Understanding Big Data Spectral Clustering
Romain Couillet, Florent Benaych-Georges

To cite this version:
Romain Couillet, Florent Benaych-Georges. Understanding Big Data Spectral Clustering. IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2015) , Dec 2015, Cancun, Mexico. 10.1109/camsap.2015.7383728 . hal-01242494

HAL Id: hal-01242494
https://hal.archives-ouvertes.fr/hal-01242494
Submitted on 12 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 analysis of the former allows one to understand deeply the mechanism 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).

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 \( f : \mathbb{R} \to \mathbb{R}_+ \).

There are multiple reasons (see e.g., [1]) to expect that the number of exploitable eigenvectors and the influence of the kernel function really matters for clustering;

1. in the large \( n, p \) regime, only a very local aspect of the kernel function really matters for clustering;
2. there exists a critical growth regime (with \( p \) and \( n \)) of the \( \mu_i \)'s and \( C_i \)'s for which spectral clustering leads to non-trivial misclassing probability;
3. 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 the operator norm for matrices. The vector \( 1_m \in \mathbb{R}^m \) stands for the vector filled with ones. The operator \( D(v) = D(\{u_k\}_{k=1}^p) \) is the diagonal matrix having \( u_1, \ldots, u_k \) as its diagonal elements. The Dirac mass at \( x \) is \( \delta_x \).

II. MODEL AND THEORETICAL RESULTS

Let \( x_1, \ldots, x_n \in \mathbb{R}^p \) be independent vectors with \( x_{n_1 + \ldots + n_{\ell-1} + 1}, \ldots, x_{n_1 + \ldots + n_\ell} \in C_\ell \) for each \( \ell \in \{1, \ldots, k\} \), where \( n_0 = 0 \) and \( n_1 + \ldots + n_k = n \). Class \( C_a \) 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 \( \mu_i \)'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 form the exact regime under which spectral clustering is a non trivial problem.

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

1. For \( \mu^0 \triangleq \sum_{i=1}^k \frac{n_i}{n} \mu_i \) and \( \mu_i^0 = \mu_i - \mu^0, ||\mu_i^0|| = O(1) \)
2. For \( C^0 \triangleq \sum_{i=1}^k \frac{n_i}{n} C_i \) and \( C_i^0 = C_i - C^0, ||C_i|| = O(1) \)
3. \( \frac{2}{p} \text{tr} C^0 \) converges as \( n \to \infty \) to \( \tau > 0 \).

The value \( \tau \) is important since \( \frac{1}{p} ||x_i - x_j||^2 \overset{\text{ap}}{\to} \tau \) uniformly on \( i \neq j \) in \( \{1, \ldots, n\} \).

\(^1\)Couillet’s work is supported by RMT4GRAPH (ANR-14-CE28-0006).
\(^2\)As shall be seen below, the (non conventional) division by \( p \) here is the proper norm scale in the large \( n, p \) regime.
We now define the kernel function as follows.

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

Then we introduce the kernel matrix

\[
K \triangleq \left\{ \frac{1}{p} \left\| x_i - x_j \right\|^2 \right\}_{i,j=1}^n.
\]

From the previous remark on \( \tau \), note that all non-diagonal elements of \( K \) tend to \( f(\tau) \) and thus \( K \) can be point-wise developed using 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 this singularity, we shall instead study the matrix

\[
L' \triangleq nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - \frac{D^{\frac{1}{2}}1_n1_n^TD^{\frac{1}{2}}}{n^2D1_n} - \frac{D^{\frac{1}{2}}1_n1_n^TD^{\frac{1}{2}}}{n^2D1_n} = \frac{(D^{\frac{1}{2}}1_n)_a(D^{\frac{1}{2}}1_n)_b}{(D1_n)_a(D1_n)_b}
\]

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 \( \| L' - \hat{L} \| \to 0 \), where \( \hat{L} \) follows a tractable random matrix model. Before introducing the latter, we need the following fundamental deterministic element notations

\[
M \triangleq [\mu_1, \ldots, \mu_k] \in \mathbb{R}^{p \times k},
\]

\[
t \triangleq \left\{ \frac{1}{\sqrt{p}} \text{tr} C_a \right\}_{a=1}^k \in \mathbb{R}^k,
\]

\[
T \triangleq \left\{ \frac{1}{p} \text{tr} C_a C_b \right\}_{a,b=1}^k \in \mathbb{R}^{k \times k},
\]

\[
J \triangleq [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k},
\]

\[
P \triangleq I_n - \frac{1}{n}1_n1_n^T \in \mathbb{R}^{n \times n}
\]

where \( j_a \in \mathbb{R}^n \) is the canonical vector of class \( C_a \), defined by \( (j_a)_i = \delta_{x_i, c_a} \), and the random element notations

\[
W \triangleq [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n},
\]

\[
\Phi \triangleq \frac{1}{\sqrt{p}} WTM \in \mathbb{R}^{n \times k},
\]

\[
\Psi \triangleq \frac{1}{p} \left\{ \| w_i \|^2 - E[\| w_i \|^2] \right\}_{i=1}^n \in \mathbb{R}^n.
\]

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

\[
\left\| L' - \hat{L} \right\| \xrightarrow{P} 0
\]

where \( \hat{L} \) is given by

\[
\hat{L} = \frac{-2f'(\tau)}{f(\tau)} \left[ \frac{PW^TWP}{p} + UBUT \right] + \frac{2f'(\tau)}{f(\tau)} F(\tau) I_n
\]

It is equivalent to study \( \hat{L}' \) or \( \hat{L} \) 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 \( \hat{L}' \).

\[
G_z = h(\tau, z)I_k + \left( h(\tau, z)M^T \left[ I_p + \sum_{j=1}^k c_j g_j(z) c_j \right] - \frac{h(\tau, z)P}{2f(\tau)} \right) T + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) \Gamma(z)
\]

and the case \( f'(\tau) = 0 \) is obtained by extension by continuity (in the limit \( f'(\tau) = 0 \) 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 \( UBUT \). When \( f'(\tau) = 0 \), \( \hat{L} \) is even a simpler 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 constituted by the vectors \( j_i \), while \( B \) contains the information about the inter-class mean-deviations through \( M \), and about the inter-class covariance deviations through \( T \) and \( I \). As such, the aforementioned isolated eigenvalue-eigenvector pairs are expected to correlate to the canonical class basis \( J \) and all the more so that \( M, T, I \) 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 \( \frac{1}{p}PW^TWP \) does not produce itself isolated eigenvalues (and thus, that the isolated eigenvalues of \( \hat{L}' \) are solely due to \( UBUT \)).

**Assumption 3 (Spike control):** Letting \( \lambda_1(C_a) \geq \ldots \geq \lambda_p(C_a) \) be the eigenvalues of \( C_a \), for each \( a \), as \( n \to \infty \),

\[
\frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_a)} \xrightarrow{P} \nu_a, \text{ with support } \text{supp}(\nu_a), \text{ and }
\]

\[
\max_{1 \leq i \leq p} \text{dist}(\lambda_i(C_a), \text{supp}(\nu_a)) \to 0.
\]

**Theorem 2 (Isolated eigenvalues):** Let Assumptions 1–3 hold and define the \( k \times k \) matrix

\[
G_z = h(\tau, z)I_k + \left( h(\tau, z)M^T \left[ I_p + \sum_{j=1}^k c_j g_j(z) c_j \right] - \frac{h(\tau, z)P}{2f(\tau)} \right) T + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) \Gamma(z)
\]

\[
\text{and the case } f'(\tau) = 0 \text{ is obtained by extension by continuity.}
\]
where

\[
h(\tau, z) = 1 + \left(5f'(\tau) \over 8f(\tau) - f''(\tau) \over 2f(\tau)\right) \sum_{i=1}^{k} c_i g_i(z) \over p \cdot \text{tr} C_i^2
\]

\[
\Gamma(z) = D \left\{ e_c g_a(z) \right\}_{a=1}^{k} - \left\{ \begin{array}{c} c_a g_a(z) c_b g_b(z) \over \sum_{i=1}^{k} c_i g_i(z) \end{array} \right\}_{a,b=1}^{k}
\]

and \(g_1(z), \ldots, g_k(z)\) are the unique solutions to the system

\[
1 \over c_0 g_a(z) = -z + 1 \over p \cdot \text{tr} C_a \left( I_p + \sum_{i=1}^{k} c_i g_i(z) C_i \right)^{-1}.
\]

Let \(\rho\), away from the eigenvalue support of \(\frac{1}{2} P W^T W P\), be such that \(h(\tau, \rho) \neq 0\) and \(G_\rho\) has a zero eigenvalue of multiplicity \(m_\rho\). Then there exists \(m_\rho\) eigenvalues of \(L\) asymptotically close to

\[
-2 f'(\tau) \over f(\tau)^{p} + f(0) - f(\tau) + \tau f'(\tau) \over f(\tau).
\]

Let us now turn to the more interesting result concerning the eigenvectors. This result is divided in two subsequent formulas, concerning respectively the eigenvector \(D^{1/2} 1_n\) associated with the eigenvalue \(n\) of \(L\), and the remaining (more interesting) eigenvectors associated with the eigenvalues exhibited in Theorem 2.

**Proposition 1 (Eigenvector \(D^{1/2} 1_n\)):** Let Assumptions 1–2 hold. Then

\[
\frac{D^{1/2} 1_n}{\sqrt{1_n D 1_n}} = {1 \over \sqrt{n}} + \frac{1}{n \sqrt{c_0}} \left[ f'(\tau) \tau 1_n \right]_{a=1}^{k} \over 2f(\tau) \left\{ \begin{array}{c} 1 \over \text{tr} C_a \left( I_p + \sum_{i=1}^{k} c_i g_i(z) C_i \right)^{-1} \over \varphi + o_p(1) \end{array} \right\}
\]

for some \(\varphi \sim N(0, I_n)\).

**Theorem 3 (Eigenvector projections):** Let Assumptions 1–3 hold. Let also \(\lambda_1^\tau, \ldots, \lambda_{k+m}\) be isolated eigenvalues of \(L\) all converging to \(\rho\) as per Theorem 2 and \(\Pi\), the projector on the eigenspace associated to these eigenvalues. Then, with the notations of Theorem 2,

\[
\frac{1}{n} J^T \Pi \rho J = -\Gamma(\rho) \sum_{i=1}^{m} {h(\tau, \rho)(V_{r,\rho} i (V_{i,\rho})^T \over (V_i, \rho) G_{ii, \rho} (V_{r,\rho}) + o_p(1)
\]

where \(V_{r,\rho} \in \mathbb{C}^{k \times m}\) and \(V_{i,\rho} \in \mathbb{C}^{k \times m}\) are sets of right and left eigenvectors of \(G_\rho\) associated with the eigenvalue \(\rho\), and \(G_i\) is the derivative of \(G_\rho\) along \(z\) taken for \(z = \rho\).

Proposition 1 provides an accurate characterization of the eigenvector \(D^{1/2} 1_n\), which conveys clustering information based on the difference in covariance traces (through \(t\)) mainly. 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

\[
1 \over n \cdot \text{tr} D(c^{-1}) J^T \Pi \rho J \in [0, m]\]
evaluates the alignment between \(\Pi\) and \(J\), thus providing a first hint on the expected performance of spectral clustering. A second interest of Theorem 3 is that, for eigenvectors \(u\) of \(L\) of multiplicity one (so \(\Pi = u u^T\)), the diagonal elements of \(\frac{1}{n} D(c^{-1}) \frac{J^T \Pi \rho J}{D(c^{-1})}\) provide the squared mean values of the successive first \(j_1\), then next \(j_2\), etc., elements of \(u\). The off-diagonal elements of \(\frac{1}{n} D(c^{-1}) \frac{J^T \Pi \rho J}{D(c^{-1})}\) then allow to decide on the signs of \(u^T j_i\) for each \(i\). These pieces of information are again crucial to estimate the expected performance of spectral clustering.

However, the statements of Theorems 2 and 3 are difficult to interpret from the onset. 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 \(\mu_i\)’s are equal and \(C_i = (1 + \gamma_i^2) \beta I_p\).

Assume first that \(C_i = \beta I_p\) for all \(i\). Then, letting \(\ell\) be an isolated eigenvalue of \(\beta I_p + M D(c) M^T\), if

\[
|\ell - \beta| > \beta \sqrt{c_0}
\]

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

\[
\rho = -2 f'(\tau) \over f(\tau)^{p} + f(0) - f(\tau) + \tau f'(\tau) \over f(\tau)
\]

Besides, we find that

\[
1 \over n J^T \Pi \rho J = (1 - {c_0 \gamma^2 \over (\beta - \ell)^2} D(c) M^T Y \rho M D(c) + o_p(1)
\]

where \(Y_\rho \in \mathbb{R}^{p \times m}\) are the eigenvectors of \(\beta I_p + M D(c) M^T\) associated with eigenvalue \(\ell\).

Aside from the very simple result in itself, note that the choice of \(f\) is (asymptotically) irrelevant here. Note also that \(M D(c) M^T\) 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 \(C_i = (1 + \gamma_i^2) \beta I_p\) for some \(\gamma_1, \ldots, \gamma_k \in \mathbb{R}\) fixed, and we shall denote \(\gamma = (\gamma_1, \ldots, \gamma_k)^T\). Then, if the separability condition (2) is met, we now find after calculus that there exists at most one isolated eigenvalue in \(L\) (beside \(n\)) again equal to

\[
\rho = -2 f'(\tau) \over f(\tau)^{p} + f(0) - f(\tau) + \tau f'(\tau) \over f(\tau)
\]

but for \(\ell = \beta \gamma^2 \left[ {f'(\tau) \over f(\tau)^{p}} - \left( f(\tau) \right)^{p} \right] (2 + \sum_{i=1}^{k} \gamma_i c_i^2)\). Moreover,

\[
1 \over n J^T \Pi \rho J = {1 - c_0 \beta^2 \gamma^2 \over 2 + \sum_{i=1}^{k} \gamma_i c_i^2} D(c) \gamma \gamma^T D(c) + o_p(1).
\]

If the separability condition is not met, then there is no isolated eigenvalue beside \(n\).

We note here the importance of an appropriate choice of \(f\). Note also that \(1 \over n D(c^{-1}) \frac{J^T \Pi \rho J}{D(c^{-1})}\) is proportional to \(D(c^{-1}) \gamma \gamma^T D(c^{-1})\) and thus the eigenvector aligns strongly to
The entries of $D(c^{\frac{1}{2}})^\gamma$ 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 60000 MNIST images. The matrix $L$ is constructed based on $f(x) = \exp(-x/2)$.

Figure 2 shows that the eigenvalues of both $L'$ 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 theoretically approximated (signed) diagonal values of $\frac{1}{\gamma} D(c^{\frac{1}{2}})^\gamma J^T \Pi_p J D(c^{\frac{1}{2}})$, which also show an extremely accurate match between theory and practice. Here, the $k$-means algorithm applied to the four displayed eigenvectors has a correct clustering rate of $\simeq 86\%$.

Introducing a $-10\text{dB}$ random additive noise to the same MNIST data 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).

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.

REFERENCES