Random matrix asymptotics of inner product kernel spectral clustering

Hafiz Tiomoko Ali, Abla Kammoun, Romain Couillet

RANDOM MATRIX ASYMPTOTICS OF INNER PRODUCT KERNEL SPECTRAL CLUSTERING

*Hafiz Tiomoko Ali**, *Abla Kammoun*†, *Romain Couillet**

*CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
†King Abdullah University of Science and Technology, Saudia Arabia.

ABSTRACT

We study in this article the asymptotic performance of spectral clustering with inner product kernel for Gaussian mixture models of high dimension with numerous samples. As is now classical in large dimensional spectral analysis, we establish a phase transition phenomenon by which a minimum distance between the class means and covariances is required for clustering to be possible from the dominant eigenvectors. Beyond this phase transition, we evaluate the asymptotic content of the dominant eigenvectors thus allowing for a full characterization of clustering performance. However, a surprising finding is that in some particular scenarios, the phase transition does not occur and clustering can be achieved irrespective of the class means and covariances. This is evidenced here in the case of the mixture of two Gaussian datasets having the same means and arbitrary difference between covariances.

Index Terms— Spectral clustering, inner product kernels, random matrices, random matrix theory.

I. INTRODUCTION

One of the most important tasks in unsupervised machine learning is clustering where a set of objects is grouped in similarity classes [1]. Clustering is mainly performed using a (weighted or unweighted) graph describing the similarities between these objects. When the graph nodes are themselves the objects of interest, the problem is known as community detection on graphs [2]; otherwise the construction of the graph adjacency matrix $\mathbf{K}$ is based on a kernel operator f and the similarity between items $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by $K_{ij} = f(\mathbf{x}_i, \mathbf{x}_j)$, often taken under the form $K_{ij} = f\left(\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ or $K_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ for some function f [3]. One of the prominent methods for clustering from $\mathbf{K}$, known as spectral procedures [4] consists in performing a Principal Component Analysis (PCA) on the dominant eigenvectors (presumably containing all the useful information about the data) of the symmetric normalized Laplacian matrix $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$ (with $\mathbf{D}$ the degree matrix)[1].

In the era of big data, important amounts of data have to be processed, the dimensions of which usually scale with their number. Consistency of spectral clustering of numerous data but with finite fixed dimension has been shown in [5]. However, the behavior of spectral methods in the simultaneously high dimensional-numerous data regime can be strikingly different from the low dimensional-numerous data scenario. In particular, it was shown in [6] through a deeper study of the Laplacian matrix $\mathbf{L}$ with the Euclidean norm based kernel similarity $K_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, that in the aforementioned high dimensional regime, the performances of spectral clustering for Gaussian mixture vectors only depend on a local behavior of the kernel function as a result of a concentration of measure effect of the kernel matrix entries. This in turn changes the classical viewpoint on optimal choices of kernel functions.

We study in this article spectral clustering with the other common kernel similarity $K_{ij} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$ which induces simpler and more interpretable insights than the previously studied kernel similarity $K_{ij} = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. The behavior of the eigenvalues and eigenvectors of data driven kernel matrices being inaccessible, we consider as in [6] a data model composed of Gaussian mixtures. As shown in [6], this is not an undesirable model since an extremely close fit in performances is obtained for real datasets (in [6] with the MNIST database) when compared to Gaussian mixture inputs generated using the same empirical means and covariances as the real data. As in [6], we exhibit a phase transition below which spectral clustering is not better than random guess and beyond which non trivial performances can be obtained. For a Gaussian mixture model, this is to say that a minimum distance between means and covariances is required to obtain non vanishing correct classification rates. However, a surprising finding of our study is that in some specific scenarios, under a carefully chosen kernel function, this phase transition is always reachable in the sense that it is possible to recover the data classes from the dominant eigenvectors even for arbitrary small differences between clusters means and covariances.

Notation: Vectors are denoted with lowercase boldface letters and matrices by boldbace uppercase letters. The norm $\|.\|$ stands for the Euclidean norm for vectors and the operator norm for matrices. The vector $\mathbf{1}_n \in \mathbb{R}^n$ stands for the vector filled with ones. The Dirac mass is $\boldsymbol{\delta}_x$.

II. MODEL AND MAIN RESULTS

Consider n independent data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ belonging to a mixture of k Gaussian distributions $\mathcal{C}_1, \ldots, \mathcal{C}_k$ such that for $\mathbf{x}_i \in \mathcal{C}_a$, $\mathbf{x}_i = \boldsymbol{\mu}_a + \sqrt{p}\mathbf{w}_i$, for some $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and $\mathbf{w}_i \sim \mathcal{N}(0, p^{-1}\mathbf{C}_a)$, with $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ non negative definite. We assume without loss of generality that the vectors are ordered by classes i.e., $\mathbf{x}_{n_1+\cdots+n_{a-1}+1}, \ldots, \mathbf{x}_{n_1+\cdots+n_a} \in \mathcal{C}_a$ for $a = 1, \ldots, k$.

We assume that both p and n grow large at the same rate. Assuming that the data are well separated, i.e. the differences in means $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|$ and the differences in covariances $\|\mathbf{C}_a - \mathbf{C}_b\|$ between clusters is sufficiently large, any classical spectral clustering with a basic kernel function should be capable of separating the data asymptotically without error. It is thus interesting to understand the appropriate kernel choices capable of separating the data in more challenging scenarios. To this end, we consider the following assumptions for which clustering is not asymptotically trivial.

Assumption 1 (Growth rate)**.** *As* $n \to \infty$, $p/n \to c_0 > 0$, $\frac{n_a}{n} \to c_a > 0$. *Furthermore,*

1) *For* $\boldsymbol{\mu}^\circ = \sum_{a=1}^k c_a\boldsymbol{\mu}_a$ *and* $\boldsymbol{\mu}_a^\circ = \boldsymbol{\mu}_a - \boldsymbol{\mu}^\circ$, $\|\boldsymbol{\mu}_a^\circ\| = \mathcal{O}(1)$.
2) *For* $\mathbf{C}^\circ = \sum_{a=1}^k c_a\mathbf{C}_a$ *and* $\mathbf{C}_a^\circ = \mathbf{C}_a - \mathbf{C}^\circ$, $\|\mathbf{C}_a\| = \mathcal{O}(1)$ *and* $\operatorname{tr}\mathbf{C}_a^\circ = \mathcal{O}(\sqrt{p})$.

The work of R. Couillet and H. Tiomoko Ali is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

[1]The symmetric normalized Laplacian is usually defined in the literature as $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$ but using the eigenvectors corresponding to smallest eigenvalues of $\mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$ is equivalent to using the eigenvectors corresponding to the largest eigenvalues of $\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$.

3) $\frac{1}{p}\operatorname{tr}\mathbf{C}^\circ$ *converges to* $\tau > 0$.

For subsequent use, we introduce the following notations

$$\begin{aligned}\mathbf{M} &\triangleq [\boldsymbol{\mu}_1^\circ, \cdots, \boldsymbol{\mu}_k^\circ] \in \mathbb{R}^{p\times k}\\ \mathbf{T} &\triangleq \left\{\frac{1}{p}\operatorname{tr}\mathbf{C}_a^\circ\mathbf{C}_b^\circ\right\}_{a,b=1}^k\\ \mathbf{W} &\triangleq [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{p\times n}\\ \mathbf{J} &\triangleq [\mathbf{j}_1, \cdots, \mathbf{j}_k] \in \mathbb{R}^{n\times k}\\ \mathbf{P} &\triangleq \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \in \mathbb{R}^{n\times n}\end{aligned}$$

with $\mathbf{j}_a \in \mathbb{R}^n$ the canonical vector of cluster $\mathcal{C}_a$ defined by $(j_a)_i = \boldsymbol{\delta}_{\mathbf{x}_i\in\mathcal{C}_a}$.

We shall work with the inner product kernel similarity matrix defined as

$$\mathbf{K} \triangleq \left\{f\left(\frac{(\mathbf{x}_i^\circ)^\mathsf{T}\mathbf{x}_j^\circ}{p}\right)\right\}_{i,j=1}^n$$

with $\mathbf{x}_i^\circ = \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and f statisfying the following conditions. Note that, here $\mathbf{K}$ takes a more general form i.e., it might not necessarily be positive semi-definite.

Assumption 2 (On the kernel function)**.** *The kernel function f is three-times continuously differentiable in a neighborhood of 0 with $f(0) > 0$ and $f'(0) \neq 0$.*

Following the popular Ng-Weiss-Jordan method [7], our objective is to precisely characterize the eigenvalues and eigenvectors of the Laplacian matrix $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$ with $\mathbf{D} = \operatorname{diag}(\mathbf{K}\mathbf{1}_n)$ in order to get insights on the performances of the clustering problem. As the matrix $\mathbf{L}$ has highly dependent entries, we use the technique in [6] to find an equivalent random matrix for which the spectral analysis is more accessible. Under the model and assumptions defined above, we notice that all the off diagonal elements of $\mathbf{K}$ converge asymptotically to $f(0)$ and the diagonal elements to $f(\tau)$. This means that at the first order, $\mathbf{K}$ is a rank one matrix not containing any class information which may suggest that spectral clustering will not perform better than random guess. However, pushing to next orders by Taylor expanding the individual elements around their limiting points allows to recover the class means and covariances information. This is the main motivation behind Assumption 2 allowing to expand the function f.

The vector $\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n$ is a trivial eigenvector of $\mathbf{L}$ associated with the eigenvalue 1 and can be shown not to contain any information about the classes. We shall thus remove its eigenspace from $\mathbf{L}$ to study the other eigenvalues which are unknown so far. We thus study instead the matrix

$$\mathbf{L}' = n\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} - \frac{\mathbf{D}^{\frac{1}{2}}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{\mathbf{1}_n^\mathsf{T}\mathbf{D}\mathbf{1}_n}\right)$$

which has the same eigenvalues and eigenvectors as $\mathbf{L}$ but for the pair $(1, \mathbf{D}^{\frac{1}{2}}\mathbf{1}_n)$ eigenpair of $\mathbf{L}$ which becomes $(0, \mathbf{D}^{\frac{1}{2}}\mathbf{1}_n)$ for $\mathbf{L}'$. The matrix $\mathbf{L}'$ having dependent entries, we proceed to the Taylor expansion of the individual entries to get a Taylor approximation of the whole matrix by controlling the different matrix norms by orders of magnitude and only keeping non-vanishing terms. All calculus made, we obtain the following equivalent approximation of the normalized Laplacian matrix.

Theorem 1. *Let Assumption 1 and 2 hold true. Let $\widehat{\mathbf{L}}'$ be given by:*

$$\widehat{\mathbf{L}}' = \frac{f'(0)}{f(0)}\left(\mathbf{P}\mathbf{W}^T\mathbf{W}\mathbf{P} + \mathbf{V}\mathbf{B}\mathbf{V}^T\right) + F(\tau)\mathbf{P}$$

where

$$\begin{aligned}\mathbf{B} &= \begin{bmatrix}\mathbf{M}^T\mathbf{M} + \frac{f''(0)}{2f'(0)}\mathbf{T} & \mathbf{I}_k\\ \mathbf{I}_k & 0\end{bmatrix}\\ \mathbf{V} &= \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{P}\mathbf{W}^T\mathbf{M}\right]\end{aligned}$$

and $F(\tau) = \frac{f(\tau)-f(0)-\tau f'(0)}{f(0)}$. Then,

$$\left\|\mathbf{L}' - \widehat{\mathbf{L}}'\right\| \xrightarrow{\text{a.s.}} 0.$$

As a direct consequence of Theorem 1, the eigenvalues of $\mathbf{L}'$ are linearly mapped to the eigenvalues of $\mathcal{L} = \frac{f(0)}{f'(0)}\mathbf{L}' + F(\tau) = \mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P} + \mathbf{V}\mathbf{B}\mathbf{V}^T$ with the same eigenvectors. As far as spectral clustering is concerned, we can therefore focus in the sequel on $\mathcal{L}$ instead of $\mathbf{L}'$. Interestingly, $\mathcal{L}$ is equivalent to a spiked random matrix of an information ($\mathbf{V}\mathbf{B}\mathbf{V}^T$) plus noise ($\mathbf{P}\mathbf{W}^\mathsf{T}\mathbf{W}\mathbf{P}$) type [8]. Classical spike random matrix analysis [8] suggest that whenever the noise energy dominates the information energy, no information can be retrieved, but there exists a phase transition (point where the information becomes predominant) beyond which the information can be retrieved with success rate better than random guess. We see in particular that when $f'(0) = 0$, $\mathcal{L}$ is essentially the low rank matrix $\mathbf{J}\mathbf{T}\mathbf{J}^\mathsf{T}$ containing only information about the clusters covariances meaning that we cannot discriminate the data upon their means. To avoid the latter limitation, we focus rather on the case $f'(0) \neq 0$ thus motivating Assumption 2. From a spectral theory point of view, the eigenvalues of spiked random matrices concentrate in bulks (representing the noise) but for a few isolated ones (which can be at either side of the bulks) and the eigenvectors associated with those isolated eigenvalues are correlated to the eigenspace of the information matrix as long as the energy in the latter is sufficiently large. A precise look at Theorem 1 allows us to expect that the eigenvectors associated with isolated eigenvalues will be correlated to $\mathbf{J}$ (containing the cluster indicator vectors) all the more that there is sufficient energy in the matrices $\mathbf{M}$ and $\mathbf{T}$ (containing statistical information about the classes). From a practical point of view, beyond this phase transition, a spectral clustering algorithm using the isolated eigenvectors should be able to recover the classes with performances better than chance. The following result provides a precise characterization of the isolated eigenvalues of the normalized Laplacian matrix associated to an inner product kernel matrix.

Theorem 2 (Isolated eigenvalues)**.** *Let Assumption 1 and 2 hold true. For $z \in \mathbb{C}$, define the $k\times k$ matrix $\mathbf{G}_z$ as:*

$$\mathbf{G}_z = \left[\frac{f''(0)}{2f'(0)}\mathbf{T} + \mathbf{M}^T\left(\mathbf{I}_p + \sum_{a=1}^k c_a g_a(z)\mathbf{C}_a\right)^{-1}\mathbf{M}\right]\boldsymbol{\Gamma}_z + \mathbf{I}_k$$

where

$$\boldsymbol{\Gamma}_z = \operatorname{diag}\left\{c_a g_a(z)\right\}_{a=1}^k - \left\{\frac{c_a g_a(z) c_b g_b(z)}{\sum_{i=1}^k c_i g_i(z)}\right\}_{a,b=1}^k$$

and $g_1(z), \ldots, g_k(z)$ are the unique solutions with $Im[g_i(z)] > 0$ when $Im[z] > 0$, to the system

$$\frac{1}{c_0 g_a(z)} = -z + \frac{1}{p}\operatorname{tr}\mathbf{C}_a\left(\mathbf{I}_p + \sum_{a=1}^k c_a g_a(z)\mathbf{C}_a\right)^{-1}.$$

Let ρ be at a macroscopic distance from the eigenvalue support of $\mathbf{P}\mathbf{W}^T\mathbf{W}\mathbf{P}$ and be such that $\mathbf{G}_\rho$ has a zero eigenvalue with multiplicity m_ρ. Then, there exist $\lambda_j \geq \cdots \geq \lambda_{j+m_\rho-1}$ eigenvalues of $\mathcal{L}$ such that:

$$|\lambda_{j+i} - \rho| \xrightarrow{\text{a.s.}} 0.$$

The result in Theorem 2 is difficult to interpret since the functions $g_a(z)$ are defined through an implicit equation. To get more insights on the different positions of the isolated eigenvalues and on the performances of kernel spectral clustering in the regime under study, we consider in the next section a simple case where the expressions are more amenable to interpretation.

III. SPECIAL CASE

We study in this section scenarios of practical interest from which new insights about kernel spectral clustering are derived. We assume that the data vectors are drawn from a mixture of two well balanced Gaussian datasets (i.e., $c_1 = c_2$) with means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ respectively and random positive definite covariances $\mathbf{C}_1$, $\mathbf{C}_2$ identically distributed, unitarily invariant and hence asymptotically free (see [9] for discussion and asymptotic freeness) such that the empirical distribution of their eigenvalues converge to a common law ν. This is the case for instance when $\mathbf{C}_1$ and $\mathbf{C}_2$ are two random independent and identically distributed Wishart matrices. Under this setting, $g_1(z) = g_2(z) = g(z)$ with

$$g(z) = \left(-zc + \frac{c}{2} \int \frac{t}{1+\frac{1}{2}g(z)t} \nu \boxplus \nu(\mathrm{d}t) \right)^{-1}$$

the Stietjes transform [10] of a probability measure with compact support $\mathcal{S}$ and $\nu \boxplus \nu$ is the additive free convolution of ν with itself.

Focusing on the location of isolated eigenvalues (hereafter called spikes), the limiting spikes ρ satisfy from Theorem 2 $\det(\mathbf{G}_\rho) = 0$ with

$$\mathbf{G}_\rho = \mathbf{I}_2 + g(\rho) \left(\frac{f''(0)}{8f'(0)} \mathbf{T} + \frac{\mathbf{M}^T\mathbf{M}}{4} \int \frac{[\nu \boxplus \nu(\mathrm{d}t)]}{1+\frac{1}{2}g(\rho)t} \right) \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

This expression follows from using the fact that the $\mathbf{C}_a$'s are drawn from unitarily invariant distributions so that $\mathbf{M}^T \left(\mathbf{I}_p + \frac{g(\rho)}{2} \sum_{a=1}^k \mathbf{C}_a \right)^{-1} \mathbf{M} - \int \frac{1}{1+\frac{1}{2}g(\rho)t} \nu \boxplus \nu(\mathrm{d}t) \mathbf{M}^T\mathbf{M} \xrightarrow{\text{a.s.}} 0$ from the trace Lemma (see e.g., Theorem 3.12 in [11]). The limiting spikes are thus the ρ's for which the rank one matrix $\mathbf{G}_\rho - I_2$ has eigenvalue -1. This is equivalent to saying that the non zero eigenvalue $\mathrm{tr}(\mathbf{G}_\rho - I_2)$ of the latter rank-one matrix is exactly -1. After calculations, we thus get that the limiting isolated eigenvalues ρ should satisfy

$$\theta g(\rho) + \delta m_{\nu \boxplus \nu}(-\frac{2}{g(\rho)}) + 1 = 0 \tag{1}$$

where

$$\theta = \frac{f''(0)}{8f'(0)} \frac{1}{p} \mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$$
$$\delta = \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2}{2}$$

and $m_{\nu \boxplus \nu}$ is the Stieltjes transform of $\nu \boxplus \nu$.

Several interesting insights can be readily extracted from (1). Let us first consider separately the limiting cases of $\delta = 0$ (equal means across classes) and $\mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = 0$ (equal covariances across classes). When $\delta = 0$, it follows from (1) that a spike appears at location ρ outside the bulk once θ is set to $-\frac{1}{g(\rho)}$. From this we deduce that, quite surprisingly, any location ρ outside of the eigenvalues bulk can asymptotically be the value of some spike. This can be made possible by choosing the kernel function f so that $\theta = -\frac{1}{g(\rho)}$. Since we are placing ourselves in the case where class means are equal, such a finding goes against the usually encountered phase transition phenomenon by which we expect that a minimum distance between the class means and covariances is required to allow isolated eigenvalues showing up. This, however, should only occur theoretically when p and n are quite large so that a good match is obtained between the inner-product kernel Laplacian matrix and its random equivalent; simulations for finite not too large p and n suggest otherwise. On the other hand, if $\mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 = 0$ suggesting equal covariances across classes, it appears that the choice of the kernel has asymptotically no impact on clustering performances, as it is asymptotically inconsequential to the spikes location or their appearance. The relevant parameter for clustering tasks is the distance between the means, captured by the parameter δ. From (1), it entails that clustering is asymptotically possible only when $-\frac{1}{\delta}$ belongs to the set $\left\{ m_{\nu \boxplus \nu}(-\frac{2}{g(\rho)}), \rho \notin \mathcal{S} \right\}$.

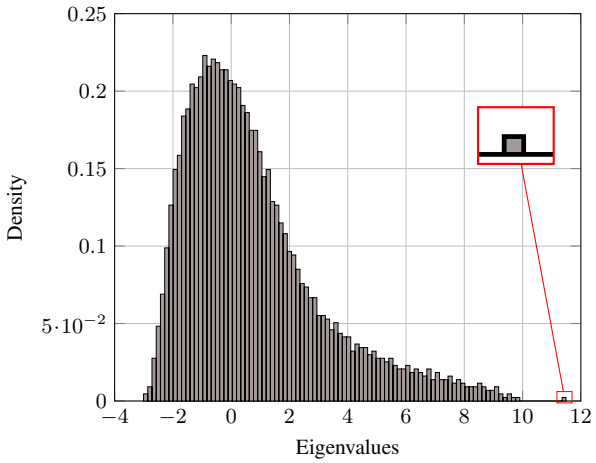

Fig. 1: Histogram of the eigenvalues of $\mathcal{L}$ with $p = 1000$, $n = 3000$, $k = 2$, $c_1 = c_2$, $\mathbf{C}_1, \mathbf{C}_2$ i.i.d from Wishart distribution with parameters $(\frac{1}{p}\mathbf{I}_p, p, 2p)$, $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 = 0.5$, $f = 1 + 2x + 20x^2$: A right-hand side spike appears when θ is largely positive.

Having studied these two limiting cases, let us now focus on the general case $\delta \neq 0$ and $\mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 \neq 0$. Equation (1) shows that a spike at location ρ should satisfy:

$$m_{\nu \boxplus \nu}\left(-\frac{2}{g(\rho)} \right) = -\frac{1}{\delta} - \frac{\theta g(\rho)}{\delta} \tag{2}$$

which can be geometrically interpreted as the intersection between the graph of $g \mapsto m_{\nu \boxplus \nu}(-\frac{2}{g})$ with the line $g \mapsto -\frac{1}{\delta} - \frac{\theta g}{\delta}$ where g belongs to $\{g(\rho), \quad \rho \notin \mathcal{S}\}$. It is noteworthy to mention that θ can take any value in $\mathbb{R}$ by acting upon the kernel function. Hence, for any ρ outside the support $\mathcal{S}$, it is always possible to tune θ so that a spike at location ρ appears. This is a major piece of information that can be used in practice to favor the appearance of the most-informative spikes for clustering tasks.

In the same vein, our analysis shows that contrary to what practitioners are used to assuming, a spike can appear either on the left or on the right-hand sides of the main bulk. This observation becomes all the more interesting that in many realistic situations only left-hand side spikes carrying clustering information appear. In such circumstances, algorithms that rely on the largest eigenvalues of kernel Laplacian matrices ($\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$) would yield very low performances. Let us further investigate the role of θ on that aspect. Since $\rho \to g(\rho)$ is a Stieltjes transform, $g(\rho) \to 0^+$ as $\rho \to -\infty$ and $g(\rho) \to 0^-$ at $\rho \to +\infty$ and investigating (2), we can expect large negative θ to be consistent with the appearance of left-hand side spikes while large positive θ should allow the appearance of right-hand side spikes. To validate this statement, we represent in Figure 1 and Figure 2 the histograms of the eigenvalues of the kernel random matrix $\mathcal{L}$ when $\mathbf{C}_1$ and $\mathbf{C}_2$ follow Wishart distribution $W_p(\frac{1}{2p}\mathbf{I}_p, p, 2p)$ where $p = 1000$, $n = 3000$, $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 = 0.5$ with the polynomial kernel functions $f(x) = 1 + 2x + 20x^2$ and $f(x) = 1 + 2x - 15x^2$. $\mathbf{C}_1$ and $\mathbf{C}_2$ are thus independent, identically distributed and unitarily invariant, hence asymptotically free.

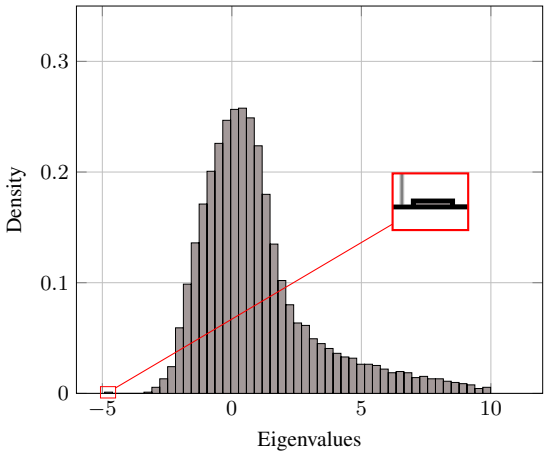

Fig. 2: Histogram of the eigenvalues of $\mathcal{L}$. Same setting as Figure 1 but for $f = 1+2x-15x^2$: A left-hand side spike appears when θ is largely negative.

When a spike appears at location ρ, the associated eigenvector $\hat{\mathbf{u}}_\rho$ will look like noisy step functions. By using statistic interchangeability within the classes, we can write

$$\hat{\mathbf{u}} = \alpha_1 \frac{\mathbf{j}_1}{\sqrt{n_1}} + \alpha_2 \frac{\mathbf{j}_2}{\sqrt{n_2}} + \sigma_1 \boldsymbol{\omega}_1 + \sigma_2 \boldsymbol{\omega}_2 \tag{3}$$

where $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are unit norm vectors supported respectively on the indices of class $\mathcal{C}_1$ and $\mathcal{C}_2$ and are orthogonal to respectively $\mathbf{j}_1$ and $\mathbf{j}_2$. The scalars α_1, α_2 capture the alignment of the eigenvector to the class step vectors $\mathbf{j}_1$ and $\mathbf{j}_2$ while the σ_1 and σ_2 can be seen as the class standard deviations of the eigenvector fluctuations around $\frac{\mathbf{j}_1}{\sqrt{n_1}}$ and $\frac{\mathbf{j}_2}{\sqrt{n_2}}$. Assume $\delta = 0$, and consider the eigenvector $\hat{\mathbf{u}}_\rho$ associated with a spike at location ρ outside the limiting support $\mathcal{S}$. As said, this suggests that θ is related to ρ through the relation $\theta = -\frac{1}{g(\rho)}$. The following result shows that the scalars $(\alpha_i)_{i=1}^2$ ans $(\sigma_i)_{i=1}^2$ associated with that eigenvector can be approximated as follows.

Theorem 3. *Assume $\delta = 0$. Let ρ be an isolated eigenvalue of $\mathcal{L}$ and $\hat{\mathbf{u}}_\rho$ its associated eigenvector decomposed as* (3)*. Then, for $\theta = -\frac{1}{g(\rho)}$,*

$$(\alpha_i)^2 = \frac{1}{2} - \frac{c_0}{2} \int \frac{t^2}{(2\theta - t)^2} \nu \boxplus \nu(dt) + o(1)$$

$$(\sigma_i)^2 = \frac{1}{2} \left[1 - \int \frac{c_0 t^2}{(2\theta - t)^2} \nu \boxplus \nu(dt) \right]$$

$$\times \left[\left(1 - c_0 - \int \frac{4c_0\theta}{t - 2\theta} \nu \boxplus \nu(dt) - \int \frac{4c_0\theta^2}{(t-2\theta)^2} \nu \boxplus \nu(dt) \right)^{-1} - 1 \right] + o(1).$$

We validate Theorem 3 by representing in Figure 3 the eigenvector associated with the left hand side eigenvalue of $\mathcal{L}$ when $f = 1+2x-20x^2$ and $n = 3000, p = 1000$. We note a good match between the theoretical findings with the empirical ones, showing the potential of our results in characterizing the statistical behavior of spectral clustering.

This characterization can be very useful in practice. Consider the situation in which clustering is performed based on the eigenvector $\hat{\mathbf{u}}_\rho$. Assuming Gaussian fluctuations on the individual entries of the eigenvector, we can compute using Theorem 3 the asymptotic clustering error probability, which interestingly does not depend on how close the covariance matrices are, but on the location of the spike ρ related to θ through $\theta = -\frac{1}{g(\rho)}$. As such, one can investigate whether the clustering performance could be enhanced

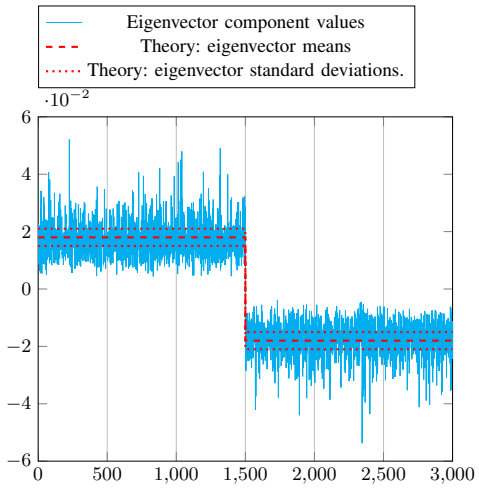

Fig. 3: Eigenvector corresponding to the leftmost eigenvalue of $\hat{L}$ in Figure 2. Index of the eigenvector (in the x-axis) are ordered by classes and the y-axis represent the values corresponding to each index.

by optimizing over the most-informative spikes' locations that are associated with low clustering error probabilities. Such a direction will be investigated in the future.

IV. CONCLUDING REMARKS

Our large dimensional spectral analysis of Laplacian kernel random matrices yields the surprising finding that one can choose in certain scenarios an appropriate kernel function allowing to get the most informative eigenvector for clustering purposes. In addition, assuming the classification of two Gaussian datasets with the same means and arbitrary random covariances independent and identically distributed, this kernel function could be chosen in such a way that it allows to get an outlying eigenvalue and the minimum achievable clustering error rate using the corresponding eigenvector. We believe that such findings can be useful in practice in particular for the choice of the eigenvector to use for classification which is not always the one associated to the largest eigenvalue. A natural extension of this work is to push forward this analysis for other cases of practical interests in order to get a comprehensive understanding of kernel spectral clustering.

V. REFERENCES

[1] M Emre Celebi and Kemal Aydin, *Unsupervised Learning Algorithms*, Springer, 2016.

[2] Santo Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[3] Rocco Langone, Raghvendra Mall, Carlos Alzate, and Johan AK Suykens, "Kernel spectral clustering and applications," in *Unsupervised Learning Algorithms*, pp. 135–161. Springer, 2016.

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

[5] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

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

[7] Andrew Y Ng, Michael I Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[8] Florent Benaych-Georges and Raj Rao Nadakuditi, "The singular values and vectors of low rank perturbations of large

rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[9] Fumio Hiai and Dénes Petz, *Asymptotic freeness almost everywhere for random matrices*, University of Aarhus. Centre for Mathematical Physics and Stochastics (MaPhySto)[MPS], 1999.

[10] Walid Hachem, Philippe Loubaton, Jamal Najim, et al., "Deterministic equivalents for certain functionals of large random matrices," *The Annals of Applied Probability*, vol. 17, no. 3, pp. 875–930, 2007.

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