Statistical mechanics of learning multiple orthogonal signals: Asymptotic theory and fluctuation effects

D. C. Hoyle*
North West Institute for Bio-Health Informatics, University of Manchester, School of Medicine, Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom

M. Rattray†
School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, United Kingdom
(Received 18 May 2006; published 9 January 2007)

The learning of signal directions in high-dimensional data through orthogonal decomposition or principal component analysis (PCA) has many important applications in physics and engineering disciplines, e.g., wireless communication, information theory, and econophysics. The accuracy of the orthogonal decomposition can be studied using mean-field theory. Previous analysis of data produced from a model with a single signal direction has predicted a retarded learning phase transition below which learning is not possible, i.e., if the signal is too weak or the data set is too small then it is impossible to learn anything about the signal direction or magnitude. In this contribution we show that the result can be generalized to the case where there are multiple signal directions. Each nondegenerate signal is associated with a retarded learning transition. However, fluctuations around the mean-field solution lead to large finite size effects unless the signal strengths are very well separated. We evaluate the one-loop contribution to the mean-field theory, which shows that signal directions are indistinguishable from one another if their corresponding population eigenvalues are separated by $O(N^{-\tau})$ with exponent $\tau > \frac{1}{3}$, where N is the data dimension. Numerical simulations are consistent with the analysis and show that finite size effects can persist even for very large data sets.

DOI: 10.1103/PhysRevE.75.016101 PACS number(s): 02.50.Sk, 05.90.+m

I. INTRODUCTION

The techniques of statistical physics have been applied to the study of many different statistical learning methods [1]. One of the most popular of these methods is principal component analysis (PCA), where one projects high-dimensional data onto a subspace of lower dimension chosen to maximize the variance of the projected data. If the data contains some intrinsically low-dimensional structure, then PCA is a useful way to uncover that structure. The principal components are a set of orthogonal vectors defining the axes of the subspace. Given a data set consisting of p, N-dimensional mean-centered vectors ξ_μ, $\mu=1,\ldots,p$, the principal components are the eigenvectors of the sample covariance matrix $\hat{C}=p^{-1}\Sigma_\mu \xi_\mu \xi_\mu^T$, that have the largest eigenvalues, i.e., those directions in the data space along which there is the greatest variation. Orthogonal decompositions of this type play a fundamental role in many areas of physics, engineering, and statistics, e.g., recent applications include wireless communication, information theory, and econophysics [2–5].

Methods from statistical physics have previously been applied to the problem of determining how much data is required to uncover genuine structure in the data. For data produced from a model including a single signal direction, it has been observed that there is a retarded learning phase transition below which learning is impossible [6–9]. If there is insufficient data, or if the signal strength is too weak, then nothing can be learned about the signal direction. For data produced by a model with multiple signal directions, we have observed similar transition behavior in the eigenvalue spectrum of the sample covariance matrix [10]. This suggests that multiple nondegenerate signals each obey a retarded learning transition behavior similar to that observed in the one signal case. However, this is only a conjecture since the behavior of the eigenvectors cannot be determined from the spectrum.

In this contribution we extend the analysis of PCA learning to the case where the data is generated by a model with multiple orthogonal signal directions. We confirm that the signal directions do follow a similar behavior to the case of a single signal direction, but we also observe larger finite size effects than those seen in the analysis of the eigenvalue spectrum. We therefore investigate fluctuations around the mean-field theory. In particular, we determine how close the signal magnitudes have to be in order for the signals to be effectively degenerate. These effectively degenerate signals turn out to be the source of the observed large finite size effects.

The paper is structured as follows. In the next section we describe the data model and review the relevant background literature. In Sec. III we present the leading order asymptotic theory, which confirms that the multiple signal case is a straightforward generalization of the single signal case. In Sec. IV we present numerical simulations that are consistent with the theory but show that finite size effects can be significant even for very large systems. In Sec. V we study fluctuations around the leading order asymptotics. We conclude with a discussion in Sec. VI.

*URL: www.nibhi.org.uk. Electronic address: david.hoyle@manchester.ac.uk
†URL: www.cs.man.ac.uk/~magnus. Electronic address: magnus@cs.man.ac.uk

II. MODEL AND BACKGROUND

The sample data vectors $\{\xi_\mu\}_{\mu=1}^p$ are considered to be drawn from the zero mean Gaussian distribution

$$P(\xi) = (2\pi)^{-N/2}(\det \boldsymbol{C})^{-1/2} \exp\left(-\frac{1}{2}\xi^T \boldsymbol{C}^{-1}\xi\right), \qquad (1)$$

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

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

where $A_m \geq 0 \forall m$, and we assume an ordering $A_1 > A_2 > \cdots > A_S$, so that A_1 represents the strongest signal strength. The isotropic case corresponds to $A_m \equiv 0 \forall m$. The sample covariance matrix $\hat{\boldsymbol{C}} = p^{-1}\Sigma_\mu \xi_\mu \xi_\mu^T$. It is common to form the sample covariance from the centred data matrix, i.e., centered to mean zero. Since our population has zero mean we choose to study, for simplicity, the behavior of the sample covariance defined as $\hat{\boldsymbol{C}} = p^{-1}\Sigma_\mu \xi_\mu \xi_\mu^T$. Simulations that we perform will also adopt this definition and we expect that any conclusions we drawn will be equally valid for a sample covariance defined from the centred data. We are interested in the observed distribution of eigenvectors of $\hat{\boldsymbol{C}}$ when N is large but finite, which can often be usefully approximated by that found in the limit $N \to \infty$ with $p = \alpha N$, for some fixed α.

The behavior of PCA when one symmetry breaking direction $\boldsymbol{B}$ is present, with signal strength A, has been widely studied using replicas in the context of unsupervised learning [8,9], where one considers the overlap $\boldsymbol{J} \cdot \boldsymbol{B}$ between $\boldsymbol{B}$ and the leading principal component $\boldsymbol{J}$ determined from the sample covariance matrix $\hat{\boldsymbol{C}}$. The order parameter R^2 is the expectation value of $(\boldsymbol{J} \cdot \boldsymbol{B})^2$, over the ensemble of different data sets, and provides a suitable means of characterizing the expected accuracy of the first principal component $\boldsymbol{J}$ in representing the true signal direction $\boldsymbol{B}$. 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 $\alpha > A^{-2}$ [6–9]. The problem of PCA batch learning when more than one symmetry breaking (signal) direction is present has not been extensively studied. Indeed it has been speculated by Watkin and Nadal [6] that, for a similar (but not identical) distribution to that in Eq. (1), replica analysis of maximal-variance learning with multiple signal directions present is problematic and requires replica symmetry breaking. PCA learing with multiple symmetry breaking signal directions has been studied in the context of on-line learning by Biehl and Schlösser [11] and Schlösser *et al.* [12]. Within the on-line learning scenario one often focuses on the approach to accurate learning of the true signal directions from an increasing number of training examples, and so typically one has $\alpha \geq 1$. This should be contrasted to the small sample size ($\alpha < 1$) batch learning scenario considered in this paper. More recently this work has been extended by Bunzmann *et al.* to include PCA as a prior stage for improving the performance of artificial neural network training [13].

Associated with the eigenvectors of the sample covariance $\hat{\boldsymbol{C}}$ are the corresponding eigenvalues λ, which indicate the importance of the various principal components in representing the data set [10]. For the case where $\boldsymbol{C}$ contains one symmetry breaking direction one observes a phase transition in the eigenvalue spectrum of $\hat{\boldsymbol{C}}$ at $\alpha_c = A^{-2}$, thus coinciding not unsurprisingly with the retarded learning transition observed in the order parameter R^2. Below α_c the spectrum is identical to that obtained from the isotropic case $\boldsymbol{C} = \sigma^2 \boldsymbol{I}$. Above α_c the bulk of the sample covariance spectrum is still identical to that for the isotropic case, but with a single eigenvalue (the largest) clearly separated from the bulk. When $\boldsymbol{C}$ contains $S > 1$ (orthogonal) symmetry breaking directions we observe a series of phase transitions at $\alpha = A_m^{-2}, m = 1, 2, \ldots, S$, with each time a single eigenvalue separating from the upper edge of the bulk of the sample covariance eigenvalue spectrum. Given this correspondence in transition point location for the one symmetry breaking direction scenario, the eigenvalue spectrum analysis would suggest a series of retarded learning transitions at $\alpha = A_m^{-2}, m = 1, 2, \ldots, S$, when using multiple principal components to learn multiple symmetry breaking directions of the population covariance $\boldsymbol{C}$. It is this aspect of PCA learning that we investigate in this paper.

The results summarized above represent the leading order asymptotic analysis, i.e., as $N \to \infty$, with $\alpha = p/N$ fixed. At finite N, sampling variation could lead to the largest eigenvalues of $\hat{\boldsymbol{C}}$ not being as well separated as suggested by the asymptotic analysis. At small finite values of N, learning of signal directions would similarly require greater separation of population eigenvalues than for larger values of N. Johnstone [14] has extended the seminal work of Tracy and Widom [15] to show that the standard deviation of the largest eigenvalue λ_1 of $\hat{\boldsymbol{C}}$ scales as $N^{-2/3}$ when $\boldsymbol{C}$ is isotropic, instead of $N^{-1/2}$ that one might expect from a standard central limit argument. The distribution of λ_1 when $\boldsymbol{C}$ contains symmetry breaking directions has been shown by Hoyle and Rattray [10], numerically, to be similar in shape (up to location and scale transformations) to that for an isotropic $\boldsymbol{C}$. Recent work by Baik *et al.* [16] has revealed that above the retarded learning transition the standard deviation of λ_1 scales as $N^{-1/2}$, even when the largest population eigenvalue is degenerate. This suggests that, irrespective of the issue of retarded learning, population eigenvalues must be separated by at least $cN^{-2/3}$ for some constant $c > 0$, otherwise they will be effectively degenerate from the viewpoint of the sample covariance $\hat{\boldsymbol{C}}$. Population covariance eigenvalues separated by less than this are associated with signal directions that cannot be distinguished as principal components. However, although learning of the actual signal directions is intimately linked to the sample covariance eigenvalues, this does not guarantee that separation of population eigenvalues by $cN^{-2/3}$ is a sufficient condition for learning the signal directions. Indeed, we will show that greater separation between the population eigenvalues is necessary in order for PCA to correctly distinguish signal directions.

In this paper we extend the leading order asymptotic analysis of PCA learning performed by Reimann *et al.* [8] to the scenario where $\boldsymbol{C}$ contains multiple symmetry breaking

directions. For illustrative purposes we analyse the one principal component case and multiple principal components case separately, in Sec. III A and III B, respectively. In Sec. V we analyze fluctuation effects to determine the effects of finite N on learning the signal directions.

III. LEADING ORDER THEORY

Given p, N-dimensional pattern vectors ξ_μ, $\mu=1,\ldots,p$, PCA aims to find an orthonormal set of vectors $\boldsymbol{J}_1,\boldsymbol{J}_2,\ldots,\boldsymbol{J}_T$ that represent the data set $\{\xi_\mu\}_{\mu=1}^p$ as accurately as possible. Typically we require a set $\{\boldsymbol{J}_i\}_{i=1}^T$ that is of lower complexity than the original data, i.e., $T<p$. In this paper we do not address the question of model selection and so take $T\leqslant S$, i.e., the true signal dimensionality S is assumed to be known for the purposes of analysis. We choose the principal components $\{\boldsymbol{J}_i\}_{i=1}^T$ so that the loss in using these to represent the original data is minimized. Thus the set $\{\boldsymbol{J}_i\}_{i=1}^T$ is determined by maximizing the projection of the data $\{\xi_\mu\}_{\mu=1}^p$ onto the principal components. We therefore seek to maximize

$$\sum_{i=1}^{T}\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J}_i)^2,\quad \boldsymbol{J}_i\cdot\boldsymbol{J}_{i'}=\delta_{ii'}, \tag{3}$$

with respect to the set $\{\boldsymbol{J}_i\}_{i=1}^T$. This is done by considering the low-temperature limit, $\beta\to\infty$ of an ensemble of principal components which has the partition function

$$Z=\int\prod_{i=1}^{T}d\boldsymbol{J}_i\delta(\boldsymbol{I}_T-\boldsymbol{K})\exp\left(\beta\sum_{i=1}^{T}\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J}_i)^2\right), \tag{4}$$

where the matrix $\boldsymbol{K}$ has elements $K_{ii'}=\boldsymbol{J}_i\cdot\boldsymbol{J}_{i'}$. As commented by Urbanczik, our analysis of the behavior of Z in the low-temperature limit actually involves performing an analytic continuation from small β to large β [17]. The overlap $\boldsymbol{J}_i\cdot\boldsymbol{B}_m$ provides a measure of the accuracy of the principal component $\boldsymbol{J}_i$ in representing the signal $\boldsymbol{B}_m$. Since the exponent in Eq. (4) is invariant under the symmetry transformation $\boldsymbol{J}_i\to-\boldsymbol{J}_i$, then in the absence of any terms that break this reflection symmetry the expectation of $\boldsymbol{J}_i\cdot\boldsymbol{B}_m$ will be zero. However, we can consider the quantity $(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2$ to study the accuracy of PCA. This quantity is considered to be self-averaging, so that as $N\to\infty$ its value for a single instance of a data set will be close to its ensemble average over $\{\xi_\mu\}_{\mu=1}^p$. The expectation value $\langle(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2\rangle_\xi$ can be evaluated by introducing source terms in the partition function, which also serve to formally break the reflection symmetry, i.e.,

$$Z(\{h_{mi}\})=\int\prod_{i=1}^{T}d\boldsymbol{J}_i\delta(\boldsymbol{I}_T-\boldsymbol{K})\times\exp\left(\beta\sum_{i=1}^{T}\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J}_i)^2+\beta\sum_{m,i}h_{mi}\boldsymbol{J}_i\cdot\boldsymbol{B}_m\right), \tag{5}$$

$$\langle(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2\rangle_\xi=\lim_{h_{mi}\to0^+}\lim_{\beta\to\infty}\left[\left\langle\beta^{-2}\frac{\partial^2\ln Z(\{h_{mi}\})}{\partial h_{mi}^2}\right\rangle_\xi+\left\langle\left(\beta^{-1}\frac{\partial\ln Z(\{h_{mi}\})}{\partial h_{mi}}\right)^2\right\rangle_\xi\right]. \tag{6}$$

The term $\partial^2\ln Z(\{h_{mi}\})/\partial h_{mi}^2$ represents the variance of $\boldsymbol{J}_i\cdot\boldsymbol{B}_m$ for a given data set $\{\xi_\mu\}_{\mu=1}^p$, and so will vanish in the limit $\beta\to\infty$. Therefore we concentrate on evaluating

$$\lim_{h_{mi}\to0^+}\lim_{\beta\to\infty}\left\langle\left(\beta^{-1}\frac{\partial\ln Z(\{h_{mi}\})}{\partial h_{mi}}\right)^2\right\rangle_\xi. \tag{7}$$

In the presence of the source term we expect $\boldsymbol{J}_i\cdot\boldsymbol{B}_m$ to be self-averaging, and so in the limit $N\to\infty$ the expectation value above has the same value as

$$\left(\lim_{h_{mi}\to0^+}\lim_{\beta\to\infty}\left\langle\beta^{-1}\frac{\partial\ln Z(\{h_{mi}\})}{\partial h_{mi}}\right\rangle_\xi\right)^2. \tag{8}$$

The expectation value in Eq. (8) is nonzero in the limit $\beta\to\infty$ even at finite values of N, due to the presence of the reflection symmetry breaking source terms, and can be easily studied through analysis of the partition function Z, thereby giving the behavior of $\lim_{N\to\infty}\langle(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2\rangle_\xi$. The average of $\ln Z$ over data sets $\{\xi_\mu\}_{\mu=1}^p$ is performed through the replica trick of using the representation

$$\ln x=\lim_{n\to0}\frac{(x^n-1)}{n}. \tag{9}$$

Consequently we determine the typical behaviour of the orthonormal set $\boldsymbol{J}_1,\boldsymbol{J}_2,\ldots,\boldsymbol{J}_T$, by evaluating the replica partition function

$$\mathcal{Z}=\langle Z^n(\{h_{mi}\})\rangle_\xi=\left\langle\left[\int\prod_{i,\nu}d\boldsymbol{J}_i^\nu\prod_{\nu=1}^{n}\delta(\boldsymbol{I}_T-\boldsymbol{K}^\nu)\exp\left(\beta\sum_{i=1}^{T}\sum_{\nu=1}^{n}\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J}_i^\nu)^2+\beta\sum_{m,i,\nu}h_{mi}\boldsymbol{J}_i^\nu\cdot\boldsymbol{B}_m\right)\right]\right\rangle_\xi, \tag{10}$$

where the matrix elements $K_{ii'}^\nu=\boldsymbol{J}_i^\nu\cdot\boldsymbol{J}_{i'}^\nu$. From the replica partition function $\mathcal{Z}$ an effective potential Γ can be calculated in standard fashion through the appropriate Legendre transform [18]. The behavior of $\left\langle\frac{\partial\ln Z}{\partial h_{mi}}\right\rangle_\xi$ can be determined from the location of the minimum of this effective potential. For nearly degenerate signals we find, at finite N, fluctuations strongly affect the asymptotic value of $\lim_{h_{mi}\to0^+}\langle\boldsymbol{J}_i\cdot\boldsymbol{B}_m\rangle_\xi$, and so we calculate next to leading order corrections to the effective potential for $\langle\boldsymbol{J}_i\cdot\boldsymbol{B}_m\rangle_\xi$ in order to determine the asymptotic behavior of $\langle(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2\rangle_\xi$. In Secs. III A and III B we consider the evaluation of Γ to leading order in N (tree level) for the case of one principal component and multiple principal components respectively. In Sec. V we consider the next to leading order (one-loop correction) to the effective potential for the one principal component case. From the replica representation of $\ln x$ in Eq. (9) it is clear we need to evaluate $\ln\mathcal{Z}$, and hence Γ, only up to $O(n)$ contributions

since only these are required to determine $\langle(\boldsymbol{J}_i\cdot\boldsymbol{B}_m)^2\rangle_\xi$. In fact the leading contribution is $O(n\beta)$, and so in all cases the effective potential Γ is only evaluated up to $O(n\beta)$ contributions.

A. One principal component

Initially we restrict ourselves to considering just one principal component, i.e., $T=1$, which we denote simply by $\boldsymbol{J}$. Overlap of $\boldsymbol{J}$ with the various signal directions defines order parameters $R_m^2=\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle_\xi$. In the limit $N\to\infty$ estimation of $\mathcal{Z}$ proceeds along standard lines using replicas [1,8], details of which are given in Appendix A. The values of R_m^2 are determined by locating extrema of the replica-symmetric effective potential for the order parameters $\{R_m^2\}$ (see Appendix A). To leading order in N (tree level) and leading order in $n\beta$ this is

$$\Gamma_{\text{tree,1PC}}(x,\{R_m\}_{m=1}^S)=-\frac{1}{2}Nn\beta\sigma^2\left\{\frac{1}{x}\left[1-\sum_{m=1}^S R_m^2\right]+\frac{2\alpha}{1-2x}\left[1+\sum_{m=1}^S A_m R_m^2\right]\right\}. \quad (11)$$

Here $x=\beta\sigma^2(1-q)$ with q being the order parameter representing the replica symmetric expectation value of the overlap between the different replica fields, i.e., $q=\langle\boldsymbol{J}_\nu\cdot\boldsymbol{J}_{\nu'}\rangle_\xi$, $\forall\nu'\neq\nu$. The subscript 1 PC denotes the one principal component case. Extrema, with respect to x, of Eq. (11) are located at $x=x_0(\{R_m\}_{m=1}^S)$ and satisfy

$$x_0=\frac{1}{2}\left(1-\sum_{m=1}^S R_m^2\right)^{1/2}\left[\left(1-\sum_{m=1}^S R_m^2\right)^{1/2}\pm\alpha^{1/2}\left(1+\sum_{m=1}^S A_m R_m^2\right)^{1/2}\right]^{-1}. \quad (12)$$

For the one symmetry breaking direction case the results of Reimann *et al.* [8] are obtained on taking the positive root in Eq. (12). Thus in general we shall take the positive root. Substituting Eq. (12) into Eq. (11) yields

$$\Gamma_{\text{tree,1PC}}[x_0(\{R_m\}_{m=1}^S),\{R_m\}_{m=1}^S]=-Nn\beta\sigma^2t^2(\{R_m\}_{m=1}^S), \quad (13)$$

where

$$t=\left(1-\sum_{m=1}^S R_m^2\right)^{1/2}+\left(\alpha\left[1+\sum_{m=1}^S A_m R_m^2\right]\right)^{1/2}. \quad (14)$$

With $t>0$, stationary points of Eq. (13) satisfy $\partial t/\partial R_m=0$. It is straight forward to verify that one has a stationary point at $R_m=0,\forall m$, and a set of stationary points $R_m^2=[\alpha A_m^2-1][A_m(1+\alpha A_m)]^{-1}$, $R_{m'}=0$, $\forall m'\neq m$, iff $\alpha>A_m^{-2}$. Evaluating $t^2(\{R_m\}_{m=1}^S)$ at this series of stationary points, one finds that for $\alpha>A_1^{-2}$ the stationary point at $R_1^2=[\alpha A_1^2-1][A_1(1+\alpha A_1)]^{-1}$, $R_{m'}=0$, $\forall m'\neq 1$ is always dominant. For $\alpha<A_1^{-2}$ the stationary point at $R_m=0,\forall m$ has the largest value of t^2. Thus the asymptotic theory predicts a behavior for the first principal component that is identical to that for the one symmetry breaking direction case [8,19], namely,

$$R_1^2=\begin{cases}0, & \alpha<A_1^{-2}\\ (\alpha A_1^2-1)/[A_1(1+\alpha A_1)], & \alpha>A_1^{-2},\end{cases}$$

$$R_m^2=0,\quad m=2,\ldots,S. \quad (15)$$

As one would intuitively expect, at least in the asymptotic limit $N\to\infty$, the presence of additional signal directions does not affect the accuracy of the determination of the first principal component.

B. Multiple principal components

We now consider the case where we seek T principal components $\boldsymbol{J}_1,\boldsymbol{J}_2,\ldots,\boldsymbol{J}_T$. These are obtained as an orthonormal set of vectors that span as much of the variance within the sample data set as possible. Thus we seek to maximize

$$\sum_{i=1}^T\sum_{\mu=1}^p(\xi_\mu\cdot\boldsymbol{J}_i)^2,\quad \boldsymbol{J}_i\cdot\boldsymbol{J}_{i'}=\delta_{ii'}. \quad (16)$$

There is a $T!$-fold degeneracy due to all possible labelings of the principal directions, but we choose $i=1,\ldots,T$ to be ordered by decreasing variance so that they correspond to the usual meanings of principal components. We determine the behaviour of the orthonormal set $\boldsymbol{J}_1,\boldsymbol{J}_2,\ldots,\boldsymbol{J}_T$, by evaluating the ensemble average

$$\left\langle\ln\left[\int\prod_{i=1}^T d\boldsymbol{J}_i\,\delta(\boldsymbol{I}-\boldsymbol{K})\exp\left(\beta\sum_{i=1}^T\sum_{\mu=1}^p(\xi_\mu\cdot\boldsymbol{J}_i)^2\right)\right]\right\rangle_\xi. \quad (17)$$

The expectation value is evaluated using the replica method and after some algebra we find the appropriate tree level effective potential (to leading order in $n\beta$) is

$$\Gamma_{\text{tree}}=-\frac{1}{2}Nn\beta\sigma^2\left\{\text{tr}\left[\boldsymbol{X}^{-1}\left(\boldsymbol{I}-\sum_{m=1}^S\mathcal{R}_m\mathcal{R}_m^T\right)\right]+2\alpha\text{tr}\left[(\boldsymbol{I}-2\boldsymbol{X})^{-1}\left(\boldsymbol{I}+\sum_{m=1}^S A_m\mathcal{R}_m\mathcal{R}_m^T\right)\right]\right\}, \quad (18)$$

where $\mathcal{R}_m$ and $\boldsymbol{X}$ are defined by

$$\boldsymbol{X}=\beta\sigma^2(\boldsymbol{I}-\boldsymbol{q}), \quad (19)$$

$$(\mathcal{R}_m)_i=R_{mi}, \quad (20)$$

$$(\boldsymbol{q})_{ii'}=q_{ii'}, \quad (21)$$

with $q_{ii'}$ the order parameter for the replica symmetric overlap between replicas for the i and i' principal components, i.e., $q_{ii'}=\langle\boldsymbol{J}_i^\nu\cdot\boldsymbol{J}_{i'}^{\nu'}\rangle_\xi$, $\forall\nu,\nu'\neq\nu$. Similarly R_{mi} is the order parameter for the replica symmetric overlap between the mth symmetry breaking direction $\boldsymbol{B}_m$ and the ith principal com-

ponent, i.e., $R_{mi}=\langle \boldsymbol{J}_i^{\nu}\cdot\boldsymbol{B}_m\rangle_{\xi}, \forall\, \nu$. We assume that replicas for different principal components are orthogonal, so that $\boldsymbol{q}$ and thus $\boldsymbol{X}$ are diagonal. The effective potential in Eq. (18) then decomposes as

$$\Gamma_{\text{tree}}=\sum_{i=1}^{T}\Gamma_{\text{tree,1PC}}(x_i,\{R_{mi}\}_{m=1}^{S}), \tag{22}$$

where $x_i=(\boldsymbol{X})_{ii}=\beta\sigma^2(1-q_{ii})$ and $\Gamma_{\text{tree,1PC}}(x,\{R_{mi}\}_{m=1}^{S})$ is the effective potential function given by Eq. (11) for the one principal component problem. Since for multiple principal components the effective potential decomposes to a sum of noninteracting terms, we can immediately express the learning curve behavior in terms of the learning curve behavior obtained from the analysis of the first principal component. Thus the asymptotic theory predicts that for $i=1,2,\ldots,T \leqslant S$,

$$R_{ii}^2=\begin{cases}0, & \alpha<A_i^{-2},\\ (\alpha A_i^2-1)/[A_i(1+\alpha A_i)], & \alpha>A_i^{-2},\end{cases} \tag{23}$$

$$R_{mi}^2=0, \quad \forall\, m\neq i, \quad m=1,2,\ldots,S.$$

As α increases a series of second order retarded learning phase transitions occurs at $\alpha=A_i^{-2}$, $i=1,2,\ldots,T$, thus coinciding precisely with the series of phase transitions observed in the asymptotic eigenvalue spectrum of the sample covariance matrix $\hat{\boldsymbol{C}}$ [10]. On passing through the transition point $\alpha=A_i^{-2}$ the order parameter R_{ii}^2 increases from zero.

IV. SIMULATION

Since the asymptotic theory for multiple principal components indicates that the learning curve for the Tth principal component is determined in isolation from the other principal components, we have performed simulations of only the multiple principal component case. Figure 1(a) shows learning curves when $\boldsymbol{C}$ contains two symmetry breaking directions, with signal strengths $A_1^2=20$ and $A_2^2=10$. This leads to transitions points at $\alpha=0.05$ and $\alpha=0.1$ for the first and second principal components, respectively. For these simulations we have set $N=3200$ and $\sigma^2=1$. The overlap values R_{11}^2 and R_{22}^2 shown are averages estimated from 1000 sample covariance matrices $\hat{\boldsymbol{C}}$ formed from data drawn from the distribution given in Eq. (1). The presence of retarded learning transitions for each of the principal components is clearly present.

Figure 1(b) shows learning curves for a smaller system size $N=200$ with all other parameters as in Fig. 1(a). For the smaller system finite size effects are clearly evident, particularly near the transition points predicted by the asymptotic theory in Eq. (23). However, Fig. 1(c) demonstrates that the expectation value for R_{11}^2 does converge towards the theoretical asymptotic value as N is increased. At smaller values of N the discrepancy, due to finite size corrections, between the asymptotic theory and simulation is more marked. For small finite systems learning the underlying signal directions is obviously a harder task and we would expect that this is exacerbated for signals with similar strengths, i.e., similar values of A. Thus how well two orthogonal signal can be resolved may be dependent on the system size N as well as the separation between signal strengths. Figure 1(d) shows theoretical and simulation learning curves for a population covariance that contains three signal directions. Although the first two signal strengths are identical to those in Fig. 1(a) and the retarded learning transition points are equally spaced, the effect of the third signal direction on the learning curve for the second principal component is very apparent when comparing Figs. 1(a) and 1(d). This is an issue we address in the next section.

V. FLUCTUATION EFFECTS

For the theoretical analysis given in Sec. III A and III B we have taken the eigenvalues of $\boldsymbol{C}$ to be nondegenerate, so that above the retarded learning transition at $\alpha=A_m^{-2}$ the signal direction $\boldsymbol{B}_m$ is learnt to some degree by the mth principal component $\boldsymbol{J}_m$. We can consider for illustrative purposes the leading eigenvalue of $\boldsymbol{C}$ being g-fold degenerate, i.e., $A_1=A_2=\cdots=A_g\equiv\bar{A}$. In the asymptotic limit $N\to\infty$ we would expect that the first principal component is invariant to rotations within the subspace spanned by the degenerate population eigenvectors. The effective potential in Eq. (11) is a function of $\Sigma_{m=1}^{g}R_m^2$, and so one has a subspace of stationary points of Eq. (11) given by

$$\sum_{m=1}^{g}R_m^2=\frac{\alpha\bar{A}^2-1}{\bar{A}(1+\alpha\bar{A})}\equiv\Sigma_1^{(0)}. \tag{24}$$

Taking an average of R_m^2 over the surface of a g-dimensional sphere of radius $\sqrt{\Sigma_1^{(0)}}$ gives $g^{-1}\Sigma_1^{(0)}$. From this one would expect the expectation value $\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle_{\xi}$ to be $g^{-1}\Sigma_1^{(0)}$.

One may question the likelihood of encountering a situation where the population covariance eigenvalues are precisely degenerate. Of more realistic interest is the case where two (or more) eigenvalues of $\boldsymbol{C}$ are similar in value. For illustration purposes we can consider the two symmetry breaking direction case, i.e., $S=2$. Above the appropriate retarded learning transitions, for sufficiently large N any difference $A_1-A_2>0$ will result in the first principal component $\boldsymbol{J}_1$ learning the signal direction $\boldsymbol{B}_1$. This leads us naturally to consider the behavior of R_1,R_2 when $\Delta A\equiv A_1-A_2\sim N^{-\tau}, \tau>0$ as $N\to\infty$. For large values of τ we would expect to approach the degenerate scenario with $R_m^2\to\frac{1}{2}\Sigma_1^{(0)}$, while for smaller values of τ we would expect the decrease in ΔA to be sufficiently slow enough that the increase in N allows for $\boldsymbol{B}_1$ to be learned without confusion by $\boldsymbol{B}_2$, i.e., $R_1^2\to\Sigma_1^{(0)}$ and $R_2^2\to 0$ as $N\to\infty$.

When the population covariance $\boldsymbol{C}$ has a rotational invariance we expect this to be reflected in the saddle point structure of the partition function given in Eq. (10). Consequently the Hessian will be nearly singular for the near-degenerate scenario. Under these circumstances critical- like fluctuations will mean that the Hessian provides a contribution to the effective potential similar in magnitude to that given by the leading order (tree level). The effect of a nearly degenerate

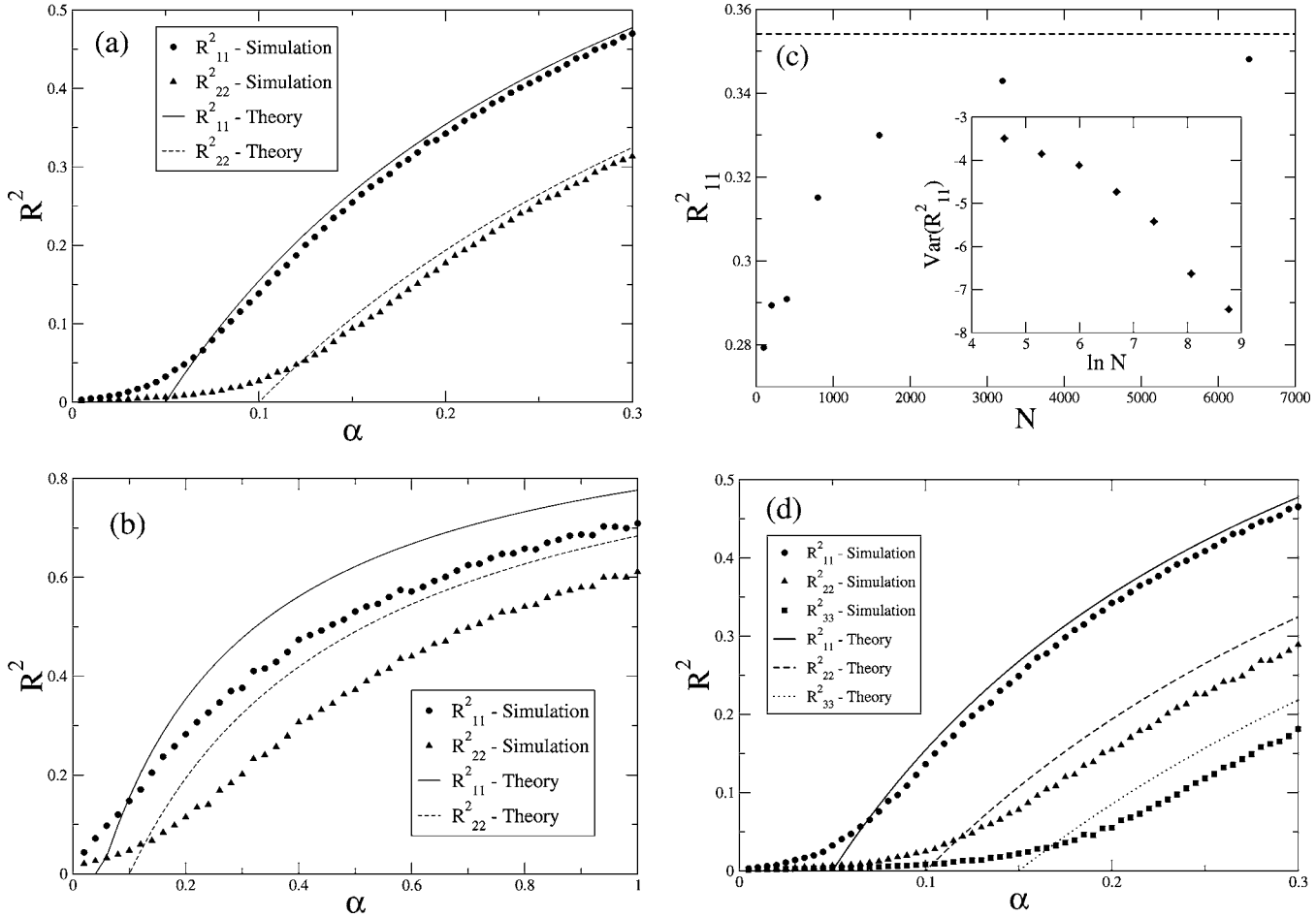

FIG. 1. (a) Learning curves, at fixed $N=3200$, for the first two principal components. The population covariance contains two symmetry breaking directions, with $A_1^2=20, A_2^2=10$, and we have set $\sigma^2=1$. Simulation values (solid symbols) are averages over 1000 data sets. The solid and dashed lines represent the theoretical results given by Eq. (23). Standard errors of the simulation averages are less than the size of the plotted symbols. (b) Learning curves, at fixed $N=200$, for the first two principal components. All other parameters as the same as for Fig. 1(a). (c) Convergence to the asymptotic value of R_{11}^2, at fixed $\alpha=0.2$. The asymptotic value predicted by Eq. (23) is denoted by the horizontal dashed line, while the solid symbols represent simulation averages over 1000 data sets. Standard errors of the simulation averages are less than the size of the plotted symbols. The inset shows simulation estimates of the sample variance for R_{11}^2. (d) Learning curves, at fixed $N=3200$, for the first three principal components. The population covariance contains three symmetry breaking directions, with $A_1^2=20$, $A_2^2=10$, $A_3^2=20/3=1/0.15$. Other parameters and simulation settings are as for (a).

population covariance can be determined by considering the fluctuation contribution to the effective potential. The relevant expression for the replica symmetric effective potential, to leading order in $n\beta$ (see Appendix B), is $\Gamma=\Gamma_{\text{tree}}+\Gamma_{\text{fluc}}$ with

$$\Gamma_{\text{tree}} = -Nn\beta\sigma^2 t^2, \tag{25}$$

$$\Gamma_{\text{fluc}} = \frac{n}{2}\beta\sigma^2\left[\sum_m C(A_m)d_m^{-1} - \sum_m \frac{C(A_m)R_m^2 f^2(A_m)d_m^{-2}}{1+\sum_{m''} R_{m''}^2 f^2(A_{m''})d_{m''}^{-1}}\right], \tag{26}$$

where

$$t = \left(1-\sum_{m=1}^{S} R_m^2\right)^{1/2} + \left(\alpha\left[1+\sum_{m=1}^{S} A_m R_m^2\right]\right)^{1/2} \tag{27}$$

and

$$d_m = t\left[\left(1-\sum_m R_m^2\right)^{-1/2} - \alpha^{1/2}A_m\left(1+\sum_m A_m R_m^2\right)^{-1/2}\right]. \tag{28}$$

The functions $f(A_m)$ and $C(A_m)$ are defined by Eqs. (B9) and (B11) in Appendix B. Note that these expressions are valid for all $S\geqslant 2$. The behavior of the order parameters $\{R_m\}_{m=1}^S$ is determined by minima of Γ. Setting $A_m=\bar{A}+\Delta A_m$ we consider the limit $\Delta A_m \to 0$. Small values of ΔA_m will lead to the fluctuation contribution in Eq. (26) being significant at finite values of N. The relative sizes of ΔA_m and N then become important in determining the behavior of R_m^2. To consider the

finite size effects of nearly degenerate signals we set $\Delta A_m = \Delta \tilde{A}_m N^{-\tau}$. In the limit $N \to \infty$ the functions $f(A_m)$ and $C(A_m)$ tend to finite nonzero limiting values that depend only on $\bar{A}$.

We separate out the asymptotic values of the order parameters and so decompose minima of Γ as $R_m^2 = R_{m,0}^2 + \Delta S_m$, where $\{R_{m,0}\}_{m=1}^S$ is any set of values which satisfies $\Sigma_m R_{m,0}^2 = \Sigma_1^{(0)}$. We have $\Delta S_m \to 0$ as $N \to \infty$, and so write $\Delta S_m = \Delta \tilde{S}_m N^{-\delta}$. Similarly, for fixed $\{\Delta \tilde{A}_m\}_{m=1}^S$ we have $d_m \to 0$. If we write $d_m = \tilde{d}_m N^{-\delta_d}$ then $\partial \Gamma_{\text{fluc}} / \partial R_m$ is dominated by derivatives with respect to d_m, and is $O(N^{-1+2\delta_d})$. Noting that $d_m = -R_m^{-1} \partial t^2 / \partial R_m$ then for $R_m^2 > 0$ we can express $\partial \Gamma_{\text{tree}} / \partial R_m$ and the dominant contribution to $\partial \Gamma_{\text{fluc}} / \partial R_m$ in terms of $\{d_m\}$. From this we immediately have that $\delta_d = \frac{1}{3}$. For $\tau < \frac{1}{3}$ the tree level contribution to $\partial \Gamma / \partial R_m$ is dominant over the fluctuation contribution and can be expanded to leading order as

$$-2Nn\beta\sigma^2 t_0 R_{m,0} \left[\frac{\alpha^{1/2} \Delta A_m}{(1+\bar{A}\Sigma_1^{(0)})^{1/2}} + \frac{1}{2} \frac{\sum_{m'} \Delta S_{m'}}{(1-\Sigma_1^{(0)})^{3/2}} - \frac{\alpha^{1/2} \bar{A} \sum_{m'} (\bar{A} \Delta S_{m'} + \Delta A_{m'} R_{m',0}^2)}{2(1+\bar{A}\Sigma_1^{(0)})^{3/2}} \right] + O(N^{-2\min(\tau,\delta)}). \tag{29}$$

From Eq. (29) we can see that since the second and third terms in $R_m^{-1} \partial \Gamma_{\text{tree}} / \partial R_m$ are independent of m then $\partial \Gamma_{\text{tree}} / \partial R_m = 0, \forall m$ iff $R_{m,0} \equiv 0$ for all but one value of m. Which of the asymptotic values $R_{m,0}^2$ is nonzero is as yet undetermined. The tree level effective potential is given by

$$\Gamma_{\text{tree}} = -Nn\beta\sigma^2 \left[t_0 + N^{-\tau} \alpha^{1/2} \frac{\sum_m \Delta \tilde{A}_m R_{m,0}^2}{(1+\bar{A}\Sigma_1^{(0)})^{1/2}} + O(N^{-2\min(\tau,\delta)}) \right]^2, \tag{30}$$

where $t_0 = t(\bar{A}, \{R_{m,0}\})$. Consequently we are free to minimize the tree level effective potential with respect to $R_{m,0}$ subject to the constraint $\Sigma_m R_{m,0}^2 = \Sigma_1^{(0)}$. The largest positive value of $\Delta \tilde{A}_m$ will occur at $m=1$ and so, as $N \to \infty$, Eq. (30) will be minimized by setting $R_{1,0}^2 = \Sigma_1^{(0)}$, and $R_{m,0}^2 = 0, \forall m > 1$. So even for an asymptotically degenerate population covariance, if $\tau < \frac{1}{3}$ the leading signal direction will be learnt with the same efficiency as for a nondegenerate population covariance. This is hardly surprising given that for $\tau < \frac{1}{3}$ and finite N we are minimizing Γ by minimizing Γ_{tree}, as we would for the asymptotic theory. The sum of leading order corrections $\Sigma_m \Delta S_m$ is then determined from Eq. (29) by balancing the $O(\Delta A_1)$ contribution in Eq. (29), to give $\delta = \tau$.

For $\tau > \frac{1}{3}$ the dominant contribution to $\partial \Gamma / \partial R_m$ is $O(N^{-1/3})$ from the fluctuation term. This is balanced by the second and third terms in Eq. (29), and so for this scenario $\delta = \frac{1}{3}$. Other terms from $\partial \Gamma_{\text{tree}} / \partial R_m$ and $\partial \Gamma_{\text{fluc}} / \partial R_m$ that are not constant over m are smaller than $O(N^{-\tau})$, and so the $O(\Delta A_m)$ term in Eq. (29) cannot be balanced for $\tau > \frac{1}{3}$, even though it is subdominant. Consequently for $\tau > \frac{1}{3}$ we have that $\partial \Gamma / \partial R_m = 0$ does not admit a perturbative solution about $\{R_{m,0}^2\}$. For $\tau > \frac{1}{3}$ we can find a minimum of Γ by constraining all R_m^2 to a common identical value, R^2, and minimizing with respect to R^2. Clearly for this minimum we have $R^2 \to g^{-1} \Sigma_1^{(0)}$. The leading order contribution to Γ comes from the tree level and is $-Nn\beta\sigma^2 t_0^2$, with next to leading contributions from both fluctuations and the tree level being $O(N^{-1/3})$. Consequently, the asymptotic behavior of the overlap order parameters, deduced from our analysis of the effective potential Γ, is summarized below.

$$\lim_{N\to\infty} R_1^2 = \begin{cases} \Sigma_1^{(0)}, & \tau < \frac{1}{3} \\ g^{-1}\Sigma_1^{(0)}, & \tau > \frac{1}{3} \end{cases},$$

$$\lim_{N\to\infty} R_m^2 = \begin{cases} 0, & \tau < \frac{1}{3} \\ g^{-1}\Sigma_1^{(0)}, & \tau > \frac{1}{3} \end{cases}, \quad m \geqslant 2. \tag{31}$$

Figure 2 shows simulation results for the two symmetry breaking direction case, i.e., $S=2$. We have set $\bar{A} = \frac{1}{2}(\sqrt{20} + \sqrt{10})$, $\sigma^2 = 1$ and $\alpha = 0.2$ so that we are above the retarded learning transition point for $\bar{A}$. Figure 2(a) shows convergence with N of R_{11}^2 for different values of τ, while Figure 2(b) shows convergence of R_{12}^2 for different values of τ. For small values of τ, e.g., $\tau = 0.1$, the convergence of R_{11}^2 towards the asymptotic value $\Sigma_1^{(0)}$ is clear. Despite the approach to a twofold degenerate population covariance the stronger signal direction $\boldsymbol{B}_1$ is learnt without confusion by $\boldsymbol{B}_2$. For intermediate values of N we in fact have $R_{11}^2 > \Sigma_1^{(0)}$ since $A_1 > \bar{A}$ for these intermediate values of N. For larger values of τ, e.g., $\tau = 0.5$ and $\tau = 1$ we have R_{11}^2 converging towards $\frac{1}{2} \times \Sigma_1^{(0)}$.

Although we have not explicitly given the analysis of fluctuations for the multiple principal component case the intuitive picture presented here suggests that, due to the orthogonality of the signals and the decomposition, the conclusions presented will still be valid for multiple principal components and when we have a population covariance containing more than one near-degenerate subspace. Figure 3 shows simulation results for a population covariance with four symmetry breaking directions. The first two signals are nearly degenerate with $A_1 - A_2 \sim N^{-1}$ and $\frac{1}{2}(A_1 + A_2) = \frac{1}{2}(\sqrt{20} + \sqrt{10})$, while the third and fourth signals are nearly degenerate with $A_3 - A_4 \sim N^{-0.1}$ and $\frac{1}{2}(A_3 + A_4) = \frac{1}{2}(\sqrt{1/0.3} + \sqrt{1/0.35})$. From our analysis of the one principal component problem we would expect the first principal component $\boldsymbol{J}_1$ to overlap with the signal directions $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ equally well, i.e., $R_{11}^2 \simeq R_{22}^2 \simeq R_{12}^2 \simeq R_{21}$ as $N \to \infty$. In contrast learning of the third signal direction by the third principal

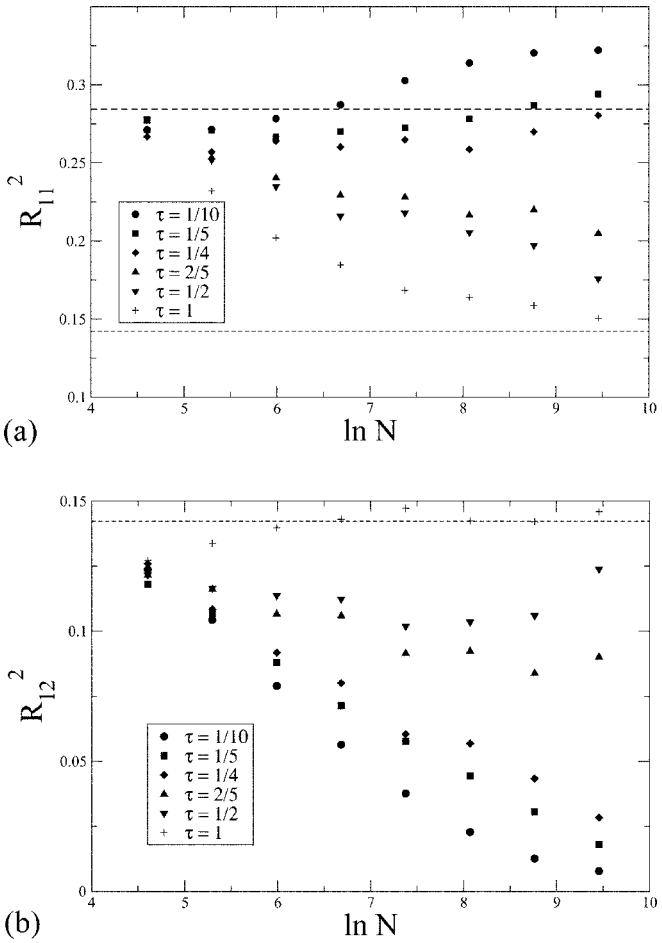

FIG. 2. Plots of overlaps R^2 between the first principal component and the signal directions for different system sizes N. We have fixed $\alpha=0.2$ and set $\sigma^2=1$. The population covariance contains two signal directions with similar signal strengths $A_1=\bar{A}+\Delta A, A_2=\bar{A}-\Delta A$. The signal strength separation is an increasingly weak function of N, i.e., $\Delta A \sim N^{-\tau}, \tau>0$. The solid symbols show simulation averages for different values of τ. Simulation averages are over 1000 data sets, except for the largest value of N for which simulation averages are over 100 data sets. (a) Plot of R_{11}^2. The upper dashed line shows the asymptotic value of R_{11}^2 predicted by Eq. (23) for $A_1=\bar{A}$, while the lower dashed line is drawn at half this value. (b) Plot of R_{12}^2. The dashed line is drawn at half the asymptotic value of R_{11}^2 predicted by Eq. (23) for $A_1=\bar{A}$.

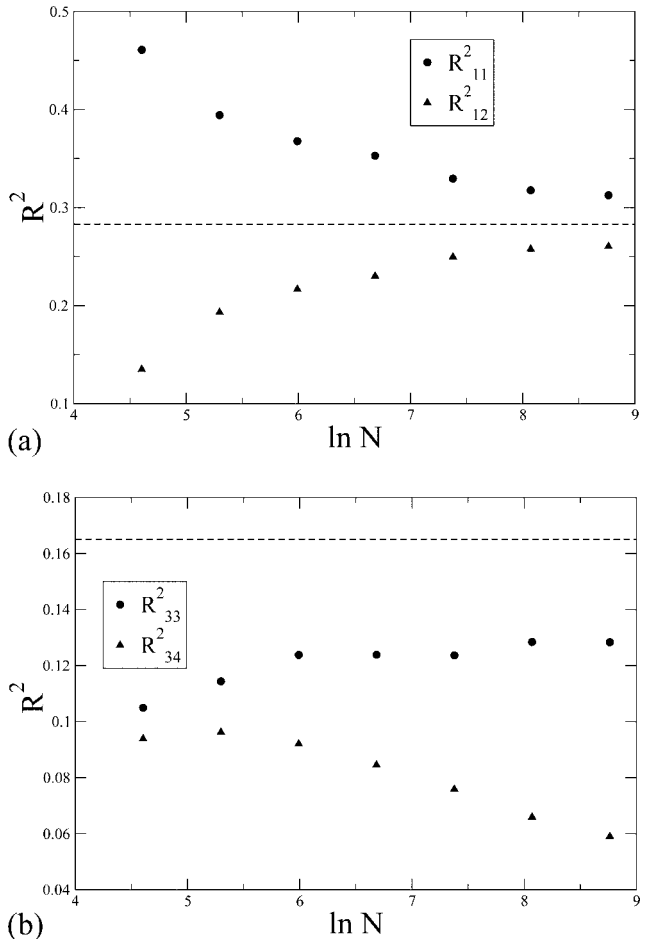

FIG. 3. Plot of R^2 for the first and third principal components. We have fixed $\alpha=0.5$ and set $\sigma^2=1$. The population covariance $\boldsymbol{C}$ contains four signal directions. A_1 and A_2 are of similar strength, with $A_1-A_2\sim N^{-1}$, while A_3 and A_4 are of similar strength, with $A_3-A_4\sim N^{-0.1}$. The solid symbols show averages over 1000 simulated data sets for different values of N. (a) shows R_{11}^2 and R_{12}^2. The dashed line shows the common predicted asymptotic value of R_{11}^2 and R_{12}^2. (b) shows R_{33}^2 and R_{34}^2. The dashed line shows the predicted asymptotic value of R_{33}^2.

component should be unaffected by the presence of the fourth signal direction, i.e., $R_{33}^2=(\alpha A_3^2-1)/[A_3(1+\alpha A_3)]$ and $R_{44}\to 0$, as $N\to\infty$. From Fig. 3(a) it is clear that the presence of the second signal component $\boldsymbol{B}_2$ is affecting the accuracy of the first principal component $\boldsymbol{J}_1$ in representing the first signal component $\boldsymbol{B}_1$. In contrast Fig. 3(b) shows that with increasing N the fourth signal component $\boldsymbol{B}_4$, has a decreasing effect upon the accuracy of the third principal component $\boldsymbol{J}_3$.

VI. DISCUSSION AND CONCLUSIONS

In this paper we have utilized ideas and techniques from statistical physics to understand the accuracy with which signals, or patterns, within data are detected. Specifically, our aim was the study of the principal component learning curves for high dimensional data. The learning curve for each principal component displays a separate phase transition that coincides with the phase transition in the sample covariance eigenspectrum, whereby a single eigenvalue detaches from the bulk spectrum [10]. That the series of phase transitions in the learning curve and eigenspectrum coincide is hardly surprising. Equally the fact that the form of the learning curves are determined by a single functional form, with only the value of A_m differing, ultimately stems from the orthogonal decomposition nature of PCA.

A number of the findings that have been determined using statistical physics techniques have begun to be put on a more rigorous mathematical and statistical footing. For example, Péché [20] demonstrates the universality of the Marčenko and Pastur form for the bulk spectrum [21] when the signal strengths $\{A_m\}_{m=1}^S$ are sufficiently weak. Similarly Paul [22]

has also derived the existence of retarded learning transitions for multiple principal components within the symmetry breaking model considered in this paper, while Baik *et al.* [16] study the distribution of the largest sample covariance eigenvalue in the symmetry breaking direction model when the signal strengths are large. Baik and Silverstein [23] determine the eigenvalue spectrum of a population covariance equivalent to that studied here, confirming the results obtained by Hoyle and Rattray using statistical physics techniques [10].

Simulation results presented in Fig. 2 focus on the two symmetry breaking direction case. Naively one might assume that when the distributions of the largest two sample covariance eigenvalues effectively overlap, then the two signal directions will be effectively indistinguishable (degenerate) and the degree to which they are learnt, as measured by the order parameters R_{11}^2 and R_{22}^2, will be reduced. Although for the isotropic case the distribution of the largest sample covariance eigenvalue, λ_1 is known to follow a Tracy-Widom distribution, with its characteristic $N^{-2/3}$ scaling of the standard deviation (in contrast to a $N^{-1/2}$ scaling expected from a standard central limit argument), the work of Baik *et al.* [16] indicates that above the corresponding retarded learning transitions an $N^{-1/2}$ scaling is appropriate, suggesting that the distributions of λ_1 and λ_2 will certainly overlap when $A_1 - A_2 \sim N^{-1/2}$. Critical-like fluctuations that result from the near degeneracy may renormalize this exponent and effective overlap between λ_1 and λ_2 will occur for larger values of $A_1 - A_2$. However, for the near degenerate scenario $A_1, A_2 \to \bar{A}$, both λ_1 and λ_2 are attempting to learn the same population eigenvalue $\sigma^2(1+\bar{A})$, and so the approaching degeneracy need not cause significant added difficulty in learning the population eigenvalue. In contrast, in the limit $A_1, A_2 \to \bar{A}$ the actual signal directions will be defined only up to an arbitrary rotation within the degenerate subspace, resulting in greater variance of the fluctuations of R_{11}^2 and R_{22}^2. This is essentially the mechanism deduced in Sec. V. If $A_1 - A_2 \sim N^{-\tau}$ and $\tau > \frac{1}{3}$, then as we approach the thermodynamic limit $N \to \infty$, N does not grow fast enough to suppress fluctuations within the subspace spanned by the nearly degenerate signal directions, and so the leading principal components become oblivious to the weak structure within the data. This conclusion is valid irrespective of the asymptotic value $\bar{A}$ of the signal strength provided we are above the retarded learning transition, i.e., $\alpha > \bar{A}^{-2}$. Ultimately, if ΔA is the separation of two population eigenvalues, then unless $\Delta A \gg N^{-1/3}$, the two population eigenvalues will be effectively degenerate from the point of view of the sample data and the sample covariance. This relatively weak decrease with N means that even in what we may naively think of as large systems, for which we believe the asymptotic theory to be accurate, the effect of fluctuations can be marked. The effect can be demonstrated by contrasting the behavior of R_{22}^2 in Figs. 1(a) and 1(d). In Fig. 1(a) the top two signals are well separated ($A_1 - A_2 \simeq 1.31$) and simulation results follow closely the theoretical learning curves. However, in Fig. 1(d) the addition of a third signal which is closer to the second signal ($A_2 - A_3 \simeq 0.58$) increases noticeably the discrepancy between simulation and asymptotic theory for R_{22}^2, even though the retarded learning transition points are equally spaced in α and N is large ($N=3200$).

How much insight do we gain from analysis of fluctuation effects in the leading principal component? We have not given here the full analysis of fluctuation effects for multiple principal components since it is considerably more involved. Although the theoretical analysis given in Sec. V concentrates on the behavior of the leading principal component, we would expect the conclusions to be valid for the multiple principal component case. Indeed the simulation results presented in Fig. 3 suggest that this is the case. The impact of finite size effects for multiple principal components is also likely to extend beyond the batch learning scenario with $\alpha < 1$ that we have primarily discussed here. Significant finite size effects will have repercussions for online training scenarios, even though by definition online training considers an increasing number of presented training examples and therefore studies the approach to accurate learning of the true signal directions from a potentially large number of training examples ($\alpha \gg 1$) [11,12]. In the initial stages of online learning the principal components have yet to learn the true signal directions. Any one particular principal component will represent all the true signal directions almost equally poorly. The system thereby displays a permutation symmetry. As training progresses this permutation symmetry is broken, with individual principal components specializing in learning a particular signal direction. One would suspect that this dynamical symmetry breaking is intricately linked to the presence of retarded learning phase transitions in the batch learning scenario. Such an observation has been commented on by Bunzmann *et al.* who used a PCA decomposition of training data as a prior input stage to an artificial neural network, with training of the principal components and neural network performed simultaneously online [13]. The putative link between dynamical symmetry breaking in online learning and retarded learning phase transitions in batch learning emphasizes the importance of the finite-size effects studied within this paper. Critical-like fluctuations arising from nearly degenerate signals that lead to renormalization of the leading order mean-field theory will most likely also lead to critical slowing down of online learning, with an extended "plateau" stage in the PCA learning curve within the online learning scenario considered by Biehl and Schlösser [11]. The work of Bunzmann *et al.* highlights that PCA may be used as a prior step to more sophisticated learning algorithms, or implicitly within those learning algorithms. Understanding how orthogonal decompositions of sample data perform is then a key component in understanding the performance of these composite learning algorithms.

ACKNOWLEDGMENTS

D.C.H. would like to acknowledge the MRC (U.K.) for financial support. We would like to thank Mr. David Nuttall for help with some of the larger, computationally intensive calculations.

APPENDIX A

One principal component. Here we set $T=1$ and denote the single principal component by $\boldsymbol{J}$. We wish to evaluate the ensemble average

$$\left\langle \ln \int d\boldsymbol{J}\,\delta(1-\|\boldsymbol{J}\|^2)\exp\left(\beta\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J})^2+\sum_{m=1}^{S}h_m\boldsymbol{J}\cdot\boldsymbol{B}_m\right)\right\rangle_\xi . \tag{A1}$$

Ultimately we take the limit $h_m\to 0^+$. The leading order (tree level) expression for the effective potential can be determined from the saddle point structure of the appropriate partition function in the absence of h_m. Therefore, for convenience we shall drop the source terms from subsequent expressions. We start with the replica partition function

$$\mathcal{Z}=(2\pi)^{-N_p/2}(\det \boldsymbol{C})^{-\frac{p}{2}}\int\prod_{\mu=1}^{p}d\xi_\mu \times\exp\left(-\frac{1}{2}\sum_{\mu}\xi_\mu^T\boldsymbol{C}^{-1}\xi_\mu\right)\int\prod_{\nu=1}^{n}d\boldsymbol{J}^\nu\prod_{\nu=1}^{n}\delta(1-\|\boldsymbol{J}^\nu\|^2) \times\exp\left(\beta\sum_{\mu,\nu}(\boldsymbol{J}^\nu\cdot\xi_\mu)^2\right). \tag{A2}$$

We can rewrite $\exp[\beta(\boldsymbol{J}^\nu\cdot\xi_\mu)^2]$ as $(1/\sqrt{2\pi})\int dx_{\nu\mu}\times\exp(-\frac{1}{2}x_{\nu\mu}^2+\beta^{1/2}x_{\nu\mu}\boldsymbol{J}^\nu\cdot\xi_\mu)$. After integrating over ξ_μ and $x_{\nu\mu}$ we obtain

$$\mathcal{Z}=(2\pi)^{-(1/2)N_n}\int\prod_{\nu=1}^{n}d\boldsymbol{J}^\nu\prod_{\nu=1}^{n}\delta(1-\|\boldsymbol{J}^\nu\|^2)(\det M)^{-p/2}, \tag{A3}$$

where (for $\boldsymbol{C}=\sigma^2\boldsymbol{I}+\sigma^2\Sigma_m A_m\boldsymbol{B}_m\boldsymbol{B}_m^T$),

$$M_{\nu\nu'}=\delta_{\nu\nu'}-2\beta\sigma^2\left(q_{\nu\nu'}+\sum_{m=1}^{S}A_mR_m^\nu R_m^{\nu'}\right),$$

$$R_m^\nu=\boldsymbol{J}^\nu\cdot\boldsymbol{B}_m,$$

$$q_{\nu\nu'}=\boldsymbol{J}^\nu\cdot\boldsymbol{J}^{\nu'}. \tag{A4}$$

Here we have abused the notation by using R_m^ν to represent the overlap between the νth replica of the first principal component $\boldsymbol{J}_\nu$ and the signal direction $\boldsymbol{B}_m$, while we also use R_m^2 to represent the expectation value $\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle_\xi$. However, it will always be apparent in which context we are using the notation. The integrations over $\boldsymbol{J}^\nu$ are performed in terms of integrations over R_m^ν and $q_{\nu\nu'}$, i.e.,

$$\int\prod_{\nu}d\boldsymbol{J}^\nu\prod_{\nu=1}^{n}\delta(1-\|\boldsymbol{J}^\nu\|^2) =\int\prod_{\nu}d\boldsymbol{J}^\nu\prod_{\nu=1}^{n}\delta(1-\|\boldsymbol{J}^\nu\|^2)\int\prod_{\nu}dR_m^\nu\delta(R_m^\nu-\boldsymbol{J}^\nu\cdot\boldsymbol{B}_m) \times\int\prod_{\nu'>\nu}dq_{\nu\nu'}\delta(q_{\nu\nu'}-\boldsymbol{J}^\nu\cdot\boldsymbol{J}^{\nu'}). \tag{A5}$$

We rewrite δ functions in terms of their Fourier representations. After integrating over $\boldsymbol{J}^\nu$ and Fourier variables, and retaining only terms in the exponent of the integrand that are extensive in N and depend upon $\{q_{\nu\nu'}\}$ or $\{R_m^\nu\}$, we obtain

$$\mathcal{Z}\simeq\int\prod_{\nu}dR_\nu\int\prod_{\nu'>\nu}dq_{\nu\nu'}\exp\left[N\left(\frac{1}{2}\operatorname{tr}\ln\boldsymbol{L}-\frac{\alpha}{2}\operatorname{tr}\ln\boldsymbol{M}\right)\right], \tag{A6}$$

where $L_{\nu\nu'}=L_{\nu'\nu}=q_{\nu\nu'}-\Sigma_m^S R_m^\nu R_m^{\nu'}$. We now look for stationary points of the exponent of the integrand in Eq. (A6). Assuming replica symmetry for such saddle points we put,

$$R_m^\nu=R_m,\quad\forall\ \nu,$$

$$q_{\nu\nu'}=q,\quad\forall\ \nu,\quad\nu'\neq\nu,\quad q_{\nu\nu}=1,\quad\forall\ \nu. \tag{A7}$$

Putting $x=\beta\sigma^2(1-q)$ then, to leading order in $n\beta$ the exponent of the integrand becomes

$$\frac{1}{2}Nn\beta\sigma^2\left\{\frac{1}{x}\left[1-\sum_{m=1}^{S}R_m^2\right]+\frac{2\alpha}{1-2x}\left[1+\sum_{m=1}^{S}A_mR_m^2\right]\right\}. \tag{A8}$$

In the limit $N\to\infty$ the integral in Eq. (A6) is dominated by stationary points of Eq. (A8) and to leading order $\ln\mathcal{Z}$ is simply the value of the exponent in Eq. (A8). At this order (tree level) the effective potential is identical in form to the negative of the saddle point exponent, but with R_m^2 now representing the order parameter $\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle_\xi$, rather than an integration variable.

Multiple Principal Components. For multiple principal components ($T>1$) the appropriate expectation value is

$$\left\langle \ln\int\prod_{i=1}^{T}d\boldsymbol{J}_i\,\delta(\boldsymbol{I}_T-\boldsymbol{K})\exp\left(\beta\sum_{i=1}^{T}\sum_{\mu=1}^{p}(\xi_\mu\cdot\boldsymbol{J}_i)^2\right)\right\rangle_\xi , \tag{A9}$$

where $K_{ii'}=\boldsymbol{J}_i\cdot\boldsymbol{J}_{i'}$. Introducing replicas for the principal components $\boldsymbol{J}_1,\boldsymbol{J}_2,\ldots,\boldsymbol{J}_T$ we obtain the replica partition function

$$\mathcal{Z}=(2\pi)^{-N_p/2}(\det \boldsymbol{C})^{-\frac{p}{2}}\int\prod_{\mu=1}^{p}d\xi_\mu \times\exp\left(-\frac{1}{2}\sum_{\mu=1}^{p}\xi_\mu^T\boldsymbol{C}^{-1}\xi_\mu\right)\int\prod_{i=1}^{T}\prod_{\nu=1}^{n}d\boldsymbol{J}_i^\nu\prod_{\nu=1}^{n}\delta(\boldsymbol{I}_T-\boldsymbol{K}^\nu) \times\exp\left(\beta\sum_{i=1}^{T}\sum_{\mu=1}^{p}\sum_{\nu=1}^{n}(\boldsymbol{J}_i^\nu\cdot\xi_\mu)^2\right). \tag{A10}$$

Here, the matrix $\boldsymbol{K}^\nu$ has elements $(\boldsymbol{K}^\nu)_{ii'}=\boldsymbol{J}_i^\nu\cdot\boldsymbol{J}_{i'}^{\nu'}$. Following the same approach as before it is an easy matter to confirm that

$$\mathcal{Z}\simeq\int\prod_{m,i,\nu}dR_{mi}^\nu\prod_{i,i'}\prod_{\nu\neq\nu'}dQ_{ii'}^{\nu\nu'}\exp\left(\frac{N}{2}(\operatorname{tr}\ln\boldsymbol{L}-\alpha\operatorname{tr}\ln\boldsymbol{M})\right), \tag{A11}$$

where $\boldsymbol{M}=\boldsymbol{I}-2\beta\boldsymbol{P}$ and

$$P_{ii'}^{\nu\nu'} = \sigma^2 Q_{ii'}^{\nu\nu'} + \sigma^2 \sum_m A_m R_{mi}^{\nu} R_{mi'}^{\nu'},$$

$$L_{ii'}^{\nu\nu'} = Q_{ii'}^{\nu\nu'} - \sum_m R_{mi}^{\nu} R_{mi'}^{\nu'},$$

$$Q_{ii'}^{\nu\nu'} = \boldsymbol{J}_i^{\nu} \cdot \boldsymbol{J}_{i'}^{\nu'}. \tag{A12}$$

Here $R_{mi}^{\nu} = \boldsymbol{J}_i^{\nu} \cdot \boldsymbol{B}_m$ represents the overlap of the νth replica of the ith principal component, $\boldsymbol{J}_i$ with the mth symmetry breaking direction $\boldsymbol{B}_m$. Under an assumption of that dominant saddle points of the integrand in Eq. (A11) are replica symmetric we put

$$R_{mi}^{\nu} = R_{mi}, \quad \forall\ \nu,$$

$$Q_{ii'}^{\nu\nu'} = q_{ii'}, \quad \forall\ \nu \neq \nu'. \tag{A13}$$

To leading order we obtain the exponent of the integrand for a replica symmetric configuration as

$$\frac{1}{2} Nn\beta\sigma^2 \left\{ \mathrm{tr}\left[\boldsymbol{X}^{-1} \left(\boldsymbol{I} - \sum_m \mathcal{R}_m \mathcal{R}_m^T \right) \right] + 2\alpha \mathrm{tr}\left[(\boldsymbol{I} - 2\boldsymbol{X})^{-1} \left(\boldsymbol{I} + \sum_m A_m \mathcal{R}_m \mathcal{R}_m^T \right) \right] \right\}, \tag{A14}$$

where $\mathcal{R}_m$ and $\boldsymbol{X}$ are defined by

$$\boldsymbol{X} = \beta\sigma^2 (\boldsymbol{I} - \boldsymbol{q}), \tag{A15}$$

$$(\mathcal{R}_m)_i = R_{mi}, \tag{A16}$$

$$(\boldsymbol{q})_{ii'} = q_{ii'}. \tag{A17}$$

APPENDIX B

For simplicity and illustrative purposes we only analyze fluctuation effects for the one principal component case. The one-loop contribution to the effective potential is given by evaluating $\ln\det(-\boldsymbol{H})$ for the Hessian $\boldsymbol{H}$ of the tree level saddle point and simply replacing the saddle point value of R_m^2 with the order parameter $\langle (\boldsymbol{J}\cdot\boldsymbol{B}_m)^2 \rangle_{\xi}$. Since we use the same notation R_m^2 to represent the integration variable and the order parameter we have only to express $\ln\det(-\boldsymbol{H})$, evaluated at the saddle point, entirely in terms of the saddle point value of R_m. We start from exponent of Eq. (A6)

$$\frac{N}{2} \mathrm{tr} \ln \boldsymbol{L} - \frac{N\alpha}{2} \mathrm{tr} \ln \boldsymbol{M}. \tag{B1}$$

We simplify the notation by using $\gamma = (\nu, \nu') = (\gamma_1, \gamma_2)$ to represent an ordered pair of replicas $\nu' > \nu$. Denoting fluctuations, in an obvious notation, as δR_m^{ν} and $\delta q_{\nu,\nu'} = \delta q_{\gamma}$ we can expand Eq. (B1) about a replica-symmetric saddle point to obtain (to second order),

$$\frac{1}{2} Nn\beta\sigma^2 \left\{ \frac{1}{x} \left[1 - \sum_{m=1}^{S} R_m^2 \right] + \frac{2\alpha}{1-2x} \left[1 + \sum_{m=1}^{S} A_m R_m^2 \right] \right\} + \frac{N}{2} \sum_{\gamma,\gamma'} \delta q_{\gamma} \boldsymbol{H}_{\gamma\gamma'}^{(qq)} \delta q_{\gamma'} + N \sum_{\gamma,m',\nu'} \delta q_{\gamma} \boldsymbol{H}_{\gamma m',\nu'}^{(qR)} \delta R_{m'}^{\nu'} + \frac{N}{2} \sum_{m,m',\nu,\nu'} \delta R_m^{\nu} \boldsymbol{H}_{m,m',\nu\nu'}^{(RR)} \delta R_{m'}^{\nu'}. \tag{B2}$$

We have decomposed the Hessian into the various contributions from (i) just fluctuations $\delta q_{\nu,\nu'}$, (ii) just fluctuations δR_m^{ν}, (iii) cross terms involving both $\delta q_{\nu\nu'}$ and δR_m^{ν}. For brevity, we do not give here the expressions for the matrices $\boldsymbol{H}^{(qq)}, \boldsymbol{H}^{(qR)}, \boldsymbol{H}^{(RR)}$, since they are easily evaluated. The Gaussian fluctuations in δq_{γ} are easily integrated out to yield

$$\frac{1}{2} Nn\beta\sigma^2 \left\{ \frac{1}{x} \left[1 - \sum_{m=1}^{S} R_m^2 \right] + \frac{2\alpha}{1-2x} \left[1 + \sum_{m=1}^{S} A_m R_m^2 \right] \right\} + \frac{N}{2} \sum_{m,m',\nu,\nu'} \delta R_m^{\nu} \delta R_{m'}^{\nu'} [\boldsymbol{H}^{(RR)} - (\boldsymbol{H}^{(qR)})^T (\boldsymbol{H}^{(qq)})^{-1} \boldsymbol{H}^{(qR)}]_{m,m',\nu\nu'} - \frac{1}{2} \mathrm{tr} \ln(-\boldsymbol{H}^{(qq)}) - \frac{n(n-1)}{4} \ln \frac{N}{2\pi}. \tag{B3}$$

The matrix $\boldsymbol{H}^{(qq)}$ has the form

$$H_{\gamma\gamma'}^{(qq)} = C_0 + C_1 \delta_{\gamma_1\gamma_1'} \delta_{\gamma_2\gamma_2'} + C_2 (\delta_{\gamma_1\gamma_1'} + \delta_{\gamma_2\gamma_2'} + \delta_{\gamma_1\gamma_2'} + \delta_{\gamma_2\gamma_1'}), \tag{B4}$$

and so is of the form considered by de Almeida and Thouless [1,24]. Eigenvalues of $\boldsymbol{H}^{(qq)}$ are then easily determined. Explicitly we find to $O(n\beta)$,

$$\ln\det(-\boldsymbol{H}^{(qq)}) = n\beta\sigma^2 \left[\frac{\left(1 - \sum_m R_m^2\right)^2}{x^4} - 16\alpha \frac{\left(1 + \sum_m A_m R_m^2\right)^2}{(1-2x)^4} \right] \times \left[\frac{\left(1 - \sum_m R_m^2\right)}{x^3} + 8\alpha \frac{\left(1 + \sum_m A_m R_m^2\right)}{(1-2x)^3} \right]^{-1} \times [1 + O(\beta^{-1})]. \tag{B5}$$

It can be verified that the inverse of $\boldsymbol{H}^{(qq)}$ is also of the form given in Eq. (B4) above. Substituting the form in Eq. (B4) into Eq. (B3) gives $\boldsymbol{H}^{(RR)} - (\boldsymbol{H}^{(qR)})^T (\boldsymbol{H}^{(qq)})^{-1} \boldsymbol{H}^{(qR)} \equiv -\boldsymbol{U}$, with $\boldsymbol{U}$ of the form

$$\boldsymbol{U} = \boldsymbol{W} \delta_{\nu\nu'} + \boldsymbol{1}_n \boldsymbol{1}_n^T \boldsymbol{Y}, \tag{B6}$$

where $\boldsymbol{1}_n$ represents an n-dimensional vector in replica space with components all equal to 1. Finally integrating out Gaussian fluctuations in δR_m^{ν} gives a contribution of

$$-\frac{1}{2}\mathrm{tr}\ln N\boldsymbol{U}+\frac{Sn}{2}\ln 2\pi = -\frac{n}{2}\left(\mathrm{tr}\ln \boldsymbol{W}+\mathrm{tr}[\boldsymbol{W}^{-1}\boldsymbol{Y}]+S\ln\frac{N}{2\pi}\right)+O(n^2). \tag{B7}$$

Using notation $\Sigma_1=\Sigma_m R_m^2$ and $\Sigma_2=\Sigma_m A_m R_m^2$, one explicitly finds

$$W_{mm'}=\beta\sigma^2\delta_{mm'}\left[\frac{1}{x}-\frac{2\alpha A_m}{1-2x}\right]+8\beta\sigma^2 R_m R_{m'} f(A_m)f(A_{m'})+O(n), \tag{B8}$$

where the function $f(A_m)$ is given by

$$f(A_m)=\left[\frac{1}{x^2}+\frac{4\alpha A_m}{(1-2x)^2}\right]\left[\frac{1-\Sigma_1}{x^3}+\frac{8\alpha(1+\Sigma_2)}{(1-2x)^3}\right]^{-1/2} \tag{B9}$$

and

$$Y_{mm'}=\beta^2\sigma^4\delta_{mm'}C(A_m,x,\Sigma_1,\Sigma_2)+\beta^2\sigma^4 g(A_m,A_{m'},x,\Sigma_1,\Sigma_2)R_m R_{m'}+O(n), \tag{B10}$$

with $C(A_m,x,\Sigma_1,\Sigma_2)$ given by

$$C(A_m,x,\Sigma_1,\Sigma_2)=-\left[\frac{4\alpha A_m(1+\Sigma_2)}{(1-2x)^2}+\frac{1-\Sigma_1}{x^2}\right]. \tag{B11}$$

Similarly we decompose $g(A_m,A_{m'},x,\Sigma_1,\Sigma_2)=g_0-16(g_1+g_2)/g_4-16g_3/g_4^2$ where,

$$g_0=\frac{1}{x^2}-\frac{4\alpha A_m A_{m'}}{(1-2x)^2}, \tag{B12}$$

$$g_1=\left[\frac{1-\Sigma_1}{x^3}-\frac{8\alpha A_m(1+\Sigma_2)}{(1-2x)^3}\right]\left[\frac{1}{x^2}+\frac{4\alpha A_{m'}}{(1-2x)^2}\right], \tag{B13}$$

$$g_2=\left[\frac{1}{x^2}+\frac{4\alpha A_m}{(1-2x)^2}\right]\left[\frac{1-\Sigma_1}{x^3}-\frac{8\alpha A_{m'}(1+\Sigma_2)}{(1-2x)^3}\right], \tag{B14}$$

$$g_3=\left[\frac{1}{x^2}+\frac{4\alpha A_m}{(1-2x)^2}\right]\left[\frac{1}{x^2}+\frac{4\alpha A_{m'}}{(1-2x)^2}\right]\left[\frac{(1-\Sigma_1)^2}{x^4}-\frac{16\alpha(1+\Sigma_2)^2}{(1-2x)^4}\right], \tag{B15}$$

$$g_4=\left[\frac{1-\Sigma_1}{x^3}+\frac{8\alpha(1+\Sigma_2)}{(1-2x)^3}\right]. \tag{B16}$$

In order to affect the leading order asymptotic calculation we require a contribution of order $O(n\beta)$. Such terms only come from $n\mathrm{tr}\boldsymbol{W}^{-1}\boldsymbol{Y}$ in (B7). The inverse matrix $\boldsymbol{W}^{-1}$ is easily evaluated as,

$$[\boldsymbol{W}^{-1}]_{mm'}=\beta^{-1}\sigma^{-2}\left[\delta_{mm'}d_m^{-1}-\frac{R_m R_{m'}f(A_m)f(A_{m'})d_m^{-1}d_{m'}^{-1}}{1+\sum_{m''}R_{m''}^2 f^2(A_{m''})d_{m''}^{-1}}\right]+O(n), \tag{B17}$$

where $d_m=x^{-1}-2\alpha A_m(1-2x)^{-1}$. Evaluating $n\mathrm{tr}\boldsymbol{W}^{-1}\boldsymbol{Y}$ gives

$$n\mathrm{tr}\boldsymbol{W}^{-1}\boldsymbol{Y}=n\beta\sigma^2\left[\sum_m C(A_m)d_m^{-1}-\sum_m\frac{C(A_m)R_m^2 f^2(A_m)d_m^{-2}}{1+\sum_{m''}R_{m''}^2 f^2(A_{m''})d_{m''}^{-1}}+\sum_m R_m^2 g(A_m,A_m)d_m^{-1}-\sum_{m,m'}\frac{g(A_m,A_{m'})R_m^2 R_{m'}^2 f(A_m)f(A_{m'})d_m^{-1}d_{m'}^{-1}}{1+\sum_{m''}R_{m''}^2 f^2(A_{m''})d_{m''}^{-1}}\right]+O(n^2). \tag{B18}$$

Source terms in Eq. (10) for the evaluation of $\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle_\xi$ do not affect the saddle point equation (12) and so we can simply substitute into d_m, C,f, and g the tree-level expression for x in terms of the set of order parameters $\{R_m\}$, i.e., insert the positive root of Eq. (12). For d_m this gives

$$d_m=t\left[\left(1-\sum_m R_m^2\right)^{-1/2}-\alpha^{1/2}A_m\left(1+\sum_m A_m R_m^2\right)^{-1/2}\right], \tag{B19}$$

where

$$t=\left(1-\sum_{m=1}^{S}R_m^2\right)^{1/2}+\left(\alpha\left[1+\sum_{m=1}^{S}A_m R_m^2\right]\right)^{1/2}. \tag{B20}$$

The (one-loop) fluctuation contribution to the determination of the order parameter $\langle(\boldsymbol{J}\cdot\boldsymbol{B}_m)^2\rangle$ then comes from differentiating Eq. (B18) with respect to R_m. We are specifically interested in the near degenerate scenario, for example, when two signal strengths are similar in value. Under this scenario the Hessian of Eq. (B1) will be nearly singular and we expect a divergent contribution from $\mathrm{tr}\ln\boldsymbol{U}$. We can write $A_m=\bar{A}+\Delta A_m$. In the limit $A_m\to\bar{A}\,\forall\, m$, we have that $d_m\to 0$. In this limit the contribution from $\ln\det(-\boldsymbol{H}^{(qq)})$ given in Eq. (B5) is finite, and so we no longer consider it. The dominant terms in $\lim_{n\to 0}\frac{\partial}{\partial R_m}\mathrm{tr}\boldsymbol{W}^{-1}\boldsymbol{Y}$ come from derivatives $\frac{\partial d_m^{-1}}{\partial R_m}$. We find derivatives, with respect to d_m, of the third and fourth terms in Eq. (B18) cancel to leading order in the limit $\Delta A_m\to 0$, for all values of the order parameters $\{R_m\}$. Therefore we concentrate on the first and second terms to determine the order parameter values $\{R_m\}$ in the limit $A_m\to\bar{A}$, i.e., the relevant, dominant contribution to the effective potential from fluctuations is

$$\frac{1}{2}n\beta\sigma^2\left[\sum_m C(A_m)d_m^{-1} - \sum_m \frac{C(A_m)R_m^2 f^2(A_m)d_m^{-2}}{\sum_{m''} R_{m''}^2 f^2(A_{m''})d_{m''}^{-1}}\right]. \quad \text{(B21)}$$

In contrast the tree level contribution is $O(nN\beta)$. There are also additional $O(n\beta)$ contributions from next-to-leading order terms omitted from the integration over Fourier variables in Eq. (A5). However, these omitted contributions do not explicitly depend upon the signal strengths A_m, and it is easily confirmed that the omitted terms remain finite in the degenerate limit $\Delta A_m \to 0$. Consequently they are subdominant in comparison to the contribution in Eq. (B21) and we no longer consider them.

[1] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).
[2] A. M. Tulino and S. Verdu, *Random Matrix Theory and Wireless Communication*, Vol. 1 of *Foundations and Trends in Communication and Information Theory* (Now Publishers, Boston, 2004).
[3] J. W. Silverstein and P. L. Combettes, IEEE Trans. Signal Process. **40**, 2100 (1992).
[4] R. Müller, IEEE Trans. Inf. Theory **48**, 2495 (2002).
[5] V. Plerou, P. Gopikrishnan, B. Rosenow, Luis A. Nures Amaral, T. Guhr, and H. E. Stanley, Phys. Rev. E **65**, 066126 (2002).
[6] T. L. H. Watkin and J.-P. Nadal, J. Phys. A **27**, 1899 (1994).
[7] M. Biehl and A. Mietzner, J. Phys. A **27**, 1885 (1994).
[8] P. Reimann, C. Van den Broeck, and G. J. Bex, J. Phys. A **29**, 3521 (1996).
[9] P. Reimann and C. Van den Broeck, Phys. Rev. E **53**, 3989 (1996).
[10] D. C. Hoyle and M. Rattray, Phys. Rev. E **69**, 026124 (2004).
[11] M. Biehl and E. Schlösser, J. Phys. A **31**, L97 (1998).
[12] E. Schlösser, D. Saad, and M. Biehl, J. Phys. A **32**, 4061 (1999).
[13] C. Bunzmann, M. Biehl, and R. Urbanczik, Phys. Rev. E **72**, 026117 (2005).
[14] I. M. Johnstone, Ann. Stat. **29**, 295 (2001).
[15] C. A. Tracy and H. Widom, Commun. Math. Phys. **177**, 727 (1996).
[16] J. Baik, G. Ben Arous, and S. Péché, Ann. Probab. **33**, 1643 (2005).
[17] R. Urbanczik, Europhys. Lett. **64**, 564 (2003).
[18] C. Itzykson and J.-M. Drouffe, *Statistical Field Theory* (Cambridge University Press, Cambridge, 1989), Vol. 1.
[19] D. C. Hoyle and M. Rattray, Europhys. Lett. **62**, 117 (2003).
[20] S. Péché, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (2003).
[21] V. A. Marčenko and L. A. Pastur, Math. USSR. Sb. **1**, 507 (1967).
[22] D. Paul (unpublished).
[23] J. Baik and J. W. Silverstein J. Multivariate Anal. (to be published).
[24] J. R. L. de Almeida and D. Thouless, J. Phys. A **11**, 983 (1978).