Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering

Thibault Lesieur, Caterina de Bacco, Jess Banks, Florent Krzakala, Cris Moore, Lenka Zdeborová

Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering

Thibault Lesieur [†], Caterina De Bacco[‡] , Jess Banks[‡], Florent Krzakala[*], Cris Moore[‡] and Lenka Zdeborová†

† Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France
‡ Santa Fe Institute, 1399 Hyde Park Road Santa Fe, New Mexico 87501, USA
∗ Laboratoire de Physique Statistique, CNRS, PSL Universités & Ecole Normale Supérieure,
Sorbonne Universités et Université Pierre & Marie Curie, 75005, Paris, France.

Abstract— **We consider the problem of Gaussian mixture clustering in the high-dimensional limit where the data consists of m points in n dimensions, $n, m \to \infty$ and $\alpha = m/n$ stays finite. Using exact but non-rigorous methods from statistical physics, we determine the critical value of α and the distance between the clusters at which it becomes information-theoretically possible to reconstruct the membership into clusters better than chance. We also determine the accuracy achievable by the Bayes-optimal estimation algorithm. In particular, we find that when the number of clusters is sufficiently large, $r > 4 + 2\sqrt{\alpha}$, there is a gap between the threshold for information-theoretically optimal performance and the threshold at which known algorithms succeed.**

Clustering m points in n-dimensional space is a ubiquitous problem in statistical inference and data science. It is especially challenging in the (practically relevant) context of high-dimensional statistics when the dimension n is large (so that there are many degrees of freedom) but the number of points m is only linear in n (so that the available information is limited). An important special case is the problem of clustering points generated by Gaussian mixture models (GMM). These are probabilistic models where all the data points are generated from a mixture of a *finite* number r of Gaussian distributions.

In this paper, we are interested in understanding the fundamental limits on our ability to cluster points generated from GMM in the high-dimensional regime, when $m, n \to \infty$ while $m/n = \alpha$ is finite. We consider both a *computational* and an *information theoretic* viewpoint, and wish to answer the following questions: Given data generated by a GMM, (i) what is the best possible estimate, information-theoretically, of the parameters of the Gaussian mixtures, and of the individual assignments of the points? (ii) Are such optimal estimates computationally feasible in practice by polynomial-time algorithms? (iii) How do standard, widely used, methods such as principal component analysis [1, 2] compare with these optimal predictions?

We adress these questions by taking advantage of the recent burst of activity in the related problem of low-rank matrix factorization, using the cavity method from statistical physics [3] and the associated approximate message passing (AMP) algorithm [4, 5]. We shall not attempt at mathematical rigor here, but it is worth noting that recent progress on closely related problems [6–8] is likely transferable to the present situation.

I. MODEL AND SETTING

Consider data generated by a GMM with r clusters: for each $k \in \{1, \ldots, r\}$ we draw each coordinate of $V_k \in \mathbb{R}^n$ from a Gaussian of zero mean and unit variance. We then generate $m = \alpha n$ points $x_1, \ldots, x_m \in \mathbb{R}^n$ independently as follows: for each j, choose $t_j \in \{1, \ldots, r\}$ uniformly at random, and set $x_j = \sqrt{\rho/n} V_{t_j} + U_j$ where $U_j \in \mathbb{R}^n$ and has mean 0 and variance Δ in each coordinate. Here ρ is a parameter playing the role of the signal to noise ratio and the $\sqrt{\rho/n}$ factor ensures that the displacement of the centers is of the same order as the fluctuation of a high-dimensional Gaussian point around the surface of a sphere of radius n. We consider the limit $n \to \infty$, while $\alpha, \rho, \Delta = O(1)$.

Using this generative model makes the clustering problem a non-trivial one: the clusters overlap significantly; assigning points to clusters without errors is impossible and it is also impossible to recover exactly the parameters of the GMM. We thus ask instead what is the best possible overlap with the ground-truth clusters. It is instrumental to view this problem as noisy low-rank matrix factorization. Specifically, let V denote the $n \times r$ matrix whose kth column is V_k, and let S be the $m \times r$ matrix where $S_{jk} = 1$ if $t_j = k$ and 0 otherwise. We denote by $v_i, s_j \in \mathbb{R}^r$ the ith and jth rows of V and S; note that v_i is the vector of the ith coordinate of the cluster centroids. Then the observed data X is an $n \times m$ matrix

$$X = \sqrt{\frac{\rho}{n}} V S^\top + U, \qquad (1)$$

where U is a Gaussian i.i.d. $n \times m$ random matrix with mean 0 and variance Δ. Thus X is a noisy observation of a matrix $V S^\top$ of rank r. We shall use without loss of generality $\Delta = 1$, as other values of Δ simply cause a rescaling of the signal-to-noise ratio ρ.

Assuming the data is generated by model (1), we want to know how well we can reconstruct the true cluster assignments $\{t_j\}$ and cluster centers $\{V_k\}$. To answer these questions, we use Bayesian inference. The generative model for the data is

$$P(X, S, V) = P(S)P(V)\, P(X|S, V)$$
$$= \prod_{i=1}^{n} P_v(v_i) \prod_{j=1}^{m} P_s(s_j) \prod_{i=1}^{n} \prod_{j=1}^{m} e^{-\frac{1}{2\Delta}\left(X_{ij} - \frac{\sqrt{\rho}\, v_i^\top s_j}{\sqrt{n}}\right)^2},$$

where P_v and P_s are the probability distributions from which the cluster centers v_i and assignments s_j are drawn. We take P_v to be Gaussian with mean zero and covariance matrix I_r. Since the labels t_j are uniform, P_s is the uniform distribution over the r canonical basis vectors

$$P_v(v) = \frac{1}{(2\pi)^{r/2}} \exp\left(\frac{-\|v\|_2^2}{2}\right), \qquad (2)$$

$$P_s(s) = \frac{1}{r} \sum_{i=1\cdots r} \delta\left(s - \vec{e_i}\right). \qquad (3)$$

Using Bayes' rule to compute the posterior probability of S and V given the data X, one gets

$$P(S,V|X) = \frac{1}{P(X)} \prod_{i=1}^{n} P_v(v_i) \prod_{j=1}^{m} P_s(s_j)$$
$$e^{-\frac{1}{2\Delta} \sum_{ij} \left(X_{ij} - \frac{\sqrt{\rho} v_i^\top s_j}{\sqrt{n}}\right)^2}. \qquad (4)$$

We denote the ground truth cluster centers and assignments used to generate X as V_0 and S_0. We assume that the correct prior distributions (2) and (3) are known to the algorithm and analysis; specifically, the correct number of clusters r and the parameters ρ, Δ. We call this the *Bayes-optimal* setting. The question of what happens when the model is incorrect is also interesting, but in this article we let is aside. We note, however, that taking the limit $\rho \to \infty$ and $\Delta \to 0$ in the Bayesian setting would correspond to the minimization problem that the k-means algorithm [9] is trying to solve. The objective of *soft k-means* corresponds instead to the Bayesian setting in the limit $\rho \to 0$ while retaining a nonzero noise level Δ. Spectral methods, on the other hand, are equivalent to Bayesian inference where one abandons the hard constraint that the labels t_j are discrete, and effectively replace P_s with a Gaussian prior.

The above Bayesian approach converts the problem of estimating an assignment of points to clusters that maximizes the number of correctly assigned points to the problem of computing marginal probabilities of the posterior distribution. The mean squared error in estimating V_0 from X is minimized by using the (marginalized) conditional expectation of V given the data [10]. For estimation of S_0 we aim to minimize the number of mis-classifications. The corresponding Bayes-optimal estimator $\hat{S}_{\text{MaxProb}}$ is constructed by assigning each data point to the cluster that, according to the marginals of (4), it is most likely to belong to. Given $\hat{S}_{\text{MaxProb}}$, the *overlap* between the estimated cluster assignments and the ground truth is defined as

$$\text{ErrorRate} = \frac{1}{m} \text{Tr}\left(\hat{S}_{\text{MaxProb}}^\top S_0\right) \qquad (5)$$

$$\text{Overlap} = \frac{\text{ErrorRate} - 1/r}{1 - 1/r}. \qquad (6)$$

The ErrorRate is the percentage of correctly assigned points. This rate is 1 when the reconstruction is perfect. When taking the assignment at random the ErrorRate is $1/r$. The normalization in (6) is introduced in order to have a quantity that varies between 0 meaning no better reconstruction that chance, to 1 meaning perfect reconstruction.

The marginal posterior probabilities of the v_i and s_j are, however, difficult to compute in general. This is because the interaction terms involving the observations X_{ij} couple these variables together. This is where we turn to techniques from the statistical physics of disordered systems which allow marginalization of joint distributions such as (4).

II. RELATION TO PREVIOUS WORKS

Study of the Bayes-optimal estimation in Gaussian mixture clustering in the regime of finite α is restricted to the statistical physics literature. Results is statistic are either for $\alpha \to \infty$ or for algorithms (such as spectral ones) that are suboptimal in the present setting. Existing literature that treats the Bayes-optimal setting either considered the problem only algorithmically [5] or analyzes it only for two clusters, $r = 2$, [11–14]. The main contribution of this paper is to extend the analysis to the general (finite) number of clusters and realize that the case with several clusters behaves considerably differently from $r = 2$.

Our approach relies on Approximate Message Passing (AMP) algorithm and its theoretical analysis, a large part of which was born in the statistical physics, along with the cavity and replica approach [3, 15, 16]. In the present case, the AMP algorithm is very close to the TAP equations [17]. AMP for Gaussian mixture clustering is a case of AMP for low-rank matrix factorization as has been written in [4, 5]. State evolution (SE) [6, 18, 19] is a theoretical technique, closely related to the cavity method [15], that allows to exactly characterize the behavior of the AMP algorithm in the limit of infinite system size. A number of works on low-rank matrix factorization used recently this approach [20–22]. The AMP, the state evolution and associated Bethe free energy were presented recently in a rather generic form in [22] and we use these results extensively here in the specific case of GMM clustering and analyze the associated phase transitions and phase diagrams.

Proving the results presented here in full generality, for a general case of low-rank matrix factorizations, is a task that has attracted a lot of attention recently. In particular the replica prediction in the symmetric rank-1 case has been proven rigorously in a number of situations [7, 20, 21, 23] and a very generic proof now exists [8]. This strongly strengthens the claim that the replica predictions for the non-symmetric matrices (which we use here) are exact, despite the conjecture being still open.

III. SUMMARY OF MAIN RESULTS

By analyzing the Bayes-optimal estimation in the Gaussian mixture clustering we find that depending on the values of ρ, r and α (without loss of generality we consider from now on $\Delta = 1$), the problem appears in one of the following three phases:

- **EASY:** The theoretically optimal reconstruction is better than chance both for the clustering problem and for finding the centroids and the AMP algorithm described in this article is able to reach this performance in the limit of large systems.

- **IMPOSSIBLE:** The theoretically optimal reconstruction does not perform better than chance. The matrix X contains no exploitable information on the assignment of each point. For the centroids the best estimate is given by taking the mean of all the data points. No algorithm can exist to recover the assignment of the points better than chance.
- **HARD:** The theoretically optimal reconstruction gives a result that is better than chance both for the clustering problem and for finding the centroids. *However*, without using prior knowledge on the assignment of the points or on the centroids, the AMP algorithm is not able to converge toward its optimal fixed point. In fact, the AMP fixed point reached from an uninformed initialization does not provide any information on the assignment of the points. From the current knowledge, it is plausible that this phase is computationally hard not only for AMP but for all polynomial algorithms.

Depending on the number of clusters (that is, the rank r) and the sample-to-dimension ratio α we identified a boundary between two regimes at

$$r_c = 4 + 2\sqrt{\alpha}. \tag{7}$$

- If the number of clusters is small enough, $r < r_c$ there are only two phases in the problem as the signal-to-noise parameter ρ increases: The impossible and the easy one. This among others means that in the large size limit we are always able to reach Bayesian optimal reconstruction performance using the AMP algorithm. These two phases are separated by a sharp impossible/easy phase transition when

$$\rho_c = \frac{r}{\sqrt{\alpha}}. \tag{8}$$

The problem is easy when $\rho > \rho_c$ and impossible otherwise.
- If instead the number of clusters is large enough, $r > r_c$ the three phases above are observed when decreasing ρ.
  - When $\rho > \rho_c = r/\sqrt{\alpha}$ we are in the **EASY** phase.
  - When $\rho_c > \rho > \rho_{\text{IT}}$ we are in the **HARD** phase.
  - When $\rho_{\text{IT}} > \rho$ we are in the **IMPOSSIBLE** phase.

IT here stands for information theoretic. We also compute the asymptotic behavior of ρ_{IT} for a large number of clusters and found that

$$\rho_{\text{IT}}(r, \alpha) = 2\sqrt{\frac{r \log r}{\alpha}} (1 + o_r(1)), \tag{9}$$

This means that for r large the HARD phase is very broad. Note that this last asymptotic result has been recently proven rigorously in [24] using the first and second moment methods, another indication of the correctness of the cavity/replica assumptions.

In what follows, we give statistical physics justifications for these claims, based on the analysis of AMP and on the interpretation of the Bethe free energy (related to the mutual information) as the exact one. We hope this will motivate further rigorous studies in this direction.

IV. APPROXIMATE MESSAGE PASSING (AMP)

We recall the AMP algorithm here using the notations of [22]. Let us define two *denoising* functions $f_v(A, B) \in \mathbb{R}^{r \times 1}$ and $f_s(A, B) \in \mathbb{R}^{r \times 1}$ as the means of the probability distributions

$$\frac{1}{\mathcal{Z}_v(A, B)} P_v(v) \exp\left(B^\top v - \frac{v^\top A v}{2}\right), \tag{10}$$

$$\frac{1}{\mathcal{Z}_s(A, B)} P_s(s) \exp\left(B^\top s - \frac{s^\top A s}{2}\right), \tag{11}$$

where $\mathcal{Z}_v(A, B)$ and $\mathcal{Z}_s(A, B)$ are normalization factors. Here B is an r dimensional column vector and A is an $r \times r$ matrix. For GMM clustering priors (2-3), we thus find

$$f_v(A, B) = (I_r + A)^{-1} B, \tag{12}$$

$$f_{s,k}(A, B) = \frac{\exp(B_k - A_{kk}/2)}{\sum\limits_{l=1,\cdots,r} \exp(B_l - A_{ll}/2)}. \tag{13}$$

Taking the derivative of $f_v(A, B)$ and $f_s(A, B)$ with respect to B yields the covariance matrices of (10) and (11), and we thus define

$$f_v'(A, B) = \left(\frac{\partial f_v}{\partial B}\right) \in \mathbb{R}^{r \times r}, \tag{14}$$

$$f_s'(A, B) = \left(\frac{\partial f_s}{\partial B}\right) \in \mathbb{R}^{r \times r}. \tag{15}$$

Algorithm 1 AMP for clustering mixtures of Gaussians

Input: Data $\mathbf{X}$ and initial condition $\hat{v}^{\text{init}}, \hat{s}^{\text{init}}$.

$$\begin{aligned}
\forall t &\in \{0, 1\}, \\
\hat{v}_i^t &\leftarrow \hat{v}_i^{\text{init}}, \ \hat{s}_j^t \leftarrow \hat{s}_j^{\text{init}}, \\
\sigma_{v,i}^t &\leftarrow 0^{\mathbb{R}^{r \times r}}, \ \sigma_{s,j}^t \leftarrow 0^{\mathbb{R}^{r \times r}}
\end{aligned}$$

For $t \geq 1$ compute

$$\forall i \in [1; n], B_{v,i}^t \leftarrow \sqrt{\frac{\rho}{n}} \sum_{j=1\cdots m} X_{ij} \hat{s}_j^t - \frac{\alpha \rho}{m} \sum_{j=1\cdots m} \sigma_{s,j}^t \hat{v}_i^{t-1}$$

$$A_v^t \leftarrow \frac{\alpha \rho}{m} \sum_{j=1\cdots m} \hat{s}_j^t \hat{s}_j^{t\top}$$

$$\forall i \in [1; n], \hat{v}_i^t \leftarrow f_v\left(A_v^t, B_{v,i}^t\right), \ \sigma_{v,i}^t \leftarrow f_v'\left(A_v^t, B_{v,i}^t\right)$$

$$\forall j \in [1; m], B_{s,j}^t \leftarrow \sqrt{\frac{\rho}{n}} \sum_{i=1\cdots n} X_{ij} \hat{v}_i^t - \frac{\rho}{n} \sum_{i=1\cdots n} \sigma_{v,i}^t \hat{s}_j^t$$

$$A_v^t \leftarrow \frac{\rho}{n} \sum_{i=1\cdots n} \hat{v}_i^t \hat{v}_i^{t\top}$$

$$\forall j \in [1; m], \hat{s}_j^{t+1} \leftarrow f_s\left(A_s^t, B_{s,j}^t\right), \ \sigma_{s,j}^{t+1} \leftarrow f_s'\left(A_s^t, B_{s,j}^t\right)$$

Here $\forall i \in [1; n] B_{v,i}^t \in \mathbb{R}^{r \times 1}$, $\forall j \in [1; m] B_{s,j}^t \in \mathbb{R}^{r \times 1}$ and $A_v^t \in \mathbb{R}^{r \times r}, A_s^t \in \mathbb{R}^{r \times r}$.
Iterate till convergence.

V. STATE EVOLUTION (SE)

The SE is a method for tracking the evolution of an AMP algorithm along iterations in the large system-size limit. A physics derivation of these SE equations can be found in

[22], rigorous proof follows from [6]. We first introduce the following order parameters

$$M_v^t = \frac{1}{n} \sum_{i=1\cdots n} \hat{v}_i^t v_{0,i}^\top, \qquad (16)$$

$$M_s^t = \frac{1}{m} \sum_{j=1\cdots m} \hat{s}_j^t s_{0,j}^\top, \qquad (17)$$

where $\hat{v}_i^t$ $\hat{s}_j^t$ are the estimators of the posterior means of the variables v_i and s_j at time t defined by

$$\hat{v}_i^t = f_v(A_v^t, B_{v,i}^t), \qquad (18)$$

$$\hat{s}_j^{t+1} = f_s(A_s^t, B_{s,j}^t). \qquad (19)$$

The $r \times r$ matrices M describe the overlap of the estimators with the ground truth solution. The SE reads [22]

$$M_v^t = \mathbb{E}_{W,v_0} \left[f_v \left(\alpha \rho M_s^t, \alpha \rho M_s^t v_0 + \sqrt{\alpha \rho M_s^t} W \right) v_0^\top \right], \quad (20)$$

$$M_s^{t+1} = \mathbb{E}_{W,s_0} \left[f_s \left(\rho M_v^t, \rho M_v^t s_0 + \sqrt{\rho M_v^t} W \right) s_0^\top \right], \quad (21)$$

where W is an r-dimensional Gaussian variable with zero mean and unit variance in each coordinate, and v_0 and s_0 are random variables distributed according to P_v (2) and P_s (3) respectively.

A fixed point of the SE equations is a local extremum of the so-called Bethe free energy ϕ_{B} [22]

$$\phi_{\mathrm{B}}(M_v, M_s) = \frac{\alpha \rho \mathrm{Tr}(M_v M_s^\top)}{2}$$
$$- \mathbb{E}_{W,v_0} \left[\log \mathcal{Z}_v(\alpha \rho M_s, \alpha \rho M_s v_0 + \sqrt{\alpha \rho M_s} W) \right]$$
$$- \alpha \mathbb{E}_{W,s_0} \left[\log \mathcal{Z}_s(\rho M_v, \rho M_v s_0 + \sqrt{\rho M_v} W) \right], \quad (22)$$

where the quantities $\mathcal{Z}_v$ and $\mathcal{Z}_s$ are the normalization factors from (10) and (11). A stable fixed point of the SE equations above is a local minimum of $\phi_{\mathrm{B}}(M_v, M_s)$ as can be seen by taking partial derivatives with respect to M_v and M_s.

It follows from the replica theory of statistical physics that if there are multiple stable fixed points of (20-21), only the one with minimal free energy ϕ_{B} corresponds to the performance of the Bayes-optimal estimation. On the other hand the performance reached by the AMP algorithm corresponds to the fixed point of SE with largest error.

VI. Analysis for Mixture of Gaussians

We analyze now the state evolution with the f_v and f_s functions corresponding to the GMM (12-13) for general number of clusters r. The difficulty here lies in the fact that there is in general no analytic expression for updating equations (20-21). We observe that the following form of the order parameters is conserved under the update of SE equations for the

$$M_s^t = \frac{I_r b_s^t}{r} + (1 - b_s^t) \frac{J_r}{r^2} \qquad (23)$$

$$M_v^t = b_v^t I_r + b_{v,J}^t \frac{J_r}{r}, \qquad (24)$$

where $(b_s^t, b_v^t, b_{v,J}^t) \in [0;1]^3$, and I_r and J_r are respectively the identity matrix and the $r \times r$ matrix filled with 1. Having

$(b_s^t, b_v^t, b_{v,J}^t) = (1,1,0)$ would mean that we have achieved perfect reconstruction of the ground truth, while $(b_s^t, b_v^t) = (0,0)$ means that we are not able to extract any information from the matrix X beyond the average of the k clusters V_k.

Rewriting eqs. (20-21) using (12-13) and (23-23) we get the following SE equations for the GMM

$$b_v^t = \frac{b_s^t \rho}{\frac{r}{\alpha} + b_s^t \rho}, \qquad b_s^{t+1} = \mathcal{M}_r \left(b_v^t \rho r \right), \qquad (25)$$

where

$$\mathcal{M}_r(x) = \frac{1}{r-1} \left[r \int \frac{e^{\frac{x}{r} + u_1 \sqrt{\frac{x}{r}}}}{e^{\frac{x}{r} + u_1 \sqrt{\frac{x}{r}}} + \sum_{i=2}^{r} e^{u_i \sqrt{\frac{x}{r}}}} \prod_{i=1}^{r} \mathcal{D}u_i - 1 \right]. \qquad (26)$$

Where the u_i are Gaussian variables of mean 0 and unit variance. The equations close on b_v and b_s and do not depend on the term $b_{v,J}$. We can combine these to obtain a single update equation for the scalar variable b_s^t

$$b_s^{t+1} = \mathcal{M}_r \left(b_s^t \frac{\rho^2}{\frac{1}{\alpha} + \frac{\rho b_s^t}{r}} \right). \qquad (27)$$

The function $\mathcal{M}_r$, including the expansions for small ρ and for large r, has been studied previously in [22]. From a numerical point of view $\mathcal{M}_r(x)$ can be effectively computed with a Monte Carlo scheme if one takes advantage of the permutation symmetry of the Gaussian variables.

A. Analysis of the phase transitions

Note that $b_s = \mathcal{M}_r(0) = 0$ is always a fixed point that we will call *uninformative*. By expanding $\mathcal{M}_r(x)$ around 0 one gets:

$$b_s^{t+1} = \frac{\alpha b_s^t \rho^2}{r^2} + \frac{\alpha^2 b_s^{t^2}}{2} \left[r - 4 - \frac{2r}{\rho} \right] \frac{\rho^4}{r^4}. \qquad (28)$$

We are interested in when the uninformative fixed point becomes numerically unstable. From the expansion of (28) one deduces that this occurs when

$$\rho > \rho_c = \frac{r}{\sqrt{\alpha}}. \qquad (29)$$

Looking at the second derivative of (28), we deduce that when the uninformative fixed point becomes unstable, the second derivative is proportional to $r - 4 - 2\sqrt{\alpha}$; if this is negative then this means that another stable fixed point appears close to 0 for $\rho > \rho_c$. If on the other hand the second derivative is positive when ρ increases and crosses ρ_c, then the new stable fixed point will not be close to zero and we see a discontinuous jump in the MSE achieved by initializing the iteration close to the uninformative fixed point. This phenomenon is known as a first-order transition. If one fixes the number r and α, we have a first order transition in the GMM if $r > r_c = 4 + 2\sqrt{\alpha}$.

It turns out that this sufficient criteria is also necessary. In order to analyze the SE numerically we introduce two ways to initialize the equations

- Uninformative initialization $b_s^{t=0} \approx 0$: Initializing in such a way is equivalent to assuming that we use
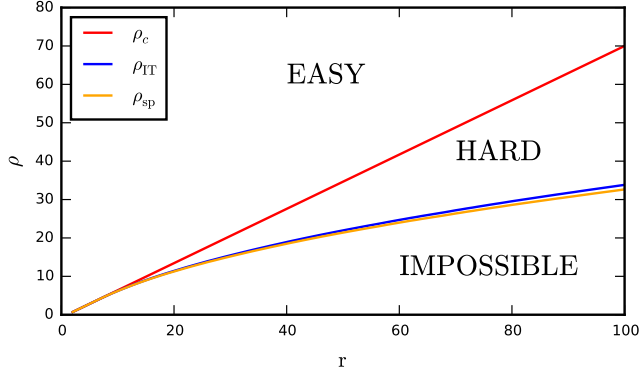
Fig. 1. The three phase transitions as a function of the number of clusters r and the signal-to-noise parameter ρ, for sample-to-dimension ratio $\alpha = 2$. In red is the algorithmic phase transition $\rho_c = r/\sqrt{\alpha}$, in blue the information theoretic threshold ρ_{IT} and in yellow the spinodal threshold ρ_{sp}. As the number of clusters r grows the gap between the algorithmic and the information-theoretic transition grows larger.

no knowledge about the ground truth signal. From an AMP point of view this means starting with means of messages close to 0. We denote b_{AMP} the reached fixed point.

- Informative initialization $b_s^{t=0} = 1$: This means initializing in the ground truth configuration. From an AMP point of view this corresponds to start with messages equal to the solution. Of course, in reality the algorithm would not have access to the solution; but here we apply this initialization in order to track the corresponding fixed point. We denote b_{inf} the reached fixed point.

For $r < 4 + 2\sqrt{\alpha}$ iterations with these two initializations lead to the same fixed point and the situation looks qualitatively as in Fig. 2 for two clusters. For $\rho < \rho_c$ we reach the uninformative fixed point $b_s = 0$, whereas for $\rho > \rho_c$ we reach a fixed point with positive overlap with the ground truth, $b_s > 0$.

If $r > 4 + 2\sqrt{\alpha}$, then depending on the value of ρ four different situations happen:

- $\rho > \rho_c$: Here $b_{\text{AMP}} = b_{\text{inf}} > 0$: The AMP algorithm is able to reach the information theoretically optimal reconstruction. This is called the **EASY** phase.
- $\rho_c > \rho > \rho_{\text{IT}}$: Here $b_{\text{inf}} > b_{\text{AMP}} = 0$, i.e. there are two fixed points to the SE equations and b_{inf} leads to a lower free energy (22). The AMP algorithm starting from the uninformative initialization is not able to reach the information theoretically optimal performance. Nevertheless it is possible, although in exponential time, to find a better fixed point with lower free energy. This is called the **HARD** phase.
- $\rho_{\text{IT}} > \rho > \rho_{\text{sp}}$: Here still $b_{\text{inf}} > b_{\text{AMP}} = 0$, but now b_{AMP} has a lower free energy (22). The AMP algorithm starting from the informative state is able to reach a good fixed point correlated with the solution. However, finding this fixed point without prior knowledge of the solution is information theoretically impossible: it is hidden among an exponential number of other

fixed points of AMP that have similar likelihood. This is called the information-theoretically **IMPOSSIBLE** phase.

- $\rho_{\text{sp}} > \rho$: Here $b_{\text{AMP}} = b_{\text{inf}} = 0$, i.e. there is only one fixed point to the SE equations. This is also the information-theoretically **IMPOSSIBLE** phase.

To compute these transitions ρ_{IT} and ρ_{sp} numerically we consider one value of $b_s = \mathcal{M}_r(x)$ and ask what is the value of ρ such that $b_s = \mathcal{M}_r(x)$ is a fixed point. Using (27) the answer is

$$\rho(x, r) = \frac{x}{2r} + \sqrt{\frac{x^2}{4r^2} + \frac{x}{\alpha \mathcal{M}_r(x)}} \, . \qquad (30)$$

The spinodal transition ρ_{sp} is the minimum value of ρ for which a fixed point other than 0 exists. We can estimate this by minimizing $\rho(x, r)$ (30) with respect to x and at fixed r. The information-theoretic transition ρ_{IT} is obtained by expressing the difference in the free energy between $b_s = 0$ and $\mathcal{M}_r(x)$ at a given ρ and then requiring this quantity to be 0. It is possible to express the Bethe free energy using $\mathcal{M}_r$. If one integrates the gradient of ϕ_{B} along the path $g(u)$ defined by

$$\forall u \in [0, \mathcal{M}_r(x)], g(u) = \left(u, u \frac{\rho(x)^2/r}{\frac{1}{\alpha} + \frac{\rho(x)u}{r}} \right), \qquad (31)$$

after integrating by parts

$$\phi_{\text{B}}(0, \rho(x, \alpha, r), \alpha; r) \qquad (32)$$
$$- \quad \phi_{\text{B}}(\mathcal{M}_r(x), \rho(x, \alpha, r), \alpha, r)$$
$$= \quad \alpha \frac{r-1}{2r^2} \left[\int_0^x du \mathcal{M}_r(u) + \int_0^{\mathcal{M}_r(x)} du \frac{u \rho^2}{\frac{1}{\alpha} + \frac{u\rho(x)}{r}} - x \mathcal{M}_r(x) \right].$$

The information-theoretic transition is found where $b_{\text{AMP}}(r, \alpha, \rho)$ and $b_{\text{inf}}(r, \alpha, \rho)$ both have the same free energy. The behavior of ρ_c, ρ_{IT} and ρ_{sp} as a function of ρ for $\alpha = 2$ is illustrated in Fig. 1.

B. Large number of clusters

Formulas (30-32) also allow us to explore the large r limit of these solutions. From [22] we know that

$$\forall \beta > 0, \lim_{r \to \infty} \mathcal{M}_r(\beta r \log r) = 1_{\beta > 2}. \qquad (33)$$

We can compute the asymptote of ρ_{IT} and ρ_{sp} when $r \to \infty$. We can do that by setting $x = \beta r \log(r)$ and then replacing $\mathcal{M}_r$ by (33) in (30) and (32).

To get ρ_{sp} one minimizes (30) with respect to β. After some computation one gets

$$\lim_{r \to +\infty} \frac{\rho(\beta r \log(r), \alpha)}{\sqrt{r \log(r)}} = \begin{cases} +\infty, & \text{if } \beta < 2 \\ \sqrt{\beta/\alpha}, & \text{if } \beta \geq 2 \end{cases} \qquad (34)$$

Therefore in the large r limit $\rho(\beta r \log(r))$ is minimized by taking $\beta = 2$ and one gets

$$\rho_{\text{sp}}(r, \alpha) = \sqrt{\frac{2r \log r}{\alpha}}(1 + o_r(1)) \, . \qquad (35)$$

To get ρ_{IT} one finds the β such that (32) is set to 0. In order to do that we write $x_{\mathrm{IT}} = \beta_{\mathrm{IT}} r \log(r)$. We take $\beta_{\mathrm{IT}} > 2$ since we want to have ρ to be above ρ_{sp} so that there are multiple fixed point to the SE equations. For $\beta > 2$ to leading order one has

$$\rho(x = \beta r \log(r), r) = \sqrt{\frac{\beta r \log(r)}{\alpha}}(1 + o_r(1)). \quad (36)$$

Therefore,

$$\frac{\rho(x = \beta r \log(r), r)}{r} \ll \frac{1}{\alpha}. \quad (37)$$

Using this, equation (32) can be further simplified. One gets

$$\phi_{\mathrm{B}}(0, \rho(x_{\mathrm{IT}}, \alpha, r), \alpha; r) - \phi_{\mathrm{B}}(\mathcal{M}_r(x_{\mathrm{IT}}), \rho(x_{\mathrm{IT}}, \alpha, r), \alpha, r)$$
$$= 0 \approx \alpha \frac{r-1}{2r^2}\left[\int_0^{x_{\mathrm{IT}}} \mathrm{d}u \mathcal{M}_r(u) - \frac{x_{\mathrm{IT}} \mathcal{M}_r(x_{\mathrm{IT}})}{2}\right], \quad (38)$$

By setting $x_{\mathrm{IT}} = \beta_{\mathrm{IT}} r \log(r)$ in this equation and taking the $r \to \infty$ limit one gets

$$0 = \int_0^{\beta_{\mathrm{IT}}} \mathrm{d}u 1(u > 2) - \frac{\beta_{\mathrm{IT}} 1(\beta_{\mathrm{IT}} > 2)}{2}. \quad (39)$$

This is solved for $\beta_{\mathrm{IT}} = 4$, therefore one has

$$\rho_{\mathrm{IT}}(r, \alpha) = 2\sqrt{\frac{r \log(r)}{\alpha}}(1 + o_r(1)). \quad (40)$$

Thus, for large number of clusters r, the gap between the information-theoretic detectability threshold ρ_{IT} and the algorithmic threshold $\rho_c = r/\sqrt{\alpha}$ becomes large.

C. Algorithmic comparison

Fig. 2 and 3 contain numerical experiments with the AMP algorithm, and comparison with the theoretically predicted performance given by state evolution analysis. Data for both $r = 2$ and $r = 20$ clusters are presented. These two cases have qualitatively different properties. As predicted, the case with $r = 20$ clusters exhibits a first order transition: there is a sharp jump in the overlap at ρ_c. Instead the case $r = 2$ exhibits a second order transition: the overlap is continuous, only its derivative has a discontinuity.

In both Fig. 2 and 3 we also compare to the performance of the principal component analysis (PCA) performed on the matrix X. PCA is a standard spectral method to solve data clustering, one computes r leading singular vectors of X and instead of clustering m points in n dimensions, one concatenates the singular vectors into m r-dimensional vectors and clusters in the r-dimensional space which is much simpler. The overlap reached with the PCA clustering follows from a more general theory of low-rank perturbations of random matrices [2], but it can also be derived from the state evolution analysis of AMP as we present in the appendix. In this case of GMM with equal-size clusters the phase transition observed in PCA coincides with the phase transition of AMP ρ_c. Concerning the performance as measured by the overlap (6), we observe that although for two clusters the difference between the performance of
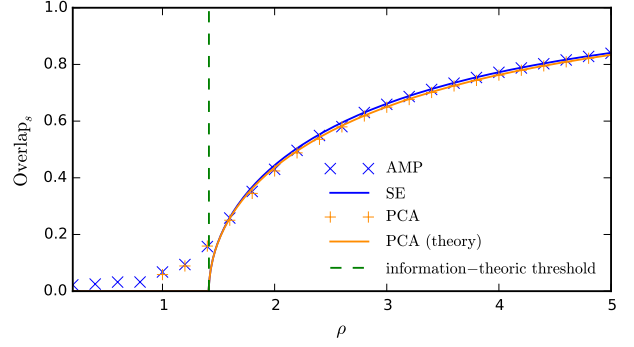
Fig. 2. Bayes-optimal overlap for clustering with $r = 2$ clusters (full lines, using State Evolution), together with the results of numerical simulations (points), for $\alpha = 2$, $n = 1000$ and $m = 2000$. The overlap is defined in (6). The information-theoretic threshold at $\rho_c \approx 1.41$ is materialized by the green dashed line. Both AMP and PCA are able to perform better than chance beyond this transition. The performance of AMP and PCA are comparable in this case.
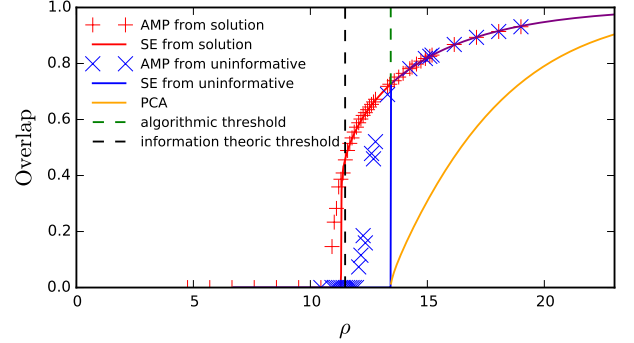
Fig. 3. Bayes-optimal overlap for clustering with $r = 20$ clusters (full lines, using State Evolution), together with the results of numerical simulations (points), for $\alpha = 2$, $n = 10000$ and $m = 20000$. In this case one observes an algorithmically hard phase, at $\rho_c \approx 14.1$ the overlap achieved by AMP in infinite size systems has a discontinuity. The algorithmic and information-theoretic thresholds are depicted by the vertical blue line and the dashed black line. Because of the discontinuous nature of the transition, the finite size effects are sizable in the hard region: the red points with non-zero overlap below ρ_c show that AMP is able to reconstruct a fraction of the $r = 20$ clusters. Increasing n decreases these effects.

the sub-optimal PCA and the Bayes-optimal AMP is hardy visible, for 20 clusters the performance of AMP close to the algorithmic transition ρ_c is considerably better.

VII. Conclusion

We analysed the problem of clustering high-dimensional data generated by the Gaussian mixture model. We computed the asymptotic accuracy of the Bayes-optimal estimator and compared it to the accuracy obtained by the approximate message-passing algorithm. We located phase transitions in both the information-theoretic and algorithmic performance.

Our main result is that, when the number of clusters is sufficiently large, eq. (7), there is a gap between the information-theoretic threshold—where it becomes possible to label points and find the cluster centers better than chance—and the computational threshold, at which polynomial-time algorithms such as PCA or AMP succeed. This suggests that, as has been conjectured for

analogous problems in statistical inference, there is a hard-but-detectable regime where clustering is information-theoretically possible, but computationally intractable.

APPENDIX I
PERFORMANCE OF PCA

Using state evolution one can analyze the distribution of the top eigenvectors of X and therefore the performance of PCA on the Gaussian mixture clustering problem. Let us write the density evolution equation the posterior measure (4) with X created using a GMM, and where we have replaced the Bayes-optimal prior on v_i and s_j with Gaussian priors

$$P_v(v_i) = \frac{\exp\left(\frac{-\|v_i\|_2^2}{2}\right)}{\sqrt{2\pi}^r}, \; P_s(s_j) = \frac{\exp\left(\frac{-r\|s_j\|_2^2}{2}\right)}{\sqrt{2\pi/r}^r}. \quad (41)$$

Let us define V_{PCA}, S_{PCA} the matrices containing the first r singular vectors of X. Where $V_{\text{PCA}} \in \mathbb{R}^{n \times r}$, $S_{\text{PCA}} \in \mathbb{R}^{m \times r}$. S_{PCA} defines m points in a r dimensional space. We expect the r dimensional points extracted from S_{PCA} to form a mixture of Gaussians. Let us write the AMP for Gaussian priors. The general AMP equations for the mismatching priors can be found in [5]. One gets

$$\hat{V}^t = \left(\sqrt{\frac{\rho}{n}} X \hat{S}^t - \rho\alpha\hat{V}^{t-1}\Sigma_s^t\right)\Sigma_v^t, \quad (42)$$

$$\Sigma_v^t = \left(I_r + \rho/n \hat{S}^{t\top} S^t\right)^{-1}, \quad (43)$$

$$\hat{S}^{t+1} = \left(\sqrt{\frac{\rho}{n}} X^\top \hat{V}^t - \rho\hat{S}^t\Sigma_v^t\right)\Sigma_s^{t+1}, \quad (44)$$

$$\Sigma_s^{t+1} = \left(rI_r + \rho/n \hat{V}^{t\top} V^t\right)^{-1}. \quad (45)$$

We aim to prove that at a fixed point $\hat{V}$ and $\hat{S}$ will be singular values of X.

After simplifications of the above equations one gets

$$\hat{V} = \sqrt{\frac{\rho}{n}} X \hat{S} \left(\Sigma_v^{-1} + \rho\alpha\Sigma_s\right)^{-1}, \quad (46)$$

$$\hat{S} = \sqrt{\frac{\rho}{n}} X^\top \hat{V} \left(\Sigma_s^{-1} + \rho\Sigma_v\right)^{-1}. \quad (47)$$

By putting this in (46) in (47) one gets

$$\hat{S} = \frac{\rho}{n} X^\top X \hat{S} \left(\Sigma_v^{-1} + \rho\alpha\Sigma_s\right)^{-1} \left(\Sigma_s^{-1} + \rho\Sigma_v\right)^{-1}. \quad (48)$$

This means that the columns of $\hat{S}$ are a linear combination of eigenvectors of $X^\top X$ that is right eigenvector of S.

We can also write the corresponding state evolution equations. One gets

$$M_v^t = \mathbb{E}_{W,v_0}\left[f_v\left(\alpha\rho M_s^t, \alpha\rho M_s^t v_0 + \sqrt{\alpha\rho M_s^t}W\right) v_0^\top\right],$$

$$Q_v^t = \mathbb{E}_{W,v_0}\left[f_v\left(\alpha\rho M_s^t, \alpha\rho M_s^t v_0 + \sqrt{\alpha\rho M_s^t}W\right) f_v(\cdots,\cdots)\right],$$

$$M_s^{t+1} = \mathbb{E}_{W,s_0}\left[f_s\left(\rho M_v^t, \rho M_v^t s_0 + \sqrt{\rho M_v^t}W\right) s_0^\top\right],$$

$$Q_s^{t+1} = \mathbb{E}_{W,s_0}\left[f_s\left(\rho M_v^t, \rho M_v^t s_0 + \sqrt{\rho M_v^t}W\right) f_s(\cdots,\cdots)\right],$$

With the Gaussian priors (41) this gives us

$$M_v^t = (I_r + \alpha\rho Q_s^t)^{-1}\alpha\rho M_s^t, \quad (49)$$
$$Q_v^t = (I_r + \alpha\rho Q_s^t)^{-2}\left(\alpha^2\rho^2 M_s^t + \alpha\rho Q_s^t\right), \quad (50)$$
$$M_s^t = (rI_r + \rho Q_v^t)^{-1}\rho M_v^t/r, \quad (51)$$
$$Q_s^t = (rI_r + \rho Q_v^t)^{-2}\left(\rho^2 M_v^t/r + \rho Q_v^t\right). \quad (52)$$

Experiments with these equations show that stable fixed point are always of the form

$$M_v^t = m_v^t R, \quad Q_v^t = m_v^t I_r, \quad (53)$$
$$M_s^t = \frac{m_s^t R}{r}, \quad Q_s^t = \frac{m_s^t I_r}{r}, \quad (54)$$

where R is some rotation matrix. The new update equation become

$$m_v^{t+1} = \frac{\rho\alpha m_s^t}{1 + \rho\alpha m_s^t}, \quad (55)$$

$$m_s^{t+1} = \frac{\rho m_v^t}{r + \rho m_v^t}. \quad (56)$$

Let us write

$$m_v = m_v^{t\to\infty} = \max\left(\frac{\alpha\rho^2 - r^2}{r + \rho\alpha}, 0\right), \quad (57)$$

$$m_s = m_s^{t\to\infty} = \max\left(\frac{\alpha\rho^2 - r^2}{r\rho\alpha + \rho^2\alpha}, 0\right), \quad (58)$$

By combining (45), (43), (53) and (54) one sees that in the large n limit Σ_v and Σ_s are proportional to the identity. Therefore using (48) one gets

$$\hat{S}a = \frac{\rho}{n} X^\top X \hat{S}, \quad (59)$$

where a is some number. This means that $\hat{S}$ is proportional to the first eigenvectors. But the AMP analysis also tells us how $\hat{S}$ will be distributed. The details of the analysis of state evolution can be found in [22] in the $XX^\top$ case

$$\hat{s}_j = f_s\left(\rho m_v I_r, \rho m_v R s_{0,j} + \sqrt{\rho m_v}W\right). \quad (60)$$

Where W is a Gaussian variable given by $\mathcal{N}(0, I_r)$. Up to a proportionality constant one has.

$$\hat{s}_j = \sqrt{\rho m_v} R s_{0,j} + W. \quad (61)$$

This means that the first r right eigenvectors of R are distributed as a mixture of Gaussians in a r dimensional space where the centers are placed at positions $\sqrt{\rho m_v} R e_k$ and the noise is a Gaussian white noise of zero mean and covariance matrix I_r. The questions now becomes how well can one cluster this mixture of Gaussians? Since we are dealing with a finite-dimensional space $r = O(1)$ and a large number of points $m \to \infty$ then we know that we should be able to learn the parameters from this mixture of Gaussian perfectly.

The problem now becomes given a point $\hat{v}_j$ what is the chance that it was created using point $v_{0,j}$ this is done by

maximizing the likelihood. To find back with what $e_k \ \hat{v}_j$ was created one needs to compute.

$$\hat{k}(\hat{v}_j) = \mathrm{argmax}\left\{ k \in [1;r], \frac{1}{r\sqrt{2\pi}}\exp\left(\frac{-\|\hat{v}_j - \sqrt{\rho m_v}\vec{e}_k\|_2^2}{2}\right)\right\}$$

or

$$\hat{k}(\hat{v}_j) = \mathrm{argmax}\left\{ k \in [1;r], \langle \hat{v}_j ; \vec{e}_k \rangle \right\}. \tag{62}$$

One need to compute with what probability $\hat{k}$ is the right k

$$P(\hat{k}(\sqrt{\rho m_v}\vec{e}_{k_0} + W) = k_0) =$$
$$P(W_{k_0} + \sqrt{\rho m_v} > \max_{k\in[1;r], k \neq k_0}\{W_k\}).$$

This can be computed using the distribution of the maxima of random variables of zero mean and variance $r-1$ Gaussian variables.

$$P(\hat{k}(\sqrt{\rho m_v}\vec{e}_{k_0} + W) = k_0) =$$
$$\int_{-\infty}^{+\infty} \mathrm{d}u\left[1/2 - 1/2\,\mathrm{erf}\left(u - \sqrt{\rho m_v b}\right)\right]P_{\max}^{r-1}(u), \quad (63)$$

where $P_{\max}^{r-1}$ is the density probability of the maximum of $r-1$ independent Gaussian variables of mean 0 and unit variance. Where $\mathrm{erf}(x) \in [-1;1]$ is defined as $\mathrm{erf}(x) = \sqrt{\frac{2}{\pi}}\int_0^x \exp\left(\frac{-x^2}{2}\right)$.

$$P_{\max}^{r-1}(u) = \frac{r-1}{\sqrt{2\pi}}\exp(-u^2/2)\left[\frac{1+\mathrm{erf}(u)}{2}\right]^{r-2}. \tag{64}$$

Therefore one can compute the average number of errors made using PCA to cluster the data points.

$$\mathrm{ErrorRate}_{\mathrm{PCA}} = \int_{-\infty}^{+\infty}\mathrm{d}u\frac{\left[1 - \mathrm{erf}\left(u - \sqrt{\rho m_v}\right)\right]}{2}P_{\max}^{r-1}(u).$$

The mean-squared error achieved by PCA is then $r(\rho - m_v)$ and is known in the theory of low-rank perturbations of random matrices [2].

REFERENCES

[1] D. C. Hoyle and M. Rattray, "Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure," *Physical Review E*, vol. 69, no. 2, p. 026124, 2004.

[2] J. Baik, G. B. Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, vol. 33, no. 5, pp. 1643–1697, 2005.

[3] M. Mézard and A. Montanari, *Information, Physics, and Computation*, ser. Oxford Graduate Texts. OUP Oxford, 2009.

[4] S. Rangan and A. K. Fletcher, "Iterative estimation of constrained rank-one matrices in noise," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2012, pp. 1246–1250.

[5] R. Matsushita and T. Tanaka, "Low-rank matrix reconstruction and clustering via approximate message passing," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 917–925.

[6] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Information and Inference*, 2013.

[7] F. Krzakala, J. Xu, and L. Zdeborová, "Mutual information in rank-one matrix estimation," *ITW 2016, arXiv:1603.08447*, 2016.

[8] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," *NIPS 2016, arxiv:1606.04142*, 2016.

[9] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[10] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[11] T. Watkin and J.-P. Nadal, "Optimal unsupervised learning," *Journal of Physics A: Mathematical and General*, vol. 27, no. 6, p. 1899, 1994.

[12] N. Barkai and H. Sompolinsky, "Statistical mechanics of the maximum-likelihood density estimation," *Physical Review E*, vol. 50, no. 3, p. 1766, 1994.

[13] M. Biehl and A. Mietzner, "Statistical mechanics of unsupervised structure recognition," *Journal of Physics A: Mathematical and General*, vol. 27, no. 6, p. 1885, 1994.

[14] A. Buhot and M. B. Gordon, "Phase transitions in optimal unsupervised learning," *Physical Review E*, vol. 57, no. 3, p. 3326, 1998.

[15] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond*, ser. Lecture Notes in Physics. Singapore: World Scientific, 1987, vol. 9.

[16] L. Zdeborová and F. Krzakala, "Statistical physics of inference: thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.

[17] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of'solvable model of a spin glass'," *Philosophical Magazine*, vol. 35, no. 3, pp. 593–601, 1977.

[18] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[19] M. Bayati, M. Lelarge, A. Montanari *et al.*, "Universality in polytope phase transitions and message passing algorithms," *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[20] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse PCA," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 2197–2201.

[21] Y. Deshpande, E. Abbe, and A. Montanari, "Asymptotic mutual information for the binary stochastic block model," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 185–189.

[22] T. Lesieur, F. Krzakala, and L. Zdeborová, "Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015.

[23] S. Korada and N. Macris, "Exact solution of the gauge symmetric p-spin glass model on a complete graph," *Journal of Statistical Physics*, vol. 136, no. 2, pp. 205–230, 2009.

[24] J. Banks, C. Moore, R. Vershynin, and J. Xu, "Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization," *arXiv preprint arXiv:1607.05222*, 2016.