RANDOM MATRIX-IMPROVED KERNELS FOR LARGE DIMENSIONAL 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

Leveraging on recent random matrix advances in the performance analysis of kernel methods for classification and clustering, this article proposes a new family of kernel functions theoretically largely outperforming standard kernels in the context of asymptotically large and numerous datasets. These kernels are designed to discriminate statistical means and covariances across data classes at a theoretically minimal rate (with respect to data size). Applied to spectral clustering, we demonstrate the validity of our theoretical findings both on synthetic and real-world datasets (here, the popular MNIST database as well as EEG recordings on epileptic patients).

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

I. INTRODUCTION

With the advent of the big data era, a strong pressure has recently been set on the development of powerful machine learning methods to perform classification or regression tasks involving large and numerous datasets (i.e., under a "large p, large n" regime). These methods notably involve non-linear approaches such as neural networks and kernel-based algorithms which, as an aftermath of their non-linear character, are difficult to analyze. In a recent line of works initiated in [1], in the large p and n asymptotics, kernel random matrices have been explored and have led to a completely renewed understanding of kernel approaches, starting with the asymptotic performance (and sometimes inconsistency) of kernel classification and spectral clustering. This includes kernel-based (least-square) support vector machines [2], semi-supervised classification [3] and spectral clustering [4]–[6], but also neural network derivatives such as extreme learning machines [7]. The main lever to analyze the performance of kernel matrices $\mathbf{K} \in \mathbb{R}^{n\times n}$ in the large dimensional regime ($p, n \to \infty$ with $p/n \to c_0 > 0$) lies in the fact that, under appropriate (what we shall call here "asymptotically non-trivial") growth rate assumptions on the data statistics, the entries $K_{ij} = f(x_i^\mathsf{T} x_j)$ or $K_{ij} = f(\|x_i - x_j\|^2)$ of $\mathbf{K}$ tend to converge to a limiting constant, *irrespective of the data class* (when classification is concerned), thereby allowing for a study of $\mathbf{K}$ through a Taylor expansion; this gives way in particular to the possible analysis of the eigenvectors of $\mathbf{K}$ or to functionals of $\mathbf{K}$ for all large p, n. These expansions notably set forth the discriminative effect of kernel-based classification methods as they tend to emphasize (in the structure of the dominant eigenvector of $\mathbf{K}$ notably) the statistical difference between the class means and class covariances, this emphasis being strongly related to the derivatives of f at a certain location. It has been notably confirmed, both theoretically on synthetic Gaussian mixture models but also in practice on real datasets that specific choices of f are more adapted to datasets containing either pronounced differences between class means as in the case of the popular MNIST dataset [8] while others are better tuned to classes having similar means but comparatively strongly differing covariances [9].

An important side remark of [4] however emphasized the fact that most natural kernels (such as the popular radial basis function kernel, given by $K_{ij} = f(\|x_i - x_j\|^2)$ with $f(t) = \exp(-t/2)$) are incapable to discriminate between class covariances at an optimal rate. By specifically choosing f to attain such high "covariance discriminative" power (precisely by taking f such that $f'(\tau) = 0$ for τ the converging value of all $\|x_i - x_j\|^2$ for translation-invariant kernels, or such that $f'(0) = 0$ for inner-product kernels), [6] indeed demonstrates a complete change in the random matrix structure of $\mathbf{K}$ and in the overall performance of classification. However, these functions f, be they highly powerful to discriminate data with distinct covariances, have as a side effect the property of completely masking the difference between the statistical means of the classes, thus leading up to extremely poor performances in means-dominated datasets (such as with MNIST).

The objective of this article is precisely to conciliate the findings of [4], [6] by proposing a new kernel parametrization that set emphasis to both differences in statistical means and covariances of the data classes, *to their optimal discriminate rate*. To this end, we shall study inner-product kernel matrices $\mathbf{K}$ designed with $f'(0) = \mathcal{O}(1/\sqrt{p})$ (rather than $\mathcal{O}(1)$ as in [4] or zero as in [6]) for datasets arising from a Gaussian mixture model with minimal growth rate (with respect to p) in the distances between means and covariances. These minimal rate conditions are in fact optimal in the sense that Neyman-Pearson hypotheses tests with perfectly known class means and covariances cannot achieve better rates.

Notation: Vectors are denoted with boldface lowercase letters and matrices by boldface 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. $\mathbb{C}^+$ is defined to be the set of complex numbers with a positive imaginary part. The Dirac mass is $\boldsymbol{\delta}_x$.

II. MODEL AND MAIN RESULTS

Consider n independent 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 $\mathbf{C}_a \in \mathbb{R}^{p\times p}$ a nonnegative definite matrix. We assume without loss in generality that the vectors are sorted as $\mathbf{x}_{n_1+\cdots+n_{a-1}+1}, \ldots, \mathbf{x}_{n_1+\cdots+n_a} \in \mathcal{C}_a$ for $a = 1, \ldots, k$.

We assume the large dimensional random matrix regime by which both p and n grow large at the same rate. Under the described data model, conditioning on the knowledge of the statistical means and covariances of each vector $\mathbf{x}_i$, a Neyman-Pearson hypothesis test applied to the data can decide on the genuine data class with asymptotically non-trivial error (i.e., with probability of error neither 0 nor $1/k$) when the following growth rate conditions (items 1–3 below) are assumed.

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

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

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* $\operatorname{tr} \mathbf{C}_a^\circ = \mathcal{O}(\sqrt{p})$.
3) $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ$ *converges in* $[0, \infty)$.
4) $\frac{1}{p} \operatorname{tr} \mathbf{C}^\circ$ *converges to* $\tau > 0$.

Under those conditions, it was shown in [4, Remark 12] that estimating the class labels by spectral clustering on $\mathbf{K}$ does not perform better than random guess for generic f, unless the condition $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(1)$ is relaxed to to $\frac{1}{\sqrt{p}} \operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(\sqrt{p})$, or that f is chosen so that $f'(\tau) = 0$ for translation-invariant kernels or $f'(0) = 0$ for inner-product kernels. But the latter choice comes along with a complete annihilation of the class means in the spectral clustering inner workings (a setting carefully studied in [6]).

As made clear by a careful random matrix analysis, setting instead $f'(\tau) = \mathcal{O}(p^{-\frac{1}{2}})$ (or $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$) allows for a fair treatment of both class means and covariances in the classification procedure. For simplicity of exposition, we focus here on the case of inner-product kernels with $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$. We thus have the following key assumption on the kernel function design.

Assumption 2 (On the kernel function)**.** *The kernel function f is three-times continuously differentiable in a neighborhood of 0 with $f(0), f''(0), f'''(0)$ constant with p while $f'(0) = \frac{\alpha}{\sqrt{p}}$ for some $\alpha \in \mathbb{R}$. We shall also denote $\beta = \frac{1}{2} f''(0)$.*

For instance, the kernels $f(x) = \beta(x + p^{-\frac{1}{2}} \beta^{-1} \alpha)^2$ or $f(x) = e^{-\beta(x + p^{-\frac{1}{2}} \beta^{-1} \alpha)^2}$ satisfy the conditions of Assumption 2.

The kernel studied in [6] thus corresponds to the particular case of Assumption 2 with $\alpha = 0$. As for [4], we shall see that it might be considered as a limiting setting where α is arbitrarily large.

For subsequent use, we introduce the following notations

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

with $\mathbf{j}_a \in \mathbb{R}^n$ the canonical vector of cluster $\mathbf{C}_a$ defined by $(\mathbf{j}_a)_i = \boldsymbol{\delta}_{\mathbf{x}_i \in \mathcal{C}_a}$. Having specified the conditions on f, let us now define $\mathbf{K}$ as the inner-product random matrix

$$\mathbf{K} \triangleq \begin{cases} f\left(\frac{1}{p} (\mathbf{x}_i^\circ)^\mathsf{T} \mathbf{x}_j^\circ\right) & , i \neq j \\ 0 & , i = j \end{cases}$$

with $\mathbf{x}_i^\circ = \mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ and f satisfying Assumption 2. Under this parametrization, we shall successively show that the matrix $\mathbf{K}$ composed of non-linear and intricately dependent entries asymptotically behaves in a simpler "almost linear" manner. From this simplified form, the asymptotic spectral characterization of $\mathbf{K}$ will be understood, in particular its dominant eigenvector contents.

As in [1] and following-up works, the non-linearity in $\mathbf{K}$ is treated by noticing that, as $p \to \infty$, $K_{ij} \to 0$ for all $i \neq j$, thereby allowing for an entry-wise Taylor expansion of $\mathbf{K}$. The theoretical difficulty next lies in the random matrix analysis of all matrix terms arising from the Taylor expansion. The key particularity that makes the setting $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$ so fundamental is that, in [4], the terms affected by the differences $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ$ used to vanish (as a result of being absorbed by background noise) when $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ = \mathcal{O}(\sqrt{p})$; by letting $f'(0) = \mathcal{O}(p^{-\frac{1}{2}})$, the dominant background noise (but also the differences in means) are reduced and now comparable to the terms involving $\operatorname{tr} \mathbf{C}_a^\circ \mathbf{C}_b^\circ$ (as long as $\beta = \frac{1}{2} f''(0) \neq 0$). An interesting side effect is that a second noise term then arises, and leads to a peculiar phenomenon where a mixture between a Marcenko–Pastur [10] type and a semi-circle type [11] noise eigenvalue distribution is observed in the limiting spectrum of $\mathbf{K}$. Still, this complication in the "noise spectrum" paradoxically comes along with a much simplified "signal spectrum", as shown in the subsequent results.

Theorem 1. *Under Assumption 1 and 2, let $\hat{\mathbf{K}}$ be given by:*

$$\begin{aligned}
\hat{\mathbf{K}} &= \alpha \mathbf{P} \mathbf{W}^T \mathbf{W} \mathbf{P} + \beta \mathbf{P} \boldsymbol{\Phi} \mathbf{P} + \mathbf{U} \mathbf{A} \mathbf{U}^T \\
\mathbf{A} &= \begin{bmatrix} \alpha \mathbf{M}^T \mathbf{M} + \beta \mathbf{T} & \alpha \mathbf{I}_k \\ \alpha \mathbf{I}_k & 0 \end{bmatrix} \\
\mathbf{U} &= \left[\frac{\mathbf{J}}{\sqrt{p}}, \mathbf{P} \mathbf{W}^T \mathbf{M} \right] \\
\frac{\boldsymbol{\Phi}}{\sqrt{p}} &= \left\{ ((\boldsymbol{\omega}_i^\circ)^T \boldsymbol{\omega}_j^\circ)^2 \boldsymbol{\delta}_{i \neq j} \right\}_{i,j=1}^n - \left\{ \frac{\operatorname{tr}(\mathbf{C}_a \mathbf{C}_b)}{p^2} \mathbf{1}_{n_a} \mathbf{1}_{n_b}^T \right\}_{a,b=1}^k .
\end{aligned}$$

Then,

$$\left\| \sqrt{p} \left(\mathbf{P} \mathbf{K} \mathbf{P} + \left(f(0) + \tau f'(0) \right) \mathbf{P} \right) - \hat{\mathbf{K}} \right\| \xrightarrow{\text{a.s.}} 0.$$

Theorem 1 states that, up to centering and scaling, $\mathbf{K}$ is asymptotically equivalent to $\hat{\mathbf{K}}$. In particular, an immediate corollary of Theorem 1 is that both matrices asymptotically share (again, up to centering and scaling) the same eigenvalues as well as *isolated* eigenvectors (i.e., eigenvectors associated to eigenvalues found at non-vanishing distance from any other eigenvalue). We may then study the asymptotic spectral properties of $\mathbf{K}$ (and as a result, the classification performance of algorithms based on $\mathbf{K}$) through $\hat{\mathbf{K}}$.

As previously hinted at, it is first interesting to note that $\hat{\mathbf{K}}$ is the sum of i) the random matrices $\alpha \mathbf{P} \mathbf{W}^T \mathbf{W} \mathbf{P}$ (of the Marcenko–Pastur type) and $\beta \mathbf{P} \boldsymbol{\Phi} \mathbf{P}$ (of the Wigner type, as shown in [6]) having entries of order $\mathcal{O}(p^{-1})$ and of ii) a maximum rank $k-1$ matrix containing linear combinations of the class-wise step vectors $\mathbf{j}_a$ intricately scaled through the inner-products between class means ($\mathbf{M}^T \mathbf{M}$) and class covariance-products ($\mathbf{T}$). This may be identified as part of the large family of *spiked random matrix models* [12], with the particularity that the low-rank addition is not independent of the noise part and that the noise part itself is a mixture between random Wishart and random symmetric matrices.

Random matrix theory today possesses all necessary tools to assess the eigenspectrum of such spiked random matrix models. As a common denominator, their eigenvalues are usually composed of a tightly connected "bulk" of eigenvalues along with up to $k-1$ isolated eigenvalues, the eigenvectors associated with which align to some extent to the eigenvectors in $\mathbf{U}$ (and thus, importantly here, to linear combinations of the vectors $\mathbf{j}_1, \ldots, \mathbf{j}_k$).

In particular, understanding the asymptotic performance of spectral clustering demands to characterize the isolated eigenvectors of $\mathbf{K}$. For these to be asymptotically informative, their associated eigenvalues must be found away from the main eigenvalue "bulk". In the following results, we evaluate the conditions upon which this transition phenomenon (i.e., the appearance of spiked eigenvalues) between asymptotically uninformative and informative eigenvectors occurs. We start by identifying the defining equations for the eigenvalue distribution of $\mathbf{K}$.

Theorem 2 (Eigenvalues Bulk)**.** *Let Assumptions 1 hold. Then, as $p \to \infty$, the spectral distribution $\nu_n \triangleq \frac{1}{n} \sum_{i=1}^n \boldsymbol{\delta}_{\lambda_i(\hat{\mathbf{K}})}$ (with*

$\lambda_i(\mathbf{X})$ *the eigenvalues of* $\mathbf{X}$*) almost surely converges (in the weak sense of probability measures) to the probability measure* ν *defined on a compact support* $\mathcal{S}$ *and having Stieltjes transform* $m(z) = \int \frac{\nu(dt)}{t-z}$ *defined for* $z \in \mathbb{C}^+$*, as the unique solution in* $\mathbb{C}^+$ *of*

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p} \operatorname{tr} \mathbf{C}^\circ \left(\mathbf{I}_k + \frac{\alpha m(z)}{c_0} \mathbf{C}^\circ \right)^{-1} - \frac{2\beta^2}{c_0} \omega^2 m(z)$$

where $\omega = \lim_{p\to\infty} \frac{1}{p} \operatorname{tr}(\mathbf{C}^\circ)^2$.

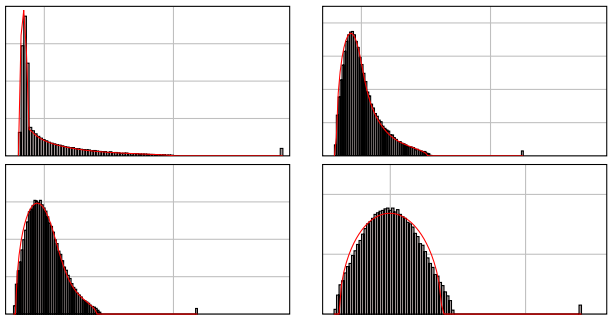

Fig. 1: Eigenvalues of $\mathbf{K}$ (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\boldsymbol{\mu}_i = 3\boldsymbol{\delta}_i$, $f(x) = \frac{1}{2}\beta \left(x + \frac{1}{\sqrt{p}} \frac{\alpha}{\beta} \right)^2$. (**Top left**): $\alpha = 8, \beta = 1$, (**Top right**): $\alpha = 4, \beta = 3$, (**Bottom left**): $\alpha = 3, \beta = 4$, (**Bottom right**): $\alpha = 1, \beta = 8$.

Figure 1 shows for different values of the parameters α and β the histogram of the eigenvalues of $\mathbf{K}$ versus the theoretical bulk ν from Theorem 2.[1]. Note that ν is indeed a mixture of the Marcenko–Pastur law (more visible when $\alpha \gg \beta$) and a Wigner semi-circle law (especially appearant as $\beta \gg \alpha$). The regime under study thus exhibits a tradeoff between the regime considered in [4] (where α is theoretically infinite and only a Marcenko–Pastur law appears in the theoretical formulas) and the regime considered in [6] (where $\alpha = 0$ and a semi-circle law is obtained).

With Theorem 2 in place, it now remains to determine the conditions under which isolated eigenvalues can be found in the spectrum of $\mathbf{K}$, i.e., eigenvalues falling outside the support $\mathcal{S}$ of the limiting measure ν. This is obtained by means of now standard random matrix techniques (see e.g., [12]) dedicated to spiked models. The main result is provided in Theorem 3 below, explicited here for simplicity in the case where $k = 2$ classes with equal number of vectors per class, i.e., for $n_1 = n_2$.

Theorem 3. *Let Assumption 1 and 2 hold and let* $\rho \in \mathbb{R} \setminus \mathcal{S}$ *be such that*

$$\frac{m(\rho)}{4c_0} (\alpha g(\rho)\delta + \beta\theta) + 1 = 0 \tag{1}$$

with $g(\rho) = \frac{1}{p} \operatorname{tr}(\mathbf{I}_p + \frac{\alpha m(\rho)}{c_0} \mathbf{C}^\circ)^{-1}$, $\delta = \|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$, *and* $\theta = \frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$.

Then, there exists λ_j *eigenvalue of* $\hat{\mathbf{K}}$ *such that*

$$|\lambda_j - \rho| \xrightarrow{\text{a.s.}} 0.$$

Any real number ρ satisfying equation (1) therefore corresponds to the (almost sure) limit of some eigenvalue of $\mathbf{K}$ (again, up to a shift and scaling). This equation in general has a solution for sufficiently large differences in class means, through the Euclidean norm distance δ, or in class covariances, through the Frobenius norm distance θ. This induces a detectability phase transition depending on the values of the pair (δ, θ). Thus, for sufficiently large δ or θ and appropriately set α, β, Equation 1 has a solution, which implies the presence of an isolated eigenvalue outside $\mathcal{S}$ and to a corresponding eigenvector "aligned to some extent" to the canonical class vectors $\mathbf{j}_a$'s. Specifically, for every isolated eigenvalue λ of $\mathbf{K}$, the associated eigenvector $\mathbf{u}_\lambda$ can be written as a linear combination of the class canonical vectors added to residual noise. Since the data are statistically interchangeable within the classes, we can write

$$\mathbf{u}_\lambda = \eta_1 \frac{\mathbf{j}_1}{\sqrt{n_1}} + \eta_2 \frac{\mathbf{j}_2}{\sqrt{n_2}} + \sigma_1 \boldsymbol{\omega}_1 + \sigma_2 \boldsymbol{\omega}_2 \tag{2}$$

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 orthogonal to respectively $\mathbf{j}_1$ and $\mathbf{j}_2$. The scalars η_1, η_2 can be seen as the empirical averages of the eigenvector entries in class $\mathcal{C}_1$ and $\mathcal{C}_2$ while the σ_1 and σ_2 represent the class standard deviations of the eigenvector fluctuations around $\eta_1 \frac{\mathbf{j}_1}{\sqrt{n_1}}$ and $\eta_2 \frac{\mathbf{j}_2}{\sqrt{n_2}}$. Intuitively, the larger $|\eta_1 - \eta_2|$ the more the separation between eigenvector entries mapped to $\mathcal{C}_1$ and those mapped to $\mathcal{C}_2$ and thus the better the clustering performance. A precise analysis of the limiting values of those parameters (similar to the approach in [4]) leads to the following result.

Theorem 4 (Isolated eigenvector)**.** *Let* λ *be an isolated eigenvalue of* $\hat{\mathbf{K}}$ *with almost sure limit* ρ*, and* $\mathbf{u}_\lambda$ *its associated eigenvector decomposed as* (2)*. Then, for both* $a = 1$ *and* $a = 2$

$$(\eta_a)^2 = \frac{m(\rho)^2}{2m'(\rho)} \frac{1}{1 - \frac{m(\rho)^2}{4m'(\rho)} \frac{\alpha g'(\rho)}{c_0} \delta} + o(1)$$

where $m(\rho)$ *and* $g(\rho)$ *are defined in Theorem 2 and* $m'(\rho)$*,* $g'(\rho)$ *are their respective first derivatives.*

Under this model (i.e., for $k = 2$ with $n_1 = n_2$), the limiting structure of eigenvector u_λ is quite symmetric, as seen through the fact that $\eta_1 = -\eta_2 + o(1)$. This in particular immediately implies that $\sigma_1^2 = \sigma_2^2 + o(1) = \frac{1}{2} - \eta_1^2 + o(1)$.

Such a symmetric model can be for example obtained by letting $\mathbf{C}_a = \mathbf{I}_p + \sqrt{\frac{\theta}{2\kappa}} p^{-5/4} \mathbf{W}_a \mathbf{W}_a^\mathsf{T}$ for some $\kappa > 0$, and with $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{p \times \kappa p}$ two independent random matrices having i.i.d. $\mathcal{N}(0,1)$ entries so that $\frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2 \xrightarrow{\text{a.s.}} \theta$,. In this case, the asymptotic correct classification $\mathbb{P}_c(\alpha, \beta)$ obtained by clustering eigenvector u_λ based on the signs of its entries satisfies

$$\mathbb{P}_c(\alpha, \beta) - Q\left(\sqrt{\frac{\eta^2}{1 - \eta^2}} \right) \xrightarrow{\text{a.s.}} 0 \tag{3}$$

where $\frac{1}{\eta^2} = \frac{2m'(\rho)}{m(\rho)^2} (1 - \frac{m(\rho)^2}{4m'(\rho)} \frac{\alpha g'(\rho)}{c_0} \delta)$.

As an illustration, Figure 2 depicts the limiting values for $\mathbb{P}_c(\alpha, \beta)$ as per (3) for different values of $\frac{\alpha}{\beta}$ and as a function of δ and θ. The figure strongly sets forth the importance of a proper choice of α, β depending on the specifics of the classification task, i.e., either means-dominant or covariance-dominant. In particular, as previously anticipated, a large value for $\frac{\alpha}{\beta}$ yields better performances in means-dominant discriminative tasks (bottom of Figure 2); conversely, small values of $\frac{\alpha}{\beta}$ are adapted to covariance-dominant tasks (top of Figure 2).

Upon anticipation of the most discriminative attribute of the data at hand, our results therefore provide an instructive direction to appropriate kernel choice. In supervised or semi-supervised learning tasks, δ and θ can be estimated through appropriate (random matrix-based) estimators, thereby helping in the choice of appropriate values for α and β. An application of this principle is performed in the subsequent section on real datasets.

[1] Obtained from the inverse formula $\nu(dt) = \frac{1}{\pi} \lim_{\epsilon \to 0} \Im[m(t + i\epsilon)] dt$

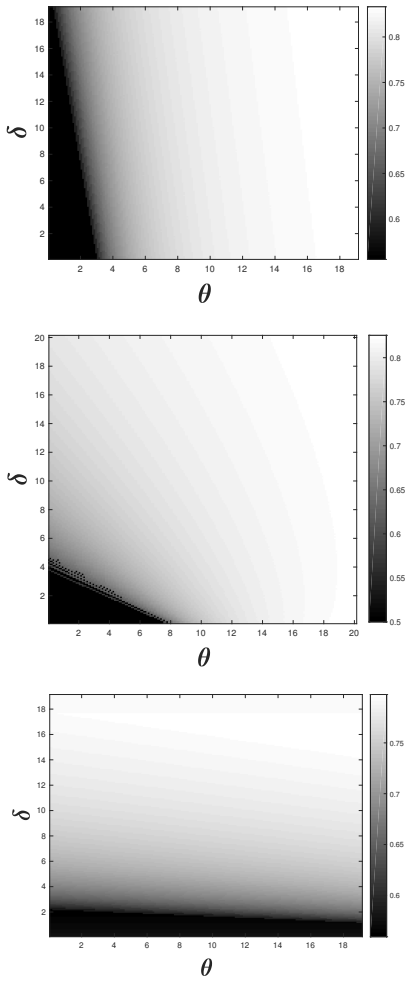

Fig. 2: $\frac{p}{n} = \frac{1}{2}$, $k = 2$, $c_1 = c_2$, $\boldsymbol{\mu}_i = \delta \boldsymbol{\delta}_i$, $\delta \in [1:20]$, $\mathbf{C}_1, \mathbf{C}_2$ as in the symmetric setting with $\theta \in [1:20]$, $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$. Probability of correct recovery for different settings $\frac{\alpha}{\beta} = \frac{1}{8}$ (top), $\frac{\alpha}{\beta} = 1$ (Middle), $\frac{\alpha}{\beta} = 8$ (Bottom), a function of δ (x-axis) and θ (y-axis).

III. APPLICATIONS

Our study has so far provided theoretical results for Gaussian mixture models, notably emphasizing the appropriateness of a kernel having first derivative scaling as $\mathcal{O}(p^{-\frac{1}{2}})$ with the data size p. In this section, we demonstrate that these findings are confirmed when applied to realistic datasets. The first dataset under consideration is the popular MNIST database of handwritten digits [8]. In this dataset, the classes (the different digits) are evidently more discriminative in means than in covariances, as confirmed by Table I. The second dataset is the epileptic EEG database from [9] which consists of five sets (A to E), each containing $p = 100$ single EEG channel segments of $23.6s$ each. Sets A and B report measures on 5 healthy volunteers and sets $C - E$ on 5 epileptic patients; each set is composed of 4097 samples. This dataset demonstrates more variations in the class covariances as shown again in Table I.

For both examples, kernel spectral clustering is performed on the dominant eigenvector of a subset of two classes, using k-means (rather than the eigenvector entry signs) to discriminate the classes. The results are depicted in Figure 3 for the MNIST data and Figure 4 for the EEG data. A clear observation is that extremely poor performances are obtained in the MNIST case for $\frac{\alpha}{\beta} \simeq 0$ while conversely extremely good performances are found on EEG for that setting, as was anticipated. Yet, note that the optimal value of $\frac{\alpha}{\beta}$ for the MNIST case does not demand that $\beta \to 0$; rather, an optimal value for $\frac{\alpha}{\beta}$ is found within the range $[0, 10]$, thereby suggesting that the differences in covariance are also exploited to some extent.

Table I: Class means and class covariances differences for some real datasets.

DATASETS	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{p}\operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MINIST (DIGITS $1, 7$)	612.7	71.1
MINIST (DIGITS $3, 6$)	441.3	39.9
MINIST (DIGITS $3, 8$)	212.3	23.5
EEG (SETS A, E)	2.4	10.9

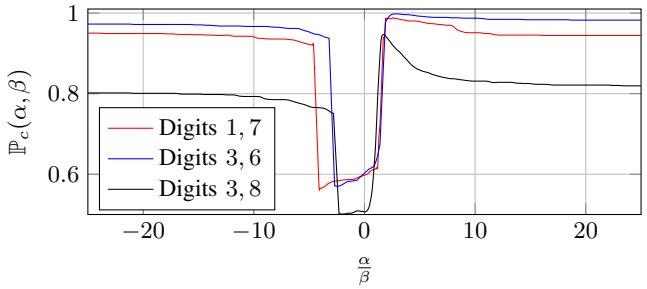

Fig. 3: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$.

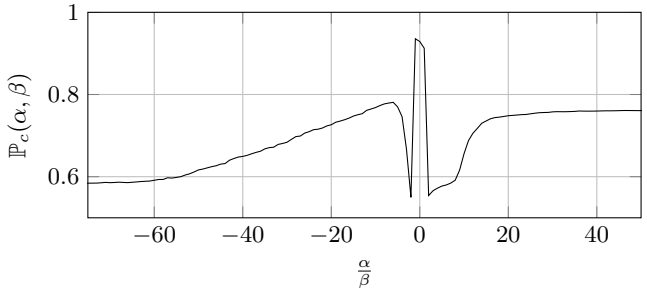

Fig. 4: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$.

IV. CONCLUSION

By relying on recent random matrix advances in big data machine learning, the article has introduced a new kernel function model for kernel-based statistical learning methods. The kernel relies heavily on the need to balance the weight carried by the statistical means and covariances in the data classes. While our results are based on idealistic Gaussian mixture models, simulations on realistic databases confirm the importance of such a kernel choice. Yet, the method only accounts so far for the performances obtained when optimally fine-tuning the hyperparameters (denoted α and β here and directly related to the first derivatives of the kernel function f) and does not provide a clear recipe for performing such a fine-tuning offline. This, and a much more comprehensive setting (accounting for more than two classes and more generic means and covariances models) shall be discussed in an extended version of the present study.

V. REFERENCES

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

[2] Zhenyu Liao and Romain Couillet, "A large dimensional analysis of least squares support vector machines," *arXiv preprint arXiv:1701.02967*, 2017.

[3] Xiaoyi Mai and Romain Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *arXiv preprint arXiv:1711.03404*, 2017.

[4] 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.

[5] Hafiz TIOMOKO ALI and Romain COUILLET, "Improved spectral community detection in large heterogeneous networks," *(submitted to) Journal of Machine Learning Research*, 2017.

[6] Romain Couillet and Abla Kammoun, "Random matrix improved subspace clustering," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 90–94.

[7] Cosme Louart, Zhenyu Liao, and Romain Couillet, "A random matrix approach to neural networks," *arXiv preprint arXiv:1702.05419*, 2017.

[8] Yann LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[9] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, vol. 64, no. 6, pp. 061907, 2001.

[10] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur, "Distribution of eigenvalues for some sets of random matrices," *Matematicheskii Sbornik*, vol. 114, no. 4, pp. 507–536, 1967.

[11] Eugene P Wigner, "Characteristic vectors of bordered matrices with infinite dimensions i," in *The Collected Works of Eugene Paul Wigner*, pp. 524–540. Springer, 1993.

[12] 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.