Two-way kernel matrix puncturing: towards resource-efficient PCA and spectral clustering

Romain Couillet [1 2] **Florent Chatelain** [1] **Nicolas Le Bihan** [1]

Abstract

The article introduces an elementary cost and storage reduction method for spectral clustering and principal component analysis. The method consists in randomly "puncturing" both the data matrix $X \in \mathbb{C}^{p\times n}$ (or $\mathbb{R}^{p\times n}$) and its corresponding kernel (Gram) matrix K through Bernoulli masks: $S \in \{0,1\}^{p\times n}$ for X and $B \in \{0,1\}^{n\times n}$ for K. The resulting "two-way punctured" kernel is thus given by $K = \frac{1}{p}[(X \odot S)^{\mathsf{H}}(X \odot S)] \odot B$. We demonstrate that, for X composed of independent columns drawn from a Gaussian mixture model, as $n, p \to \infty$ with $p/n \to c_0 \in (0, \infty)$, the spectral behavior of K – its limiting eigenvalue distribution, as well as its isolated eigenvalues and eigenvectors – is fully tractable and exhibits a series of counter-intuitive phenomena. We notably prove, and empirically confirm on various real image databases, that it is possible to drastically puncture the data, thereby providing possibly huge computational and storage gains, for a virtually constant (clustering or PCA) performance. This preliminary study opens as such the path towards rethinking, from a large dimensional standpoint, computational and storage costs in elementary machine learning models.

1. Introduction

The ever-increasing tremendous amounts of data that machine learning algorithms now need to face start to tip the scale towards a major computational and storage resource bottleneck. In such fields as astrophysics with the recent SKA radiotelescope or Internet data mining, the collected data are simply too large to be stored and must therefore be processed in real-time before being discarded altogether. In parallel, even if those data could be stored, algorithm complexities beyond linear can in general not be afforded. This is already a problem for as elementary methods as principal component analysis (PCA) or spectral clustering – both related to Gram matrix eigenvector retrieval.

Evidently, numerous works have proposed various directions of cost-efficient methods for PCA and spectral clustering. For instance, the line of works (Johnstone & Lu, 2009; Cai et al., 2013; Deshpande & Montanari, 2014) provides a series of *sparse PCA* methods by assuming that the principal components are sparse: the main gain arises from automatically selecting the reduced set of covariates having largest amplitude. More recently, inspired by statistical physics, (Zhong et al., 2020) proposes an *empirical Bayes* version of PCA, by setting a (non-Gaussian) product measure prior on the principal components: (Zhong et al., 2020) in particular obtains (in simulations) a thousand-fold reduction in the number of data necessary to maintain equal performance with respect to standard PCA. Yet, the most popular methods to handle large dimensional PCA fall into the realm of *dimensionality reduction* and *random projections* (Freund et al., 2007) which, one way or another, also require prior knowledge on the sought principal components to avoid dramatic performance losses. Similar ideas have been devised for spectral clustering, such as hierarchical clustering (Murtagh & Contreras, 2012).

But these works all exploit strong structural prior on the data (e.g., a prior on principal components) to reduce the effective data dimension, and in general only operate on one dimension – either the data size or number.

As for mitigating storage constraints, clustering can be performed in a streaming manner, as proposed in (Keriven et al., 2018) by means of a *data sketching* approach. This approach however looses much discriminating power in not effectively "comparing" all raw data and thus fails to compete against spectral methods. Stochastic gradient descent in deep neural networks also performs clustering in a non-spectral manner by "streaming" in small data batches (Bottou, 1991), but these algorithms only converge after multiple epochs, meaning that the data must be stored for later reuse. More

[*]Equal contribution [1]GIPSA-lab, CNRS, Grenoble-INP, University Grenoble-Alps [2]CentraleSupélec, University Paris Saclay. Correspondence to: Romain Couillet <romain.couillet@gipsa-lab.grenoble-inp.fr>, Florent Chatelain <florent.chatelain@gipsa-lab.grenoble-inp.fr>, Nicolas Le Bihan <nicolas.lebihan@gipsa-lab.grenoble-inp.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

conventionally, since the addition of new data induce successive rank-1 perturbations of the sample covariance, iterative perturbation methods based on the Sherman-Morrison formula can be exploited (Engel et al., 2004), however here again at the cost of full data storage.

To cope with these limitations, the present article introduces a new random data sparsification method which trades off storage and computational cost reduction against performance. The proposed *two-way puncturing* approach consists in random Bernoulli deletions of entries (i) of the data matrix $X = [x_1, \ldots, x_n] \in \mathbb{C}^{p\times n}$ (the indices of non-zero entries differing across data) and (ii) of the Gram (sample covariance $\frac{1}{n}XX^{\mathsf{H}}$ or kernel $\frac{1}{p}X^{\mathsf{H}}X$) matrix, generically resulting in the kernel matrix model

$$K = \left\{ \frac{1}{p}(X \odot S)^{\mathsf{H}}(X \odot S) \right\} \odot B \in \mathbb{C}^{n\times n} \qquad (1)$$

for random independent Bernoulli $S \in \{0,1\}^{p\times n}$ and (symmetric) $B \in \{0,1\}^{n\times n}$, with respective parameters ε_S and $\varepsilon_B \in (0,1]$. Small values of ε_S reduce the storage size of X and the cost of the inner-product evaluation $x_i^{\mathsf{H}}x_j$, while small values of ε_B reduce the number of inner-product calculus in K and the subsequent processing of the sparsified matrix K. The approach follows after our preliminary work (Zarrouk et al., 2020), restricted to $S = 1_p 1_n^{\mathsf{T}}$ (or equivalently $\varepsilon_S = 1$) and to a simpler model for X, which already revealed that, contrary to intuition, the puncturing procedure in general *does not affect the structure* of the estimated eigenvectors (thus principal components in PCA or data classes in clustering). This conclusion still holds true here. More surprisingly, the analysis also demonstrates that there exist well-defined regimes – in terms of the ratio p/n and puncturing intensities ε_S and ε_B – for which the *PCA performance is virtually unaltered*. In particular, for equivalent levels of sparsity (in terms of resulting computational costs), we confirm here the finding of (Zarrouk et al., 2020) according to which the performance of PCA or spectral clustering on K largely overtakes the performance of the possibly more natural subsampling alternative.[1] This result is recalled in Figure 1 for $\varepsilon_S = 1$ and $\varepsilon_B \equiv \varepsilon$.

Our main findings may be summarized as follows:

1. for data x_i arising from a Gaussian mixture model $\sum_{\ell=1}^k \pi_\ell \mathcal{N}(\mu_\ell, I_n)$ (resp., a Gaussian measure $\mathcal{N}(0, C)$ with $C = I_p + R$ and R of low rank), we show that K has a limiting eigenvalue distribution following a *variation of the popular Marčenko-Pastur and semi-circle laws*; upon conditions on the eigenvalues of the matrix $\{\sqrt{\pi_i\pi_j}\mu_i^{\mathsf{T}}\mu_j\}_{i,j=1}^k$ (resp., of the matrix R), a phase transition phenomenon occurs beyond which some eigenvalues of K isolate, and their associated eigenvectors correlate to the population eigenvectors;

2. the quantities p/n, ε_S, and ε_B modulate the storage-and-computational cost versus (PCA or spectral clustering) performance trade-off; in particular, for small $\varepsilon_S, \varepsilon_B$, the performance only depends on $\varepsilon_S^2\varepsilon_B\frac{p}{n}$;

3. for small p/n ratios (i.e., for huge amounts of data), the performance of PCA and spectral clustering *plateaus* for a large range of values of ε_B (with ε_S fixed), before suffering a sharp avalanche phenomenon for ε_B below a certain threshold: this in particular indicates that *intensive puncturing (and thus complexity and storage reduction) almost comes for free* in this regime;

4. simulations on Fashion-MNIST and BigGAN generated images qualitatively (and partially quantitatively) confirm our theoretical findings, justifying the possibility to drastically reduce computational cost with virtually no impairment on classification performance.

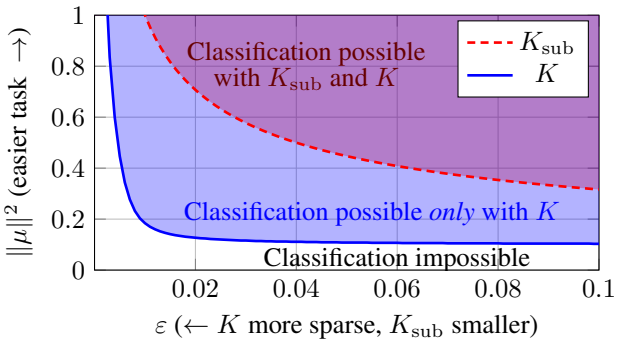

Figure 1. Phase transition diagram of spectral clustering for puncturing matrix K with $\varepsilon_S = 1$ and $\varepsilon_B \equiv \varepsilon$ versus subsampling $K_{\rm sub} \in \mathbb{C}^{n\varepsilon\times n\varepsilon}$. Here for $x_i \sim \frac{1}{2}\mathcal{CN}(\mu, I_p) + \frac{1}{2}\mathcal{CN}(-\mu, I_p)$, and $n/p = 100$ in the large n, p limit. Solid and dashed lines indicate theoretical phase transitions. ***The puncturing approach largely overtakes the subsampling method.***

Supplementary material and codes. The proofs of our main results are deferred to the supplementary material. All codes to reproduce our figures are available in the gitlab repository https://gricad-gitlab.univ-grenoble-alpes.fr/chatelaf/two-way-kernel-matrix-puncturing.

2. The two-way puncturing model

Before relating our study to principal component analysis and spectral clustering, we first formalize the model under study in a generic (and thus abstract) manner.

[1] Subsampling consists here in performing PCA or spectral clustering on n/ε subsets of the data, each of size εn, for some $\varepsilon \in (0,1]$ a multiple of $1/n$, before merging the n/ε results (which for simplicity we assume here comes at no cost).

2.1. Abstract model

Let $X \in \mathbb{C}^{p\times n}$ be a random matrix satisfying the following assumptions.[2]

Assumption 1 (Data model)**.**

$$X = Z + P$$

in which $Z_{ij} \sim \mathcal{CN}(0,1)$ are independent, and $P \in \mathbb{C}^{p\times n}$ is a rank-k matrix for some integer k.

Also define the binary *puncturing* matrices $S \in \{0,1\}^{p\times n}$ and $B \in \{0,1\}^{n\times n}$ as follows.

Assumption 2 (Puncturing matrices)**.** *Let*

- $S_{ij} \in \{0,1\}$ *be Bernoulli random variables with mean ε_S, independent across i, j;*
- $B_{ij} = B_{ji} \in \{0,1\}$ *be Bernoulli random variables with mean ε_B, independent across $i > j$;*
- $B_{ii} = b \in \{0,1\}$ *be deterministic and fixed.*

Besides, matrices S, B, and X are mutually independent.

Our objective is to study the spectral properties of the random matrix model (1). Specifically, we determine the limiting spectrum as well as the existence and characterization of isolated eigenvalues (i.e., away from the limiting spectrum and referred to as *spikes*) and their associated eigenvectors, in the limit of large p, n. To this end, the following growth rate assumptions are requested.

Assumption 3 (Large p, n asymptotics)**.** *As $n \to \infty$,*

$$p/n \to c_0 \in (0,\infty)$$

and there exists a decomposition $P = LV^{\mathsf{H}}$ of P with $V \in \mathbb{C}^{n\times k}$ isometric (i.e., $V^{\mathsf{H}}V = I_k$) and

$$\frac{1}{n}L^{\mathsf{H}}L \to \mathcal{L}$$

for some deterministic matrix $\mathcal{L} \in \mathbb{C}^{k\times k}$. In particular, the eigenvalues of $\mathcal{L}$ are the limiting k non-trivial eigenvalues of $\frac{1}{n}P^{\mathsf{H}}P$. Besides,

$$\limsup_n \max_{\substack{1\le i\le n\\ 1\le j\le k}} \{\sqrt{n}V_{ij}^2\} = 0.$$

The condition $p/n \to c_0 \in (0,\infty)$ translates the practical fact that both the dimension and number of data are large and commensurable. The convergence $(1/n)L^{\mathsf{H}}L \to \mathcal{L}$ with $P = LV^{\mathsf{H}}$ is merely technical: the decomposition $P = LV^{\mathsf{H}}$ can always be ensured by singular value decomposition, and the convergence to $\mathcal{L}$ is mostly for technical convenience. In effect, the only stringent condition is that $\limsup_n \max_{i,j} \sqrt{n}V_{ij}^2 = 0$: while naturally satisfied for spectral clustering (the V_{ij}'s are the normalized binary class indicators), for PCA this demands that the principal components be *delocalized*, i.e., not sparse.

[2] All results are provided in $\mathbb{C}$ but are equally valid in $\mathbb{R}$.

2.2. PCA and spectral clustering

The model (1) specializes to principal component analysis and spectral clustering.

Spectral clustering. Letting $P = MJ^{\mathsf{T}}$, where $M = [\mu_1, \ldots, \mu_k] \in \mathbb{C}^{p\times k}$ and $J = [j_1, \ldots, j_k] \in \{0,1\}^{n\times k}$ with $[j_\ell]_i = \delta_{\{\mathbb{E}[x_i]=\mu_\ell\}}$ for some $n_1, \ldots, n_k$, X models a k-class Gaussian mixture model with $x_i \sim \sum_{a=1}^k \pi_a \mathcal{CN}(\mu_a, I_p)$ and $n_a/n \to \pi_a$ almost surely as $n \to \infty$. Further assuming that

$$n_\ell/n \to \pi_\ell = [\pi]_\ell \in (0,\infty)$$
$$\mathcal{D}_\pi^{\frac{1}{2}} M^{\mathsf{H}} M \mathcal{D}_\pi^{\frac{1}{2}} \to \mathcal{M}$$

where $\mathcal{D}_\pi = \operatorname{diag}(\{\pi_i\}_{i=1}^k)$, we get that $P = (MD_n^{\frac{1}{2}})(JD_n^{-\frac{1}{2}})^{\mathsf{H}}$ with $D_n = \operatorname{diag}(\{n_i\}_{i=1}^k)$, for which

$$(JD_n^{-\frac{1}{2}})^{\mathsf{T}}(JD_n^{-\frac{1}{2}}) = I_k, \quad \frac{1}{n}(MD_n^{\frac{1}{2}})^{\mathsf{H}}(MD_n^{\frac{1}{2}}) \to \mathcal{M}$$

thereby satisfying Assumptions 1–3, for $\mathcal{L} = \mathcal{M}$. Under this setting, $\frac{1}{p}X^{\mathsf{H}}X$ is (the elementary version of) a kernel random matrix used in machine learning as the base ingredient for kernel-based classification methods. In particular, the eigenvectors associated with the dominant eigenvalues of $\frac{1}{p}X^{\mathsf{H}}X$ are the base elements of the popular *(kernel) spectral clustering* algorithm (Von Luxburg, 2007).

Principal component analysis. Letting instead $P = \tilde{Z}A^{\mathsf{H}}$ with $A \in \mathbb{C}^{n\times k}$ deterministic and $\tilde{Z} \in \mathbb{C}^{p\times k}$ random with i.i.d. $\mathcal{CN}(0,1)$ entries, independent of Z, we get

$$X^{\mathsf{H}} = \begin{bmatrix} I_n & A \end{bmatrix} \begin{bmatrix} Z^{\mathsf{H}} \\ \tilde{Z}^{\mathsf{H}} \end{bmatrix}$$

which is a matrix with $\mathcal{CN}(0, I_n + AA^{\mathsf{H}})$ independent columns, so that $\frac{1}{p}X^{\mathsf{H}}X$ is a sample covariance matrix for the p *rows*[3] of X of dimension n; the dominant eigenvectors of $\frac{1}{p}X^{\mathsf{H}}X$ are therefore the principal components of the popular *principal component analysis* method. Further requesting A to have spectral decomposition $A = USV^{\mathsf{H}}$, where $S \in \mathbb{R}_+^{k\times k}$ satisfies $S^{\mathsf{H}}S \to \mathcal{S}$ deterministic, one gets that $P = (\tilde{Z}US)V^{\mathsf{H}}$ with $V^{\mathsf{H}}V = I_k$ and

$$\frac{1}{n}(\tilde{Z}US)^{\mathsf{H}}(\tilde{Z}US) \to \mathcal{S}$$

again satisfying Assumption 3 for $\mathcal{L} = \mathcal{S}$.

3. Main results

As per standard random matrix methods, the technical approach to study the limiting spectrum of K consists in characterizing the *resolvent* matrix

$$Q(z) = (K - zI_n)^{-1}$$

[3] One must be careful here that standard notations of n and p are reversed under this setting.

defined for $z \in \mathbb{C} \setminus \{\lambda_i\}_{i=1}^n$ with λ_i the eigenvalues of K. Specifically, the *spectral measure* $\nu_n \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$ of K relates to the *Stieltjes transform* $m_n(z) \equiv \int (t-z)^{-1}\nu_n(dt) = \frac{1}{n}\mathrm{tr}Q(z)$, while the *eigenvector* $\hat{u}_i \in \mathbb{C}^n$ associated to eigenvalue $\lambda_i(K)$ relates to the Cauchy-integral $\hat{u}_i\hat{u}_i^\mathsf{H} = \frac{-1}{2\pi\imath}\oint_{\Gamma_{\lambda_i}} Q(z)dz$ for Γ_{λ_i} a small positively oriented complex contour circling around λ_i only.

3.1. Limiting spectral behavior

Our core technical result provides a said *deterministic equivalent* for the random matrix $Q(z)$, from which the limiting behavior of the eigenvalues and eigenvectors of K follows.

Theorem 1 (Deterministic equivalent for Q)**.** *Under Assumptions 1–3, let $z \in \mathbb{C}$ be away from the limsup of the union of the supports of $\nu_1, \nu_2, \ldots$. Then, as $n \to \infty$,*

$$Q(z) \leftrightarrow m(z)\left[I_n + \frac{c_0^{-1}\varepsilon_S^2\varepsilon_B m(z)}{1+\varepsilon_B\varepsilon_S c_0^{-1}m(z)} V\mathcal{L}V^\mathsf{H}\right]^{-1}$$

where $m(\cdot)$ is the unique Stieltjes transform solution to

$$z = \varepsilon_S b - \frac{1}{m(z)} - c_0^{-1}\varepsilon_B\varepsilon_S^2 m(z) + \frac{c_0^{-2}\varepsilon_B^3\varepsilon_S^3 m(z)^2}{1+c_0^{-1}\varepsilon_B\varepsilon_S m(z)}$$

and the notation $A \leftrightarrow B$ indicates that, for any linear functional $u : \mathbb{C}^{n\times n} \to \mathbb{R}$ of bounded infinity norm, $u(A-B) \to 0$ almost surely as $n \to \infty$.

One must understand the theorem as follows: since $Q(z)$ encapsulates the structural spectral information about K, this information is fully determined (in the large n, p limit)

(i) by the scalars ε_S, ε_B, c_0 and b; these mostly impact the *shape* of the limiting spectrum in defining $m(\cdot)$) and modulate the "noise level" of the eigenvectors (from the factor preceding $V\mathcal{L}V^\mathsf{H}$ in the expression of $Q(\cdot)$);

(ii) by the rank-k matrix $V\mathcal{L}V^\mathsf{H}$; this matrix defines the "average" behavior of the dominant eigenvectors of K: these eigenvectors are simply "isotropic noisy versions" of linear combinations of the columns of V. That is, mapped to the applications in Section 2.2, noisy versions of either the class canonical vectors j_a's or of the genuine PCA vector.

As an immediate – and possibly quite surprising – consequence, the dominant eigenvectors of K are, up to extra *homogeneous noise*, the same as those of $P^\mathsf{H}P = \mathbb{E}[\frac{1}{p}X^\mathsf{H}X] - I_n$. The proposed two-way puncturing algorithm therefore *does not affect spectral algorithms* as the structure of the retrieved eigenvectors is maintained.

Let us now quantify these so far qualitative statements. As a first corollary of Theorem 1, with probability one,

$$\frac{1}{n}\mathrm{tr}Q(z) \equiv m_n(z) \to m(z)$$

which implies, according to random matrix theory, that

$$\nu_n \equiv \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i} \to \nu$$

almost surely, where ν is the unique probability measure having Stieltjes transform $m(z)$ (i.e., $m(z) = \int (t-z)^{-1}\nu(dt)$). It thus suffices to solve the defining equation for $m(z)$ in Theorem 1 to estimate the limiting spectral distribution ν of K.[4] Figure 2 indeed confirms the correspondence between the empirical (finite n, p) spectrum ν_n of K versus the estimated limit ν.

Remark 1 (Sitting between Marčenko-Pastur and Wigner)**.** *Not surprisingly, when $\varepsilon_B = 1$ and $b = 1$, $K = \frac{1}{p}(X \odot S)^\mathsf{H}(X \odot S)$ with $X \odot S$ a matrix with i.i.d. entries of zero mean and variance ε_S^2, so that ν falls back onto the popular Marčenko-Pastur distribution (Marčenko & Pastur, 1967) (up to an ε_S scale). Precisely, for $z' = z/\varepsilon_S$ and $\tilde{m}(z) = \int (t/\varepsilon_S - z)^{-1}\nu(dt)$ (i.e., the Stieltjes transform of the limiting measure of the λ_i/ε_S), the canonical equation of $m(z)$ in Theorem 1 becomes*

$$z' = 1 - \frac{1}{\tilde{m}(z')} - \frac{c_0^{-1}\tilde{m}(z')}{1+c_0^{-1}\tilde{m}(z')}$$

which is the defining Stieltjes transform equation of the Marčenko-Pastur law. The more interesting small ε_B setting is treated in Section 3.3 and gives rise to a Wigner semi-circle limit instead (Wigner, 1958). As such, through the values $\varepsilon_S, \varepsilon_B$, the limiting spectral measure ν continuously moves from the Marčenko-Pastur to the Wigner semi-circle laws. Figure 2 illustrates this observation: the shape of ν is simultaneously reminiscent of both laws.

3.2. Phase transition and dominant eigenvectors

The limiting Stieltjes transform $m(z)$ determines the "macroscopic" behavior of the spectrum ν_n of K, but does not provide the position of its isolated eigenvalues and even less the shape of the associated eigenvectors. To this end, a deeper investigation of the deterministic equivalent of $Q(z)$ is needed. Our next result provides this analysis.

Theorem 2 (Phase transition, isolated eigenvalues and eigenvectors)**.** *Define the functions*

$$F(t) = t^4 + \frac{2}{\varepsilon_S}t^3 + \frac{1}{\varepsilon_S^2}\left(1 - \frac{c_0}{\varepsilon_B}\right)t^2 - \frac{2c_0}{\varepsilon_S^3}t - \frac{c_0}{\varepsilon_S^4}$$

$$G(t) = \varepsilon_S b + c_0^{-1}\varepsilon_B\varepsilon_S(1+\varepsilon_S t) + \frac{\varepsilon_S}{1+\varepsilon_S t} + \frac{\varepsilon_B}{t(1+\varepsilon_S t)}$$

and $\Gamma \in \mathbb{R}$ be the largest real solution to $F(\Gamma) = 0$. Further denote $\ell_1 > \ldots > \ell_{\bar{k}}$ the $\bar{k} \leq k$ distinct eigenvalues of $\mathcal{L}$ of respective multiplicities $L_1, \ldots, L_{\bar{k}}$, and

[4] The measure ν is practically retrieved from $m(\cdot)$ by using the inverse formula $\nu(dt) = \lim_{y\downarrow 0}\frac{1}{\pi}\Im[m(t+\imath y)]dt$.

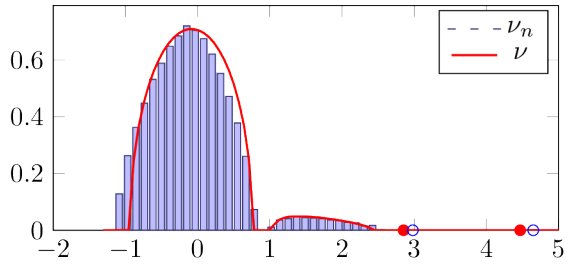

Figure 2. Eigenvalue distribution ν_n of K versus limit measure ν, for $p = 200$, $n = 4\,000$, $x_i \sim .4\mathcal{N}(\mu_1, I_p) + .6\mathcal{N}(\mu_2, I_p)$ for $[\mu_1^\mathsf{T}, \mu_2^\mathsf{T}]^\mathsf{T} \sim \mathcal{N}(0, \frac{1}{p}\left[\begin{smallmatrix} 10 & 5.5 \\ 5.5 & 15 \end{smallmatrix}\right] \otimes I_p)$; $\varepsilon_S = .2$, $\varepsilon_B = .4$, $b = 1$. Sample vs theoretical spikes in blue vs red circles. ***The two "humps" remind the semi-circular and Marčenko-Pastur laws.***

$\Pi_1, \ldots, \Pi_{\bar{k}} \in \mathbb{R}^{k\times k}$ the projectors on their respective associated eigenspaces. Similarly denote $(\lambda_1, \hat{v}_1), \ldots, (\lambda_n, \hat{v}_n)$ the eigenvalue-eigenvector pairs of K in descending order and gather the first k eigenvectors under the isometric matrices $\hat{\mathcal{V}}_1 = [\hat{v}_1, \ldots, \hat{v}_{L_1}]$ up to $\hat{\mathcal{V}}_{\bar{k}} = [\hat{v}_{k-L_{\bar{k}}+1}, \ldots, \hat{v}_k]$.

Then, for $i \in \{1, \ldots, \bar{k}\}$ and for all $j \in \{L_1 + \ldots + L_{i-1} + 1, \ldots, L_1 + \ldots + L_i\}$,

$$\lambda_j \to \rho_i \equiv \begin{cases} G(\ell_i) & , \text{ if } \ell_i > \Gamma \\ G(\Gamma) & , \text{ if } \ell_i \leq \Gamma \end{cases}$$

almost surely, and

$$\hat{\mathcal{V}}_i \hat{\mathcal{V}}_i^\mathsf{H} \leftrightarrow \zeta_i V \Pi_i V^\mathsf{H}, \text{ for } \zeta_i = \begin{cases} \frac{F(\ell_i)\varepsilon_S^3}{\ell_i(1+\varepsilon_S\ell_i)^3} & , \ell_i > \Gamma \\ 0 & , \ell_i \leq \Gamma \end{cases}$$

with the notation '$\leftrightarrow$' introduced in Theorem 1. In particular, if the ℓ_i's have unit multiplicities with associated population eigenvectors v_i, then

$$|v_i^\mathsf{H} \hat{v}_i|^2 \to \zeta_i, \quad i = 1, \ldots, k.$$

To best understand the theorem, suppose that $P = lv^\mathsf{H}$ is a rank-1 matrix with $\|v\|^2 = 1$ and $\|l\|^2/n = \ell$. Then, if $\ell > \Gamma$, with Γ the largest solution to $F(\Gamma) = 0$ – this threshold *only depending on ε_S, ε_B and c_0* –, the spectrum of K exhibits an isolated eigenvalue λ, the eigenvector $\hat{v}$ of which *aligns* to v: i.e., $|\hat{v}^\mathsf{H} v|^2 \to \zeta > 0$. Otherwise, if $\ell < \Gamma$, the largest eigenvalue λ of K remains "stuck" in the limiting bulk of eigenvalues of K and $|\hat{v}^\mathsf{H} v|^2 \to 0$ (i.e., the eigenvector $\hat{v}$ does not carry any information on v: PCA and spectral clustering both fail in this scenario). Figure 3 illustrates the limiting (squared) alignment ζ as a function of ℓ.

In the more general setting where P is a rank-k matrix, possibly with multiplicities, the theorem specifies the conditions on ε_B, ε_S and c_0 under which the dominant eigenvectors of K remain correlated (and to which extent) to the population eigenspaces. This characterization is of tremendous importance to assess the exact performance of PCA and spectral clustering under the double-puncturing cost reduction. Figure 3 illustrates Theorem 2 in a clustering setting.

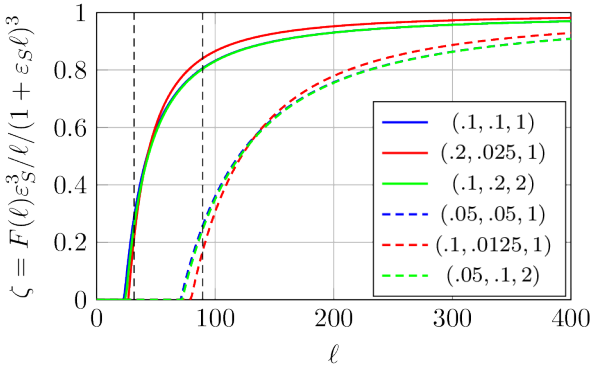

Figure 3. Illustration of Theorem 2: asymptotic sample-population eigenvector alignment for $\mathcal{L} = \ell \in \mathbb{R}$, as a function of the "information strength" ℓ. Various values of $(\varepsilon_S, \varepsilon_B, c_0)$ indicated in legend. Black dashed lines indicate the limiting (small $\varepsilon_S, \varepsilon_B$) phase transition threshold $\Gamma = (\varepsilon_S^2 \varepsilon_B c_0^{-1})^{-\frac{1}{2}}$. ***As $\varepsilon_S, \varepsilon_B \to 0$, performance curves coincide when $\varepsilon_B \varepsilon_S^2 c_0^{-1}$ is constant (plain versus dashed set of curves).***

An important quantity of Theorem 2 is the function F, which intervenes both to establish the condition under which informative isolated eigenvalues are found in the spectrum of K, thereby defining the *phase transition* threshold for the population eigenvalue ℓ_i (through $F(\ell_i) = 0$), and to evaluate the corresponding empirical eigenvector(s) quality through $\zeta_i = F(\ell_i)\varepsilon_S^3/(\ell_i(1+\varepsilon_S\ell_i)^3)$ (which is zero right at the phase transition threshold). The phase transition determines which values of the tuple $(\varepsilon_S, \varepsilon_B, c_0, \ell_i)$ coincide with the emergence of an isolated eigenvalue in the spectrum of K associated to the population eigenvalue ℓ_i, and thus to the actual feasibility of PCA or spectral clustering.

Assume now that $c_0 \ll 1$ (i.e., $n \gg p$) and that ε_B and ℓ_i are kept fixed and away from zero. Then, in the expression of $F(\ell_i)$, $1 \gg c_0/\varepsilon_B$ so that, in the first order, $F(\ell_i)$ is independent of ε_B. This quite importantly implies that the "function" $\varepsilon_S : \varepsilon_B \mapsto \varepsilon_S(\varepsilon_B)$ such that $F(\ell_i) = 0$ is mostly flat for a range of non-small values of ε_B. This behavior is confirmed in Figure 4 (left display). Also, since ζ_i would also marginally depend on ε_B, the eigenvector quality is also the same for a wide range of ε_B. The major consequence of this remark is that, for $c_0 \ll 1$, ε_B can be taken quite small without affecting the quality of the dominant eigenvectors: *puncturing through B does not affect the PCA or spectral clustering performance and thus almost comes for free!*

Conversely, still for $c_0 \ll 1$, for ε_S fixed and away from zero, we find that, at the phase transition,

$$\varepsilon_B \simeq c_0/(1+\varepsilon_S\ell_i)^2.$$

As such, the reverse function $\varepsilon_B(\varepsilon_S)$ is quite different from $\varepsilon_S(\varepsilon_B)$: it mostly behaves as $1/\varepsilon_S^2$ so that, in order not to

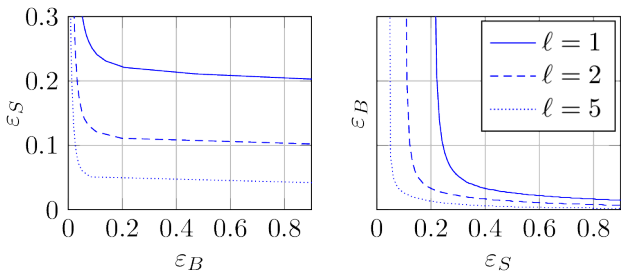

Figure 4. Phase transition curves $F(\ell) = 0$ for $\mathcal{L} = \ell \in \mathbb{R}$ and varying values of ℓ, for $c_0 = .05$. Above each phase transition curve, a spike eigenvalue is found away from the support of ν. ***For large ℓ, a wide range of ε_B's (resp.ε_S) is admissible at virtually no performance loss. Here, also, sparser B matrices are more effective than sparser S matrices.***

loose performance, increased sparsification through S must come along with reduced sparsification through B.

Of utmost interest though is the case where c_0 and ℓ_i are fixed (although, as we will see, ℓ_i^2/c_0 must be large), and where both Bernoulli parameters ε_B and ε_S assume small values. This scenario is all the more relevant that Theorems 1–2 and their corollaries take on simple and intuitive forms. This setting is discussed next.

3.3. Small ε_B, ε_S limit

Letting $z' = \sqrt{c_0/\varepsilon_B\varepsilon_S^2}(z - \varepsilon_S b)$, we obtain, in the limit of small ε_B and ε_S, that

$$z' = -1/m_0(z') - m_0(z') + o\left((\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}} m_0(z')\right)$$

with $m_0(z) = (c_0^{-1}\varepsilon_B\varepsilon_S^2)^{\frac{1}{2}} m((c_0^{-1}\varepsilon_B\varepsilon_S^2)^{\frac{1}{2}} z + \varepsilon_S b)$, i.e., for ν the measure associated to $m(z)$, m_0 is the Stieltjes transform of the measure $\nu(t/(c_0^{-1}\varepsilon_B\varepsilon_S^2)^{\frac{1}{2}})$.

This is the defining equation of Wigner's semi-circle law (Wigner, 1958) centered at $\varepsilon_S b$ and with edges $\varepsilon_S b \pm 2(c_0^{-1}\varepsilon_B\varepsilon_S^2)^{\frac{1}{2}}$.

Similarly, assuming $\ell_i > \Gamma$, and letting $\rho_i' = (\rho_i - \varepsilon_S b)/c_0^{-1}\varepsilon_B\varepsilon_S^{2\frac{1}{2}}$ and $\ell_i' = \ell_i(\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}}$, we find, after first order Taylor expansion, the spike equation $m_0(\rho_i') = -1/\ell_i' + o((\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}})$, or equivalently

$$\rho_i' = \ell_i' + 1/\ell_i' + o\left((\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}}\right)$$

which is the classically known isolated eigenvalue from the deformed Wigner random matrix (Pastur & Shcherbina, 2011, Chapter 2.2). The scaling of ℓ_i into ℓ_i' importantly indicates that, for a non-trivial spike to emerge, the eigenvalue ℓ_i of $\mathcal{L}$ must scale like $O((c_0\varepsilon_B^{-1}\varepsilon_S^{-2})^{\frac{1}{2}})$.

In practical terms, these results show that (i) for spectral clustering to be feasible (but non-trivial), the inter-class distance $\|\mu_a - \mu_b\|^2$ must scale like $\sqrt{c_0/(\varepsilon_B\varepsilon_S^2)}$, and (ii) for PCA, the eigenvalues of the principal components must scale like $\sqrt{c_0/(\varepsilon_B\varepsilon_S^2)}$.

As for the alignment of eigenspaces, it is given by

$$\hat{\mathcal{U}}_i\hat{\mathcal{U}}_i' \leftrightarrow \left(1 - 1/(\ell_i')^2 + o\left((\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}}\right)\right) V\Pi_i V^{\mathsf{H}}$$

which, again, is a classical result in the deformed Wigner random matrix model. Setting the alignment to zero, this result also provides a much simpler value for the phase transition threshold $\ell_i = \Gamma$ of Theorem 2 (in the limit of small $\varepsilon_S, \varepsilon_B$) which corresponds to $\ell_i' \simeq 1$, or equivalently

$$\Gamma \simeq 1/(\varepsilon_B\varepsilon_S^2 c_0^{-1})^{\frac{1}{2}}.$$

Remark 2 (Trading off ε_B, ε_S and c_0)**.** *As a consequence of the results above, it appears that, for small values of ε_B, ε_S and c_0^{-1}, the spectral behavior (eigenvalues and eigenvectors) of K is unaltered so long that $\varepsilon_B\varepsilon_S^2 c_0^{-1}$ is constant. For instance, doubling n is equivalent to doubling ε_B or multiplying ε_S by $\sqrt{2}$. This is confirmed by Figure 3 in which the two sets of plain or dashed curves, corresponding to constant $\varepsilon_B\varepsilon_S^2 c_0^{-1}$, almost coincide.*

It is important to further note that, unlike ε_B, ε_S is squared in the expression $\varepsilon_B\varepsilon_S^2 c_0^{-1}$ due to the fact that, denoting $S = [s_1, \ldots, s_n]$, the inner products $(x_i \odot s_i)^{\mathsf{H}}(x_j \odot s_j)$, for all $i \neq j$, involve on average ε_S^2 terms (since $\frac{1}{p}\mathbb{E}[s_i^{\mathsf{T}} s_j] = \varepsilon_S^2$).

One must be careful not to confuse the findings of Section 3.2 on *non-small* ε_B according to which $\varepsilon_B \in (0, 1]$ has a marginal impact on performance (and thus that intensive puncturing comes for free), to the present results which on the opposite indicate that *for small* ε_B, more intensive puncturing decreases the performance. Both regimes are very different as Figure 4 clearly indicates.

4. Practical consequences: the storage/complexity performance trade-off

The main interest of the two-way puncturing approach lies in its effective computational and storage cost reductions, while maintaining high performance levels. As a follow-up of Remark 2, puncturing through the matrix S can be traded off by puncturing through B, and vice-versa, with, we will see, varying effects on storage and computational costs.

4.1. Storage and computation costs

Computing K. For $B_{ij} = 1$, evaluating K_{ij} comes at average cost of $\mathbb{E}[\sum_{\ell=1}^p S_{i\ell}S_{\ell j}] = \varepsilon_S^2$ products. As a result, the whole matrix K, with an average $\sum_{i,j=1}^n \mathbb{E}[B_{ij}] = \varepsilon_B n^2$ (if $b = 1$, and $\varepsilon_B(n-1)^2$ if $b = 0$) non-zero entries, has $O(n^2 p\varepsilon_S^2\varepsilon_B)$ theoretical computation cost.

Storage data. In terms of storage, if one wishes to maintain the data information $X \odot S$ for further (non-kernel related) use, the net gain is a factor ε_S on average (for a net storage of $\varepsilon_S pn$ values). If instead only the matrix K is of relevance for future use, then the storage is restricted to $\varepsilon_B n(n-1)/2 + n$ values when $b = 1$ (accounting for symmetry) or $\varepsilon_B n(n-1)/2$ values when $b = 0$.

Spectral methods. When it comes to spectral methods (PCA or spectral clustering), one needs to retrieve the (few) dominant eigenvectors of K. Using a power method on K to sequentially iterate over each eigenvector is in general optimal and comes at a cost of $O(\varepsilon_B n^2)$, where the $O(\cdot)$ notation encompasses the number of iterations required for convergence (which depends on the spectral gap between isolated eigenvalues and thus does not scale with n in our setting). This is a gain of order ε_B over no puncturing.

Yet, when $p \ll n$, to evaluate the dominant eigenpairs of $X^{\mathsf{H}}X$, it is more efficient in practice to proceed to a singular value decomposition of the $n \times p$ matrix X^{H}, again via a power method. When operating the Hadamard product with B though, this strategy cannot be put in place as $X^{\mathsf{H}}X \odot B$ is in general of full rank n. It is thus in this case beneficial to divert the sparsity into letting $\varepsilon_B = 1$ and $\varepsilon_S \ll 1$ so to be able to run a singular vector decomposition over the very sparse matrix $(X \odot S)^{\mathsf{H}}$.

Remark 3 (Cache issues)**.** *The computational costs reported in this section are provided in terms of net number of product operations, irrespective of computer architecture or implementation. But computing the entries of the Gram matrix $X^{\mathsf{H}}X$ can be advantageously performed "block-wise" by caching vectors in sequences of blocks and computing the corresponding subblocks of $X^{\mathsf{H}}X$. This powerful trick cannot be performed on the two-way punctured matrix K which, due to the randomness in S and B, is not organized in blocks. In practice, we observed that the cost of systematically retrieving the x_i's by pairs from remote memory is not outbalanced by the gains in net number of products. Improved software designs are thus required to overtake this practical limitation.*

4.2. Application: large data clustering

As a telling application of our results, let us consider the spectral clustering setting described in Section 2.2.

4.2.1. Synthetic data

We first let $x_1, \dots, x_n \in \mathbb{R}^p$ arise from a synthetic two-class Gaussian mixture with $n = 4\,000$ and $p = 2\,000$. Two puncturing approaches are compared: (i) reducing the cost of the inner products $x_i^{\mathsf{T}}x_j$ using a 5-fold ($\varepsilon_S = .2$ while $\varepsilon_B = 1$) random puncturing of the data vectors x_i, versus (ii) a 25-fold puncturing of the matrix $\frac{1}{p}X^{\mathsf{T}}X$ ($\varepsilon_B = .04$ while $\varepsilon_S = 1$). Figure 5 depicts (for a setting detailed in caption) the matrices K, their spectra and second dominant eigenvector $\hat{v}_2$ ($\hat{v}_1$ is not discriminating in this setting, due to $P^{\mathsf{H}}P$ having a dominant all-ones eigenvector). The reported scenario is interesting in that we purposely took $\varepsilon_B \varepsilon_S^2 c_0^{-1}$ constant in both cases; as such, while the matrices K and their spectra dramatically differ, eigenvector $\hat{v}_2$ is essentially the "same" in both matrices. This first confirms the theory but most importantly defies the natural intuition that so different matrices cannot possibly give rise to the same eigenvector structure and quality.

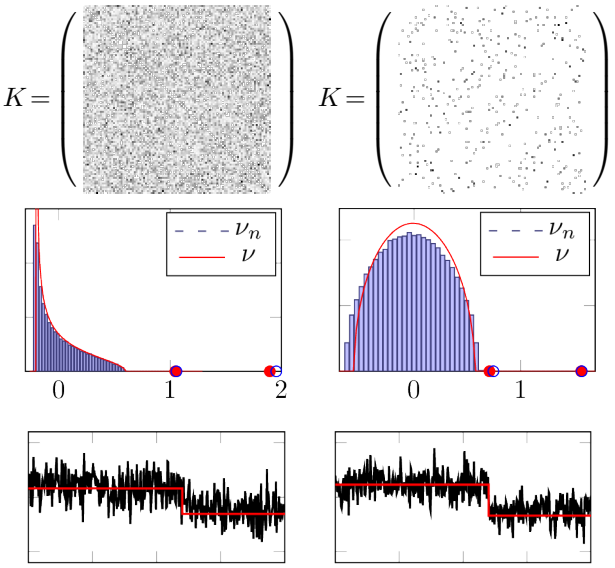

Figure 5. Two-way punctured matrices K for **(left)** $(\varepsilon_S, \varepsilon_B) = (.2, 1)$ or **(right)** $(\varepsilon_S, \varepsilon_B) = (1, .04)$, with $c_0 = \frac{1}{2}$, $n = 4\,000$, $p = 2\,000$, $b = 0$. Clustering setting with $x_i \sim .4\mathcal{N}(\mu_1, I_p) + .6\mathcal{N}(\mu_2, I_p)$ for $[\mu_1^{\mathsf{T}}, \mu_2^{\mathsf{T}}]^{\mathsf{T}} \sim \mathcal{N}(0, \frac{1}{p}\left[\begin{smallmatrix} 20 & 12 \\ 12 & 30 \end{smallmatrix}\right] \otimes I_p)$. **(Top)** first 100×100 absolute entries of K (white for zero); **(Middle)** spectrum of K, theoretical limit, and isolated eigenvalues; **(Bottom)** second dominant eigenvector $\hat{v}_2$ of K against theoretical average in red. ***As confirmed by theory, although (top) K is dense for $\varepsilon_B = 1$ and sparse for $\varepsilon_B = .04$ (96% empty) and (middle) the spectra strikingly differ, (bottom) since $\varepsilon_S^2 \varepsilon_B c_0^{-1}$ is constant, the eigenvector alignment $|\hat{v}_2^{\mathsf{T}} v_2|^2$ is the same in both cases.***

In the very symmetric setting of two classes of equal sizes ($n/2$ elements per class) and opposed statistical means (i.e., with $x_i \sim .5\mathcal{N}(\mu, I_p) + .5\mathcal{N}(-\mu, I_p)$), only one spike population eigenvalue is non-zero and $v = v_1$ is known: its normalized entries belong to $\{\pm\frac{1}{\sqrt{n}}\}$ (indeed, here $\mathcal{M} = \frac{1}{2}\|\mu\|^2\left[\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right]$, the eigenvalues of which equal $\|\mu\|^2$ and 0 with respective eigenvectors $[1, -1]$ and $[1, 1]$). By symmetry, the random entries of the sample eigenvector $\hat{v} \equiv \hat{v}_1$ are asymptotically centered on $\pm\sqrt{\zeta/n}$ with variance asymptotically equal to $(1 - \zeta)/n$ for $\zeta \equiv \zeta_1$ provided by Theorem 2 (with $\ell_1 = \|\mu\|^2$). Related random matrix studies (e.g., (Kadavankandy & Couillet, 2019) for

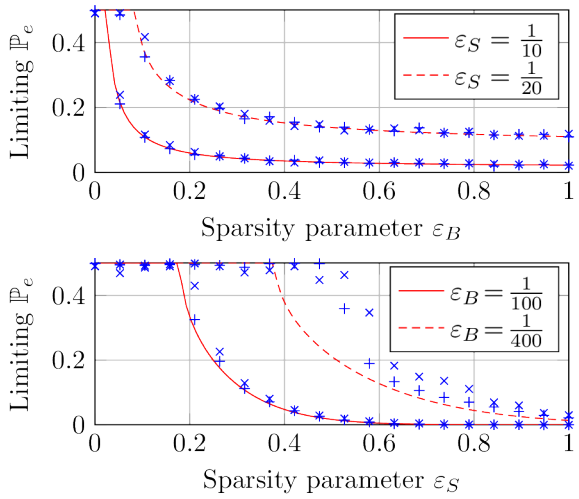

Figure 6. Limiting probability of error of spectral clustering of $\mathcal{N}(\pm\mu, I_p)$ with equal class sizes on K: as a function of ε_B for fixed $\ell = \|\mu\|^2 = 50$ **(top)**, and ε_S for fixed $\ell = 50$ **(bottom)**. Simulations (single realization) in markers for $p = n = 4\,000$ ($\times$) and $p = n = 8\,000$ ($+$). ***Very good fit between theory and practice for not too small*** $\varepsilon_S, \varepsilon_B$.

$\varepsilon_S = \varepsilon_B = 1$) have shown that the fluctuations of the entries of $\hat{v}$ are asymptotically Gaussian and pairwise independent; this suffices to justify that the asymptotic classification error $\mathbb{P}_e$ incurred by spectral clustering is given by:

$$\mathbb{P}_e = \frac{1}{n}\sum_{i=1}^{n} \delta_{\{\text{sign}([\hat{v}]_i[v]_i)<0\}} \to Q\left(\sqrt{\zeta/(1-\zeta)}\right)$$

almost surely, where $Q(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-u^2/2}du$ is the Gaussian tail function, and the (arbitrary) signs of $v, \hat{v}$ are chosen such that $0 \le P_e \le \frac{1}{2}$. Figure 6 depicts the limiting error for various values of $(\varepsilon_S, \varepsilon_B, c_0, \ell)$. Despite ε_B and ε_S being particularly in this setting, the simulations show a strong fit between theory and practice, even for not so large values of n.

Remark 4 (How large should n, p be in practice?)**.** *It is well established in random matrix theory that limiting results can be obtained at speeds up to $O(1/\sqrt{pn}) = O(1/n)$. We may in particular show here that $\mathbb{P}_e = Q(\sqrt{\zeta/(1-\zeta)}) + O(1/n)$. As a consequence, our practical predictions are already accurate for quite small values of n.*

This being said, the $O(1/n)$ term hides constants, particularly depending on $\varepsilon_S, \varepsilon_B$ which cannot be taken too small. As a rule of thumb, $1/\varepsilon_S, 1/\varepsilon_B$ must remain small compared to p, n.[5] *This last remark explains in passing the disrupted behavior of Figure 6-(bottom) for too small ε_B.*

[5]If not, as discussed in the article concluding remarks, K falls into a "sparse regime" no longer supported by the present random matrix analysis.

Figure 7. Examples of BigGAN-generated images, '`collie`' dog instances (**top row**), '`tabby`' cat instances (**bottom row**).

4.2.2. Resilience to real-world images

To practically confirm our theoretical findings, we next apply the two-way puncturing kernel to vectors x_i arising from a two-class mixture ('`tabby`' cats versus '`collie`' dogs; see Figure 7) of the (globally centered and scaled) $p = 4\,096$-VGG features of randomly BigGAN-generated images (Brock et al., 2018). The results are for varying ε_B and either fixed ε_S or ε_S set such that $\varepsilon_S^2\varepsilon_B = 5\cdot 10^{-4}$. The simulation depicted in Figure 8 corroborates the presence of a performance "plateau" and a significant reduction of the transition value of ε_B (from .05 to .015) when n (and thus $1/c_0$) increases fourfold. This supports the theoretical performance of the central display in Figure 6. Maintaining $\varepsilon_S^2\varepsilon_B$ constant pushes this plateau further down to smaller values of ε_B until the method breaks. The same conclusion can be drawn on non-pretreated $p = 784$-dimensional real word images from the Fashion-MNIST dataset, as shown in Figure 9.

More interestingly, as shown in Figure 10, while for $\varepsilon_B = \varepsilon_S = 1$ the eigenvalues of K for the GAN images spread far from the theoretical Marčenko-Pastur limit,[6] for $\varepsilon_B, \varepsilon_S \ll 1$, the empirical spectrum is very close to the predicted (uncorrelated vector) limit: this strongly suggests that intensive puncturing has the effect to "decorrelate" data. This remark has the powerful advantage to improve the theoretical tractability of these preprocessed data. More surprisingly, for both small or large $\varepsilon_B, \varepsilon_S$, despite the general spectrum mismatch, the anticipated dominant eigenvalue position and eigenvector behavior are extremely good, making it still possible to predict clustering performance with good accuracy. The same conclusions apply to the Fashion-MNIST dataset (see figures in the gitlab repository).

[6]This may at first be thought to follow from strong feature covariance (thus not close to I_p), but it turns out that in-sample correlation is even stronger as the VGG-features of the produced GAN images appear to have a very low variability.

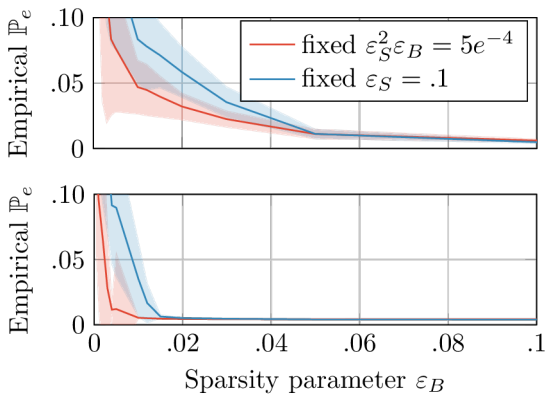

Figure 8. Empirical classification errors for 2-class (balanced) BigGAN-generated images (‘tabby’ vs ‘collie’), with $n = 2\,500$ (**top**) and $n = 10\,000$ (**bottom**). ***Theoretically predicted “plateau”-behavior observed for all ε_B not too small.***

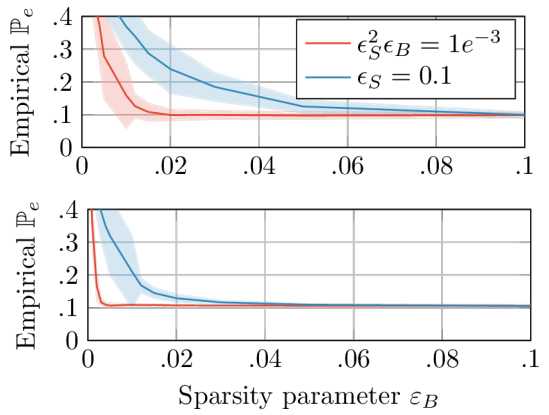

Figure 9. Empirical classification errors for 2-class (balanced) MNIST-fashion images (‘trouser’ vs ‘pullover’), with $n = 512$ (**top**) and $n = 2048$ (**bottom**). ***Similar “plateaus” as predicted by the theory and observed in Figure 8.***

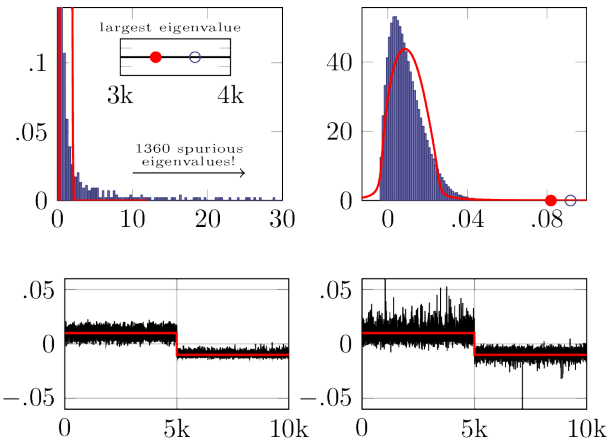

Figure 10. Sample vs limiting spectra and dominant eigenvector of K for 2-class GAN images (tabby vs collie); **(left)** $\varepsilon_S = \varepsilon_B = 1$ (error rate: $\mathbb{P}_e = .004$); **(right)** $\varepsilon_S = 0.01$, $\varepsilon_B = 0.2$ ($\mathbb{P}_e = .011$). ***Surprisingly good fit between sample and predicted isolated eigenvalue/eigenvector in all cases; as for spectral measure, significant prediction improvement as $\varepsilon_S, \varepsilon_B \to 0$.***

5. Concluding remarks

A fundamental conclusion of the article, confirmed on practical data, is that drastic computation and storage reduction can be theoretically achieved while virtually incurring no loss in PCA or spectral clustering. This follows from the peculiar behavior of (doubly) punctured kernel and sample covariance matrices K. As shown in an enlarging spectrum of articles, the large dimensional behavior of Q has immediate further implications to the performance behavior of many machine learning algorithms, ranging from support vector machines (Kammoun & Alouini, 2020; Huang, 2017) to semi-supervised graph inference (Mai & Couillet, 2018), transfer and multi-task learning (Tiomoko et al., 2020), random feature maps (Liao & Couillet, 2018b; Pennington & Worah, 2019), or neural network dynamics (Liao & Couillet, 2018a; Advani et al., 2020), to cite a few. As such, the article, rather than providing a ready-to-use method for fast unsupervised learning, really lays the theoretical ground to a systematic cost and storage reduction approach to a host of learning algorithms.

On the downside though, following up on Remark 3, the effective software libraries for sparse matrix operations (which heavily rely on block-sparsity) are far from optimal when compared to efficient dense matrix operations, and thus demand a profound treatment to ensure that our claimed computational cost improvements are truly met in practice. This is not a negligible aspect of the puncturing framework which we shall investigate in greater depth in the future.

Another critical aspect lies in the request that $\varepsilon_B, \varepsilon_S = O(1)$ with respect to p, n, thereby *not allowing for truly sparse* K. For more severe puncturing, random matrix theory fails to provide accurate predictions and, worse, the optimal phase transition threshold is no longer met by clustering from K but from more elaborate matrices (such as proposed by statistical physicists (Krzakala et al., 2013; Dall’Amico et al., 2019)). Pushing towards sparser models therefore demands a dramatic change of theoretical standpoint.

Acknowledgment

Couillet’s work is supported by the ANR-MIAI LargeDATA chair at University Grenoble-Alpes (ANR-19-P3IA-0003), and the HUAWEI-GIPSA LarDist project.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Bottou, L. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Cai, T. T., Ma, Z., Wu, Y., et al. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6): 3074–3110, 2013.

Dall'Amico, L., Couillet, R., and Tremblay, N. Revisiting the bethe-hessian: improved community detection in sparse heterogeneous graphs. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*, pp. 4039–4049, 2019.

Deshpande, Y. and Montanari, A. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pp. 2197–2201. IEEE, 2014.

Engel, Y., Mannor, S., and Meir, R. The kernel recursive least-squares algorithm. *IEEE Transactions on signal processing*, 52(8):2275–2285, 2004.

Freund, Y., Dasgupta, S., Kabra, M., and Verma, N. Learning the structure of manifolds using random projections. In *NIPS*, volume 7, pp. 59. Citeseer, 2007.

Huang, H. Asymptotic behavior of support vector machine for spiked population model. *The Journal of Machine Learning Research*, 18(1):1472–1492, 2017.

Johnstone, I. M. and Lu, A. Y. Sparse principal components analysis. *arXiv preprint arXiv:0901.4392*, 2009.

Kadavankandy, A. and Couillet, R. Asymptotic gaussian fluctuations of spectral clustering eigenvectors. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 694–698. IEEE, 2019.

Kammoun, A. and Alouini, M.-S. On the precise error analysis of support vector machines. *arXiv preprint arXiv:2003.12972*, 2020.

Keriven, N., Bourrier, A., Gribonval, R., and Pérez, P. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7(3): 447–508, 2018.

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

Liao, Z. and Couillet, R. The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, pp. 3072–3081. PMLR, 2018a.

Liao, Z. and Couillet, R. On the spectrum of random features maps of high dimensional data. In *International Conference on Machine Learning*, pp. 3063–3071. PMLR, 2018b.

Mai, X. and Couillet, R. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *The Journal of Machine Learning Research*, 19(1):3074–3100, 2018.

Marčenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

Murtagh, F. and Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1): 86–97, 2012.

Pastur, L. A. and Shcherbina, M. *Eigenvalue distribution of large random matrices*. Number 171. American Mathematical Soc., 2011.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124005, 2019.

Tiomoko, M., Couillet, R., and Tiomoko, H. Large dimensional analysis and improvement of multi task learning. *arXiv preprint arXiv:2009.01591*, 2020.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Wigner, E. P. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pp. 325–327, 1958.

Zarrouk, T., Couillet, R., Chatelain, F., and Le Bihan, N. Performance-complexity trade-off in large dimensional statistics. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2020.

Zhong, X., Su, C., and Fan, Z. Empirical bayes pca in high dimensions. *arXiv preprint arXiv:2012.11676*, 2020.