Understanding Big Data Spectral Clustering

Romain Couillet, Florent Benaych-Georges

Understanding Big Data Spectral Clustering

Romain Couillet*, Florent Benaych-Georges†
*CentraleSupélec – LSS – Université ParisSud, Gif sur Yvette, France
†MAP 5, UMR CNRS 8145 – Université Paris Descartes, Paris, France.

***Abstract*—This article introduces an original approach to understand the behavior of standard kernel spectral clustering algorithms (such as the Ng–Jordan–Weiss method) for large dimensional datasets. Precisely, using advanced methods from the field of random matrix theory and assuming Gaussian data vectors, we show that the Laplacian of the kernel matrix can asymptotically be well approximated by an analytically tractable equivalent random matrix. The study of the latter unveils the mechanisms into play and in particular the impact of the choice of the kernel function and some theoretical limits of the method. Despite our Gaussian assumption, we also observe that the predicted theoretical behavior is a close match to that experienced on real datasets (taken from the MNIST database).**[1]

I. Introduction

Letting $x_1, \ldots, x_n \in \mathbb{R}^p$ be n data vectors, kernel spectral clustering consists in a variety of algorithms designed to cluster these data in an unsupervised manner by retrieving information from the leading eigenvectors of (a possibly modified version of) the so-called kernel matrix $K = \{K_{ij}\}_{i,j=1}^n$ with e.g., $K_{ij} = f(\|x_i - x_j\|/p)$ for some (usually decreasing) $f : \mathbb{R}_+ \to \mathbb{R}_+$. There are multiple reasons (see e.g., [1]) to expect that the aforementioned eigenvectors contain information about the optimal data clustering. One of the most prominent of those was put forward by Ng–Jordan–Weiss in [2] who notice that, if the data are ideally well split in k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ that ensure $f(\|x_i - x_j\|/p) = 0$ if and only if x_i and x_j belong to distinct classes, then the eigenvectors associated with the $k-1$ smallest eigenvalues of $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, $D \triangleq \mathcal{D}(K1_n)$, live in the span of the canonical class-wise basis vectors. In the non-trivial case where such a separating f does not exist, one would thus expect the leading eigenvectors to be instead perturbed versions of indicator vectors. We shall precisely study the matrix $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ in this article.

Nonetheless, despite this conspicuous argument, very little is known about the performance of kernel spectral clustering in actual working conditions. In particular, to the authors' knowledge, there exists no contribution addressing the case of arbitrary p and n. In this article, we propose a new approach consisting in assuming that both p and n are large, and exploiting recent results from random matrix theory. Our method is inspired by [3] which studies the asymptotic distribution of the eigenvalues of K for i.i.d. vectors x_i. We generalize here [3] by assuming that the x_i's are drawn from a mixture of k Gaussian vectors having means $\mu_1, \ldots, \mu_k$ and covariances $C_1, \ldots, C_k$. We then go further by studying the resulting model and showing that $L = D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ can be approximated by a matrix of the so-called spiked model type [4], [5], that is a matrix with clustered eigenvalues and a few isolated outliers. Among other results, our main findings are:

- in the large n, p regime, only a very local aspect of the kernel function really matters for clustering;
- there exists a critical growth regime (with p and n) of the μ_i's and C_i's for which spectral clustering leads to non-trivial misclustering probability;
- we precisely analyze elementary toy models, in which the number of exploitable eigenvectors and the influence of the kernel function may vary significantly.

On top of these theoretical findings, we shall observe that, quite unexpectedly, the kernel spectral algorithms behave similar to our theoretical findings on real datasets. We precisely see that clustering performed upon a subset of the MNIST (handwritten figures) database behaves as though the vectorized images were extracted from a Gaussian mixture.

Notations: The norm $\|\cdot\|$ stands for the Euclidean norm for vectors and operator norm for matrices. The vector $1_m \in \mathbb{R}^m$ stands for the vector filled with ones. The operator $\mathcal{D}(v) = \mathcal{D}(\{v_a\}_{a=1}^k)$ is the diagonal matrix having $v_1, \ldots, v_k$ (scalar or vectors) down its diagonal. The Dirac mass at x is $\boldsymbol{\delta}_x$. Almost sure convergence is denoted $\xrightarrow{\text{a.s.}}$, and convergence in distribution $\xrightarrow{\mathcal{D}}$.

II. Model and Theoretical Results

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be independent vectors with $x_{n_1+\cdots+n_{\ell-1}+1}, \ldots, x_{n_1+\cdots+n_\ell} \in \mathcal{C}_\ell$ for each $\ell \in \{1, \ldots, k\}$, where $n_0 = 0$ and $n_1 + \cdots + n_k = n$. Class $\mathcal{C}_a$ encompasses data $x_i = \mu_a + w_i$ for some $\mu_a \in \mathbb{R}^p$ and $w_i \sim \mathcal{N}(0, C_a)$, with $C_a \in \mathbb{R}^{p \times p}$ nonnegative definite.

We shall consider the large dimensional regime where both n and p grow simultaneously large. In this regime, we shall require the μ_i's and C_i's to behave in a precise manner. As a matter of fact, we may state as a first result that the following set of assumptions forms the exact regime under which spectral clustering is a non trivial problem.

Assumption 1 (Growth Rate): As $n \to \infty$, $\frac{p}{n} \to c_0 > 0$, $\frac{n_a}{n} \to c_a > 0$ (we will write $c = [c_1, \ldots, c_k]^\mathsf{T}$). Besides,

1) For $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$ and $\mu_a^\circ = \mu_a - \mu^\circ$, $\|\mu_a^\circ\| = O(1)$
2) For $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$ and $C_a^\circ = C_a - C^\circ$, $\|C_a\| = O(1)$ and $\operatorname{tr} C_a^\circ = O(\sqrt{n})$.
3) As $p \to \infty$, $\frac{2}{p} \operatorname{tr} C^\circ \to \tau > 0$.

The value τ is important since $\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$ uniformly on $i \neq j$ in $\{1, \ldots, n\}$.

[1]Couillet's work is supported by RMT4GRAPH (ANR-14-CE28-0006).

We now define the kernel function as follows.

Assumption 2 (Kernel function): Function f is three-times continuously differentiable around τ and $f(\tau) > 0$.

Having defined f, we introduce the kernel matrix as

$$K \triangleq \left\{ f\left(\frac{1}{p}\left\|x_i - x_j\right\|^2\right)\right\}_{i,j=1}^{n}.$$

From the previous remark on τ, note that all non-diagonal elements of K tend to $f(\tau)$ and thus K can be point-wise developed using a Taylor expansion. However, our interest is on (a slightly modified form of) the Laplacian matrix

$$L \triangleq nD^{-\frac{1}{2}}KD^{-\frac{1}{2}}$$

where $D = \mathcal{D}(K1_n)$ is usually referred to as the degree matrix. Under Assumption 1, L is essentially a rank-one matrix with $D^{\frac{1}{2}}1_n$ for leading eigenvector (with n for eigenvalue). To avoid technical difficulties, we shall study the equivalent matrix[2]

$$L' \triangleq nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n} \tag{1}$$

which we shall show to have all its eigenvalues of order $O(1)$.

Our main technical result shows that there is a matrix $\hat{L}'$ such that $\|L' - \hat{L}'\| \xrightarrow{\text{a.s.}} 0$, where $\hat{L}'$ follows a tractable random matrix model. Before introducing the latter, we need the following fundamental deterministic element notations[3]

$$\begin{aligned}
M &\triangleq [\mu_1^\circ, \ldots, \mu_k^\circ] \in \mathbb{R}^{p\times k}\\
t &\triangleq \left\{\frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ\right\}_{a=1}^{k} \in \mathbb{R}^k\\
T &\triangleq \left\{\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ\right\}_{a,b=1}^{k} \in \mathbb{R}^{k\times k}\\
J &\triangleq [j_1, \ldots, j_k] \in \mathbb{R}^{n\times k}\\
P &\triangleq I_n - \frac{1}{n}1_n 1_n^\mathsf{T} \in \mathbb{R}^{n\times n}
\end{aligned}$$

where $j_a \in \mathbb{R}^n$ is the canonical vector of class $\mathcal{C}_a$, defined by $(j_a)_i = \boldsymbol{\delta}_{x_i\in\mathcal{C}_a}$, and the random element notations

$$\begin{aligned}
W &\triangleq [w_1, \ldots, w_n] \in \mathbb{R}^{p\times n}\\
\Phi &\triangleq \frac{1}{\sqrt{p}}W^\mathsf{T}M \in \mathbb{R}^{n\times k}\\
\psi &\triangleq \frac{1}{p}\left\{\|w_i\|^2 - \mathrm{E}[\|w_i\|^2]\right\}_{i=1}^{n} \in \mathbb{R}^n.
\end{aligned}$$

Theorem 1 (Random Matrix Equivalent): Let Assumptions 1 and 2 hold and L' be defined by (1). Then, as $n \to \infty$,

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

where $\hat{L}'$ is given by

$$\hat{L}' \triangleq \frac{-2f'(\tau)}{f(\tau)}\left[\frac{PW^\mathsf{T}WP}{p} + UBU^\mathsf{T}\right] + \frac{2f'(\tau)}{f(\tau)}F(\tau)I_n$$

with $F(\tau) = \frac{f(0)-f(\tau)+\tau f'(\tau)}{2f'(\tau)}$ and

$$\begin{aligned}
U &\triangleq \left[\frac{1}{\sqrt{p}}J, \Phi, \psi\right]\\
B &\triangleq \begin{bmatrix} B_{11} & I_k - 1_k c^\mathsf{T} & \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t\\ I_k - c1_k^\mathsf{T} & 0_{k\times k} & 0_{k\times 1}\\ \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)t^\mathsf{T} & 0_{1\times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\end{bmatrix}\\
B_{11} &= M^\mathsf{T}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)}T + \frac{p}{n}F(\tau)1_k 1_k^\mathsf{T}
\end{aligned}$$

and the case $f'(\tau) = 0$ is obtained by extension by continuity (in the limit $f'(\tau)B$ being well defined as $f'(\tau) \to 0$).

From a mathematical standpoint, excluding the identity matrix, when $f'(\tau) \neq 0$, $\hat{L}'$ follows a spiked random matrix model, that is its eigenvalues congregate in bulks but for a few isolated eigenvalues, the eigenvectors of which align to some extent to the eigenvectors of UBU^T. When $f'(\tau) = 0$, $\hat{L}'$ is merely a small rank matrix. In both cases, the isolated eigenvalue-eigenvector pairs of $\hat{L}'$ are amenable to analysis.

From a practical aspect, note that U is notably constituted by the vectors j_a, while B contains the information about the inter-class mean deviations through M, and about the inter-class covariance deviations through t and T. As such, the aforementioned isolated eigenvalue-eigenvector pairs are expected to correlate to the canonical class basis J and all the more so that M, t, T have sufficiently strong norm.

From the point of view of the kernel function f, note that, if $f'(\tau) = 0$, then M vanishes from the expression of $\hat{L}'$, thus not allowing spectral clustering to rely on differences in means. Similarly, if $f''(\tau) = 0$, then T vanishes, and thus differences in "shape" between the covariance matrices cannot be discriminated upon. Finally, if $\frac{5f'(\tau)}{8f(\tau)} = \frac{f''(\tau)}{2f'(\tau)}$, then differences in covariance traces are seemingly not exploitable.

Before introducing our main results, we need the following technical assumption which ensures that $\frac{1}{p}PW^\mathsf{T}WP$ does not in general produce itself isolated eigenvalues (and thus, that the isolated eigenvalues of $\hat{L}'$ are solely due to UBU^T).

Assumption 3 (Spike control): With $\lambda_1(C_a) \geq \ldots \geq \lambda_p(C_a)$ the eigenvalues of C_a, for each a, as $n \to \infty$, $\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{\delta}_{\lambda_i(C_a)} \xrightarrow{\mathcal{D}} \nu_a$, with support $\operatorname{supp}(\nu_a)$, and

$$\max_{1\leq i\leq p}\operatorname{dist}(\lambda_i(C_a), \operatorname{supp}(\nu_a)) \to 0.$$

Theorem 2 (Isolated eigenvalues[4]): Let Assumptions 1–3 hold and define, for $z \in \mathbb{R}$, the $k \times k$ matrix

$$G_z = h(\tau, z)I_k + D_{\tau,z}\Gamma_z$$

[2]It is clearly equivalent to study L' or L that have the same eigenvalue-eigenvector pairs but for the pair $(n, D^{\frac{1}{2}}1_n)$ of L turned into $(0, D^{\frac{1}{2}}1_n)$ for L'.

[3]Capital M stands here for *means* while t, T account for vector and matrix of *traces*, P for a projection matrix (onto the orthogonal of $1_n 1_n^\mathsf{T}$).

[4]Again here, the case $f'(\tau) = 0$ is obtained by extension by continuity.

where

$$h(\tau,z) = 1 + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)\sum_{i=1}^k c_i g_i(z)\frac{2}{p}\operatorname{tr} C_i^2$$

$$D_{\tau,z} = h(\tau,z)M^\mathsf{T}\left[I_p + \sum_{j=1}^k c_j g_j(z) C_j\right]^{-1} M - h(\tau,z)\frac{f''(\tau)}{f'(\tau)}T + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^\mathsf{T}$$

$$\Gamma_z = \mathcal{D}\left\{c_a g_a(z)\right\}_{a=1}^k - \left\{\frac{c_a g_a(z) c_b g_b(z)}{\sum_{i=1}^k c_i g_i(z)}\right\}_{a,b=1}^k$$

and $g_1(z),\ldots,g_k(z)$ are, for well chosen z, the unique solutions to the system

$$\frac{1}{c_0}\frac{1}{g_a(z)} = -z + \frac{1}{p}\operatorname{tr} C_a\left(I_p + \sum_{i=1}^k c_i g_i(z) C_i\right)^{-1}.$$

Let ρ, away from the eigenvalue support of $\frac{1}{p}PW^\mathsf{T}WP$, be such that $h(\tau,\rho)\neq 0$ and G_ρ has a zero eigenvalue of multiplicity m_ρ. Then there exists m_ρ eigenvalues of L asymptotically close to

$$-2\frac{f'(\tau)}{f(\tau)}\rho + \frac{f(0)-f(\tau)+\tau f'(\tau)}{f(\tau)}.$$

We now turn to the more interesting result concerning the eigenvectors. This result is divided in two formulas, concerning (i) the eigenvector $D^{\frac{1}{2}}1_n$ associated with the eigenvalue n of L and (ii) the remaining eigenvectors associated with the eigenvalues exhibited in Theorem 2.

Proposition 1 (Eigenvector $D^{\frac{1}{2}}1_n$): Let Assumptions 1–2 hold true. Then, for some $\varphi\sim\mathcal{N}(0,I_n)$, almost surely,

$$\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\mathsf{T} D 1_n}} = \frac{1_n}{\sqrt{n}} + \frac{1}{n\sqrt{c_0}}\left[\frac{f'(\tau)}{2f(\tau)}\left\{t_a 1_{n_a}\right\}_{a=1}^k + \mathcal{D}\left\{\sqrt{\frac{2}{p}\operatorname{tr}(C_a^2)}1_{n_a}\right\}_{a=1}^k\varphi + o(1)\right].$$

Theorem 3 (Eigenvector projections): Let Assumptions 1–3 hold. Let also $\lambda_j^p,\ldots,\lambda_{j+m_\rho-1}^p$ be isolated eigenvalues of L all converging to ρ as per Theorem 2 and Π_ρ the projector on the eigenspace associated to these eigenvalues. Then,

$$\frac{1}{p}J^\mathsf{T}\Pi_\rho J = -\Gamma(\rho)\sum_{i=1}^{m_\rho}\frac{h(\tau,\rho)(V_{r,\rho})_i(V_{l,\rho})_i^\mathsf{T}}{(V_{l,\rho})_i^\mathsf{T}G_\rho'(V_{r,\rho})_i} + o(1)$$

almost surely, where $V_{r,\rho},V_{l,\rho}\in\mathbb{C}^{k\times m_\rho}$ are sets of right and left eigenvectors of G_ρ associated with the eigenvalue zero, and G_ρ' is the derivative of G_z along z taken for $z=\rho$.

From Proposition 1, we get that $D^{\frac{1}{2}}1_n$ is centered around the sum of the class-wise vectors $t_a j_a$ with fluctuations of amplitude $\frac{2}{p}\operatorname{tr}(C_a^2)$. As for Theorem 3, it states that, as p,n grow large, the alignment between the isolated eigenvectors of L and the canonical class-basis $j_1,\ldots,j_k$ tends to be deterministic in a theoretically tractable manner. In particular, the quantity

$$\operatorname{tr}\left(\frac{1}{n}\mathcal{D}(c^{-\frac{1}{2}})J^\mathsf{T}\Pi_\rho J\mathcal{D}(c^{-\frac{1}{2}})\right)\in[0,m_\lambda]$$

evaluates the alignment between Π_ρ and the canonical class basis, thus providing a first hint on the expected performance of spectral clustering. A second interest of Theorem 3 is that, for eigenvectors $\hat{u}$ of L of multiplicity one (so $\Pi_\rho=\hat{u}\hat{u}^\mathsf{T}$), the diagonal elements of $\frac{1}{n}\mathcal{D}(c^{-\frac{1}{2}})J^\mathsf{T}\Pi_\rho J\mathcal{D}(c^{-\frac{1}{2}})$ provide the squared mean values of the successive first j_1, then next j_2, etc., elements of $\hat{u}$. The off-diagonal elements of $\frac{1}{n}\mathcal{D}(c^{-\frac{1}{2}})J^\mathsf{T}\Pi_\rho J\mathcal{D}(c^{-\frac{1}{2}})$ then allow to decide on the signs of $\hat{u}^\mathsf{T}j_i$ for each i. These pieces of information are crucial to estimate the expected performance of spectral clustering.

However, the statements of Theorems 2 and 3 are difficult to interpret as they stand. These become more explicit when applied to simpler scenarios and allow one to draw interesting conclusions. This is the target of the next section.

III. Special cases

In this section, we apply Theorems 2 and 3 to the cases where: (i) $C_i=\beta I_p$ for all i, with $\beta>0$, (ii) all μ_i's are equal and $C_i=(1+\frac{\gamma_i}{\sqrt{p}})\beta I_p$.

Assume first that $C_i=\beta I_p$ for all i. Then, letting ℓ be an isolated eigenvalue of $\beta I_p + M\mathcal{D}(c)M^\mathsf{T}$, we get that, if

$$|\ell-\beta|>\beta\sqrt{c_0} \tag{2}$$

then the matrix L has an eigenvalue (asymptotically) equal to

$$-\frac{2f'(\tau)}{f(\tau)}\left(\frac{\ell}{c_0}+\beta\frac{\ell}{\ell-\beta}\right)+\frac{f(0)-f(\tau)+\tau f'(\tau)}{f(\tau)} \tag{3}$$

Besides, we find that

$$\frac{1}{n}J^\mathsf{T}\Pi_\rho J = \left(\frac{1}{\ell}-\frac{c_0\beta^2}{\ell(\beta-\ell)^2}\right)\mathcal{D}(c)M^\mathsf{T}\Upsilon_\rho\Upsilon_\rho^\mathsf{T}M\mathcal{D}(c)+o(1)$$

almost surely, where $\Upsilon_\rho\in\mathbb{R}^{p\times m_\rho}$ are the eigenvectors of $\beta I_p + M\mathcal{D}(c)M^\mathsf{T}$ associated with eigenvalue ℓ.

Aside from the very simple result in itself, note that the choice of f is (asymptotically) irrelevant here. Note also that $M\mathcal{D}(c)M^\mathsf{T}$ plays an important role as its eigenvectors rule the behavior of the eigenvectors of L used for clustering.

Assume now instead that for each i, $\mu_i=\mu$ and $C_i=(1+\frac{\gamma_i}{\sqrt{p}})\beta I_p$ for some $\gamma_1,\ldots,\gamma_k\in\mathbb{R}$ fixed, and we shall denote $\gamma=[\gamma_1,\ldots,\gamma_k]^\mathsf{T}$. Then, if condition (2) is met, we now find after calculus that there exists *at most one* isolated eigenvalue in L (beside n) again equal in the limit to (3) but now for $\ell=\beta^2\left(\frac{5f'(\tau)}{8f(\tau)}-\frac{f'(\tau)}{2f(\tau)}\right)\left(2+\sum_{i=1}^k c_i\gamma_i^2\right)$. Moreover,

$$\frac{1}{n}J^\mathsf{T}\Pi_\rho J = \frac{1-c_0\frac{\beta^2}{(\beta-\ell)^2}}{2+\sum_{i=1}^k c_i\gamma_i^2}\mathcal{D}(c)\gamma\gamma^\mathsf{T}\mathcal{D}(c)+o_\mathrm{P}(1).$$

If (2) is not met, there is no isolated eigenvalue beside n. We note here the importance of an appropriate choice of f. Also observe that $\frac{1}{n}\mathcal{D}(c^{-\frac{1}{2}})J^\mathsf{T}\Pi_\rho J\mathcal{D}(c^{-\frac{1}{2}})$ is proportional to

Fig. 1. Samples from the MNIST database, without and with -10dB noise.

$\mathcal{D}(c^{\frac{1}{2}})\gamma\gamma^\mathsf{T}\mathcal{D}(c^{\frac{1}{2}})$ and thus the eigenvector aligns strongly to $\mathcal{D}(c^{\frac{1}{2}})\gamma$ itself. Thus the entries of $\mathcal{D}(c^{\frac{1}{2}})\gamma$ should be quite distinct to achieve good clustering performance.

IV. Simulations

We complete this article by demonstrating that our results, that apply in theory only to Gaussian x_i's, show a surprisingly similar behavior when applied to real datasets. Here we consider the clustering of $n = 3\times 64$ vectorized images of size $p = 784$ from the MNIST training set database (numbers 0, 1, and 2, as shown in Figure 1). Means and covariance are empirically obtained from the full set of 60 000 MNIST images. The matrix L is constructed based on $f(x) = \exp(-x/2)$.

Figure 2 shows that the eigenvalues of L' and $\hat{L}'$, both in the main bulk and outside, are quite close to one another (precisely $\|L' - \hat{L}'\|/\|L'\| \simeq 0.11$). As for the eigenvectors (displayed in decreasing eigenvalue order), they are in an almost perfect match, as shown in Figure 3. In the latter is also shown in thick (blue) lines the theoretical approximated (signed) diagonal values of $\frac{1}{n}\mathcal{D}(c^{-\frac{1}{2}})J^\mathsf{T}\Pi_\rho J\,\mathcal{D}(c^{-\frac{1}{2}})$, which also show an extremely accurate match to the empirical class-wise means. Here, the k-means algorithm applied to the four displayed eigenvectors has a correct clustering rate of $\simeq 86\%$.

Introducing a -10dB random additive noise to the same MNIST data (see images in Figure 1) brings the approximation error down to $\|L' - \hat{L}'\|/\|L'\| \simeq 0.04$ and the k-means correct clustering probability to $\simeq 78\%$ (with only two theoretically exploitable eigenvectors instead of previously four).

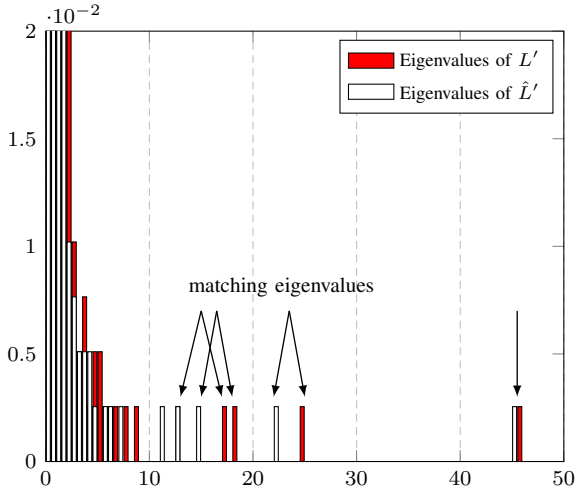

Fig. 2. Eigenvalues of L' and $\hat{L}'$, MNIST data, $p = 784$, $n = 192$.

V. Concluding remarks

The random matrix analysis of kernel matrices constitutes a first step towards a precise understanding of the underlying mechanism of kernel spectral clustering. Our first theoretical findings allow one to already have a partial understanding of the leading kernel matrix eigenvectors on which clustering is based. Notably, we precisely identified the (asymptotic) linear combination of the class-basis canonical vectors around which the eigenvectors are centered. Currently on-going work aims at studying in addition the fluctuations of the eigenvectors around the identified means. With all these informations, it shall then be possible to precisely evaluate the performance of algorithms such as k-means on the studied datasets.

This innovative approach to spectral clustering analysis, we believe, will subsequently allow experimenters to get a clearer picture of the differences between the various classical spectral clustering algorithms (beyond the present Ng–Jordan–Weiss algorithm), and shall eventually allow for the development of finer and better performing techniques, in particular when dealing with high dimensional datasets.

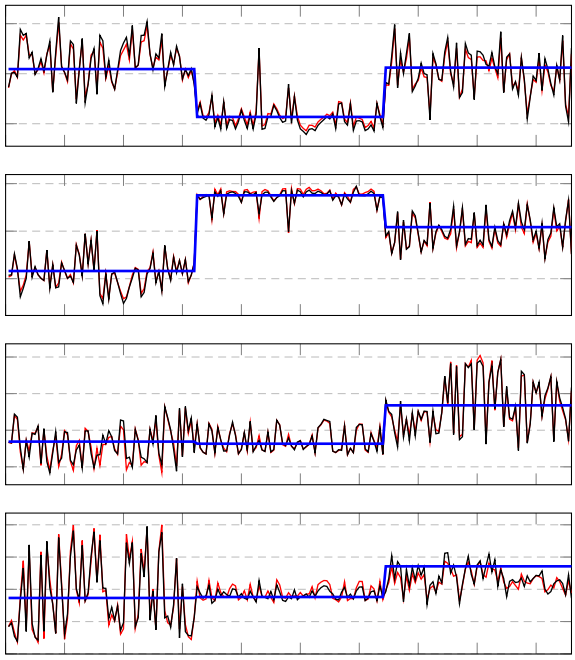

Fig. 3. Leading four eigenvectors of L (red) versus $\hat{L}$ (black) and theoretical class-wise means (blue); MNIST data.

References

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

[2] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, vol. 14, pp. 849–856, 2001.

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

[4] F. Benaych-Georges and R. R. 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.

[5] F. Chapon, R. Couillet, W. Hachem, and X. Mestre, "The outliers among the singular values of large rectangular random matrices with additive fixed rank deformation," *Markov Processes and Related Fields*, vol. 20, pp. 183–228, 2014.