Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization

Jess Banks Cristopher Moore Roman Vershynin Nicolas Verzelen Jiaming Xu

***Abstract*—We study the problem of detecting a structured, low-rank signal matrix corrupted with additive Gaussian noise. This includes clustering in a Gaussian mixture model, sparse PCA, and submatrix localization. Each of these problems is conjectured to exhibit a sharp information-theoretic threshold, below which the signal is too weak for any algorithm to detect. We derive upper and lower bounds on these thresholds by applying the first and second moment methods to the likelihood ratio between these "planted models" and null models where the signal matrix is zero. For sparse PCA and submatrix localization, we determine this threshold exactly in the limit where the number of blocks is large or the signal matrix is very sparse; for the clustering problem, our bounds differ by a factor of $\sqrt{2}$ when the number of clusters is large. Moreover, our upper bounds show that for each of these problems there is a significant regime where reliable detection is information-theoretically possible but where known algorithms such as PCA fail completely, since the spectrum of the observed matrix is uninformative. This regime is analogous to the conjectured 'hard but detectable' regime for community detection in sparse graphs.**

I. Introduction

Many problems in machine learning, signal processing, and statistical inference have a common, unifying goal: reconstruct a low-rank signal matrix observed through a noisy channel. This framework can encompass a wide range of tasks as we vary the channel and low-rank signal, but we focus here on the case where the noise is additive and Gaussian, and the signal is relatively weak in comparison to the noise. To be precise, suppose we are given an $m \times n$ data matrix

$$X = M + W \quad \text{with} \quad M = \frac{\mathsf{snr}}{\sqrt{n}} UV^\dagger \,, \tag{1}$$

J. Banks is with the Department of Mathematics, University of California, Berkeley, Berkeley, CA, jess.m.banks@berkeley.edu. C. Moore is with the Santa Fe Institute, Santa Fe, NM, moore@santafe.edu. R. Vershynin is with the Department of Mathematics, University of Michigan, Ann Arbor, MI, romanv@umich.edu. N. Verzelen is with UMR 729 MISTEA, INRA, Montpellier, nicolas.verzelen@inra.fr. J. Xu is with the Krannert School of Management, Purdue University, West Lafayette, IN xu972@purdue.edu.

where snr is a fixed parameter characterizing the signal-to-noise ratio, $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ are generated from some known prior distribution independent of n, and $W \in \mathbb{R}^{m\times n}$ is a noise matrix whose entries are independent Gaussians with unit variance. We will refer to this as the *planted model*: it consists of a noisy observation X of a signal matrix M of rank k, and may possess additional structure through the priors on U and V.

Given the observed matrix X, the problem of interest is to reconstruct M, or at least detect that it exists. For simplicity, we will work in the Bayes-optimal case where model parameters such as the true rank and signal-to-noise ratio are known to the estimators. In the low signal-to-noise ratio regime we consider, exact reconstruction of M is fundamentally impossible (see §2 for more details). Instead, we focus on the following two tasks: first, detecting that the signal M exists, i.e., telling with high probability whether X was indeed generated by the planted model as opposed to a *null model* where $M = 0$ and X consists only of noise; and second, reconstructing M to some accuracy better than chance. We define these tasks formally as follows.

Definition 1 (Detection)**.** *Let $\mathbb{P}(X)$ be the distribution of X in the planted model (1), and denote by $\mathbb{Q}(X)$ the distribution of X in the null model where $X = W$. A test statistic $\mathcal{T}(X)$ with a threshold ϵ achieves detection if $\lim_{n\to\infty}[\mathbb{P}(\mathcal{T}(X) < \epsilon) + \mathbb{Q}(\mathcal{T}(X) \geq \epsilon)] = 0$, so that the criterion $\mathcal{T}(X) \geq \epsilon$ determines with high probability whether X is drawn from $\mathbb{P}$ or $\mathbb{Q}$.*

Definition 2 (Reconstruction)**.** *An estimator $\widehat{M} = \widehat{M}(X)$ achieves reconstruction if $\mathbb{E}_X\|\widehat{M}\|_F^2 = O(n)$ and there exists a constant $\epsilon > 0$ such that $\lim_{n\to\infty}(1/n)\mathbb{E}_{M,X}\langle M, \widehat{M}\rangle \geq \epsilon$, where $\langle A, B\rangle = \operatorname{Tr} A^\dagger B$ denotes the matrix inner product and $\|A\|_F^2 = \langle A, A\rangle$.*

For many natural problems in this class, it is believed that there is a phase transition, i.e., a threshold value of

snr below which both tasks are information-theoretically impossible: no test statistic can distinguish the null and planted models, and no estimator can beat the trivial one $\widehat{M} = 0$. This threshold is known as the *information-theoretic* threshold and it also depends on the structure of the problem, i.e., on the priors of U and V; if this prior is more strongly structured, we expect the threshold to be lower.

We focus on three cases of (1) which arise in many applications. In *Sparse PCA*, $k = 1$ and $U = V = v$ for some vector v. We further assume that v is sparse, with a constant fraction of nonzero entries. This corresponds to the *sparse, spiked* Wigner model of [16], [29]. In *Submatrix Localization* (also known as submatrix detection and noisy clustering), $U = V$ and M contains $k \geq 2$ distinct blocks of elevated mean. This model arises in the analysis of social networks and gene expression, see e.g. [22], [11], [19]; it can also be thought of as a Gaussian version of the stochastic block model [15], [14]. Finally, in *Gaussian Mixture Clustering*, there are $k \geq 2$ clusters, and each row of M is the center of the cluster to which the corresponding data point belongs. This model has been widely studied, see, e.g., [36], [35], [21], [2].

For each of these three problems, our goal is to compute the information-theoretic threshold, and understand how it scales with the parameters of the problem: for instance, the sparsity of the underlying signal or the number of clusters. In particular, a simple upper bound on the information-theoretic threshold for each of these problems is the point at which spectral algorithms succeed, i.e., the point at which the likely spectrum of X becomes distinguishable from the spectrum of the random matrix W. The spectral thresholds for our problems are well known from the theory of Gaussian matrices with low-rank perturbations. However, based on compelling but non-rigorous arguments from statistical physics (e.g. [28], [27]), it has been conjectured that when the signal is sufficiently sparse, or its rank (the number of clusters or blocks) is sufficiently large, the information-theoretic threshold falls strictly below the spectral one.

In this paper, we prove upper and lower bounds on information-theoretic thresholds of all three problems, determining the threshold within a multiplicative constant in interesting regimes. For sparse PCA, we determine the precise threshold in the limit where the signal matrix is very sparse; similarly, for the submatrix localization problem, we determine the threshold when the number of blocks is large. For the clustering problem, our bounds differ by a factor of $\sqrt{2}$ in the limit where the number of clusters is large. Moreover, our results verify the conjecture that the information-theoretic threshold dips below the spectral one when the signal is sufficiently sparse, or when the number of clusters or blocks is sufficiently large. This corresponds to recent results [1], [5], [13], [18] showing that, in the stochastic block model, the information-theoretic detectability threshold falls below the Kesten-Stigum bound above which efficient spectral and message-passing algorithms succeed [15], [14], [30], [24], [10]. We consider this evidence for the conjecture that these problems posses a 'hard but detectable' regime where detection and reconstruction are information-theoretically possible but take at least exponential time.

Although our computations are specific to these models, our proof techniques are quite general and may be applied with mild adjustment to a broad range of similar problems. In particular, our upper bounds are derived by analyzing the generalized likelihood ratio $\max_M \mathbb{P}(X|M)/\mathbb{Q}(X)$ based on simple first moment arguments. The lower bounds are proved by showing the second moment of likelihood ratio $\mathbb{E}_{X\sim Q}[(\mathbb{P}(X)/\mathbb{Q}(X))^2] \leq C$ for a universal constant C, which further implies non-detectability in general and non-reconstruction in additive, Gaussian noise settings.

Since the initial posting of this paper as an arXiv preprint, a number of interesting papers [34], [33], [7], [26] have appeared, some extending or improving our results. Sharp lower bounds for sparse PCA were also obtained recently in [33] using a conditional second moment method similar to ours. Complete, but not explicit, characterizations of information-theoretic reconstruction thresholds were obtained in [25], [26] for sparse PCA and submatrix localization through the Guerra interpolation technique and cavity method. However, their characterization of reconstruction thresholds does not directly apply to detection.

Next we present our main results without proofs; the excluded details can be found in the full paper [6].

II. Sparse PCA

Consider the following *spiked Wigner* model, where the underlying signal is a rank-one matrix:

$$X = \frac{\lambda}{\sqrt{n}} vv^\dagger + W\,, \tag{2}$$

Here, $v \in \mathbb{R}^n$, $\lambda > 0$ and $W \in \mathbb{R}^{n\times n}$ is a Wigner random matrix with $W_{ii} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,2)$ and $W_{ij} = W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i < j$. We assume v is drawn from the sparse Rademacher prior, although many alternatives may be imposed. Specifically, for some $\gamma \in [0,1]$ the

support of v is drawn uniformly from all $\binom{n}{\gamma n}$ subsets $S \subset [n]$ with $|S| = \gamma n$ (when n is finite, we assume that γn is an integer). Once the support is chosen, each nonzero component v_i is drawn independently and uniformly from $\{\pm\gamma^{-1/2}\}$, so that $\|v\|_2^2 = n$. When γ is small, the data matrix X is a sparse, rank-one matrix observed through Gaussian noise.

One natural approach for this problem is PCA: that is, diagonalize X and use its leading eigenvector $\widehat{v}$ as an estimate of v. The threshold at which this algorithm succeeds can be computed using the theory of random matrices with rank-one perturbations [4], [32], [8]:

(1) When $\lambda > 1$, the leading eigenvalue of $X/\sqrt{n}$ converges almost surely to $\lambda + \lambda^{-1}$, and $\langle v, \widehat{v}\rangle^2$ converges almost surely to $1 - \lambda^{-2}$; thus PCA succeeds in reconstructing better than chance;
(2) When $\lambda \leq 1$, the leading eigenvalue of $X/\sqrt{n}$ converges almost surely to 2, and $\langle v, \widehat{v}\rangle^2$ converges almost surely to 0; thus PCA fails to reconstruct better than chance.

Because the leading eigenvalue of W is 2 w.h.p., detection is only possible when $\lambda > 1$. Intuitively, PCA only exploits the low-rank structure of the underlying signal, and not the sparsity of v; it is natural to ask whether one can succeed in detection or reconstruction for some $\lambda < 1$ by taking advantage of this additional structure. Through analysis of an approximate message-passing algorithm and the free energy, the following conjecture was made in statistical physics [29], [25]:

Conjecture 1. *Let the computational threshold be the minimum of λ so that reconstruction or detection can be attained in polynomial-time in n for a given γ. There exists $\gamma^* \in (0,1)$ such that*

(1) If $\gamma \geq \gamma^*$, then both the information-theoretic and computational thresholds are given by $\lambda = 1$.
(2) If $\gamma < \gamma^*$, then the computational threshold is given by $\lambda = 1$, but the information-theoretic threshold for λ is strictly smaller.

We derive the following upper and lower bounds on the information-theoretic threshold in terms of λ and γ, and confirm that the threshold is $\lambda = 1$ when γ is relatively large and falls strictly below $\lambda = 1$ when γ is sufficiently small. Throughout, we use

$$h(\gamma) = -\gamma \log \gamma - (1-\gamma)\log(1-\gamma)$$

to denote the entropy function, and $\mathcal{W}(y)$ for the root x of $x\mathrm{e}^x = y$. All our logarithms are natural.

Theorem 1. *Let*

$$\lambda^{\text{upper}} = 2\sqrt{h(\gamma) + \gamma \log 2} \tag{3}$$

$$\lambda^{\text{lower}} = \begin{cases} 1 & \gamma \geq 0.6 \\ \sqrt{2\gamma\, \mathcal{W}\left(\frac{1}{2\gamma\sqrt{\mathrm{e}}}\right)} & \mathrm{e}^{-41}/81 \leq \gamma < 0.6 \\ \sqrt{g(\gamma)} & \gamma < \mathrm{e}^{-41}/81\,, \end{cases} \tag{4}$$

where

$$g(\gamma) = 4\gamma\left(-\log\gamma - 2.1\sqrt{-2\log\gamma} - \frac{3}{2}\log\frac{3\mathrm{e}}{1-\gamma}\right).$$

Then detection and reconstruction are information-theoretically possible when $\lambda > \lambda^{\text{upper}}$ and are impossible when $\lambda < \lambda^{\text{lower}}$.

In our proof, we give tighter lower bounds, but these are analytically convenient.

Note that λ^{upper} falls below the spectral threshold $\lambda = 1$ whenever $\gamma \leq 0.054$, and λ^{lower} matches the spectral threshold whenever $\gamma \geq 0.6$. Hence, Theorem 1 proves Conjecture 1 on information-theoretic threshold, albeit without pinning down γ^* exactly. In addition, in the limit $\gamma \to 0$, both λ^{upper} and λ^{lower} give an information-theoretic threshold of

$$\lambda_c = 2\,(1 + o_\gamma(1))\,\sqrt{-\gamma\log\gamma}\,, \tag{5}$$

determining the threshold fully in the limit where the low-rank matrix is very sparse. Independent of the present work, and building on our preprint [6], Perry et al. [33] obtained the same tight threshold in this limit with a smaller error term. Previous work [12] had determined that threshold scales as $\lambda = \Theta(\sqrt{-\gamma\log\gamma})$ up to a constant factor.

In passing, we note that there is a very interesting line of work on *exact or approximate support reconstruction* for sparse PCA, i.e., estimating correctly or consistently the positions of non-zeros in v, in a regime where the size of the support is sublinear in n (see e.g., [20], [3], [9], [23], [17] and references therein). In contrast, we focus on the regime where the size of the support is linear in n, i.e., $\gamma = \Theta(1)$, and $\lambda = \Theta(1)$. In this regime it is impossible to correctly or consistently estimate the support of v, and hence we instead focus on detection and reconstruction better than chance.

III. Submatrix Localization

In the submatrix localization problem, our task is to detect within a large Gaussian matrix a small block or blocks with atypical mean. Let $\sigma : [n] \to [k]$ be a *balanced* partition, i.e. one for which $|\sigma^{-1}(t)| = n/k$ for all $t \in [k]$, chosen uniformly from all such partitions. This

terminology will recur throughout the paper. Construct a $n \times n$ matrix Y such that $Y_{i,j} = \mathbf{1}_{\sigma(i)=\sigma(j)}$. In the planted model,

$$X = \frac{\mu}{\sqrt{n}} \left(Y - \frac{1}{k} \mathbb{J} \right) + W \,, \tag{6}$$

where W is again a Wigner matrix and $\mathbb{J}$ is the all-ones matrix. In the null model, $X = W$. The subtraction of $\mathbb{J}/k$ centers the signal matrix so that $\mathbb{E}X = 0$ in both the null and planted models. In the planted model, $(\mu/\sqrt{n})\,(Y - \mathbb{J}/k)$ is a rank-$(k-1)$ matrix with the largest $(k-1)$ eigenvalues all equal to $\mu\sqrt{n}/k$, making X a Wigner matrix with a rank-$(k-1)$ additive perturbation. Matrices of this type exhibit the following spectral phase transition:

(1) When $\mu > k$, the k leading eigenvalues of $X/\sqrt{n}$ converge to $\mu/k + k/\mu$ almost surely;
(2) When $\mu \leq k$, the k leading eigenvalues of $X/\sqrt{n}$ converge to 2 almost surely.

Hence, it is possible to detect the presence of the additive perturbation from the spectrum of X alone when $\mu > k$. We prove the following upper and lower bounds on the information-theoretic threshold:

Theorem 2. *Let*

$$\mu^{\text{upper}} = 2k\sqrt{\frac{\log k}{k-1}} \tag{7}$$

$$\mu^{\text{lower}} = \begin{cases} 2 & k = 2 \\ k\sqrt{\frac{2\log(k-1)}{k-1}} & 3 \leq k \leq e^{22^4} \\ 2\sqrt{k\log k - 11k\log^{3/4}(k)} & k > e^{22^4} \,. \end{cases} \tag{8}$$

Then detection and reconstruction are information-theoretically possible when $\mu > \mu^{\text{upper}}$ and impossible when $\mu < \mu^{\text{lower}}$.

Note that μ^{upper} dips below the spectral threshold $\mu = k$ when $k \geq 11$, indicating a regime where standard spectral methods fail but detection is information-theoretically possible. Also, Theorem 2 proves the conjecture in [28] that as $k \to \infty$, the information-theoretic threshold is given by $\mu = 2\sqrt{k \log k}$.

Previous work in *submatrix detection* and *localization*, also known as *noisy biclustering*, mostly focuses on finding a single submatrix, see, e.g., [22], [11], [13], [19] and the references therein. In our setting, $\mu = \Theta(1)$ and $k = \Theta(1)$, so it is impossible to consistently estimate the support and we instead resort to detection and reconstruction better than the chance.

IV. Gaussian Mixture Clustering

Finally, we study a model of clustering with limited data in high dimension. Let $v_1, ..., v_k$ be independently and identically distributed as $\mathcal{N}\left(0, k/(k-1)\,\mathbb{I}_{n,n}\right)$, and define $\overline{v} = (1/k)\sum_s v_s$ to be their mean. The scaling of the expected norm of each v_s with k ensures that $\mathbb{E}\|v_s - \overline{v}\|_2^2 = n$ for all $1 \leq s \leq k$. For a fixed parameter $\alpha > 0$, we then generate $m = \alpha n$ points $x_i \in \mathbb{R}^n$ which are partitioned into k clusters of equal size by a balanced partition $\sigma : [n] \to [k]$, again chosen uniformly at random from all such partitions. For each data point i, let $\sigma_i \in [k]$ denote its cluster index, and generate x_i independently according to Gaussian distribution with mean $\sqrt{\rho/n}\,(v_{\sigma_i} - \overline{v})$ and identity covariance matrix, where $\rho > 0$ is a fixed parameter characterizing the cluster separation. We can put this in the form of model (1) by constructing an $n \times k$ matrix $V = [v_1, ..., v_k]$, an $m \times k$ matrix S with $S_{i,t} = \mathbf{1}_{\sigma_i = t}$, and setting

$$X = \sqrt{\frac{\rho}{n}} \left(S - \frac{1}{k} \mathbb{J}_{m,k} \right) V^\dagger + W, \tag{9}$$

where $W_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. In the null model, there is no cluster structure and $X = W$. The subtraction of $\mathbb{J}_{m,k}/k$ once again centers the signal matrix so that $\mathbb{E}X = 0$ in both models. The following spectral phase transition follows from the celebrated BBP phase transition [4], [31]:

(1) When $\rho\sqrt{\alpha} > k - 1$, then the largest eigenvalue of $(1/m)X^\dagger X$ converges to $(1 + \frac{\rho}{k-1})(1 + \frac{k-1}{\rho\alpha})$ almost surely;
(2) When $\rho\sqrt{\alpha} \leq k - 1$, then the largest eigenvalue of $(1/m)X^\dagger X$ converges to $(1 + 1/\sqrt{\alpha})^2$ almost surely.

Thus spectral detection is possible if $\rho\sqrt{\alpha} > (k-1)$.

We prove the following upper and lower bounds on the information-theoretic threshold, which differ by a factor of $\sqrt{2}$ when k is large.

Theorem 3. *Let*

$$\rho^{\text{upper}} = 2\sqrt{\frac{k\log k}{\alpha}} + 2\log k \tag{10}$$

$$\rho^{\text{lower}} = \begin{cases} \sqrt{1/\alpha} & k = 2 \\ \sqrt{\frac{2(k-1)\log(k-1)}{\alpha}} & k \geq 3 \,. \end{cases} \tag{11}$$

Then detection and reconstruction are possible when $\rho > \rho^{\text{upper}}$ and impossible when $\rho < \rho^{\text{lower}}$.

We conjecture that in the limit $k \to \infty$, the information-theoretic threshold is $\rho = 2\sqrt{k \log k/\alpha}$, but we do not find a proof. Most previous work in Gaussian

mixture clustering focuses exact or near-exact reconstruction based on PCA, see e.g., [35], [36], [2], [21]. In our setting, since ρ is a fixed constant, the cluster separation is not sufficient for exact reconstruction and we turn to detection and reconstruction better than chance. Somewhat surprisingly, we find that if the number of clusters is large, clustering is informationally possible even below the spectral phase transition threshold, and we conjecture that in this regime it is computationally hard to identify the clusters. We note that a similar "hard-but-detectable" regime has been determined empirically in [35].

REFERENCES

[1] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. arXiv 1512.09080, Dec 2015.

[2] D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Learning Theory*, pages 458–469. Springer, 2005.

[3] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, 37(5B):2877–2921, 10 2009.

[4] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

[5] J. Banks, C. Moore, J. Neeman, and P. Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 383–416, 2016.

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

[7] 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. arXiv:1606.04142, June 2016.

[8] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[9] Q. Berthet, P. Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[10] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS*, pages 1347–1357, 2015.

[11] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 11 2013.

[12] T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, 2015.

[13] Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. In *Proceedings of ICML 2014 (Also arXiv:1402.1267)*, Feb 2014.

[14] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, 2011.

[15] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

[16] Y. Deshpande and A. Montanari. Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory*, pages 2197–2201, June 2014.

[17] Y. Deshpande and A. Montanari. Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*, pages 334–342, 2014.

[18] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. arXiv 1509.07859, September 2015.

[19] B. Hajek, Y. Wu, and J. Xu. Submatrix localization via message passing. arXiv 1510.09219, October 2015.

[20] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, June 2009.

[21] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM Journal on Computing*, 38(3):1141–1156, 2008.

[22] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, 2011.

[23] R. Krauthgamer, B. Nadler, and D. Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, June 2015.

[24] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[25] F. Krzakala, J. Xu, and L. Zdeborová. Mutual information in rank-one matrix estimation. arXiv 1603.08447, March 2016.

[26] M. Lelarge and L. Miolane. Fundamental limits of symmetric low-rank matrix estimation. arXiv:1611.03888, Nov. 2016.

[27] T. Lesieur, C. D. Bacco, J. Banks, F. Krzakala, C. Moore, and L. Zdeborová. Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. *Arxiv preprint arxiv:1610.02918*, 2016.

[28] T. Lesieur, F. Krzakala, and L. Zdeborová. MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *53rd Annual Allerton Conference on Communication, Control, and Computing*, pages 680–687, Sept 2015.

[29] T. Lesieur, F. Krzakala, and L. Zdeborová. Phase transitions in sparse PCA. In *IEEE International Symposium on Information Theory*, pages 1635–1639. IEEE, 2015.

[30] E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of The 27th Conference on Learning Theory*, pages 356–370, 2014.

[31] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.

[32] S. Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.

[33] A. Perry, A. S. Wein, and A. S. Bandeira. Statistical limits of spiked tensor models. arXiv:1612.07728, Dec. 2016.

[34] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Optimality and sub-optimality of PCA for spiked random matrices and synchronization. arXiv:1609.05573, Sept. 2016.

[35] N. Srebro, G. Shakhnarovich, and S. Roweis. An investigation of computational and informational limits in Gaussian mixture clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 865–872, New York, NY, USA, 2006. ACM.

[36] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, June 2004.