2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

PERFORMANCE-COMPLEXITY TRADE-OFF IN LARGE DIMENSIONAL STATISTICS

Tayeb Zarrouk, Romain Couillet, Florent Chatelain, Nicolas Le Bihan

GIPSA-lab, University Grenoble-Alpes, Grenoble, France.

ABSTRACT

This article introduces a random matrix framework for the analysis of the trade-off between performance and complexity in a class of machine learning algorithms, under a large dimensional data $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ regime. Specifically, we analyze the spectral properties of $K \odot B \in \mathbb{R}^{n\times n}$, for a kernel random matrix $K = \{f(\frac{1}{p}\|x_i - x_j\|^2)\}_{i,j=1}^n$ upon which a sparsity mask $B \in \{0,1\}^{n\times n}$ is applied: this reduces the number of K_{ij} to evaluate, thereby reducing complexity, while weakening the power of statistical inference on K, thereby impeding performance. Assuming the data structured as $X = Z + \sqrt{n}\mu v^\mathsf{T}$ for informative vectors $\mu \in \mathbb{R}^p$, $v \in \mathbb{R}^n$, and white noise Z, we exhibit a phase transition phenomenon below which spectral methods must fail and which is a function of the sparsity structure of B. This finds immediate applications to the fundamental limits of complexity-reduced spectral clustering as well as principal component analysis.

Index Terms— Random matrix theory; large dimensional statistics; spectral clustering; PCA.

1. INTRODUCTION

The ongoing exponential increase of the volume of data and the imperious need to devise smart and autonomous algorithms to process them sets a strong pressure on (i) the scalability of machine learning techniques, and (ii) a theoretical understanding of the fundamental limitations of large dimensional data processing.

However, as they rely on non linear feature extraction and deal with complex structured data, most machine learning algorithms fail to be prone to analysis and theoretical guarantees. This is all the more exacerbated in the present large and numerous data era, and sometimes leads to disruptive algorithm behavior as the dimension p of the data increases. Recent breakthroughs in the field of random matrix theory provide renewed hope to address Challenges (i) and (ii). Assuming that the dimension p and the sample size n of data $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ are simultaneously large (formally, $n, p \to \infty$ with $p/n \to c \in (0, \infty)$), random matrix theory has recently shown that large non linear kernel $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ [1, 2, 3] and activation-function [4] matrices, which are ubiquitous in machine learning, exhibit a simple and tractable behavior in the limit; not only does this enable a thorough understanding of the algorithm performance but it also allows for the revision, the improvement, and even the entire reconsideration of machine learning intuitions and methods (e.g., [5, 6, 7]). Possibly more importantly, it has moreover been shown [8, 9] that these large n, p results are *universal* with respect to the data distribution, i.e., the theoretical results developed on Gaussian data models remain provably valid for a large range of realistic generative data models.

Couillet's work is supported by the MIAI "LargeData" chair project, the ANR DARLING project, and the HUAWEI Lardist project.

Anchored in these findings which mostly address Challenge (ii), this article investigates instead Challenge (i). Being at the core of machine learning algorithms, the above kernel matrices $K \in \mathbb{R}^{n\times n}$ are expensive to evaluate (in general of complexity $O(n^2 p)$), possibly to store, but most fundamentally to operate on (invert, extract eigenvectors, etc.). The objective of the article is to evaluate the theoretical consequences of a (possibly drastic) reduction in the number of entries of K being evaluated, in terms of algorithm performances. Specifically, by introducing a random mask B which entry-wise discards a proportion $1 - \varepsilon$ of the entries of K, the article:

- determines the limiting eigen-spectrum of the 'punctured' kernel $K \odot B$, where $K = \frac{1}{p}X^\mathsf{T}X$, $X = Z + \sqrt{n}\mu v^\mathsf{T}$, for deterministic $\mu \in \mathbb{R}^p$, $v \in \mathbb{R}^n$ of fixed norms, Z random with independent $Z_{ij} \sim \mathcal{N}(0,1)$ entries, and $B \in \{0,1\}^{n\times n}$ random with independent Bernoulli entries with parameter ε;
- identifies a phase-transition phenomenon by which: (a) if $\|\mu\|^2$ exceeds a threshold Γ, the largest eigenvalue $\hat{\lambda}$ of $K \odot B$ is isolated and its associated eigenvector $\hat{v}$ has a non trivial alignment to the vector v, (b) if not, the dominant eigenvector $\hat{v}$ of $K \odot B$ is asymptotically orthogonal to v;
- applies these findings to determine the fundamental limits of two novel algorithms: (a) 'punctured' spectral clustering in a two-class setting, (b) 'punctured' principal component analysis in a single-direction factor model; from these results unfolds a fundamental *performance-complexity trade-off* for these elementary kernel methods. This trade-off is shown to largely outperform more conventional subsampling (for (a)) and dimensionality reduction (for (b)) methods.

The remainder of the article is structured as follows. Section 2 introduces the formal model under study along with the key random matrix quantities of interest. Section 3 presents our main results, which are then concretely applied and discussed in Section 4, with a proof sketch provided in Section 5. The article closes in Section 6 on a discussion of the anticipated consequences and future investigations in this new performance-complexity trade-off line of research.

2. MODEL AND PROBLEM SETTING

Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ be a collection of n data samples of dimension p. For reasons that will be clarified in the applications of Section 4, we assume that X is modelled as

$$X = Z + \sqrt{n}\mu v^\mathsf{T}$$

where $Z \in \mathbb{R}^{p\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries, $\mu \in \mathbb{R}^p$ and $v \in \mathbb{R}^n$ are deterministic vectors with $\|\mu\|$ fixed and independent of p and n,[1] $\|v\|^2 = 1$ and $\limsup_n \max_{1\le i\le n}\{\sqrt{n}v_i^2\} = 0$.

[1] This assumption is set for simplicity of exposition. It may be relaxed to $\mu \in \mathbb{R}^p$ random independent of Z and such that $\|\mu\|$ converges almost surely to a finite limit as $p \to \infty$. See Section 4.2 for a concrete application of this generalization.

978-1-7281-6662-9/20/$31.00

We further define the symmetric binary matrix $B \in \{0,1\}^{n\times n}$, which will serve as a 'puncturing' mask applied to the canonical inner-product kernel matrix $\frac{1}{p}X^\mathsf{T}X$, with, for $1 \leq i < j \leq n$,

$$B_{ij} \sim \text{Bern}(\varepsilon)$$

and $B_{ji} = B_{ij}$, and for $1 \leq i \leq n$, $B_{ii} = b \in \{0,1\}$.

The core object under study in the present article is the *punctured kernel matrix*

$$S = \frac{1}{p}X^\mathsf{T}X \odot B.$$

From the definition of B, on average, a proportion $1-\varepsilon$ of the off-diagonal entries of S is set to zero (thus in practice not evaluated), so that $1-\varepsilon$ plays the role of a sparsity enhancer; as for its diagonal entries, they are either all maintained (if $b = 1$) or set to zero (if $b = 0$). We will later see it to be crucial for the diagonal entries B_{ii} to be either all maintained or all discarded.

The objective of the article is to provide a description of the spectral behavior, and most fundamentally (a) of the existence of a dominant isolated eigenvalue $\hat{\lambda}$ in the spectrum of S and (b) of the correlation between the eigenvector $\hat{v}$ associated to $\hat{\lambda}$ and the population vector v, as a function of the limiting ratio c, the signal-to-noise ratio $\|\mu\|^2$, and the sparsity parameter ε.

To this end, we assume both p and n are large and will, for mathematical purpose, assume that $p, n \to \infty$ in such a way that $p/n \to c \in (0,\infty)$. We further recall that $\|\mu\|$ is fixed with respect to p, n, that $\|v\| = 1$, and that $\limsup_n \max_{1\leq i\leq n}\{\sqrt{n}v_i^2\} = 0$.

Elaborating on tools from random matrix theory (in particular tools developed in [10]), the large dimensional spectral behavior of S is accessible via a thorough analysis of its *resolvent*

$$Q(z) \equiv (S - zI_p)^{-1} \tag{1}$$

defined for all $z \in \mathbb{C} \setminus \text{Sp}(S)$ with $\text{Sp}(S)$ the set of eigenvalues of S. The resolvent $Q(z)$ is as such the central object of our main technical results, delineated next.

3. MAIN RESULTS

.

3.1. Deterministic equivalent and limiting spectrum

Our main technical result provides a *deterministic equivalent* $\bar{Q}(z)$ for the resolvent $Q(z)$ defined in (1), that is, $\bar{Q}(z)$ is deterministic and such that, for all sequences of deterministic matrix $A \in \mathbb{R}^{n\times n}$ and vectors $a, b \in \mathbb{R}^n$ of bounded norms (with respect to n), with probability one,

$$\frac{1}{n}\operatorname{tr} A(Q(z) - \bar{Q}(z)) \to 0, \quad a^\mathsf{T}\left(Q(z) - \bar{Q}(z)\right)b \to 0.$$

This will be denoted $Q(z) \leftrightarrow \bar{Q}(z)$ and allows, as shown subsequently, for the transfer of most of the spectral properties of $Q(z)$ (and thus of S) to $\bar{Q}(z)$.

Theorem 3.1 (Deterministic equivalent of $Q(z)$)**.** *Under the assumptions and notations of Section 2, as $p, n \to \infty$,*

$$\begin{aligned} Q(z) \leftrightarrow \bar{Q}(z) &\equiv m(z)\left(I_n + \frac{\|\mu\|^2\varepsilon m(z)}{c+\varepsilon m(z)}vv^\mathsf{T}\right)^{-1} \\ &= m(z)I_n - \frac{\|\mu\|^2\varepsilon m(z)^2 vv^\mathsf{T}}{c+\varepsilon m(z)(1+\|\mu\|^2)} \end{aligned} \tag{2}$$

where $m(z)$ is the Stieltjes transform (i.e., $m(z) = \int (t-z)^{-1}\nu(dt)$) of the almost sure limiting spectral measure $\nu = \lim_n \frac{1}{n}\sum_{\lambda\in\text{Sp}(S)}\delta_\lambda$ of S, and is the unique complex analytic solution to

$$z = b - \frac{1}{m(z)} - \frac{\varepsilon}{c}m(z) + \frac{\varepsilon^3 m(z)^2}{c(c+\varepsilon m(z))}. \tag{3}$$

Proof. A sketch of proof is given in Section 5. □

Before exploiting Theorem 3.1 to our present objective, a few remarks are in order. We may first observe that $\bar{Q}(z)$ takes the form of a perturbation of the scaled identity matrix by the scaled rank-1 matrix vv^T. As per the study of *spiked random matrix* models [11, 12], this form predicts the possible existence of an isolated dominant eigenvalue $\hat{\lambda}$ in the spectrum of S with associated eigenvector $\hat{v}$ aligned to some extent to v. This is established in section 3.2. Before this, interesting conclusions from Theorem 3.1 can be drawn in the limit where $\varepsilon \to 0$ or 1.

Remark 1 (Marčenko-Pastur and semi-circle limits). When $\varepsilon = 1$, letting $z' = z + 1 - b$, Equation (3) can be rewritten

$$z' m_b(z')^2 + (cz' + 1 - c)m_b(z') + c = 0$$

where $m_b(z') \equiv m(z' + b - 1)$ is the Stieltjes transform of the measure $\nu(\cdot + b - 1)$. We thus recover the defining equation of the Stieltjes transform of the Marčenko-Pastur distribution [13] for the variable $z + 1 - b$. In particular, the limiting measure ν has support $[(1-\sqrt{1/c})^2 + b - 1, (1+\sqrt{1/c})^2 + b - 1]$.

If instead, $\varepsilon \ll 1$, it can be shown that Equation (3) becomes

$$z - b + \frac{1}{m(z)} + \frac{\varepsilon}{c}m(z) = O_\varepsilon(\varepsilon^2) \tag{4}$$

which, letting $z' = \sqrt{c/\varepsilon}(z-b)$, leads to

$$m_{b,\varepsilon}(z')^2 + z' m_{b,\varepsilon}(z') + 1 = O_\varepsilon(\varepsilon^{\frac{3}{2}}) \tag{5}$$

where $m_{b,\varepsilon}(z') = \sqrt{\varepsilon/c}\, m(\sqrt{\varepsilon/c}z' + b)$ for $m_{b,\varepsilon}$ the Stieltjes transform of the shifted-scaled measure $\nu((\cdot - b)\sqrt{c/\varepsilon})$. We thus retrieve the defining Stieltjes transform of Wigner's semi-circle law [14]. In particular, in the first order in ε, the limiting support of ν is $[-2\sqrt{\varepsilon/c} + b, 2\sqrt{\varepsilon/c} + b]$.

Remark 1 thus predicts that the limiting measure ν of the spectrum of the punctured kernel $S = K \odot B$ evolves from the Marčenko-Pastur law, typical of the Gram matrix $X^\mathsf{T}X$ towards a semi-circle measure, typical of the symmetric matrix B with independent entries (up to symmetry). Interestingly, note that the ε-negligible term in (5) is of order $O_\varepsilon(\varepsilon^{\frac{3}{2}})$, suggesting a fast convergence towards the semi-circle behavior, as soon as ε is away from one. Figure 1 visually confirms this observation by displaying the limiting behaviors for $\varepsilon \in \{0.1, 0.5, 0.9\}$. Figure 1 also predicts the possibility, for $c < 1$, of a transitional state where the support of ν is divided in two connected components.[2]

From Remark 1, we may also already anticipate some of our subsequent findings on the punctured kernel performance in the case of stringent puncturing. Indeed, in the regime of small ε, for (4) to have a non-trivial behavior (recall that $z - cb = O(\varepsilon^{-\frac{1}{2}})$, one must

[2] Writing Equation (3) under the functional form $z(m) = b - \frac{1}{m} - \frac{\varepsilon}{c}m + \frac{\varepsilon^3 m^2}{c(c+\varepsilon m)}$, the edges of the connected components of ν are the *real* solutions to $z'(m) = 0$ which can be expressed as a polynomial of order 4: the support may then contain up to two components (i.e., 4 edges).

have $m(z) = O(\varepsilon^{-\frac{1}{2}})$. Plugged into (2), it thus comes that the ratio between the 'noise' contribution ($m(z)I_n$) and the 'information' contribution ($\frac{\|\mu\|^2 \varepsilon m(z)^2 vv^\mathsf{T}}{c+\varepsilon m(z)(1+\|\mu\|^2)}$) of the deterministic equivalent $\bar{Q}(z)$ of $Q(z)$ is only non-trivial if $\|\mu\|^2 = O(\varepsilon^{-\frac{1}{2}})$. As a consequence (see more in Section 3.2), as the puncturing becomes more severe (i.e., as $\varepsilon \to 0$), the 'energy' $\|v\|^2\|\mu\|^2 = \|\mu\|^2$ of the information matrix μv^T must increase at least as $1/\sqrt{\varepsilon}$ for it to remain visible and retrievable.

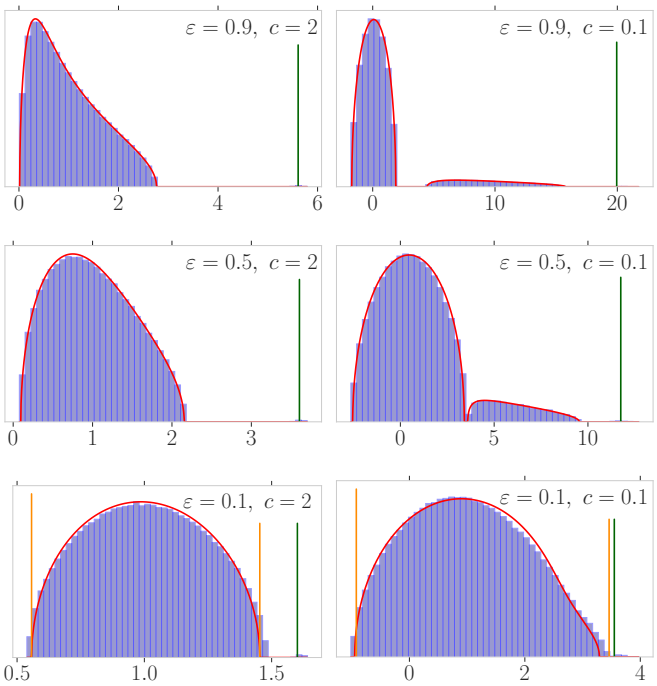

Fig. 1: Empirical eigenvalue distributions of $\frac{1}{p}X^\mathsf{T}X \odot B$, for $b = 1$, $n = 1000$ and $v = [1_{\frac{n}{2}}^\mathsf{T}, -1_{\frac{n}{2}}^\mathsf{T}]^\mathsf{T}$. (**Top** row) $\varepsilon = 0.9$; (**middle** row) $\varepsilon = 0.5$; (**bottom** row) $\varepsilon = 0.1$. (**Left** column) $p = 2000$ ($c = 2$) and $\|\mu\|^2 = 3$; (**right** column) $p = 100$ ($c = 0.1$) and $\|\mu\|^2 = 1$. (**Red line**) theoretical limiting spectral density ν computed by numerically inverting the Stieltjes transform $m(z)$ in Theorem 3.1.[3] (**Green line**) theoretical position of isolated spike as per Theorem 3.2. (**Orange lines**) small ε approximation of support edges as per Corollary 3.2.1.

Remark 2 (On the elements B_{ii}). The assumption that the elements $B_{ii} = b$ be fixed and equal, while the B_{ij}, $i \neq j$, are taken random, may surprise the reader. Letting B_{ii} random would however break the statistical 'exchangeability' of the entries Q_{ij} of the resolvent: this would specifically lead to $\bar{Q}(z)$ being of the form $D_1(z) + D_2(z)vv'D_2(z)$ for random diagonal matrices $D_1(z)$ and $D_2(z)$, each having two distinct elements (i.e., the $[D_j(z)]_{ii}$'s are the same for all B_{ii} constant). As an immediate consequence, the dominant eigenvector $\hat{v}$ of S, which we expect to faithfully recover v, would instead be a severely ('B_{ii}-wise') deformed version of v.

[3] We recall that, if $m(z) = \int(\lambda - z)d\nu(\lambda)$, then $d\nu(x) = \lim_{y\downarrow 0} \frac{1}{\pi}\mathrm{Im}[m(x + \imath y)]dx$.

As for the choice of $b \in \{0, 1\}$, note that $\bar{Q}(z)$ only depends on b through (3). As such, with the change of variable $z' = z - b$ and the fact that $m(z) = \int(\lambda - z)^{-1}d\nu(\lambda) = \int(\lambda - z')^{-1}d\nu(\lambda - b) \equiv \mathbf{m}(z')$, $\bar{Q}(z')$ when expressed as a function of z' and $\mathbf{m}(z')$ does not depend on b. This indicates that the only effect of b is to 'shift' the whole spectrum of the limiting spectral measure ν of S by a constant b . Consequently, b may be chosen arbitrarily.

Remark 3 (On the assumptions on μ, v). The asymmetry between the assumptions on $\mu \in \mathbb{R}^p$, which is taken arbitrary, and on $v \in \mathbb{R}^n$, instead constrained by $\max_i \sqrt{n}v_i^2 = o(1)$, may seem surprising. Recalling that $\frac{1}{p}(\sqrt{n}\mu v^\mathsf{T})^\mathsf{T}(\sqrt{n}\mu v^\mathsf{T}) \odot B = \frac{n}{p}\|\mu\|^2 vv^\mathsf{T} \odot B$, it appears that only v is directly affected by the puncturing imposed by B to the information matrix. If v is 'localized' (i.e., only has non-zero entries in fixed positions as with, say, $v = \frac{1}{\sqrt{2}}(1, 1, 0, \ldots, 0)^\mathsf{T}$), each random realization of B may either maintain ($B_{11} = B_{12} = B_{22} = 1$) or discard ($B_{11} = B_{12} = B_{22} = 0$) a non-trivial part of the energy of vv^T, thereby leading to a non converging (and possibly severely deleterious) behavior of the kernel statistics in the large n, p limit. This is avoided by letting v be quite 'delocalized', i.e., has no largely dominant entry.

Technically, in the course of the proof of Theorem 3.1, the matrix $\frac{n}{p}\|\mu\|^2 vv^\mathsf{T} \odot \dot{B}$, where $\dot{B} = B - \mathbb{E}[B]$, should be appended to the expression (2) of $\bar{Q}(z)^{-1}$. Noticing that $vv^\mathsf{T} \odot \dot{B} = D_v \dot{B} D_v$, where $D_v = \mathrm{diag}(v_1, \ldots, v_n)$, it then unfolds that

$$\left\| vv^\mathsf{T} \odot \dot{B} \right\| \leq \sqrt{\varepsilon(1-\varepsilon)} \max_{1 \leq i \leq n} \{\sqrt{n}|v_i|^2\} \left\| \frac{\dot{B}}{\sqrt{n\varepsilon(1-\varepsilon)}} \right\|$$

where the matrix in the right-hand side norm has i.i.d. entries of zero mean and variance $1/n$: from [14], its almost sure limiting spectrum is the semi-circle law with norm almost surely converging to 2. As such, one needs $\max_i \sqrt{n}|v_i|^2 = o(1)$ to ensure that $\|\frac{n}{p}\|\mu\|^2 vv^\mathsf{T} \odot \dot{B}\|$ vanishes. This is achieved if v is sufficiently delocalized, for instance by imposing, as we presently do, that $\max_i v_i^2 = o(n^{-\frac{1}{2}})$.[4]

Having established Theorem 3.1 and discussed some preliminary intuitions on the large dimensional behavior of the punctured kernel S, we are now in position to evaluate the exact conditions under which the signal μv^T can be recovered from S and, under these conditions, the quality of the estimation of the information vector v.

3.2. Phase transition, isolated eigenvalue and eigenvector

This section establishes (i) the condition on $\|\mu\|$ under which the largest eigenvalue $\hat{\lambda}$ of S isolates (and thus becomes informative) and, in this case, (ii) the limit ζ of the alignement $\hat{\zeta} \equiv |\hat{v}^\mathsf{T}v|^2$ between the eigenvector $\hat{v}$ associated to $\hat{\lambda}$ and the (non-trivial) population eigenvector v of the information matrix $(\mu v^\mathsf{T})^\mathsf{T}(\mu v^\mathsf{T})$.

To this end, we exploit Theorem 3.1 by noticing that, as per Cauchy's integral,

$$|\hat{v}^\mathsf{T}v|^2 = \frac{-1}{2\pi\imath}\oint_{\mathcal{C}_{\hat{\lambda}}} v^\mathsf{T}Q(z)v dz \simeq \frac{-1}{2\pi\imath}\oint_{\mathcal{C}_{\lambda}} v^\mathsf{T}\bar{Q}(z)v dz \quad (6)$$

for $\mathcal{C}_x$ a sufficiently small positively-oriented complex contour surrounding x, λ the presumably existing limit for $\hat{\lambda}$ as $p, n \to \infty$; the

[4] Note in passing that we could have let $n^{\frac{1}{4}}v$ be random, independent of B and Z, with i.i.d. entries with bounded $4 + \delta$ (for some $\delta > 0$) moment to obtain, by Markov's inequality, the same result, with high probability.

approximation only holds true if $\hat{\lambda}$ indeed remains isolated from all other eigenvalues of S. The *deterministic* expression on the right-hand side can be evaluated explicitly, leading to the following result.

Theorem 3.2 (Isolated spectrum)**.** *Define the functions*

$$F(x) = x^4 + 2x^3 + \left(1 - \frac{c}{\varepsilon}\right) x^2 - 2cx - c$$

$$G(x) = b + \frac{\varepsilon}{c}(1+x) + \frac{1}{1+x} + \frac{\varepsilon}{x(1+x)}$$

and let Γ be the largest real solution to $F(\Gamma) = 0$. Then, under the assumptions of Section 2, as $p, n \to \infty$, with probability one, the largest eigenvalue $\hat{\lambda}$ of S and its associated eigenvector $\hat{v}$ satisfy

$$\hat{\lambda} \to \lambda = \begin{cases} G(\|\mu\|^2) & , \|\mu\|^2 > \Gamma \\ G(\Gamma) & , \|\mu\|^2 \leq \Gamma \end{cases}$$

$$\hat{\zeta} \equiv |\hat{v}^\mathsf{T} v|^2 \to \zeta = \begin{cases} \frac{F(\|\mu\|^2)}{\|\mu\|^2(1+\|\mu\|^2)^3} & , \|\mu\|^2 > \Gamma \\ 0 & , \|\mu\|^2 \leq \Gamma. \end{cases}$$

Theorem 3.2 ensures the presence of an isolated dominant eigenvalue $\hat{\lambda}$ of S and a non-trivial alignment $\hat{\zeta}$ between the corresponding eigenvector $\hat{v}$ and the information vector v of the model, if and only if $\|\mu\|^2 > \Gamma$. If instead $\|\mu\|^2 \leq \Gamma$, then $\hat{\lambda}$ converges to the right-edge E_ν^+ of the support $[E_\nu^-, E_\nu^+]$ of ν, so that

$$E_\nu^+ = b + \frac{\varepsilon}{c}(1+\Gamma) + \frac{1}{1+\Gamma} + \frac{\varepsilon}{\Gamma(1+\Gamma)}.$$

Interestingly, the values of the limits λ and ζ assume explicit formulations, while the threshold Γ remains implicit (at least it takes the non-convenient form of a root of a fourth-order polynomial). In the regime $\varepsilon \ll 1$ though, the value for Γ becomes tractable.

Corollary 3.2.1 (Small ε approximation)**.** *Under the notations of Theorem 3.2, in the limit of small ε,*

$$\Gamma = \sqrt{\frac{c}{\varepsilon}} - 1 + \varepsilon + O(\varepsilon^{\frac{3}{2}}), \quad E_\nu^\pm = b \pm 2\sqrt{\frac{\varepsilon}{c}} + \frac{\varepsilon^2}{c} + O(\varepsilon^{\frac{5}{2}}).$$

The accuracy of these estimates is illustrated in Figure 1.

3.3. Performance-complexity trade-off

By discarding a proportion ε of the entries of K, the computational cost of the matrix $K \odot B$ is reduced by a factor ε when compared to the computational cost of K.

Besides, to estimate v from the leading eigenvector $\hat{v}$ of $K \odot B$, one may naturally resort to the *power method* which runs the procedure $v^{t+1} = \tilde{v}^{t+1}/\|\tilde{v}^{t+1}\|$ with $\tilde{v}^{t+1} = (K \odot B)v^t$ for all $t \geq 0$ for some arbitrary v^0 until convergence. For all large n, the product operation $(K \odot B)v^t$ has a computational cost of order $O(n^2\varepsilon)$, thereby again reduced by a factor ε.

4. APPLICATIONS

The punctured kernel analysis performed in Section 2 finds several immediate applications, which we illustrate in this section. A straightforward application is that of a 'punctured kernel spectral clustering' of large Gaussian data $x_i \sim \mathcal{N}(\pm\mu, I_p)$ in which $\sqrt{n}v_i \in \{\pm 1\}$ accounts for the sign of $\mathbb{E}[x_i]$. Less immediate, as it requires to flip the meaning of p (now the number of samples) and n (now the sample dimension) and to take (with a slight abuse) $\mu \sim \mathcal{N}(0, \frac{1}{n}\sigma_\mu^2 I_p)$, is the application to 'punctured principal component analysis' based on the p i.i.d. n-dimensional random vectors $X^\mathsf{T} \equiv \tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_p]$ (i.e., $\tilde{x}_i$ is the i-th *row* and not column of X), where $\mathbb{E}[\tilde{x}_i] = 0$ and $\mathbb{E}[\tilde{x}_i \tilde{x}_i^\mathsf{T}] = I_n + \sigma_\mu^2 vv^\mathsf{T}$.

The asymptotic performances of the punctured kernel $\frac{1}{p}X^\mathsf{T}X \odot B$ and punctured sample covariance $\frac{1}{p}\tilde{X}\tilde{X}^\mathsf{T} \odot B$ are described below, along with a comparison to the standard ε-data sampling and ε-dimensionality reduction techniques, respectively.

4.1. Kernel spectral clustering

In this application, we assume that $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ models a dataset generated by a two-class Gaussian mixture model, with $x_i \sim \mathcal{N}(v_i\mu, I_p)$. We may then write $X = Z + \sqrt{n}\mu v^\mathsf{T}$ where $\sqrt{n}v \in \{-1, 1\}^n$. We further impose that $\sum_i v_i = 0$, i.e., each class has the same size.

From the popular spectral clustering algorithm [15] with kernel $K = \{\frac{1}{p}x_i^\mathsf{T}x_j\}_{i,j=1}^n = \frac{1}{p}X^\mathsf{T}X$, the dominant eigenvector $\hat{\mathbf{v}}$ of $X^\mathsf{T}X$ is expected to be aligned to v. Besides, by the complete symmetry of the class model, the natural estimate $\hat{\mathcal{C}}_i$ of the class $\mathcal{C}_i$ of x_i is directly given by $\mathrm{sgn}(\hat{\mathbf{v}}_i)$. Working instead on the punctured kernel $\frac{1}{p}X^\mathsf{T}X \odot B$, and thus with the dominant eigenvector $\hat{v}$ rather than $\hat{\mathbf{v}}$ (both being equal when $\varepsilon = 1$ and $b = 1$), the corresponding performance of punctured spectral clustering is as follows.

Theorem 4.1 (Performance of punctured kernel spectral clustering)**.** *Let $\hat{\mathcal{C}}_i = \mathrm{sgn}(\hat{v}_i)$ be the estimated class $\mathcal{C}_i$ of vector x_i, with the eigenvector convention $v_1\hat{v}_1 > 0$. Then, with probability one,*

$$\frac{1}{n}\sum_{i=1}^n \delta_{\{\mathcal{C}_i = \hat{\mathcal{C}}_i\}} = Q\left(\sqrt{\zeta/(1-\zeta)}\right) + o(1)$$

where ζ is defined in Theorem 3.2 and $Q(x) = \frac{1}{2\pi}\int_x^\infty e^{-t^2/2}dt$.

Figure 2 illustrates Theorem 4.1 by comparing theoretical versus simulated performance for varying ε and $\|\mu\|^2$. It notably confirms the sudden drop of classification accuracy below the phase transition threshold and shows that the predicted asymptotics are already quite accurate in this moderately large $n = 200$, $p = 800$ setting.

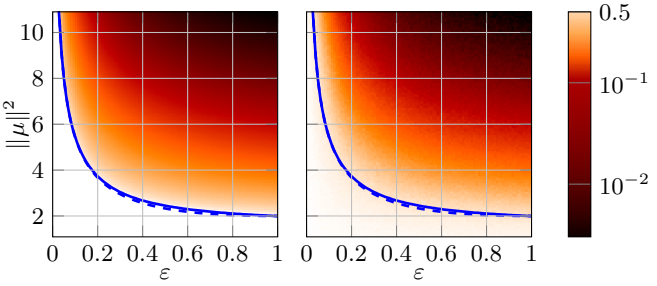

Fig. 2: Classification performance of spectral clustering as a function of ε (x-axis) and $\|\mu\|^2$ (y-axis) for $c = 4$, $n_1 = n_2 = n/2$ and $n = 200$. (**Left**) asymptotic result of Theorem 4.1; (**right**) simulations averaged over 100 Monte-Carlo runs. (**Blue**) theoretical phase transition $\|\mu\|^2 = \Gamma$ evaluated from its expression in Theorem 3.2. (**Dashed blue**) approximation $\varepsilon \ll 1$ as per Corollary 3.2.1.

An alternative approach to reduce the computational complexity of spectral clustering consists in subsampling $n_s < n$ vectors

$X_s \in \mathbb{R}^{p\times n_s}$ of the whole dataset X and performing spectral clustering on the resulting matrix $\frac{1}{p}X_s^\mathsf{T}X_s$ (without puncturing). By taking $n_s = \lceil \varepsilon n \rceil$, the complexity, using a power method, is reduced by a factor $O(\varepsilon^2)$. To reduce the complexity of an n-dimensional spectral clustering, one may then let $n_s = n/m$ ($\varepsilon = 1/m$) and perform m parallel spectral clustering, each of $1/m^2$ reduced complexity: ultimately, similar to the proposed punctured kernel method, this reduces the overall cost by a factor $\varepsilon = 1/m$.

However, the latter procedure looses the benefit of the 'redundancy' inherent to data arising from the same class, which kernel methods leverage [3]. This is quite detrimental to its performance. Indeed, the asymptotic accuracy $|\hat{v}_s^\mathsf{T} v_s|$, with $v_s \in \mathbb{R}^{n_s}$ the normalized subset of v on the n_s selected indices and $\hat{v}_s \in \mathbb{R}^{n_s}$ the dominant eigenvector of $\frac{1}{p}X_s^\mathsf{T}X_s$, follows from Theorem 3.2 by successively letting, in the theorem statement: 1) $\varepsilon = 1$ and 2) $c \to c/\varepsilon$ where $\varepsilon = n_s/n$ becomes the subsampling rate. Using the subscript s in the following to denote the subsampling case, this yields

$$F_s(x) = x^4 + 2x^3 + \left(1 - \frac{c}{\varepsilon}\right)x^2 - \frac{2cx}{\varepsilon} - \frac{c}{\varepsilon} \tag{7}$$
$$= (x+1)^2\left(x + \sqrt{\frac{c}{\varepsilon}}\right)\left(x - \sqrt{\frac{c}{\varepsilon}}\right)$$

the largest root $\Gamma_s = \sqrt{c/\varepsilon}$ of which is the classical Marčenko-Pastur spike phase transition [11]. We thus obtain

$$|\hat{v}_s^\mathsf{T} v_s|^2 \to \zeta_s = \frac{\max\{F_s(\|\mu\|^2), 0\}}{\|\mu\|^2(1+\|\mu\|^2)^3} = \frac{\max\{\|\mu\|^4 - c/\varepsilon 0\}}{\|\mu\|^2(1+\|\mu\|^2)}.$$

Note that, for $x > 0$, $F(x) - F_s(x) = c(2x+1)(\varepsilon^{-1} - 1) \geq 0$ with equality only if $\varepsilon = 1$, so that (i) $\Gamma < \Gamma_s$, i.e., the phase transition of the punctured kernel arises at lower signal-to-noise ratios $\|\mu\|^2$, and (ii) $\zeta > \zeta_s$, i.e., the asymptotic alignment is greater for the punctured kernel method. The two methods only perform equivalently when $\varepsilon = 1$, while the gain of the punctured kernel is increased in the low density ($\varepsilon \ll 1$) regime. The comparison with the phase transition derived for the punctured kernel as per Corollary 3.2.1, i.e., $\Gamma = \sqrt{c/\varepsilon} - 1 + O(\varepsilon)$, shows a gain of order 1 on the signal-to-noise ratio phase transition when ε is small. Figure 3, to be compared to Figure 2, clearly illustrates this result.

But Figure 3 reveals a more fundamental message: here for c small, the punctured kernel phase transition has a peculiar "plateau"-like shape, which strongly suggests that very harsh puncturing may be performed with almost no loss of performance.

4.2. Principal component analysis

We here let $\tilde{X} = [\tilde{x}_1, \ldots, \tilde{x}_{\tilde{n}}] \in \mathbb{R}^{\tilde{p}\times\tilde{n}}$ be a collection of $\tilde{n}$ independent random vectors, each of dimension $\tilde{p}$, such that $\tilde{x}_i \sim \mathcal{N}(0, I_{\tilde{p}} + \sigma_\mu^2 v v^\mathsf{T})$ for a given unit norm vector $v \in \mathbb{R}^{\tilde{p}}$ drawn uniformly at random in the unit sphere $\mathbb{S}^{\tilde{p}-1}$ and for some $\sigma_\mu > 0$. The objective is to perform a 'punctured principal component analysis' on $\tilde{X}$, by retrieving the dominant eigenvector of the punctured sample covariance $\frac{1}{\tilde{n}}\tilde{X}\tilde{X}^\mathsf{T} \odot B$ (with B as in the previous sections), in order to estimate the main direction v of the data covariance.

This model falls under our present analysis if we let $\tilde{p} = n$, $\tilde{n} = p$ (i.e., we exchange the roles of the number and dimension of samples), $\tilde{X} = X^\mathsf{T}$, where $X = Z + \sqrt{n}\mu v^\mathsf{T}$ with Z as in the previous sections and $\mu \in \mathbb{R}^p$ taken random independent of Z as $\mu \sim \mathcal{N}(0, \sigma_\mu^2 I_p/n)$ (note that μ and its norm $\|\mu\|$ are random and thus not fixed: yet, as per Footnote 1 in Section 2, the results of the article remain valid since $\|\mu\|^2 \to c\sigma_\mu^2 = \frac{\sigma_\mu^2}{\tilde{c}}$ almost surely, where

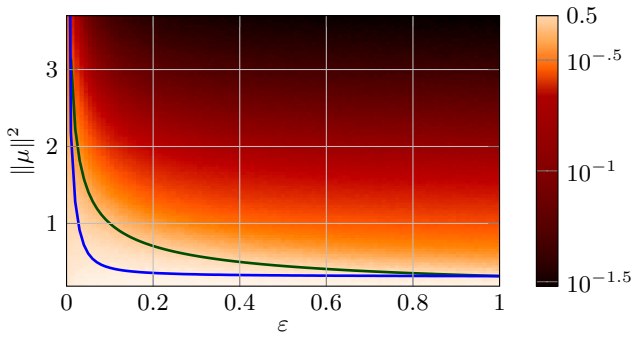

Fig. 3: Classification accuracy of ε-subsampling clustering as a function of ε (x-axis) and $\|\mu\|^2$ (y-axis) for $c = 0.1$, $n_1 = n_2 = n/2$ and $n = 1000$. Simulations averaged over 100 Monte-Carlo runs. (**Green**) phase transition of ε-subsampling ($\|\mu\|^2 = \sqrt{c/\varepsilon}$). (**Blue**) phase transition of punctured kernel (from Theorem 3.2).

$\tilde{c} \equiv c^{-1}$; similarly, v satisfies the condition $\max_i \sqrt{n} v_i^2 = o(1)$ only with high probability). In this case, being the sum of independent Gaussian vectors, $\tilde{x}_i = [Z^\mathsf{T}]_{\cdot,i} + \sqrt{\tilde{p}}\mu_i v$ is indeed (conditionally to v) a Gaussian vector with zero mean and covariance $\mathbb{E}[\tilde{x}_i\tilde{x}_i^\mathsf{T}] = I_{\tilde{p}} + \sigma_\mu^2 v v^\mathsf{T}$, as requested.

As a consequence, the performance of the punctured PCA method, estimated as the squared Euclidean distance between the population eigenvector v and the punctured PCA estimate $\hat{v}$ the dominant eigenvector of $\frac{1}{\tilde{n}}\tilde{X}\tilde{X}^\mathsf{T} \odot B = \frac{1}{p}X^\mathsf{T}X \odot B$, is given by

$$\|v - \hat{v}\|^2 = 2\left(1 - |v^\mathsf{T}\hat{v}|\right)$$
$$= 2\left(1 - \sqrt{\frac{\max(F\left(\sigma_\mu^2/\tilde{c}\right), 0)}{(\sigma_\mu^2/\tilde{c})\left(1+\sigma_\mu^2/\tilde{c}\right)^3}}\right) + o(1)$$

with $F(\cdot)$ defined in Theorem 3.2, where we implicitly assumed that the correct sign of $\hat{v}$ (i.e., the one for which $v^\mathsf{T}\hat{v} > 0$) is selected.

An alternative cost reduction technique consists in proceeding to PCA on a 'dimensionality-reduced' subset $\tilde{X}_r \in \mathbb{R}^{\tilde{p}_r\times\tilde{n}}$ of rows of $\tilde{X}$, with $\tilde{p}_r < \tilde{p}$. If we let $\tilde{p}_r/\tilde{p} = \varepsilon$, the computational cost is reduced by a factor ε^2, however with a 'sacrifice' (on average) of $1-\varepsilon$ of the eigenvector energy. Letting $v_r \in \mathbb{R}^{\tilde{p}_r}$ be the normalized corresponding subvector of v, the performance associated to the estimation of v_r by the dominant eigenvector $\hat{v}_r$ of $\frac{1}{\tilde{n}}\tilde{X}_r\tilde{X}_r^T$ satisfies

$$\|v_r - \hat{v}_r\|^2 = 2\left(1 - \sqrt{\frac{\max\{\varepsilon\sigma_\mu^4 - \tilde{c}, 0\}}{\varepsilon\sigma_\mu^2(\tilde{c}+\sigma_\mu^2)}}\right) + o(1)$$
$$\geq \|v - \hat{v}\|^2 + o(1),$$

the inequality following from the same arguments as in Section 4.1.

Figure 4 illustrates these results and shows that, while at first thoughts reducing the ratio $\tilde{p}/\tilde{n}$ by dimensionality reduction improves the detectability threshold $\Gamma_r = \sqrt{\tilde{p}/\tilde{n}} + o(1)$ by a factor $\sqrt{\tilde{p}_r/\tilde{p}} = \sqrt{\varepsilon}$, this simultaneously reduces the effective signal-to-noise ratio $\sigma_\mu^2(\tilde{n}/\tilde{p})$ by a factor ε: the performance thus drops by a factor $\sqrt{\varepsilon}$ *and* a full loss of $1-\varepsilon$ of the total (squared) energy of v is left aside. The proposed punctured PCA approach, in addition to leveraging the latter issue by estimating the full vector v, also outperforms its estimation over the mere (and intuitively simpler) estimation of v_r alone by $\hat{v}_r$.

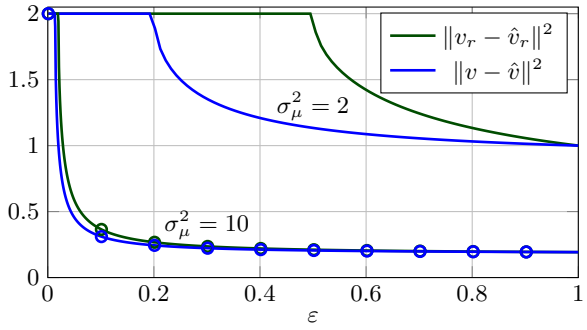

Fig. 4: Squared Euclidean distance between population eigenvectors v (and sub-vector v_r) and sample eigenvectors of punctured $\hat{v}$ versus dimensionality-reduced $\hat{v}_r$ as a function of sparsity ε (x-axis) for $\tilde{c} = 2$, $\tilde{p} = 1000$, and $\sigma_\mu^2 \in \{2, 10\}$.

5. ELEMENTS OF PROOF OF THE MAIN RESULTS

Theorem 3.1 is proved using the 'Gaussian tools' advocated in [10]. Relying on the resolvent identity $Q(z) = -\frac{1}{z} I_n + \frac{1}{z}(\frac{1}{p} X^\mathsf{T} X \odot B) Q(z)$, the method consists in finding a fixed-point relation for $\mathbb{E}[Q_{ij}(z)]$ by expanding $\mathbb{E}[[(X^\mathsf{T} X \odot B) Q(z)]_{ij}]$ thanks to Stein's lemma $\mathbb{E}[x f(x)] = \mathbb{E}[f'(x)]$ for $x \sim \mathcal{N}(0, 1)$. As opposed to the setting where B is absent (leading to the Marčenko-Pastur theorem [13]), the evaluation of $\mathbb{E}[[(X^\mathsf{T} X \odot B) Q(z)]_{ij}]$ produces terms of the type $\mathbb{E}[(X^\mathsf{T} X \odot b_\ell b_{\ell'}^\mathsf{T}) Q(z)]$, for $b_\ell, b_{\ell'}'$ columns of B, which make it difficult to 'close' the fixed-point equation. The independent Bernoulli assumption on the entries of B guarantees that these terms all have the same limiting behavior, depending on whether $\ell = \ell'$ or $\ell \neq \ell'$, which in the end allows for closing the system.

The proof of Theorem 3.2 then follows from standard arguments of spiked model analysis [16, Chp. 9], starting from Cauchy's integral formula (6). For the integral to be non-trivial, the limiting isolated eigenvalue λ (around which the complex integral is performed) must be the (unique) pole of the resolvent $\bar{Q}(z)$ in Theorem 3.1, i.e., λ is such that the denominator $c + \varepsilon m(z)(1 + \|\mu\|^2)$ evaluated at $z = \lambda$ vanishes. The associated residue is then obtained as $\lim_{z \to \lambda} (\lambda - z) v^\mathsf{T} \bar{Q}(z) v$, which follows from a first order expansion of $\bar{Q}(z)^{-1}$ around $z = \lambda$. Finally, the phase transition condition exploits the fact that, as $\|\mu\|$ decreases from ∞, $|v^\mathsf{T} \hat{v}|$ must decrease from 1 until reaching 0 at the position of the phase transition for $\|\mu\|$. This is precisely given by the equation $F(\|\mu\|^2) = 0$.

6. CONCLUSION AND DISCUSSION

Under a simplified, yet theoretically expressive data model, the article demonstrates both the validity and the convincing performance of a 'kernel puncturing' approach for spectral methods. It is in particular not a trivial result that the puncturing does not alter the structure of the dominant estimated eigenvector, which is indeed perfectly recovered in the limit $\|\mu\| \to \infty$.

This opens the path to more structured puncturing methods to optimize the performance-complexity trade-off of more advanced tasks. One may notably seek the optimal structure of a (non-identically distributed) puncturing matrix B when addressing a multi-class supervised kernel method: should one puncture proportionally to the number of elements per class? to the average distance within and across class elements? A generalized analysis of the present model surely holds the answer.

7. REFERENCES

[1] Noureddine El Karoui, "The spectrum of kernel random matrices," *The Annals of Statistics*, vol. 38, no. 1, pp. 1–50, 2010.

[2] Xiuyuan Cheng and Amit Singer, "The spectrum of random inner-product kernel matrices," *Random Matrices: Theory and Applications*, vol. 2, no. 04, pp. 1350010, 2013.

[3] Romain Couillet and Florent Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[4] Zhenyu Liao and Romain Couillet, "On the spectrum of random features maps of high dimensional data," in *Proceedings of the 35th International Conference on Machine Learning*. 2018, vol. 80, pp. 3063–3071, PMLR.

[5] Khalil Elkhalil, Abla Kammoun, Romain Couillet, Tareq Y Al-Naffouri, and Mohamed-Slim Alouini, "Asymptotic performance of regularized quadratic discriminant analysis based classifiers," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[6] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[7] Emmanuel J Candès and Pragya Sur, "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression," *The Annals of Statistics*, vol. 48, no. 1, pp. 27–42, 02 2020.

[8] Cosme Louart and Romain Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," 2019.

[9] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet, "Kernel random matrices of large concentrated data: the example of gan-generated images," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7480–7484.

[10] Leonid Andreevich Pastur and Mariya Shcherbina, *Eigenvalue distribution of large random matrices*, Number 171. American Mathematical Soc., 2011.

[11] Jinho Baik and Jack W Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of multivariate analysis*, vol. 97, no. 6, pp. 1382–1408, 2006.

[12] Florent Benaych-Georges and Raj Rao Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011.

[13] Vladimir A Marcenko and Leonid Andreevich Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457, 1967.

[14] Eugene P. Wigner, "Characteristic vectors of bordered matrices with infinite dimensions," *Annals of Mathematics*, vol. 62, no. 3, pp. 548–564, 1955.

[15] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[16] Romain Couillet and Merouane Debbah, *Random matrix methods for wireless communications*, Cambridge University Press, 2011.