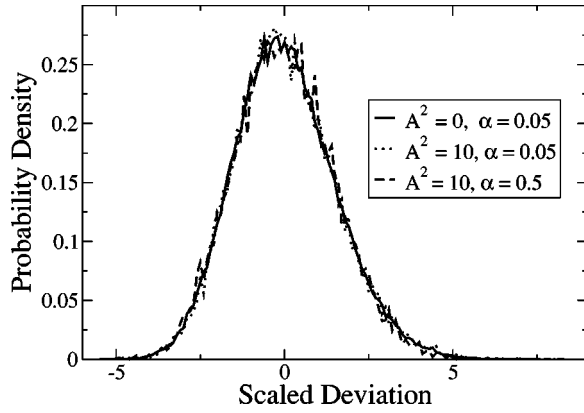
FIG. 3. Plot of distribution of rescaled values of the largest eigenvalue for $\sigma^2 = 1$, $N = 1000$, and $\alpha = 0.05 < \alpha_c = \infty$ (solid line), $\alpha = 0.05 < \alpha_c = 0.1$ (dotted line), $\alpha = 0.5 > \alpha_c = 0.1$ (dashed line).

We can also compare the distribution of the top eigenvalue λ_1 with the expected Tracy-Widom distribution when **C** is isotropic. Plotted in Fig. 3 are distributions of rescaled values of λ_1 obtained from simulation with $N = 1000$ and $\sigma^2 = 1$. For each sample we have centered the median of the sample to 0 and the median absolute deviation (MAD) to 1. Johnstone [7] shows that the variance of the largest eigenvalue of the Wishart matrix $\mathbf{X}\mathbf{X}^T$ scales as $N^{2/3}$ when **C** is isotropic. Thus we expect the variance of the largest eigenvalue of $\hat{\mathbf{C}}$ to scale as $N^{-4/3}$ when **C** is isotropic. When **C** contains a symmetry-breaking direction we cannot *a priori* exclude the possibility that any prefactors to the scaling relation for the variance of λ_1 depend upon A. To compare sample distributions of λ_1 above and below the transition point $\alpha = \alpha_c$ we have therefore applied this location and scale transformation. The solid line (obtained from a sample of 100 000 values) corresponds to $A^2 \equiv 0$ and $p = 50$, i.e., the isotropic case which we expect to correspond to the TW distribution. The dotted line (15 000 values) corresponds to $A^2 = 10$ and $p = 50$, i.e., below the critical value of $\alpha_c = 0.1$, while the dashed line (15 000 values) corresponds to $A^2 = 10$ and $p = 500$, i.e., above α_c. In all cases the probability density is estimated by a histogram with a bin width of 0.1. One can see that the differences between the three distributions of scaled values of λ_1 are small, although statistically significant when testing using the Kuiper statistic [32,33] (a variant of the Kolmogorov-Smirnov test [33]).

C. Multiple symmetry-breaking directions

Consider the case where we have S (orthogonal) symmetry-breaking directions, i.e.,

$$\mathbf{C} = \sigma^2 \mathbf{I} + \sigma^2 \sum_{m=1}^{S} A_m \mathbf{B}_m \mathbf{B}_m^T, \tag{33}$$

where $\mathbf{B}_m \cdot \mathbf{B}_{m'} = \delta_{mm'}$ and without loss of generality we have ordered the eigenvalues such that $A_1 > A_2 > \cdots > A_S > 0$. The saddle point approximation of $Z^n(\lambda)$ when multiple symmetry-breaking directions are present in **C** is straightforward and the appropriate mean-field free energy is

$$-F(q_0, x, \{R_m\}_{m=1}^{S})$$

$$= \ln x + \frac{q_0 - \sum_m R_m^2}{x} - \alpha \ln(1 - \alpha^{-1} \sigma^2 x) + \frac{\alpha \sigma^2 x}{\sigma^2 x - \alpha}$$

$$- \frac{\alpha \sigma^2 \left(q_0 + \sum_m A_m R_m^2 \right)}{\sigma^2 x - \alpha} - \lambda q_0, \tag{34}$$

where q_0 and x are defined as before and R_m is the replica symmetric overlap of the replica fields $\boldsymbol{\phi}_\nu$ with $\mathbf{B}_m$. It is easy to see from the form of the free energy in Eq. (34) that the saddle point behavior of Eq. (34) will be very similar to that in the one symmetry-breaking direction case. Specifically one has a saddle point at $R_m = 0 \forall m$. For $\alpha > A_m^{-2}$ and if $\lambda = \sigma^2 (1 + A_m)(1 + (1/\alpha A_m)) = \lambda_u(A_m)$, this extends to a line of saddle points with $|R_m| > 0, R_{m'} = 0 \forall m' \neq m$. The distribution $\rho(\lambda)$ is now approximated by

$$\rho(\lambda) = (1 - \alpha) \Theta(1 - \alpha) \delta(\lambda) + \frac{1}{N} \sum_{m=1}^{S} \delta(\lambda - \lambda_u(A_m))$$

$$\times \Theta(\alpha - A_m^{-2}) + \left(1 - N^{-1} \sum_{m=1}^{S} \Theta(\alpha - A_m^{-2}) \right)$$

$$\times \frac{\alpha}{2 \pi \lambda \sigma^2} \sqrt{\max[0, (\lambda - \lambda_{min})(\lambda_{max} - \lambda)]}, \tag{35}$$

where $\lambda_u(A)$ is given by Eq. (23). Again the set of transition points, $\alpha = A_m^{-2}$, $m = 1, \ldots, S$ correspond to phase transitions in the equivalent learning problem [34].

Simulation results again confirm the accuracy of Eq. (35). Figure 4(a) shows the top 20 eigenvalues from a sample covariance matrix generated according to Eq. (33) with three symmetry-breaking directions present. We have set $N = 2000$, $\sigma^2 = 1$, $A_1^2 = 20$, $A_2^2 = 15$, and $A_3^2 = 10$. We have also set $\alpha = 0.5$ so that α is above all the transition points. From Fig. 4(a) one can see the top three eigenvalues clearly separated from the remaining bulk, which still follows the predicted distribution of nonzero eigenvalues given by Eq. (4)—see inset. Figure 4(b) shows a plot of $\ln \Delta \lambda$ against $\ln N$, where $\Delta \lambda$ is the fractional difference between the average value $\langle \lambda_i \rangle, i = 1, 2, 3$, of the top three eigenvalues, and the theoretical value $\lambda_u(A_i)$, i.e., $\Delta \lambda_i = [\langle \lambda_i \rangle - \lambda_u(A_i)]/\lambda_u(A_i)$ [with λ_u given by Eq. (23)]. We have set $\alpha = 0.2$ but all other parameters are as in Fig. 4(a). All the average values have been estimated from simulation samples of 1000 values, and error bars in Fig. 4(b) are typically less than the size of the plotted symbols. The differences between the simulation results and the theoretical result given by Eq. (23) are small, but statistically significant. The theoretical result is derived from the leading-order asymptotic contribution to the replica partition function, and so for any finite value of p and N we
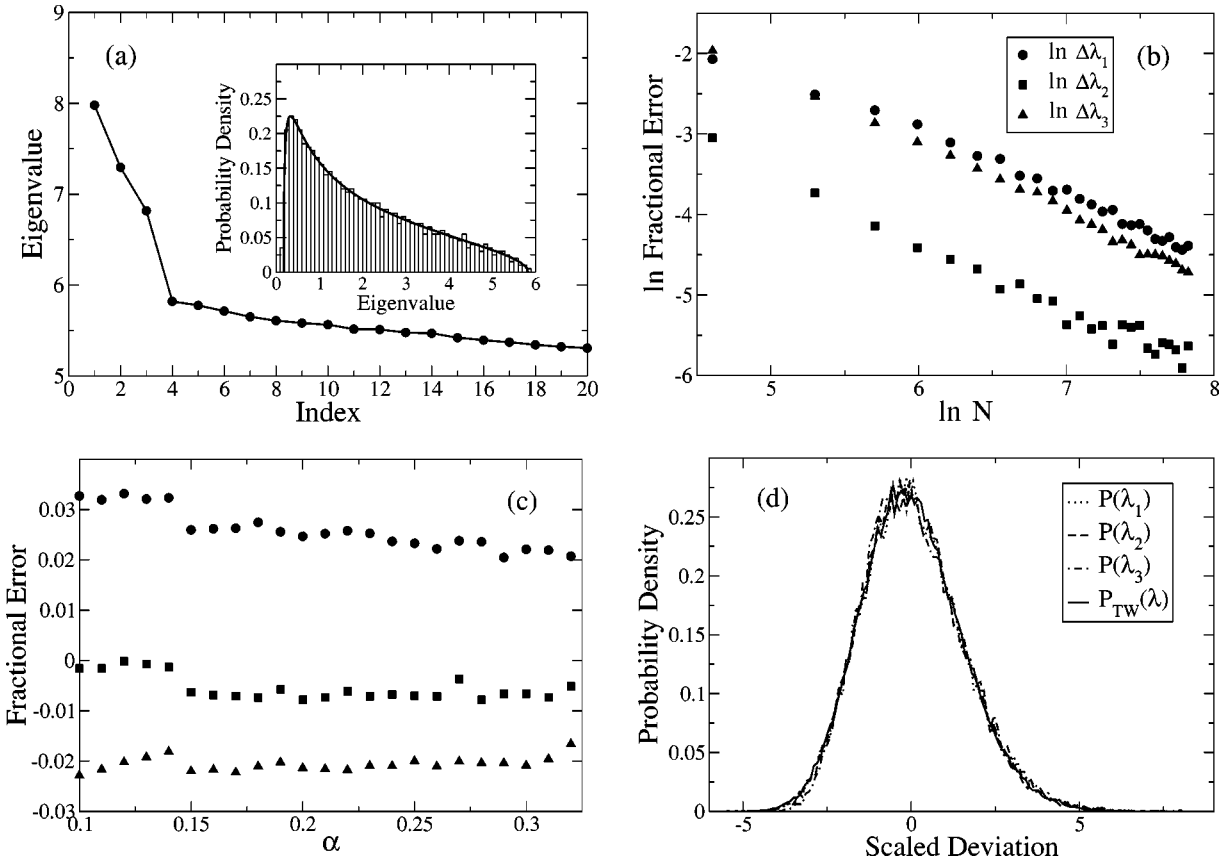
FIG. 4. (a) Plot of eigenvalue against rank for $\sigma^2=1$, $N=2000$, $\alpha=0.5$. Three symmetry-breaking directions are present in the covariance matrix **C**, with $A_1^2=20$, $A_2^2=15$, $A_3^2=10$. Eigenvalues are calculated from a single instance of the sample covariance matrix $\hat{\mathbf{C}}$. The inset shows the distribution of all nonzero eigenvalues except the largest three. The solid line corresponds to the distribution, obtained from Eq. (4), of nonzero eigenvalues when **C** is isotropic. (b) Log-Log plot, with increasing $\ln N$, of the difference between simulation values of $\langle\lambda_i\rangle$ and $\lambda_u(A_i)$ given by Eq. (23) for $i=1,2,3$. We have set $\alpha=0.2$ but all other parameter values are the same as for (a). (c) As for (b) but with $N=1000$ and α increasing. (d) Distributions of top three eigenvalues (after median and MAD location and scale transformations—see main text for details).

would expect finite size corrections. However, as expected, $|\Delta\lambda_i|$ decreases as $N\rightarrow\infty$. Figure 4(b) confirms the asymptotic accuracy of Eq. (23) at fixed α. In some applications (see, for example, Ref. [26]) we are interested in how many samples are required for good accuracy, in which case we may be more interested in fixing N and varying α. Figure 4(c) shows a plot of $\Delta\lambda_i$ against α. We have set $N=1000$, with all other parameters as in Fig. 4(a). All the expectation values have been obtained from simulation samples of 1000 values, and error bars in Fig. 4(c) correspond to ± 1 standard error. We have only plotted results for $\alpha>0.1$, i.e., the largest of the transition points. Again the differences between the simulation results and theory are small, but statistically significant.

The distributions of the largest eigenvalues would again appear to converge to the same distribution (up to location and scale transformations). Figure 4(d) shows a plot of the distributions of the largest three eigenvalues. Again median and MAD location and scale transformations have been applied for all three eigenvalues. Here we have set $N=1000$ and $\alpha=0.2$, while all other parameters are as for Fig. 4(a). The distributions are histograms of bin width 0.1, evaluated from a sample of 30 000 values. The solid line in Fig. 4(d)

represents the simulation estimate of the TW distribution, replotted from Fig. 3. The similarity between all four distributions is striking, suggesting that there is a universal limiting two-parameter family of distributions, irrespective of whether any symmetry-breaking directions are present in **C**. As for the one symmetry-breaking direction case, at these finite values of N the four different sample distributions are statistically significantly different from each other (again using the Kuiper statistic).

IV. ANALYSIS OF THE STIELTJES TRANSFORM $m_\rho(Z)$

Equivalent results to those we have derived using replicas must be implicit within the relation (6) for the Stieltjes transform, $m_\rho(z)$, of $\alpha^{-1}\rho(\lambda)$, particularly since Sengupta and Mitra [23] derived an equivalent result to Eq. (6) from a replica formulation of the resolvent of $\hat{\mathbf{C}}$. Solving for $m_\rho(z)$ in closed form is not necessarily straightforward, however it is instructive to confirm that the results derived from the replica analysis can be obtained from Eq. (6). We can start from the expression (6) for the Stieltjes transform $m_\rho(z)$. For $dH(t)=\delta(t-\sigma^2)dt$, corresponding to **C** being isotropic, the usual distribution of eigenvalues results [14]. Silverstein and
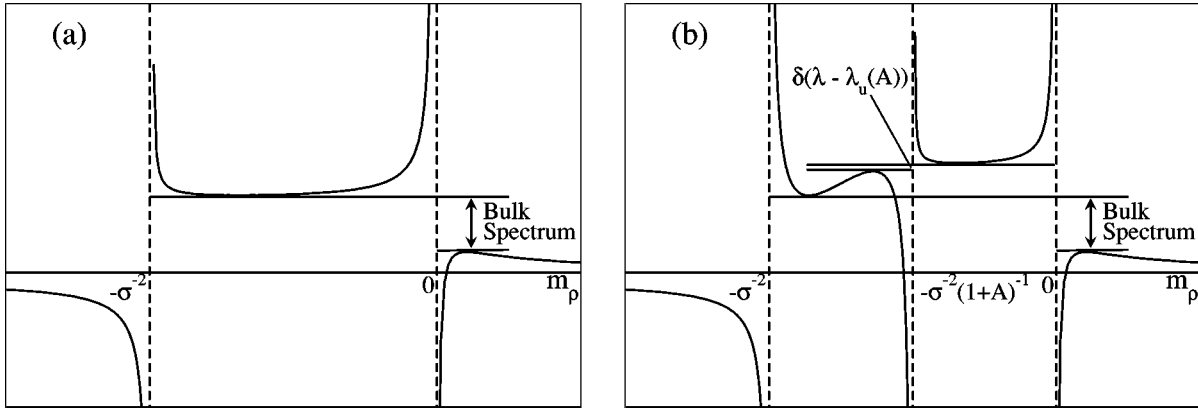
FIG. 5. (a) Schematic plot of the behavior of $z(m_\rho)$ when the true covariance matrix $\mathbf{C}$ is isotropic. (b) Schematic plot of the behavior of $z(m_\rho)$ when one symmetry-breaking direction, of strength A, is present in the true covariance matrix $\mathbf{C}$ and $\alpha > A^{-2}$.

Choi [35] show that the support of $\rho(\lambda)$ can be determined from the intervals between extrema of $z(m_\rho)$. This approach has been used, in particular by Silverstein and Combettes [36], to determine the signal component of a spectrum when a large number variables are present. For the case of $\mathbf{C}$ isotropic, if we consider $dH(t) = \delta(t - \sigma^2)dt$, a straightforward calculation yields a single interval, $[\lambda_{min}, \lambda_{max}]$, for the support of $\rho(\lambda)$. Figure 5(a) shows a schematic plot of the behavior of $z(m_\rho)$ when $\mathbf{C}$ is isotropic.

For the case of a single symmetry-breaking direction present in $\mathbf{C}$ we take $dH(t) = (1-\epsilon)\delta(t-\sigma^2)dt + \epsilon\delta(t - \sigma^2(1+A))dt$, with $\epsilon \simeq 1/N$. This gives

$$z(m_\rho) = -\frac{1}{m_\rho} + \frac{(1-\epsilon)\alpha^{-1}}{\sigma^{-2}+m_\rho} + \frac{\epsilon\alpha^{-1}}{\sigma^{-2}(1+A)^{-1}+m_\rho},$$ (36)

and stationary points satisfy

$$0 = \frac{1}{m_\rho^2} - \frac{(1-\epsilon)\alpha^{-1}}{(\sigma^{-2}+m_\rho)^2} - \frac{\epsilon\alpha^{-1}}{[\sigma^{-2}(1+A)^{-1}+m_\rho]^2}.$$ (37)

Since $\epsilon \ll 1$ we do not expect the behavior of $z(m_\rho)$ to be modified substantially in the interval $[\lambda_{min}, \lambda_{max}]$. Therefore we look for additional stationary points close to the singularity at $m_\rho = -\sigma^{-2}(1+A)^{-1}$. Setting $m_\rho = -\sigma^{-2}(1+A)^{-1} + \delta$ and expanding Eq. (37) yields

$$\delta = \pm\frac{\epsilon^{1/2}}{\sigma^2(1+A)\sqrt{(\alpha-A^{-2})}} + O(\epsilon).$$ (38)

Substituting Eq. (38) into Eq. (36) yields $z(-\sigma^{-2}(1+A)^{-1}+\delta) = \sigma^2(1+A)(1+(\alpha A)^{-1}) + O(\epsilon^{1/2})$. Thus, as $N \to \infty$, if the stationary points at $-\sigma^{-2}(1+A)^{-1}+\delta$ exist, they will define a small interval of z centered on $\lambda_u(A)$ and so define an approximate contribution of $N^{-1}\delta(\lambda - \lambda_u(A))$

to the observed distribution of eigenvalues of $\hat{\mathbf{C}}$, in agreement with the previous calculations using replicas. From Eq. (38) we see that for δ to be real requires $\alpha > A^{-2}$, in agreement with our previous calculations. Figure 5(b) shows a schematic plot of the behavior of $z(m_\rho)$ for the symmetry broken $\mathbf{C}$.

A similar perturbative analysis when $\mathbf{C}$ contains more than one symmetry-breaking direction gives a set of contributions $N^{-1}\delta(\lambda - \lambda_u(A_m)), m=1,\ldots,S$, to $\rho(\lambda)$. Again this is in agreement with our previous replica analysis of the resolvent.

The relationship (6) can be obtained with only very weak convergence conditions placed upon the elements of the data matrix $\mathbf{X}$, and we essentially require only the second moment of each of the elements of $\hat{\mathbf{C}}$ to exist. Indeed Bai [37] gives convergence rates for the Stieltjes transform for a variety of differing constraints upon the moments of the elements of $\mathbf{X}$. This suggests that the results from the perturbative analysis of the Stieltjes transform, and therefore analysis of the replica saddle point equations maybe more general, extending beyond the case studied here where the sample vectors $\boldsymbol{\xi}_\mu$ are drawn from a multivariate Gaussian distribution. To test this we consider the case where the signal and noise components of the data vectors are drawn from the same zero mean, unit variance distribution $f(x)$, i.e.,

$$P(\boldsymbol{\zeta}) = \delta\left(\|\boldsymbol{\zeta}\|^2 - \sum_{m=1}^{S}(\boldsymbol{\zeta}\cdot\mathbf{B}_m)^2\right)\prod_{m=1}^{S}\frac{1}{\sigma\sqrt{A_m}}f(\boldsymbol{\zeta}\cdot\mathbf{B}_m/\sigma\sqrt{A_m}),$$ (39)

$$P(\mathbf{e}) = \prod_{i=1}^{N}\sigma^{-1}f(e_i/\sigma),$$ (40)

where e_i is the ith component of $\mathbf{e}$ and $f(x)$ satisfies $\int f(x)dx = 0$, $\int x^2 f(x)dx = 1$. We have performed simulations using three different distributions,
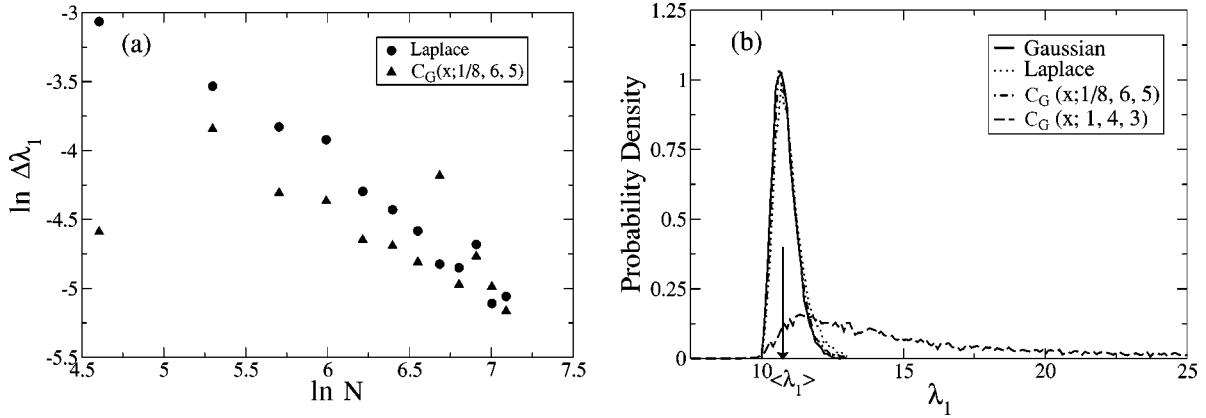
FIG. 6. (a) Log-Log plot of the fractional error $\Delta\lambda_1$ of the theoretical estimate of the top eigenvalue for the Laplace distribution (solid circles ●) and the generalized Cauchy distribution $C_G(x;\frac{1}{8},6,5)$ (up triangles ▲). Here we have fixed $A^2 = 10$, $\sigma^2 = 1$, and $\alpha = 0.2$. (b) Plot of the distribution of λ_1 for a number of different distributions $f(x)$. We have fixed $A^2 = 10$, $\alpha = 0.2$, and $N = 500$. The arrow indicates the value of $\langle\lambda_1\rangle$ given by Eq. (23).

$$f(x) = \begin{cases} \dfrac{1}{\sqrt{2}}\exp(-\sqrt{2}|x|), & \text{(i)} & \text{Laplace} \\[2mm] \dfrac{3}{2\pi\sqrt{2}}\left(1+\dfrac{1}{8}x^6\right)^{-1}, & \text{(ii)} & \text{generalized Cauchy } C_G\left(x;\dfrac{1}{8},6,5\right) \\[2mm] \dfrac{\sqrt{2}}{\pi}(1+x^4)^{-1}, & \text{(iii)} & \text{generalized Cauchy } C_G(x;1,4,3), \end{cases} \tag{41}$$

where $C_G(x;a,c,\nu) \sim (1+a|x|^c)^{-(1+\nu)/c}$ is the generalized Cauchy distribution [38,39]. For distribution (i) all moments of each element of $\hat{\mathbf{C}}$ exist, while for distribution (ii) the second moment exists, and for distribution (iii) only the first moment of each element of $\hat{\mathbf{C}}$ exists. For simplicity we have only simulated the one symmetry-breaking direction case. We have set $A^2 = 10$ ($\alpha_c = 0.1$) and $\sigma^2 = 1$. Figure 6(a) shows a log-log plot of simulation results (1000 values) for the fractional error $\Delta\lambda_1$ of the largest eigenvalue for distributions (i) and (ii). In the simulations we have fixed $\alpha = 0.2 > \alpha_c$. The fractional error is clearly decreasing with increasing N, confirming the accuracy of the asymptotic theory given by Eq. (23), and has approximately the same behavior irrespective of which distribution is used. Noticeable is the difference in $\Delta\lambda_1$ from the two distributions at the smallest value of N shown. Typical standard errors (not shown) in $\Delta\lambda_1$ for the Laplace distribution are 10–20 % of the plotted value, while for the generalized Cauchy distribution $C_G(x;\frac{1}{8},6,5)$, typical standard errors are 20–40 % of the plotted values. For the generalized Cauchy distribution $C_G(x;1,4,3)$ the mean simulation value of λ_1 is distinctly different from that predicted by Eq. (23). Figure 6(b) shows distributions of λ_1 for the different choices of $f(x)$: Gaussian (solid line), Laplace (dotted line), $C_G(x;\frac{1}{8},6,5)$ (dot-dashed line), and $C_G(x;1,4,3)$ (dashed line). We have fixed $A^2 = 10$, $\alpha = 0.2$, and $N = 500$. Distributions of λ_1 are con-

structed as histograms of 10 000 simulation values. The distribution of λ_1 obtained from the Laplace distribution and $C_G(x;\frac{1}{8},6,5)$ are similar to that obtained from the Gaussian case. However the distribution of λ_1 obtained from $C_G(x;1,4,3)$ shows a distinct heavy tail.

V. DISCUSSION AND CONCLUSIONS

Evaluation of the resolvent of the sample covariance matrix $\hat{\mathbf{C}}$ using the replica method has enabled us to determine the distribution of eigenvalues, $\rho(\lambda)$, of $\hat{\mathbf{C}}$ in the limit $N \to \infty$, with $p = \alpha N$ for some fixed value of α. Most existing studies consider the case where the elements of the data matrix $\mathbf{X}$ are i.i.d., the idea being that such a model serves to provide a null hypothesis for $\rho(\lambda)$ against which to test the largest observed eigenvalue of real data sets. However the ultimate motivation for using PCA is to apply it to data sets that have some definite structure. Thus it is instructive to consider the expected eigenvalue distribution when definite structure is genuinely present in the data, so that bias in eigenvalue estimates can be examined, and the effect of repeatedly using the i.i.d. case as a null model for lower eigenvalues determined. To this end we have used the replica method to obtain the distribution $\rho(\lambda)$ for a particular case where the elements of $\mathbf{X}$ are not i.i.d. Specifically, when the population covariance matrix $\mathbf{C}$ contains a number of symmetry-breaking directions then phase transitions in the

eigenvalue spectrum $\rho(\lambda)$ are observed. With each transition a single eigenvalue detaches from the upper edge of the bulk of the spectrum. The bulk of the distribution $\rho(\lambda)$ is still given by Eq. (4), i.e., is identical to that for the case when **C** is isotropic. This transition behavior is similar to that observed in the original replica analysis of a random matrix by Edwards and Jones [20], although in that case the phase transition was as a result of varying the mean of the Gaussian distribution from which the matrix elements were drawn.

When **C** is nonisotropic the distribution $\rho(\lambda)$ can in theory be determined from its Stieltjes transform, αm_ρ, for which an integral equation has been derived by Marčenko and Pastur [14] and also Wachter [16]. Indeed an equivalent result for the Stieltjes transform m_ρ has been derived using replicas by Sengupta and Mitra [23]. In practice, even when the eigenvalues of **C** are restricted to a finite number of localized values, the solution of the resolvent or Stieltjes transform in closed form involves solution of at least a third-degree polynomial. When **C** contains a finite number of symmetry-breaking directions, determination of the support of $\rho(\lambda)$ through a perturbative analysis of the integral equation controlling m_ρ is possible, and the results agree with those obtained from explicit solution of the saddle point equations of the replica partition function. In contrast to Sengupta and Mitra [23], explicit solution of the replica saddle point equations reveals the form of the bulk of $\rho(\lambda)$ and the isolated contributions from the symmetry-breaking directions present in **C**. However if the number of symmetry-breaking directions, S, is also extensive in N, then the saddle point evaluation of the replica partition function given in the Appendix is no longer valid. In these circumstances the distribution $\rho(\lambda)$ must be obtained by analysis of its Stieltjes transform, and one finds that for sufficiently large α the spectrum $\rho(\lambda)$ is composed exactly of a separate portion due to the noise and a separate portion due to the symmetry-breaking directions [36,40,41].

The fact that the location of the mth transition and the locus (with A_m) of the mth eigenvalue does not depend upon the values of the other parameters $\{A_{m'}\}_{m' \neq m}$ suggests that testing each of the m top eigenvalues separately against a Tracy-Widom distribution, as Johnstone [7] does, will correctly select those eigenvalues which are due to genuine symmetry-breaking directions in **C**. The distribution of the largest eigenvalue of $\hat{\mathbf{C}}$ would appear to be universal (up to a location and scale transformation), irrespective of the value of A_1.

The explicit result for the expectation value $\lambda_u(A_m) = \sigma^2(1 + A_m)[1 + (\alpha A_m)^{-1}]$ of the mth eigenvalue of $\hat{\mathbf{C}}$ (when $\alpha > A_m^{-2}$) allows us to correct for the bias due to having a finite number of samples p. Thus for those eigenvalues which we select as being not just the result of the additive noise σ^2, we can obtain a much more accurate estimate of the true underlying eigenvalue. One should note that the bias is nontrivial. The fractional error (in the limit $N \rightarrow \infty$) in the estimate of the mth largest eigenvalue is predicted to be $(\alpha A_m)^{-1}$. Even if one is above the transition point ($\alpha > A_m^{-2}$) for this eigenvalue to be detected as separated from the bulk, this only bounds the fractional error in the eigen-

value estimate to less than A_m.

Sample covariance matrix spectra also impact upon other areas in the study of learning, to which the techniques of statistical physics have been applied. For example, the eigenvalues of the sample covariance matrix $\hat{\mathbf{C}}$ determine the optimal learning rate within large linear perceptrons [27,28,31,42,43]. Using the self-averaging property of tr**G** but without specific recourse to replicas, a number of authors obtain an equivalent result to the Stieltjes transform relationship (6) [27,28,43], which can be iteratively solved to obtain the perturbative $O(N^{-1})$ corrections to the bulk density $\rho(\lambda)$. Sollich [27,28] has analyzed finite size effects in the isotropic case and derived essentially the same form as Eq. (26) for the $O(N^{-1})$ corrections. This approach has been extended by Halkjær and Winther [43] to the case when **C** contains a definite signal. A similar form to Eq. (28) for $\rho(\lambda)$ is derived, containing an isolated contribution $\delta(\lambda - \sigma^2(1 + A))$. The location of the δ function differs from $\lambda_u(A)$. However the derivation by Halkjær and Winther essentially considers the large signal limit $A \rightarrow \infty$ and so unsurprisingly agrees with $\lambda_u(A)$ only in this limit. One can confirm that in general the last term of Eq. (10) of Halkjær and Winther [43] has a pole at $\lambda = \lambda_u(A)$, and that the derivation of Halkjær and Winther can be extended to the case when **C** contains more than one symmetry-breaking direction. However to the best of our knowledge the current work represents the first derivation of Eq. (28) using a replica calculation.

APPENDIX

We wish to evaluate the ensemble average

$$\text{tr}\mathbf{G}(\lambda) = -\frac{\partial}{\partial \lambda} \langle \ln Z(\lambda) \rangle_\xi. \tag{A1}$$

We start with the replica partition function $\mathcal{Z} = (2\pi)^{Nn/2} \langle \Pi_{\nu=1}^n \exp[\frac{1}{2}\ln Z(\lambda)] \rangle_\xi$, which can be rewritten as

$$\mathcal{Z} = (2\pi)^{-Np/2}(\det \mathbf{C})^{-p/2} \int \prod_{\mu=1}^p d\boldsymbol{\xi}_\mu$$

$$\times \exp\left[-\frac{1}{2}\sum_\mu \boldsymbol{\xi}_\mu^T \mathbf{C}^{-1} \boldsymbol{\xi}_\mu\right] \int \prod_{\nu=1}^n d\boldsymbol{\phi}_\nu$$

$$\times \exp\left[-\frac{\lambda}{2}\sum_\nu \|\boldsymbol{\phi}_\nu\|^2 + \frac{1}{2p}\sum_{\nu,\mu}(\boldsymbol{\phi}_\nu \cdot \boldsymbol{\xi}_\mu)^2\right], \tag{A2}$$

from which we have

$$\text{tr } \mathbf{G}(\lambda) = -2\frac{\partial}{\partial\lambda}\lim_{n\rightarrow 0}\frac{\partial}{\partial n}\mathcal{Z} = \lim_{n\rightarrow 0}\frac{\partial}{\partial n}\left\langle \sum_\nu \|\boldsymbol{\phi}_\nu\|^2\right\rangle, \tag{A3}$$

where we have interchanged the order of differentiation to obtain the second expression, and the expectation value is with respect to the integrand in (A2). After integrating over $\boldsymbol{\xi}_\mu$ we obtain

$$\mathcal{Z} = \int \prod_{\nu=1}^{n} d\boldsymbol{\phi}_\nu \exp\left[-\frac{\lambda}{2} \sum_\nu \|\boldsymbol{\phi}_\nu\|^2 \right] (\det \mathbf{M})^{-p/2},$$

$$\text{(A4)}$$

where (for $\mathbf{C} = \sigma^2 \mathbf{I} + \sigma^2 A \mathbf{B}\mathbf{B}^T$)

$$M_{\nu\nu'} = \delta_{\nu\nu'} - \sigma^2 \alpha^{-1}(q_{\nu\nu'} + A R_\nu R_{\nu'}),$$

$$R_\nu = N^{-1/2} \boldsymbol{\phi}_\nu \cdot \mathbf{B},$$

$$q_{\nu\nu'} = N^{-1} \boldsymbol{\phi}_\nu \cdot \boldsymbol{\phi}_{\nu'}. \qquad \text{(A5)}$$

The integrations over $\boldsymbol{\phi}_\nu$ are performed in terms of integrations over R_ν and $q_{\nu\nu'}$, i.e.,

$$\int \prod_\nu d\boldsymbol{\phi}_\nu = \int \prod_\nu d\boldsymbol{\phi}_\nu \int \prod_\nu dR_\nu \delta(R_\nu - N^{-1/2}\boldsymbol{\phi}_\nu \cdot \mathbf{B})$$

$$\times \int \prod_{\substack{\nu \\ \nu' \geq \nu}} dq_{\nu\nu'} \delta(q_{\nu\nu'} - N^{-1}\boldsymbol{\phi}_\nu \cdot \boldsymbol{\phi}_{\nu'}).$$

$$\text{(A6)}$$

We rewrite δ functions in terms of their Fourier representations, with $\hat{R}_\nu$ and $\hat{q}_{\nu\nu'}$ being the Fourier variable conjugates to R_ν and $q_{\nu\nu'}$, respectively. After integrating over $\boldsymbol{\phi}_\nu$ we have

$$\mathcal{Z} = (2\pi)^{(N-n-3)n/2} \int \prod_\nu dR_\nu d\hat{R}_\nu$$

$$\times \int \prod_{\substack{\nu \\ \nu' \geq \nu}} dq_{\nu\nu'} d\hat{q}_{\nu\nu'} \exp(NT), \qquad \text{(A7)}$$

where the exponent T is given by

$$T = -\frac{\lambda}{2} \sum_\nu q_{\nu\nu} - \frac{\alpha}{2} \text{tr} \ln \mathbf{M} + i \sum_\nu R_\nu \hat{R}_\nu - \frac{1}{2} \sum_{\nu,\nu'} \hat{Q}_{\nu\nu'} q_{\nu'\nu}$$

$$- \frac{1}{2} \sum_{\nu,\nu'} \hat{R}_\nu (\hat{\mathbf{Q}}^{-1})_{\nu\nu'} \hat{R}_{\nu'} - \frac{1}{2} \text{tr} \ln \hat{\mathbf{Q}}. \qquad \text{(A8)}$$

The matrix $\hat{\mathbf{Q}}$ has elements $\hat{Q}_{\nu\nu'} = i\hat{q}_{\nu\nu'}$. To obtain the leading-order contribution to the eigenvalue density $\rho(\lambda)$ we integrate over the Fourier variables $\hat{R}_\nu$ and $\hat{q}_{\nu\nu'}$, and retain only terms in the exponent of the integrand that are extensive in N. We also drop any irrelevant prefactors This gives

$$\mathcal{Z} \approx \int \prod_\nu dR_\nu \int \prod_{\substack{\nu \\ \nu' \geq \nu}} dq_{\nu\nu'}$$

$$\times \exp\left(N\left[-\frac{\lambda}{2} \sum_\nu q_{\nu\nu} - \frac{\alpha}{2} \text{tr} \ln \mathbf{M} + \frac{1}{2} \text{tr} \ln \mathbf{L} \right] \right),$$

$$\text{(A9)}$$

where $L_{\nu\nu'} = i(q_{\nu\nu'} - R_\nu R_{\nu'})$. We now look for saddle points of the exponent of the integrand in Eq. (A9). Assuming replica symmetry for such saddle points we put

$$R_\nu = R \; \forall \; \nu,$$

$$q_{\nu\nu} = q_0 \; \forall \; \nu,$$

$$q_{\nu\nu'} = q_1 \; \forall \; \nu, \nu' \neq \nu. \qquad \text{(A10)}$$

Setting $x = q_0 - q_1$, the exponent of the integrand becomes $-\frac{1}{2} NnF(q_0, x, R)$, with

$$-F(q_0, x, R) = \ln x + \frac{q_0 - R^2}{x} - \alpha \ln(1 - \alpha^{-1}\sigma^2 x) + \frac{\alpha\sigma^2 x}{\sigma^2 x - \alpha}$$

$$- \frac{\alpha\sigma^2(q_0 + AR^2)}{\sigma^2 x - \alpha} - \lambda q_0. \qquad \text{(A11)}$$

We then seek to find extrema of (A11) with respect to R, q_0, and q_1. When more than one symmetry-breaking direction is present in $\mathbf{C}$, i.e., $\mathbf{C} = \sigma^2 \mathbf{I} + \sigma^2 \sum_m A_m \mathbf{B}_m \mathbf{B}_m^T$, it is an easy matter to confirm that

$$M_{\nu\nu'} = \delta_{\nu\nu'} - \sigma^2 \alpha^{-1}\left(q_{\nu\nu'} + \sum_m A_m R_{m,\nu} R_{m,\nu'} \right),$$

$$L_{\nu\nu'} = i\left(q_{\nu\nu'} - \sum_m R_{m,\nu} R_{m,\nu'} \right), \qquad \text{(A12)}$$

where $R_{m,\nu} = N^{-1/2} \boldsymbol{\phi}_\nu \cdot \mathbf{B}_m$. Assuming replica symmetry so that $R_{m,\nu} = R_m \; \forall \; \nu$ the exponent of the integrand in Eq. (A9) becomes $-\frac{1}{2} NnF(q_0, x, \{R_m\}_{m=1}^{S})$ with $-F(q_0, x, \{R_m\}_{m=1}^{S})$ given by Eq. (34).

Hessian for $R=0$ saddle point

The free-energy expression in Eq. (A11) has a stationary point (see main text) at $R=0, q_0 = x$ with x, given by Eq. (19). To evaluate $O(N^{-1})$ contributions to $\rho(\lambda)$ we need to evaluate the Hessian of T given by Eq. (A8). For the saddle point at $R=0$ evaluation of the Hessian $\mathbf{H}$ is straightforward and follows closely that for the perceptron problem [1]. The Hessian takes a block diagonal form

$$\mathbf{H} = \begin{pmatrix} \begin{matrix} \mathbf{H}_{Q1} \mathbf{H}_{Q2} \\ \mathbf{H}_{Q3} \mathbf{H}_{Q4} \end{matrix} & \mathbf{0} \\ \mathbf{0} & \begin{matrix} \mathbf{H}_{R1} \mathbf{H}_{R2} \\ \mathbf{H}_{R3} \mathbf{H}_{R4} \end{matrix} \end{pmatrix}. \qquad \text{(A13)}$$

The matrices $\mathbf{H}_{Q1}, \dots, \mathbf{H}_{Q4}, \mathbf{H}_{R1}, \dots, \mathbf{H}_{R4}$ are diagonal,

$$\mathbf{H}_{Q1} = \begin{pmatrix} \frac{1}{2} x^2 \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & 2x^2 \mathbf{I}_{n(n-1)/2} \end{pmatrix},$$

$$\mathbf{H}_{Q2}=\mathbf{H}_{Q3}=\begin{pmatrix}\dfrac{1}{2}\mathbf{I}_n & \mathbf{0}\\[2mm] \mathbf{0} & 2\mathbf{I}_{n(n-1)/2}\end{pmatrix},$$

$$\mathbf{H}_{Q4}=\begin{pmatrix}\dfrac{\sigma^4\alpha^{-1}}{2(1-\sigma^2\alpha^{-1}x)^2}\mathbf{I}_n & \mathbf{0}\\[4mm] \mathbf{0} & \dfrac{2\sigma^4\alpha^{-1}}{(1-\sigma^2\alpha^{-1}x)^2}\mathbf{I}_{n(n-1)/2}\end{pmatrix},$$

$$\mathbf{H}_{R1}=\frac{A\sigma^2}{1-\sigma^2\alpha^{-1}x}\mathbf{I}_n,$$

$$\mathbf{H}_{R2}=\mathbf{H}_{R3}=i\mathbf{I}_n,$$

$$\mathbf{H}_{R4}=-x\mathbf{I}_n,$$

from which we have, after some simplification, two contributions to det $\mathbf{H}$,

$$\det\mathbf{H}=4^{-n}[\alpha^{-1}(\lambda x-1)^2-1]^{n(n+1)/2}\times[1+A-A\lambda x]^n, \tag{A14}$$

where x is given by Eq. (19).

When multiple symmetry-breaking directions are present in $\mathbf{C}$ the Hessian evaluated at the saddle point corresponding to $R_m=0, m=1,\ldots,S$, still takes a simple block diagonal form and we find

$$\det\mathbf{H}=4^{-n}[\alpha^{-1}(\lambda x-1)^2-1]^{n(n+1)/2}$$
$$\times\prod_{m=1}^{S}[1+A_m-A_m\lambda x]^n. \tag{A15}$$

[1] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[2] I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).

[3] M.E. Tipping and C. Bishop, J. R. Stat. Soc. Ser. B. Methodol. **61**, 611 (1999).

[4] M.E. Tipping and C. Bishop, Neural Comput. **11**, 443 (1999).

[5] K. W. Wachter, in *Proceedings of the Ninth Interface Symposium Computer Science and Statistics*, edited by David C. Hoaglin and Roy E. Welsch (Prindle, Weber and Schmidt, Boston, 1976), p. 299.

[6] T.W. Anderson, Ann. Math. Stat. **34**, 122 (1963).

[7] I.M. Johnstone, Ann. Stat. **29**, 295 (2001).

[8] A. Soshnikov, J. Stat. Phys. **108**, 1033 (2002).

[9] J. Wishart, Biometrika **20**, 32 (1928).

[10] R.A. Janik and M.A. Nowak, J. Phys. A **36**, 3629 (2003).

[11] M. L. Mehta, *Random Matrices* (Academic Press, San Diego, 1991).

[12] Z.D. Bai, Statistica Sinica **9**, 611 (1999).

[13] P.J. Forrester, N.C. Snaith, and J.J.M. Verbaarschot, J. Phys. A **36**, R1 (2003).

[14] V.A. Marčenko and L.A. Pastur, Math. USSR. Sb. **1**, 507 (1967).

[15] A. Edelman, SIAM J. Matrix Anal. Appl. **9**, 543 (1988).

[16] K.W. Wachter, Ann. Probab. **6**, 1 (1978).

[17] C.A. Tracy and H. Widom, Commun. Math. Phys. **177**, 727 (1996).

[18] P. Reimann, C. Van den Broeck, and G.J. Bex, J. Phys. A **29**, 3521 (1996).

[19] P. Reimann and C. Van den Broeck, Phys. Rev. E **53**, 3989 (1996).

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

[21] G.J. Rodgers and A.J. Bray, Phys. Rev. B **37**, 3557 (1988).

[22] H.J. Sommers, A. Crisanti, H. Sompolinsky, and Y. Stein, Phys. Rev. Lett. **60**, 1895 (1988).

[23] A.M. Sengupta and P.P. Mitra, Phys. Rev. E **60**, 3389 (1999).

[24] G. M. Cicuta, in *Random Matrices and Their Applications*, edited by P. M. Bleher and A. R. Its, Mathematical Sciences Research Institute Publications Vol. 40 (Cambridge University Press, Cambridge, 2001), p. 95.

[25] I. Derényi, T. Geszti, and G. Györgyi, Phys. Rev. E **50**, 3192 (1994).

[26] D.C. Hoyle and M. Rattray, Europhys. Lett. **62**, 117 (2003).

[27] P. Sollich, J. Phys. A **27**, 7771 (1994).

[28] P. Sollich, in *Advances in Neural Information Processing Systems 7*, edited by G. Tesauro, D. S. Touretzky, and T. K. Leen (MIT Press, Cambridge, MA 1995), p. 207.

[29] G.S. Dhesi and R.C. Jones, J. Phys. A **23**, 5577 (1990).

[30] J.J.M. Verbaarschot and M.R. Zirnbauer, Ann. Phys. (N.Y.) **158**, 78 (1984).

[31] M. Opper, Europhys. Lett. **8**, 389 (1989).

[32] N.H. Kuiper, Proc. K. Ned. Akad. Wet., Ser. A: Math. Sci. **63**, 38 (1962).

[33] M.A. Stephens, J. R. Stat. Soc. Ser. B. Methodol. **32**, 115 (1970).

[34] D. C. Hoyle and M. Rattray (unpublished).

[35] J.W. Silverstein and S. Choi, J. Multivariate Anal. **54**, 295 (1995).

[36] J.W. Silverstein and P.L. Combettes, IEEE Trans. Signal Process. **40**, 2100 (1992).

[37] Z.D. Bai, Ann. Probab. **21**, 649 (1993).

[38] P.R. Rider, Ann. Inst. Stat. Math. **9**, 215 (1957).

[39] J.H. Miller and J.B. Thomas, IEEE Trans. Inf. Theory **18**, 241 (1972).

[40] Z.D. Bai and J.W. Silverstein, Ann. Probab. **26**, 316 (1998).

[41] Z.D. Bai and J.W. Silverstein, Ann. Probab. **27**, 1536 (1999).

[42] Y. Le Cun, I. Kanter, and S.A. Solla, Phys. Rev. Lett. **66**, 2396 (1991).

[43] S. Halkjær and O. Winther, in *Neural Information Processing Systems 9*, edited by M. Mozer, M. Jordan, and T. Petsche (MIT Press, Cambridge, MA 1997), p. 169.