Random Matrix Improved Subspace Clustering

Romain Couillet* and Abla Kammoun†
*LANEAS Group, CentraleSupélec, University of Paris-Saclay, France
†King Abdullah University of Science and Technology, Saudi Arabia.
romain.couillet@centralesupelec.fr, abla.kammoun@kaust.edu.sa

***Abstract*—This article introduces a spectral method for statistical subspace clustering. The method is built upon standard kernel spectral clustering techniques, however carefully tuned by theoretical understanding arising from random matrix findings. We show in particular that our method provides high clustering performance while standard kernel choices provably fail. An application to user grouping based on vector channel observations in the context of massive MIMO wireless communication networks is provided.**

I. Introduction

Clustering, be it for graphs (community detection) or multivariate data, is an important leg of machine learning by which one aims at grouping together alike elements in a completely unsupervised manner (that is based on no training examples). With the advent of the big data paradigm, clustering of large dimensional datasets becomes a key ingredient of machine learning. Because the underlying problem (that of devising an optimal grouping) is inherently discrete and often computationally expensive, most clustering strategies are based on relaxed optimization schemes. This is the case of *spectral kernel clustering* [1] and notably of the popular Ng–Weiss–Jordan algorithm [2], which will be our starting point here.

Precisely analyzing the performance of spectral kernel algorithms is quite challenging for two essential reasons: (i) studying generically the eigenvectors of a data-driven adjacency (or similarity) matrix is intractable, and (ii) even if made possible, the non-linear kernel function induces an involved correlation structure within the kernel matrix. To tackle this difficulty, [3] proved that kernel spectral clustering methods are consistent when the number of observations tends to infinity while the data dimensions remain fixed. This is a desirable result but that makes non desirable assumptions from a big data perspective, where the observation dimensions may be larger than their number. Recently, in [4], a novel approach was considered, whereby the size and number of samples were assumed simultaneously large *and* the data to be clustered arising from a Gaussian mixture. In this case, it is shown that, thanks to a concentration of measure effect, the kernel function can be expanded in a Taylor series, resulting in the kernel matrix being asymptotically equivalent to a well-understood random matrix model. The results of [4] notably evidence the inner mechanisms under play in spectral clustering and allow for many improvements. Besides, simulations on real datasets (MNIST database) suggest an extremely close fit in performance when compared to inputs extracted from a Gaussian mixture with same (empirical) means and covariances as the dataset.

A particularly striking finding of [4] is that, for data vectors with vanishing differences in (empirical or statistical) means across classes, a very specific kernel choice allows for a dramatic (asymptotic) performance improvement – corresponding in fact to a phase transition towards a faster convergence regime. The purpose of this article is to study this setting in depth. Precisely, we shall consider the kernel spectral clustering of a Gaussian zero mean mixture and shall focus on clustering the data upon their "normalized shape" (i.e., irrespective of their amplitude). This setting places us naturally in a *subspace clustering context*. Subspace clustering is an important branch of clustering, whereby the objective is to identify groups of data living in similar subspaces [5].

We will place ourselves in a regime where classical spectral clustering kernel choices provably fail, while our proposed approach achieves arbitrarily good performance. Our theoretical results are then appended into a novel algorithm for zero-mean subspace clustering when multiple independent copies of each sample is available. This algorithm finds concrete applications in user grouping for modern massive MIMO wireless communications [6], [7]. Practically speaking, the performances of the proposed massive MIMO method dramatically improves over alternatives [6], [8] where an extremely large number of channel observations from each user is assumed, while we obtain good performances already for very few channel observations.

II. Main Results

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be n independent vectors such that, for $n_1, \ldots, n_k$ with $\sum_i n_i = n$,

$$x_{n_1+\ldots+n_{j-1}+1}, \ldots, x_{n_1+\ldots+n_j} \sim \mathcal{N}(0, p^{-1} C_j)$$

where $C_1, \ldots, C_k \in \mathbb{R}^{p \times p}$ are nonnegative definite covariance matrices satisfying $\frac{1}{p} \operatorname{tr} C_a = 1$ for each $a = 1, \ldots, k$. The latter constraint is inconsequential for the remainder and is merely imposed for convenience. The fact that the x_i's labels are sorted per class is merely for convenience and is no more restrictive. We shall say that $x_i \in \mathcal{C}_a$ when $x_i \sim \mathcal{N}(0, p^{-1} C_a)$.

The objective is to devise an appropriate spectral clustering method to group the x_i's within their respective classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$. We recall that spectral clustering [1] consists first in building a matrix $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = \kappa(x_i, x_j)$ containing some affinity metric between vectors x_i and x_j. Here we

shall consider a radial kernel $\kappa(x_i, x_j) = f(\|x_i - x_j\|^2)$. It can then be shown by intuitive arguments that, when clustering is "not too difficult", K tends to have large isolated eigenvalues, the eigenvectors of which look like noisy step functions (when the x_i are sorted per class as above). Each plateau of the step functions is mapped to one class and thus clustering can be performed by smartly exploiting these eigenvectors. Calling $u_1, \ldots, u_\ell$ the ℓ dominant eigenvectors of K, the natural method to do this smart exploitation is to perform standard low-dimensional clustering (such as k-means or expectation maximization) on the n vectors $([u_1]_i, \ldots, [u_\ell]_i) \in \mathbb{R}^\ell$ for $i = 1, \ldots, n$. The top graph in Figure 3 at the end of this document provides an $\ell = 2$-dimensional representation of $n = 400$ vectors $([u_1]_i, [u_2]_i) \in \mathbb{R}^2$ with colors corresponding to ground truth classes.

Our precise objective is to select an appropriate function f such that the aforementioned clustering approach provides *non-trivial performances in difficult scenarios*. To this end, we shall place ourselves in a regime where $p, n \to \infty$ and shall impose growth rates of n, p and $C_1, \ldots, C_k$ in such a way that the probability of misclustering remains of order $O(1)$.

Assumption 1 (Growth Rate): As $p \to \infty$, $n/p \to c_0 \in (0, \infty)$ and, for each $a \in \{1, \ldots, k\}$, $n_a/n \to c_a \in (0, 1)$. Besides, for each $a, b \in \{1, \ldots, k\}$, denoting $C^\circ = \sum_{i=1}^k \frac{n_i}{n} C_i$ and $C_a^\circ = C_a - C^\circ$,

$$\frac{1}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \text{ converges in } [0, \infty).$$

We shall further define

$$\mathcal{T} = \left\{ \lim_{p\to\infty} \sqrt{\frac{c_a c_b}{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}$$
$$\omega = \sqrt{2} \lim_{p\to\infty} \frac{1}{p} \operatorname{tr}(C^\circ)^2.$$

The fundamental assumption here is that $\operatorname{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$. It is shown in [4] that, if $\operatorname{tr} C_a^\circ C_b^\circ = o(p)$, then it is *in general* impossible to recover the classes using the kernel approach, while non-trivial clustering probability can be achieved when $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$. However, [4, Remark 12] points to an exception to this statement for a very specific choice of the kernel. Under this choice, it is proved that asymptotically *perfect clustering* is achieved when $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$. One of our contributions here is to show that, for this very kernel option, $\operatorname{tr} C_a^\circ C_b^\circ$ can be made as small as $O(\sqrt{p})$ with non-trivial class recovery.

We make this discussion more precise below by introducing the kernel matrix of interest.

Assumption 2 (Kernel Matrix): For $x_1, \ldots, x_n \in \mathbb{R}^p$, denoting $\bar{x} = \frac{x}{\|x\|}$, let

$$K = \left\{ f(\|\bar{x}_i - \bar{x}_j\|^2) \right\}_{i,j=1}^n$$

where $f : \mathbb{R}^+ \to \mathbb{R}^+$ is a three-times differentiable function satisfying $f'(2) = 0$ and $f''(2) \neq 0$.

Examples of functions f satisfying the conditions of Assumption 2 are $f_{\text{poly}}(t) \equiv a(t-2)^2 + b$ for some $a, b > 0$ or, to anticipate evident problems of robustness to outliers using polynomials, $f_{\exp}(t) \equiv \exp(-a(t-2)^2)$ for some $a > 0$.

Letting $D = \operatorname{diag}(K 1_n)$ with $1_n \in \mathbb{R}^n$ the vector of ones, we define the (normalized centered) Laplacian matrix of K as

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}} - n \frac{D^{\frac{1}{2}} 1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n}$$

which is the central object of interest (rather than K itself) here, reminiscent of the Ng–Weiss–Jordan Laplacian [2].

Note that $D^{\frac{1}{2}} 1_n$ is the eigenvector of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ corresponding to its leading eigenvalue 1 (a well-known property of Laplacian matrices). Thus, the matrix L is (up to the scalar n) the projection of the matrix $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ on the subspace orthogonal to $D^{\frac{1}{2}} 1_n$. This projection, also exploited in [4], allows for a simplified spectral study of $nD^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ which is composed of (i) a dominant unit rank eigenspace carried by $D^{\frac{1}{2}} 1_n$ with associated eigenvalue n (thus diverging) and (ii) an $n-1$ orthogonal space with associated eigenvalues of order $O(1)$. As shown in [4], the subspace (i) does not contain any valuable information and thus the dominant eigenvector of $nD^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ asymptotically contain no clustering information. We shall therefore restrict ourselves to the study of L as defined above.

The central hypothesis of Assumption 2 is the choice of $f'(2) = 0$ without which kernel spectral clustering is asymptotically infeasible. Equipped with these intuitions, we are in position to introduce our main result which provides a *tractable random matrix equivalent for* L. This matrix is much simpler to study than L itself and will allow for a clear understanding of the eigenvectors content.

Theorem 1 (Random Matrix Equivalent): Under Assumption 1, as $n, p \to \infty$, almost surely,

$$L = \frac{f(0) - f(2)}{f(2)} P + \frac{2f''(2)}{f(2)} \left\{ \frac{1}{p} \operatorname{tr} C_a^\circ C_b^\circ \frac{1_{n_a} 1_{n_b}^\mathsf{T}}{p} \right\}_{a,b=1}^k + \frac{2f''(2)}{f(2)} P \Phi P + O_{\|\cdot\|}(p^{-\frac{1}{2}})$$

where $P = I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}$ and $\Phi \in \mathbb{R}^{n \times n}$ is defined by

$$\Phi_{ij} = \boldsymbol{\delta}_{i \neq j} \left[(x_i^\mathsf{T} x_j)^2 - \frac{1}{p^2} \operatorname{tr} C_a C_b \right]$$

with $x_i \in \mathcal{C}_a$, $x_j \in \mathcal{C}_b$, and $\|\cdot\|$ denotes the operator norm. In particular, we can evaluate, for $i \neq j$, $x_i \in \mathcal{C}_a$, $x_j \in \mathcal{C}_b$,

$$\mathrm{E}[\Phi_{ij}] = 0$$
$$\operatorname{Var}[\Phi_{ij}] = \frac{2}{p^4} (\operatorname{tr} C_a C_b)^2 + \frac{6}{p^4} \operatorname{tr}(C_a C_b)^2 = O(p^{-2}).$$

As a corollary of Theorem 1, it is easily shown that

$$\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[L - \frac{f(0) - f(2)}{f(2)} P \right] = \left\{ \frac{1}{p} \operatorname{tr} C_a^\circ C_b^\circ \frac{1_{n_a} 1_{n_b}^\mathsf{T}}{p} \right\}_{a,b=1}^k + P(\sqrt{p}\Phi) P + o_{\|\cdot\|}(1)$$

is a matrix having asymptotically the same dominant eigenvectors as those of L (with eigenvalues defined up to a mapping). As such, the eigenvectors of interest for clustering are those of $\mathcal{L}$. In the rest of the article, we shall then exclusively focus on $\mathcal{L}$ rather than L.

Note first that $\mathcal{L}$ takes the form of the sum of: (i) a random matrix with zero mean and entries of variance $O(p^{-1})$ and (ii) a maximum rank-k (in fact even $k-1$) matrix with eigenvectors made of linear combinations of the canonical class vectors $j_1, \ldots, j_k$ with $j_i = (0, \ldots, 0, 1_{n_i}^\mathsf{T}, 0, \ldots, 0) \in \mathbb{R}^n$, modulated by the class informations $\frac{1}{p} \operatorname{tr} C_a^\circ C_b^\circ$. This is a *spiked random matrix model* of the *information-plus-noise* type [9]. For such models, there exists a phase transition phenomenon by which, if the eigenvalues of matrix (ii) are large enough, $\mathcal{L}$ will contain isolated large amplitude eigenvalues with associated eigenvectors much aligned to those of (ii) itself. If instead no such large eigenvalue exists, no isolated eigenvalue is found in the spectrum of $\mathcal{L}$ and no information is exploitable for clustering. This intuitive discussion is made clear in the following result.

Theorem 2 (Eigenvalue Localization): Let Assumption 1 hold. Then, as $p \to \infty$, the eigenvalue distribution $\mu_n \equiv \frac{1}{n} \sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(\mathcal{L})}$ (with $\lambda_i(X)$ the eigenvalues of X) almost surely converges (in the weak limit of probability measures) to the probability measure μ with

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2} \sqrt{(4c_0\omega^2 - t^2)^+} dt$$

with support $\mathcal{S} = [-2\sqrt{c_0}\omega, 2\sqrt{c_0}\omega]$. Besides, for all large n almost surely, there are at most $k-1$ eigenvalues of $\mathcal{L}$ found at a macroscopic distance of $\mathcal{S}$. These are defined as follows. Let $\nu_1 \geq \ldots \geq \nu_k$ be the eigenvalues of $\mathcal{T}$. Then, for each $i \in \{1, \ldots, k\}$, if $\sqrt{c_0}|\nu_i| > \omega$, for all large n almost surely, $\mathcal{L}$ has an isolated λ_i satisfying

$$\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}$$

(since $\mathcal{T}$ has a zero eigenvalue, the inequality $\sqrt{c_0}|\nu_i| > \omega$ is not met at least once).

Theorem 2 unveils a phase transition effect by which, if $|\nu_i|/\omega$ or c_0 are large enough, then isolated eigenvalues are found in the limiting spectrum of $\mathcal{L}$. Figure 1 depicts the histogram of the eigenvalues of L versus the asymptotic semi-circle law μ, where one can observe isolated eigenvalues on the left side of the main bulk. When this occurs, the eigenvector u_i associated with the isolated eigenvalue λ_i of $\mathcal{L}$ will correlate the eigenvector with eigenvalue ν_i of $\{\frac{1}{p^2} \operatorname{tr} C_a^\circ C_b^\circ \cdot 1_{n_a} 1_{n_b}^\mathsf{T}\}_{a,b=1}^k$, this correlation being all the stronger that $|\nu_i|$ is large. As such, for sufficiently large $|\nu_i|$, the eigenvectors u_i will tend to behave like (noisy) step vectors. We make this intuition more rigorous in what follows.

Assume now that λ_i is an isolated eigenvalue of $\mathcal{L}$ as per Theorem 2 with unit multiplicity. From the statistical interchangeability of vectors x_i within each class $\mathcal{C}_a$, we may write the eigenvector u_i associated with λ_i as

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

for $j_a = [0_{n_1}^\mathsf{T}, \ldots, 0_{n_{a-1}}^\mathsf{T}, 1_{n_a}^\mathsf{T}, 0_{n_{a+1}}^\mathsf{T}, \ldots, 0_{n_k}^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^n$, w_i^a a vector of unit norm, supported on the indices of class $\mathcal{C}_a$ and orthogonal to j_a, and $\alpha_i^a \in \mathbb{R}$, $\sigma_i^a > 0$ scalars to be determined. Similarly, for two isolated eigenvalues $\lambda_i, \lambda_{i'}$ of $\mathcal{L}$, it shall be interesting to study the correlation

$$\sigma_{i,i'}^a \equiv \left(u_i^a - \alpha_i^a \frac{j_a}{\sqrt{n_a}}\right)^\mathsf{T} \left(u_{i'}^a - \alpha_{i'}^a \frac{j_a}{\sqrt{n_a}}\right)$$

with $u_i^a = \operatorname{diag}(j_a) u_i \in \mathbb{R}^n$ the restriction of u_i to its indexes in class $\mathcal{C}_a$. In particular, $(\sigma_i^a)^2 = \sigma_{ii}^a$.

These quantities characterize completely the mutual behavior of the isolated eigenvectors of $\mathcal{L}$ used for spectral clustering and thus allow for anticipating the performance of kernel spectral clustering. In the next theorem, we provide the asymptotic values of those parameters.

Theorem 3 (Expression of Eigenvectors): For each (ν_i, v_i) eigenpair of $\mathcal{T}$ with ν_i of unit multiplicity satisfying $\sqrt{c_0}|\nu_i| > \omega$ and for each $a \in \{1, \ldots, k\}$, let $\alpha_i^a = \frac{1}{n_a} u_i^\mathsf{T} j_a$ with (λ_i, u_i) the eigenpair of $\mathcal{L}$ mapped to ν_i as per Theorem 1. Then, as $p \to \infty$, under the conditions of Assumption 1, for each $a, b \in \{1, \ldots, k\}$,

$$\alpha_i^a \alpha_i^b \xrightarrow{\text{a.s.}} \left(1 - \frac{1}{c_0} \frac{\omega^2}{\nu_i^2}\right) \left[v_i v_i^\mathsf{T}\right]_{ab}.$$

Besides, for (λ_i, u_i), $(\lambda_{i'}, u_{i'})$ two such eigenpairs of $\mathcal{L}$, letting $\sigma_{ii'}^a = \frac{1}{n_a}(u_i^a - \alpha_i^a j_a)^\mathsf{T}(u_{i'}^a - \alpha_{i'}^a j_a)$, we have

$$\sigma_{ii'}^a \xrightarrow{\text{a.s.}} \boldsymbol{\delta}_{ii'} \frac{c_a}{c_0} \frac{\omega^2}{\nu_i^2}.$$

Figure 2 depicts the leading two eigenvectors under the setting of Figure 1 and the theoretical values for α_i^j and σ_i^j.

Remark 1 (Relation to Subspace Clustering): Subspace clustering consists in grouping vectors x_i in classes defined through the distance between the covariance matrices $\mathrm{E}[x_i x_i^\mathsf{T}]$ (or between their dominant subspaces). In particular, for $k = 2$ classes, such a metric may be the Frobenius norm $\operatorname{tr}(C_1 - C_2)^2$ between C_1 and C_2. It appears that this is exactly what is implemented by the proposed method. For $k \geq 3$, the method instead considers a metric based on the eigenvalues of $\{\operatorname{tr}(C_a^\circ - C_b^\circ)^2\}_{a,b=1}^k$ which is one among multiple possibilities of such distance definitions.

III. Application to Wireless Communications

The setting under study is of broad interest in massive MIMO wireless communications where, to avoid the bottleneck problem known as *pilot contamination* [7], it is necessary to smartly schedule data transmissions to users having channels belonging to as far subspaces as possible. That is, given a massive MIMO transmitter equipped with a large number p antennas serving n mobile users with respective channels

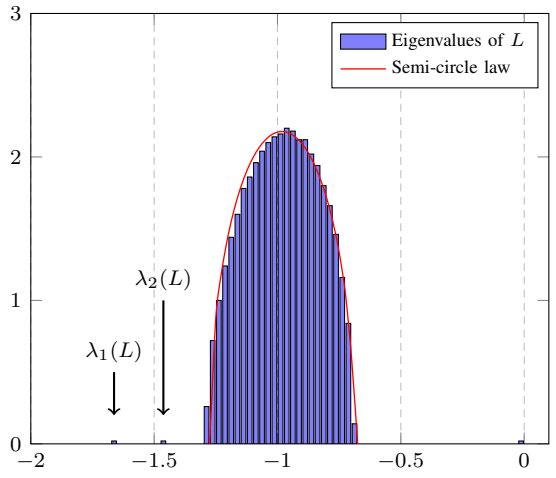

Fig. 1. Eigenvalues of L and semi-circle law, for $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^\mathsf{T}$, $W_i \in \mathbb{R}^{p \times (p/8)}$ with i.i.d. $\mathcal{N}(0,1)$ entries, $f = f_{\exp}$.

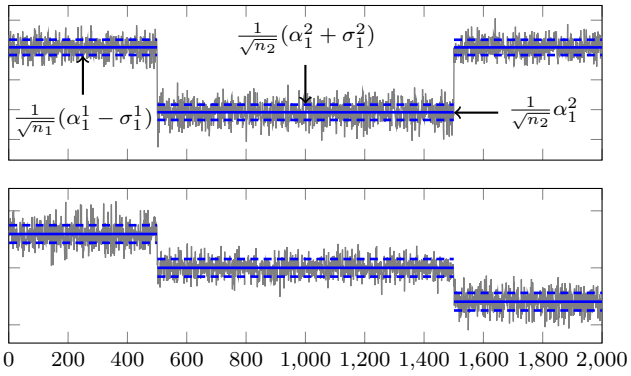

Fig. 2. Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$ as per Theorem 3. Setting identical to that of Figure 1.

$h_1, \ldots, h_n \in \mathbb{R}^p$, it is a first objective to the transmitter to group users into subsets of alike channel subspaces.[1] Here the so-called subspaces relate to the covariances of the *zero mean* random channels h_i (randomness arising from mobility and the null mean for the absence of line of sight components). This problem is traditionally handled by assuming a method to extract good estimators of each covariance $\mathrm{E}[h_i h_i^\mathsf{T}]$, see e.g., [6], [8]. For large p (traditionally ~ 400 in massive MIMO) and n, this means accessing a tremendous amount of independent channels which is unrealistic.

Instead, we propose here to use a single channel observation to meet the same results, which we then refine to several independent observations. The underlying reason for such a gain is that, while [6], [8] believe in the need to retrieve the extremely large matrices $\mathrm{E}[h_i h_i^\mathsf{T}]$, we claim that only estimating $\mathrm{tr}\, C_a^\circ C_b^\circ$ for groups a, b of users is enough, and this is easily obtained through kernel spectral clustering.

In the case of a single channel acquisition, the idea here consists in building a kernel matrix K, and its associated (modified) Laplacian $\mathcal{L}$, with $K_{ij} = f(\|\bar{h}_i - \bar{h}_j\|^2)$ for $\bar{h}_i = \frac{h_i}{\|h_i\|}$ and f with $f'(2) = 0$ and $f''(2) \neq 0$, and then to perform subspace clustering as previously introduced upon $\mathcal{L}$. When T independent copies of the channels $h_i^{(1)}, \ldots, h_i^{(T)}$ for each user i are obtained, we suggest to proceed as follows: (i) build a $Tn \times Tn$ matrix K, and subsequently $\mathcal{L}$, for the ordered channels $[h_1^{(1)}, \ldots, h_1^{(T)}, \ldots, h_n^{(1)}, \ldots, h_n^{(T)}]$ as if the channels $h_i^{(j)}$ were arising from Tn rather than n users, then (ii) to exploit the fact that the vectors $h_i^{(1)}, \ldots, h_i^{(T)}$ are mapped to a single user (and therefore a single point for the clustering problem at hand), average the dominant nT-dimensional eigenvectors u_i of $\mathcal{L}$ across the T indexes corresponding to channels of the same users, resulting into the n-dimensional vectors

$$\bar{u}_i \equiv \frac{1}{T}(I_T \otimes 1_T^\mathsf{T}) u_i \tag{1}$$

with $1_T \in \mathbb{R}^T$ the all-ones vector, and finally (iii) proceed to clustering (from k-means or expectation-maximization procedures) over the dominant vectors $\bar{u}_i$. The main effect of the folding operation (1) is to reduce the variance of the rows of matrix $[u_1, \ldots, u_n]$ by a factor T.

To simulate the performance of the method, we model the $h_i^{(j)} \in \mathbb{R}^{2p}$ as two-dimensional real vectors of the stacked real and imaginary parts of Gaussian circularly symmetric zero mean channels with covariance matrices $C_a \in \mathbb{R}^{2p \times 2p}$ for users i in class $\mathcal{C}_a$. Letting $\Gamma_a \in \mathbb{C}^{p \times p}$ be the complex representation of C_a, we consider the popular solid angular model

$$[\Gamma_a]_{ij} = \frac{1}{\Delta_+^a - \Delta_-^a} \int_{\theta^a + \Delta_-^a}^{\theta^a + \Delta_+^a} \exp\left(-2\pi\imath \frac{d}{\lambda} \sin(t)(j-i)\right) dt$$

for a linear antenna array with inter-antenna distance d and transmission wavelength λ taken here such that $d = \lambda$. The array beam focuses uniformly across the angles $[\theta^a + \Delta_-^a, \theta^a + \Delta_+^a]$. We shall take here $\Delta_-^a = -\Delta_+^a = \pi/20$ for each a, while $\theta^1 = -\pi/30$, $\theta^2 = 0$, and $\theta^3 = \pi/30$.

Figure 3 illustrates the gain of the per-user folding strategy (1) on a single random scenario (see figure caption for details). It is seen that the large spread of data points prior to folding (top figure) is significantly reduced after $T = 10$ copies folding (bottom figure). In both graphs are shown in blue the theoretical 1σ and 2σ standard deviations obtained from Theorem 3 and the T-fold reduction in variance from (1).

The performances associated with the setting of Figure 3 are provided in Figure 4, where a comparison between the classical Gaussian kernel $f(t) = \exp(-t^2)$ and our proposed improved kernel $f(t) = \exp(-(t-2)^2)$ is provided. The last step of clustering is performed either via a k-means or an expectation-maximization (EM) method (assuming Gaussian

[1] For simplicity, we assume here that channels are modeled as real vectors. When complex, real and imaginary parts can be stacked into a $2p$-dimensional vector.

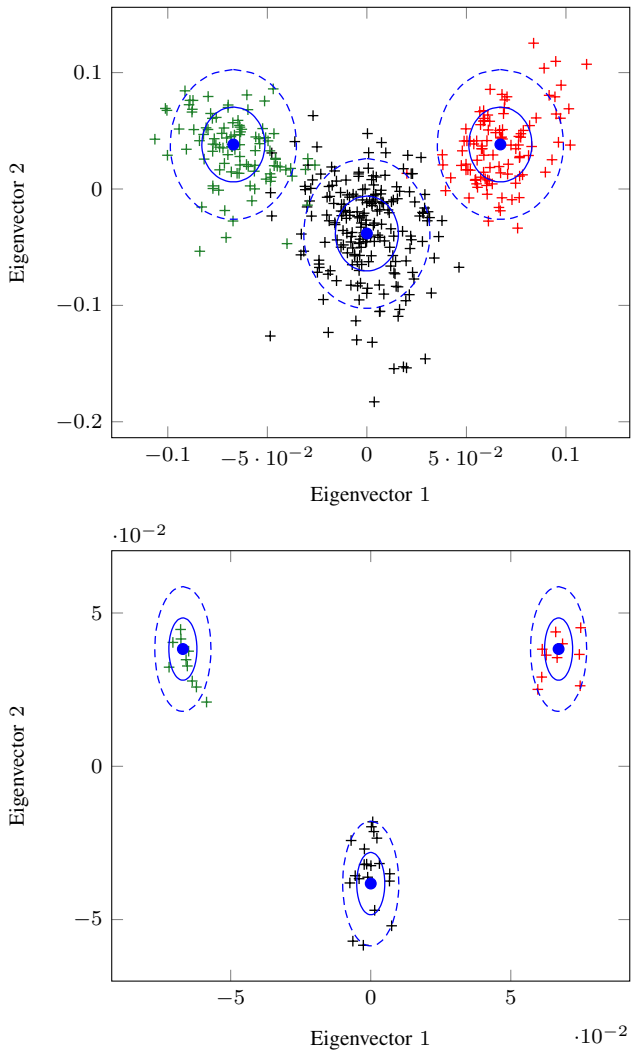

Fig. 3. Leading two eigenvectors before (left figure) and after (right figure) T-averaging. Setting: $p = 400$, $n = 40$, $T = 10$, $k = 3$, $c_1 = c_3 = 1/4$, $c_2 = 1/2$, Γ_a's built from angular spread model with angles $-\pi/30 \pm \pi/20$, $0 \pm \pi/20$, and $\pi/30 \pm \pi/20$. Kernel function $f(t) = \exp(-(t-2)^2)$.

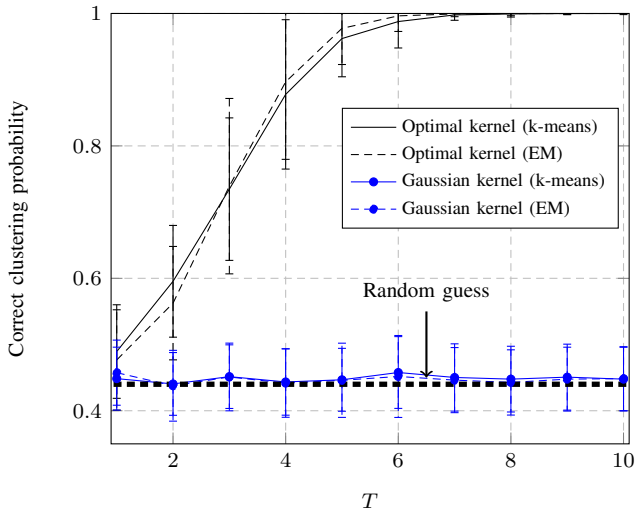

Fig. 4. Comparison of the correct clustering probability between the optimal kernel $f(x)$ (here $f(x) = \exp(-(x-2)^2)$) and the Gaussian kernel $\exp(-x^2)$, for different number of observations T per source, using the k-means or EM algorithms. Setting identical to Figure 3.

eigenvector fluctuations). As is shown, under the considered setting, irrespective of T, the Gaussian kernel does not manage to provide any efficient clustering. In comparison, the kernel choice with $f'(2) = 0$ achieves 100% clustering probability already for 8 independent channel acquisitions. As is standard in clustering, k-means operates better than EM at low correct clustering rates but is then overtaken by EM for higher rates.

IV. Concluding remarks

In this article, we have exploited the recent theoretical analysis [4] on the asymptotic performance of kernel spectral clustering to produce an improved method for *subspace clustering* of large dimensional datasets. Possibly the most striking outcome of the study is its demonstrating that smartly chosen kernel choices allow for data clustering where traditional kernels provably fail. This unfolds from a *change in regime* for the specific problem at hand when using appropriate kernels.

Generically speaking, this article may be considered as a first step towards enabling random matrix-based improvements in large dimensional machine learning problems. It is envisioned that, based again on the investigation method from [4] naturally leading to spiked equivalent models of random kernel matrices, questions such as deep theoretical understanding of semi-supervised learning and support vector machine methods for large dimensional datasets could be similarly addressed, thereby possibly leading to improved versions of these so popular methods.

V. Acknowledgements

Couillet's work is supported by the ANR RMT4GRAPH (ANR-14-CE28-0006) and the HUAWEI RMTin5G projects.

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] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," *The Annals of Statistics*, pp. 555–586, 2008.

[4] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *arXiv preprint arXiv:1510.03547*, 2016.

[5] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[6] S. Haghighatshoar and G. Caire, "Channel vector subspace estimation from low-dimensional projections," *arXiv preprint arXiv:1509.07469*, 2015.

[7] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive mimo in the ul/dl of cellular networks: How many antennas do we need?" *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 160–171, 2013.

[8] A. Adhikary and G. Caire, "Joint spatial division and multiplexing: Opportunistic beamforming and user grouping," *arXiv preprint arXiv:1305.7252*, 2013.

[9] 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.