ON INNER-PRODUCT KERNELS OF HIGH DIMENSIONAL DATA

Zhenyu Liao, Romain Couillet

CentraleSupélec, University Paris-Saclay
GIPSA-lab, University Grenoble-Alpes.

ABSTRACT

In this article we investigate the eigenspectrum of inner-product kernel matrices of the type $\sqrt{p}\mathbf{K} = \{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}_{i,j=1}^n$. Under a two-class mixture modeling of the input data $\mathbf{x}_i \in \mathbb{R}^p$, we position ourselves in the regime where the number of data n and their dimension p are both large and comparable, and show, for a wide range of kernel functions f, that the spectrum of $\mathbf{K}$ only depends on f via *three* key parameters, with only *two* of them useful in extracting the statistical structure from the data. By carefully balancing these *two* parameters, a huge gain in classification performance is observed on real-world datasets.

***Index Terms*—** High dimensional statistics, kernel methods, random matrix theory, spectral analysis.

1. INTRODUCTION

Most theoretical analyses in statistical learning are derived under the assumption that the number of available data n is overwhelmingly larger than their dimension p. Nonetheless, under the current big data paradigm, we constantly face the situation where not only the size, but also the dimension of the data, are large. Understanding the resulting impact of popular statistical learning methods when n and p are both large and comparable is becoming a growing research concern in modern statistics.

Despite a long history of successful applications (e.g., kernel PCA, locally linear embedding [1], as well as the popular Ng–Jordan–Weiss kernel spectral clustering [2]), the theoretical analysis of kernel methods in the large n, p setting has not been investigated until very recently. In the line of works [3, 4, 5], the authors considered the kernel matrix $\mathbf{K} = \{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)\}_{i,j=1}^n$ built from the (nonlinear mapping of the) inner-product between n independent Gaussian data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and studied its eigenspectrum behavior. Based on a local expansion of the nonlinear kernel function f, which follows from the "concentration" of the similarity measure $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p$ around 0, the authors showed that it is enough to study a "linearized" version of $\mathbf{K}$ to characterize its eigenstructure, for any locally smooth kernel function f. Intuitively speaking, all off-diagonal entries of $\mathbf{K}$ evaluate f around 0 so that the eigenspectrum of $\mathbf{K}$ only depends on f via the successive derivatives $f(0), f'(0)$ and $f''(0)$. More specifically, they demonstrated that the eigenvalue distribution of $\mathbf{K}$ is (asymptotically) propositional to the more tractable sample covariance matrix model $\mathbf{X}^\mathsf{T}\mathbf{X}/p$ and tends to (a rescaled version of) the popular Marčenko-Pastur law [6] in the limit $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

To better exploit the "global" information of the nonlinear function f, in [7, 8] the authors considered the inner-product kernel matrix of the type $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$ which, thanks to the fact that $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p} \to \mathcal{N}(0, 1)$ (and thus of order $O(1)$) for $i \neq j$, avoids the (asymptotic) "concentration" effect with the more natural $\sqrt{p}$ normalization. The point-wise concentration to zero being replaced by a Gaussian limit, the authors in [7, 8] resort to an orthogonal polynomial approach in place of the Taylor expansion performed in [3, 4, 5]; this in particular allows for f to be non-differentiable.

Yet, only the spectrum of the "null model" for $\mathbf{K}$ (i.e., built upon random independent measurements $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$) was addressed in [7, 8]. From a machine learning viewpoint, this provides little understanding of the optimal nonlinear function f to be used for classification tasks, with a mixture data model.

In this article, we investigate the eigenspectrum of the inner-product kernel matrix $\mathbf{K} = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})/\sqrt{p}$ under a two-class multivariate mixture model (detailed below) that captures the first ($\boldsymbol{\mu}$) and second order ($\mathbf{E}$) statistical information of the input data. We show that $\mathbf{K}$ asymptotically follows an "information-plus-noise" pattern in the sense that $\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0$ as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, for some $\tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$ with the "noise" part $\mathbf{K}_N$ characterized in [7, 8] and the "information" part $\tilde{\mathbf{K}}_I$ depending on the data statistics $\boldsymbol{\mu}, \mathbf{E}$ and the kernel function f via only two scalar parameters. Empirical evidences on the popular MNIST [9] and epileptic EEG time series data [10] establish a close match to our theoretical prediction and thus convey a strong applicative motivation for this work.

Notations: Boldface lowercase (uppercase) characters stand for vectors (matrices). The notation $(\cdot)^\mathsf{T}$ denotes the transpose operator. The norm $\|\cdot\|$ is the Euclidean norm for vectors and the operator norm for matrices, and we denote $\|\cdot\|_F$ the Frobenius norm: $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})$.

2. SYSTEM MODEL AND PRELIMINARIES

2.1. Basic settings

Consider n data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independently drawn from the following two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) mixture model:

$$\begin{cases} \mathcal{C}_1: & \mathbf{x} = \boldsymbol{\mu}_1 + (\mathbf{I}_p + \mathbf{E}_1)^{\frac{1}{2}}\mathbf{z} \\ \mathcal{C}_2: & \mathbf{x} = \boldsymbol{\mu}_2 + (\mathbf{I}_p + \mathbf{E}_2)^{\frac{1}{2}}\mathbf{z} \end{cases} \tag{1}$$

each having cardinality $n/2$, for $\boldsymbol{\mu}_a \in \mathbb{R}^p$, $\mathbf{E}_a \in \mathbb{R}^{p\times p}$, $a \in \{1, 2\}$ and random vector $\mathbf{z} \in \mathbb{R}^p$ having i.i.d. entries of zero mean, unit variance and bounded moments. To ensure the information of $\boldsymbol{\mu}_a, \mathbf{E}_a$ is neither (asymptotically) too simple nor impossible to to be exploited from the noisy data, we shall position ourselves (as in [11]) under the following assumption.

Assumption 1 (Non-trivial classification)**.** *As $n \to \infty$ we have, for $a \in \{1, 2\}$,*

Couillet's work is supported by the IDEX DataScience Chair GSTATS at University Grenoble-Alpes and the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

1. $p/n \to c \in (0, \infty)$,
2. $\|\boldsymbol{\mu}_a\| = O(1)$, $\|\mathbf{E}_a\| = O(p^{-1/4})$, $|\operatorname{tr}(\mathbf{E}_a)| = O(\sqrt{p})$ *and* $\|\mathbf{E}_a\|_F^2 = O(\sqrt{p})$.

Following [4, 7], we consider the following random inner-product kernel matrix

$$\mathbf{K} = \left\{ \delta_{i \neq j} f(\mathbf{x}_i^\mathsf{T} \mathbf{x}_j / \sqrt{p}) / \sqrt{p} \right\}_{i,j=1}^n \tag{2}$$

for some nonlinear function $f : \mathbb{R} \mapsto \mathbb{R}$ satisfying regularity conditions detailed in Assumption 2 below. As in [4, 7], the diagonal elements $f(\mathbf{x}_i^\mathsf{T} \mathbf{x}_i / \sqrt{p})$ are discarded since, under Assumption 1, $\mathbf{x}_i^\mathsf{T} \mathbf{x}_i / \sqrt{p} = O(\sqrt{p})$ which is an "improper scaling" for the evaluation by f (unlike non-diagonal ones $\mathbf{x}_i^\mathsf{T} \mathbf{x}_j / \sqrt{p} = O(1)$ for independent $\mathbf{x}_i, \mathbf{x}_j$).

In the null model where $\boldsymbol{\mu}_a = \mathbf{0}$, $\mathbf{E}_a = \mathbf{0}$ for $a = 1, 2$, we write $\mathbf{K} = \mathbf{K}_N$ with

$$[\mathbf{K}_N]_{ij} = \delta_{i \neq j} f(\mathbf{z}_i^\mathsf{T} \mathbf{z}_j / \sqrt{p}) / \sqrt{p}. \tag{3}$$

Let $\xi_p \equiv \mathbf{z}_i^\mathsf{T} \mathbf{z}_j / \sqrt{p}$. By the central limit theorem, $\xi_p \to \mathcal{N}(0, 1)$ in distribution as $p \to \infty$. As such, the entries $[\mathbf{K}_N]_{ij}$, $1 \leq i \neq j \leq n$, asymptotically behave like a family of *dependent* standard Gaussian variables to which f is applied. To assess the joint behavior of this family, some concepts in the theory of orthogonal polynomials and, in particular, of the class of Hermite polynomials for the standard Gaussian distribution [12, 13] need to be recalled.

2.2. The orthogonal polynomial framework

For a probability measure μ, we denote the set of orthonormal polynomials with respect to the scalar product $\langle f, g \rangle = \int f g d\mu$ as $\{P_l(x), l = 0, 1, \ldots\}$, obtained from the Gram-Schmidt procedure on the monomials $\{1, x, x^2, \ldots\}$ such that $P_0(x) = 1$, P_l is of degree l and $\langle P_{l_1}, P_{l_2} \rangle = \delta_{l_1 - l_2}$. By the Riesz-Fischer theorem [14, Theorem 11.43], for any function $f \in L^2(\mu)$, the set of squared integrable functions with respect to $\langle \cdot, \cdot \rangle$, one can formally expand f as

$$f(x) \sim \sum_{l=0}^{\infty} a_l P_l(x), \quad a_l = \int f(x) P_l(x) d\mu(x) \tag{4}$$

where "$f \sim \sum_{l=0}^{\infty} a_l P_l$" indicates that $\|f - \sum_{l=0}^{L} a_l P_l\| \to 0$ as $L \to \infty$ (and $\|f\|^2 = \langle f, f \rangle$).

To investigate the asymptotic behavior of $\mathbf{K}$ and $\mathbf{K}_N$ as $n, p \to \infty$, we assume the nonlinear function f can be well approximated by a polynomial for p large enough.

Assumption 2. *For each p, let $\xi_p = \mathbf{z}_i^\mathsf{T} \mathbf{z}_j / \sqrt{p}$ and $\{P_{l,p}(x),\ l \geq 0\}$ be the set of orthonormal polynomials with respect to the probability measure μ_p of ξ_p. For $f \in L^2(\mu_p)$ we have the formal expansion*

$$f(x) \sim \sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x)$$

for $a_{l,p}$ defined in (4) *we assume*

1. $\sum_{l=0}^{\infty} a_{l,p} P_{l,p}(x) \mu_p(dx)$ *converges in $L^2(\mu_p)$ to $f(x)$ uniformly over large p, i.e., for any $\epsilon > 0$ there exists L such that for all p large,*

$$\left\| f - \sum_{l=0}^{L} a_{l,p} P_{l,p} \right\|_{L^2(\mu_p)}^2 = \sum_{l=L+1}^{\infty} |a_{l,p}|^2 \leq \epsilon,$$

2. *as $p \to \infty$, $\sum_{l=1}^{\infty} |a_{l,p}|^2 \to \nu \in [0, \infty)$. Moreover, for $l = 0, 1, 2$, $a_{l,p}$ converges and we denote a_0, a_1 and a_2 their limits, respectively.*
3. *the function f is "centered" with respect to the standard Gaussian measure, i.e., $a_0 = 0$.*

Since $\xi_p \to \mathcal{N}(0, 1)$, the parameters a_0, a_1, a_2 and ν are simply (generalized) moments of the standard Gaussian measure involving f. Precisely, $a_0 = \mathbb{E}[f(\xi)]$, $a_1 = \mathbb{E}[\xi f(\xi)]$, $\sqrt{2} a_2 = \mathbb{E}[(\xi^2 - 1) f(\xi)] = \mathbb{E}[\xi^2 f(\xi)] - a_0$ and $\nu = \mathbb{E}[f^2(\xi) - a_0^2] \geq a_1^2 + a_2^2$, for $\xi \sim \mathcal{N}(0, 1)$. As we shall see in Theorem 1 and 2, the three parameters (a_1, a_2, ν) are of crucial significances in determining the eigenspectrum behavior of the kernel matrix $\mathbf{K}$.

Let us first focus on the null model $\mathbf{K}_N$. It has been shown in [7, 8] that the *empirical spectral measure* $\omega_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\mathbf{K}_N)}$ of the null model $\mathbf{K}_N$ has an asymptotically deterministic behavior (also referred to as the *limiting spectral measure* of $\mathbf{K}_N$) as $n, p \to \infty$, described as follows.

Theorem 1 (from [7, 8])**.** *Let $n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and Assumption 2 hold. Then, with probability one, the empirical spectral measure ω_n of $\mathbf{K}_N$ defined in* (3) *converges weakly to a probability measure ω. The latter is uniquely defined through its Stieltjes transform $m : \mathbb{C}^+ \to \mathbb{C}^+$, $m(z) \equiv \int (t - z)^{-1} \omega(dt)$, given as the unique solution in $\mathbb{C}^+$ of the cubic equation*[1]

$$-\frac{1}{m(z)} = z + \frac{a_1^2 m(z)}{c + a_1 m(z)} + \frac{\nu - a_1^2}{c} m(z). \tag{5}$$

Note that, by taking $a_1 = 0$ in (5) one gets $\frac{\nu}{c} m^2(z) + z m(z) + 1 = 0$ which corresponds to (a rescaled version of) the well-known Wigner semi-circle law [15]

$$\omega_{SC}(dx) = \frac{\sqrt{(4 - x)^+}}{2\pi} dx$$

with $(x)^+ \equiv \max(0, x)$. On the other hand, with $\nu = a_1^2$ (so that $a_l = 0$ for $l \geq 2$) one retrieves the popular Marčenko-Pastur law [6] (of parameter c^{-1})

$$\omega_{MP, c^{-1}}(dx) = (1 - c)^+ \delta(x) + \frac{c \sqrt{(x - a)^+ (b - x)^+}}{2\pi x} dx$$

for $a = (1 - \sqrt{c^{-1}})^2$ and $b = (1 + \sqrt{c^{-1}})^2$. Indeed, the spectral measure ω presented in Theorem 1 is a "mix" of the semicircle law ω_{SC} and the Marčenko-Pastur law $\omega_{MP, c^{-1}}$ in the sense of additive free convolution [16]

$$\omega = a_1(\omega_{MP, c^{-1}} - 1) \boxplus \sqrt{(\nu - a_1^2) c^{-1}} \omega_{SC} \tag{6}$$

as pointed out in [17], where we denote $a_1(\omega_{MP, c^{-1}} - 1)$ the law of $a_1(x - 1)$ for $x \sim \omega_{MP, c^{-1}}$ and $\sqrt{(\nu - a_1^2) c^{-1}} \omega_{SC}$ the law of $\sqrt{(\nu - a_1^2) c^{-1}} x$ for $x \sim \omega_{SC}$. From its form in (6), the limiting measure ω is of compact support and admits a density [18]. Figure 1 illustrates this mixed limiting measure. A key observation here is that, for a given a_1, having larger values for ν enlarges the support of ω as per (6) and Figure 1.

[1] $\mathbb{C}^+ \equiv \{z \in \mathbb{C},\ \Im[z] > 0\}$. We also recall that, for $m(z)$ the Stieltjes transform of a measure ω, ω can be obtained from $m(z)$ via $\omega([a, b]) = \lim_{\epsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m(x + \imath\epsilon)] dx$ for all $a < b$ continuity points of ω.

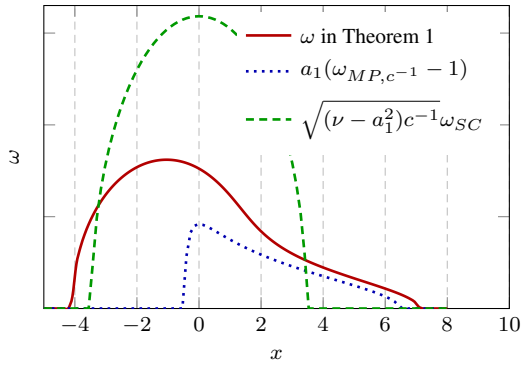


Fig. 1. Comparison between the Marčenko-Pastur law ω_{MP}, the semicircle law ω_{SC} and the limiting spectral measure ω given in Theorem 1 with $a_1 = a_2 = 1$, $\nu = 2$ and $c = 1/3$.

3. MAIN RESULTS

Built upon the spectral characterization of the "noise-only" model $\mathbf{K}_N$ in Theorem 1, the major contribution of this article is to provide an asymptotically accurate and theoretically tractable approximation matrix $\tilde{\mathbf{K}}$ for the "informative" kernel matrix $\mathbf{K}$, as detailed in the following theorem.

Theorem 2 (Asymptotic approximation of $\mathbf{K}$)**.** *Let Assumptions 1–2 hold. Then, with probability one,*

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \to 0, \quad \tilde{\mathbf{K}} = \mathbf{K}_N + \tilde{\mathbf{K}}_I$$

for $\mathbf{K}_N$ *defined in* (3) *and*

$$\tilde{\mathbf{K}}_I = \frac{a_1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T}) + \frac{a_2}{p}\mathbf{J}(\mathbf{T}+\mathbf{S})\mathbf{J}^\mathsf{T} \quad (7)$$

with $\mathbf{Z} = \left[\mathbf{z}_1, \ldots, \mathbf{z}_n\right] \in \mathbb{R}^{p\times n}$, a_1, a_2 *the coefficients (of the normalized Hermite polynomials) defined in Assumption 2 and*[2]

$$\mathbf{M} = \left[\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2\right] \in \mathbb{R}^{p\times 2}, \quad \mathbf{J} = \left[\mathbf{j}_1 \quad \mathbf{j}_2\right] \in \mathbb{R}^{n\times 2}$$
$$\sqrt{p}\mathbf{T} = \{\mathrm{tr}(\mathbf{E}_a + \mathbf{E}_b)\}_{a,b=1}^2, \quad \sqrt{p}\mathbf{S} = \{\mathrm{tr}(\mathbf{E}_a\mathbf{E}_b)\}_{a,b=1}^2$$

for $\mathbf{j}_a$ *the canonical vector of class* $\mathcal{C}_a$ *with* $[\mathbf{j}_a]_i = \delta_{\mathbf{x}_i \in \mathcal{C}_a}$.

Sketch of proof. The asymptotic spectral analysis of $\mathbf{K}$ comes in two steps: i) first, by an expansion of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ as a function of $\mathbf{z}_i, \mathbf{z}_j$ and the statistical mixture model parameters $\boldsymbol{\mu}, \mathbf{E}$, we decompose $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ (under Assumption 1) into successive orders of magnitudes with respect to p. This further allows for a Taylor expansion of $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})$ for at least three-times differentiable functions f around its dominant term $f(\mathbf{z}_i^\mathsf{T}\mathbf{z}_j/\sqrt{p})$. Then, ii) we rely on the orthogonal polynomial approach of [7] to "linearize" the resulting matrix terms $\{f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$, $\{f'(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ and $\{f''(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/\sqrt{p})\}$ (all terms corresponding to higher order derivatives have entries of $o(n^{-1})$ and thus asymptotically vanishing operator norm as $n, p \to \infty$, since $\|\mathbf{A}\| \leq n\|\mathbf{A}\|_\infty = n\max_{i,j}|\mathbf{A}_{ij}|$ for $\mathbf{A} \in \mathbb{R}^n$). Eventually, Assumption 2 is used to extend this approximation (that solely holds for differentiable functions) to arbitrary square-summable f. We leave the complete derivation of the theorem to an extended version of this article. □

[2]As a mental reminder, $\mathbf{M}$ stands for *means*, $\mathbf{T}$ accounts for the difference in *traces* of covariance matrices and $\mathbf{S}$ for the "*shapes*" of covariances.

Theorem 2, together with Theorem 1, unveils the surprising fact that, under the mixture model in (1) and the non-trivial classification condition in Assumption 1, the eigenspectrum of the kernel matrix $\mathbf{K}$ depends on the square-summable nonlinear function f only through the three parameters (a_1, a_2, ν). More precisely, the (limiting) spectral measure[3] of $\mathbf{K}$ is the same as that of $\mathbf{K}_N$ and is determined only by a_1 and ν. For the information matrix $\tilde{\mathbf{K}}_I$, both a_1 and a_2 play an important role and they control respectively the statistical information in means ($\boldsymbol{\mu}_a$) and covariances ($\mathbf{E}_a$). As a consequence, every square-summable f is asymptotically equivalent to a cubic function $\tilde{f}(x) = c_3x^3 + c_2(x^2 - 1) + c_1x$ from a kernel spectrum viewpoint, so long that the parameters (a_1, a_2, ν) of both functions match.[4]

To visually confirm the fact that the eigenspectrum of $\mathbf{K}$ remains (asymptotically) identical for functions f sharing the same (a_1, a_2, ν), we compare, for Gaussian random vectors $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, the eigenvalue distribution of $\mathbf{K}$ with the sign function $f(x) = \mathrm{sign}(x)$ in 2(a), as well as the (equivalent) cubic function $\tilde{f}(x)$ in Figure 2(b). The (top) eigenvectors associated to the largest isolated eigenvalues are plotted in 2(c), which turn out to be "noisy" versions of the class structure information $\mathbf{j}_1 - \mathbf{j}_2$. It is also interesting to remark that, even though in the setting of Figure 2 the mixtures differ in both their first and second orders ($\boldsymbol{\mu}_a$ and $\mathbf{E}_a$), one observes only *one* isolated eigenvalue in the spectrum of $\mathbf{K}$, which is here due to the fact that $a_2 = 0$.

4. NUMERICAL VALIDATION

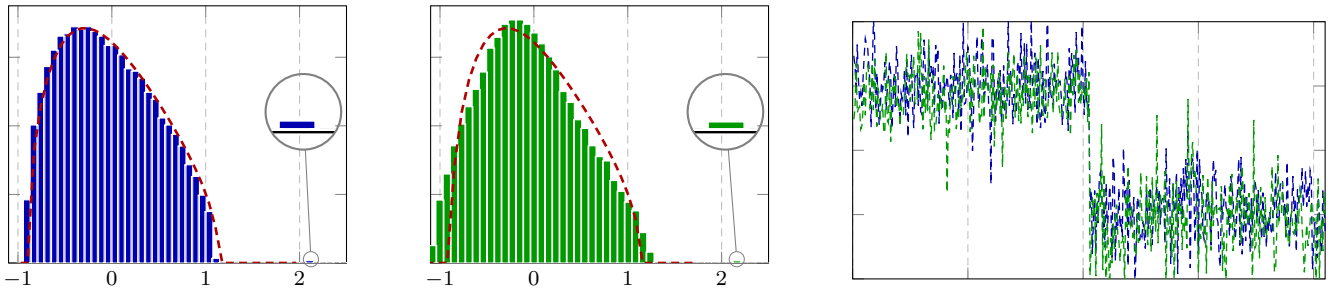
Aiming to find the optimal design for the kernel function f, a direct consequence of Theorem 1 and 2 is that, for any pair (a_1, a_2) the informative matrix $\tilde{\mathbf{K}}_I$ has the same expression, while letting $a_l \neq 0, l \geq 3$, one enlarges the support of the (limiting) eigenvalue distribution of the "noisy" matrix $\mathbf{K}_N$, as per (6). This results in a smaller *eigengap* between the informative and noisy eigenvalues, which deteriorates the performance of any spectrum-based algorithms[20, 21].

In Figure 3 we display the performance of kernel spectral clustering to separate a two-class Gaussian mixture. More precisely, we exploit the top two eigenvectors to form a two-dimensional representation for all the n data and perform a k-means clustering on the resulting 2D representation. Figure 3 shows that the classification error increases monotonously as ν increases and the minimal error rate is achieved, as expected, with $\nu = a_1^2 + a_2^2 (= 2)$ which is the minimal value possible for the given (a_1, a_2). As such, we shall constantly take $\nu = a_1^2 + a_2^2$ which imposes the function $f(x) = c_3x^3 + c_2(x^2 - 1) + c_1x$ to have $c_3 = 0$. Clearly, depending on the values of $\boldsymbol{\mu}_a$ and $\mathbf{E}_a$, the optimal choice of (c_1, c_2) (or equivalently (a_1, a_2)) varies from task to task.

We complete this article by showing that our theoretical understanding, derived from the simple mixture model in (1), generalizes well to some popular real-world datasets. We consider the classification of i) the MNIST handwritten digits (numbers 1 and 5) and ii) the epileptic EEG time series data (sets A and E) [10]. In Figure 4 we depict the kernel spectral clustering error rate as a function of the ratio a_1/a_2 under the condition $\nu = 2 = a_1^2 + a_2^2$. We see that for MNIST data, low error is obtained for large $|a_1/a_2|$; on the contrary, for EEG time series data the error rapidly decreases as a_1/a_2 gets

[3]As a finite rank perturbation of $\mathbf{K}_N$, adding the informative matrix $\tilde{\mathbf{K}}_I$ to $\mathbf{K}_N$ does not change the limiting spectral measure, see [19, Lemma 2.6].

[4]The coefficients are precisely related through $a_1 = 3c_3 + c_1$, $a_2 = \sqrt{2}c_2$ and $\nu = (3c_3 + c_1)^2 + 6c_3^2 + 2c_2^2$.



(a) With sign function (b) With cubic function (c) Top eigenvectors in Figures 2(a) and 2(b)

Fig. 2. Comparison between the eigenvalue distribution of $\mathbf{K}$ and the limiting spectral measure given in Theorem 1 (red), for $n = 2\,048$, $p = 8\,192$, $\mathbf{j}_1 = [\mathbf{1}_{n/2}; \mathbf{0}_{n/2}]$, $\mathbf{j}_2 = \mathbf{1}_n - \mathbf{j}_1$, $\boldsymbol{\mu}_1 = -[3; \mathbf{0}_{p-1}] = -\boldsymbol{\mu}_2$ and $\mathbf{E}_1 = -10\mathbf{I}_p/\sqrt{p} = -\mathbf{E}_2$. $f(x) = \text{sign}(x)$ in Figure 2(a) and $f(x) = \sqrt{(\pi-2)/(6\pi)} \cdot x^3 + (2 - \sqrt{3(\pi-2)})/\sqrt{2\pi} \cdot x$ in Figure 2(b) such that $a_1 = \sqrt{2/\pi}$, $a_2 = 0$ and $\nu = 1$ in both cases.

close to zero. This is because the first ($\boldsymbol{\mu}$) and second order ($\mathbf{E}$) information weighs in a strikingly different manner in each case. This is numerically confirmed in Table 1, where we estimate empirically the differences in means and covariances between the two classes (using all available samples in the class with the empirical mean estimator $\hat{\boldsymbol{\mu}}$ and the sample covariance matrix $\hat{\mathbf{C}}$), for both datasets.

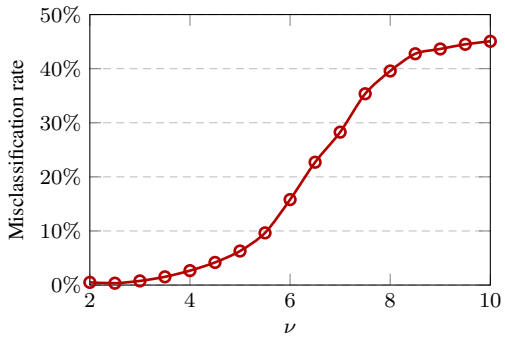


Fig. 3. Kernel spectral clustering error rate as a function of ν for cubic function such that $a_1 = a_2 = 1$. With $n = 512$, $p = 2\,048$ and the same expressions of $\mathbf{j}_1, \mathbf{j}_2, \boldsymbol{\mu}_a, \mathbf{E}_a$ as in Figure 2. Performance obtained by averaging over 50 realizations of Gaussian $\mathbf{Z}$.

Table 1. Empirical estimation of the differences in means and covariances of the MNIST and epileptic EEG datasets.

	$\|\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2\|^2$	$\|\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\|$
MNIST (number 1 versus 5)	464.17	166.35
EEG (set A versus E)	2.41	14.90

5. CONCLUDING REMARKS

Note that, although derived from a local Taylor expansion based on a Gaussian mixture model, the simpler "$\alpha - \beta$" inner-product kernel proposed in [22] and defined by $f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j/p)$ with $f'(0) = \alpha/\sqrt{p}$, $f''(0) = 2\beta$ for $(\alpha, \beta) \in \mathbb{R}^2$ asymptotically behaves similar to the present "properly scaled" inner-product kernel, yet with the additional constraint that $f(x)$ is at least three-times continuously differentiable in a neighborhood of $x = 0$. In a sense we thus extended the results in [22] to cover non-differentiable f with the conclusion that, among this large class of functions f, the second-order polynomials have the best discriminative power (not only for Gaussian mixture, but also for more generic mixture models with bounded moments). This paradoxically suggests that, as far as large dimensional data are concerned, elaborate kernels are less efficient than the simplest quadratic kernels.

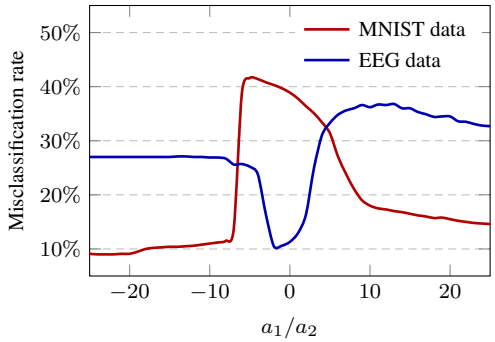


Fig. 4. Kernel spectral clustering error rate versus a_1/a_2 for quartic function $f(x) = a_2x^2 + \sqrt{2}a_1x - a_2$. With $n = 512$, $p = 784$ for MNIST and $p = 100$ for EEG data. Performance obtained by averaging over 50 runs.

As random projections are intimately related to kernels [23], the proposed orthogonal polynomial framework can also be applied to the understanding of nonlinear activation functions in the spectral analysis of large neural networks with random weights. In this respect, the recent line of works [24, 25] showed that the (limiting) eigenvalue distribution of the nonlinear Gram matrix[5] $\sigma(\mathbf{WX})^\mathsf{T}\sigma(\mathbf{WX})$ depends on the nonlinear activation function $\sigma(\cdot)$ via its two Hermite coefficients a_1 and ν, for $\mathbf{W}, \mathbf{X}$ with i.i.d. sub-Gaussian entries. It is thus of future interest to see how the introduction of informative patterns in $\mathbf{X}$ or $\mathbf{W}$ may affect this conclusion and how to extend the present analysis to deeper networks that are of more practical relevance today.

[5] With $\mathbf{W}$ the (random) network weights and $\mathbf{X}$ the input data.

6. REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

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

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

[4] ——, "On information plus noise kernel random matrices," *The Annals of Statistics*, vol. 38, no. 5, pp. 3191–3216, 2010.

[5] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[6] V. A. Marcenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, p. 457, 1967.

[7] X. Cheng and A. Singer, "The spectrum of random inner-product kernel matrices," *Random Matrices: Theory and Applications*, vol. 2, no. 04, p. 1350010, 2013.

[8] Y. Do and V. Vu, "The spectrum of random kernel matrices: universality results for rough and varying kernels," *Random Matrices: Theory and Applications*, vol. 2, no. 03, p. 1350005, 2013.

[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. 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, p. 061907, 2001.

[11] R. Couillet, Z. Liao, and X. Mai, "Classification asymptotics in the random matrix regime," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1875–1879.

[12] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1965, vol. 55.

[13] G. E. Andrews, R. Askey, and R. Roy, *Special functions*. Cambridge university press, 2000, vol. 71.

[14] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[15] E. P. Wigner, "Characteristic vectors of bordered matrices with infinite dimensions," *Annals of Mathematics*, vol. 62, no. 3, pp. 548–564, 1955.

[16] D. Voiculescu, "Addition of certain non-commuting random variables," *Journal of functional analysis*, vol. 66, no. 3, pp. 323–346, 1986.

[17] Z. Fan and A. Montanari, "The spectral norm of random inner-product kernel matrices," *Probability Theory and Related Fields*, vol. 173, no. 1-2, pp. 27–85, 2019.

[18] P. Biane, "On the free convolution with a semi-circular distribution," *Indiana University Mathematics Journal*, pp. 705–718, 1997.

[19] J. W. Silverstein and Z. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate analysis*, vol. 54, no. 2, pp. 175–192, 1995.

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

[21] A. Joseph, B. Yu *et al.*, "Impact of regularization on spectral clustering," *The Annals of Statistics*, vol. 44, no. 4, pp. 1765–1791, 2016.

[22] H. Tiomoko Ali, A. Kammoun, and R. Couillet, "Random matrix-improved kernels for large dimensional spectral clustering," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2018, pp. 453–457.

[23] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[24] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 2637–2646.

[25] L. Benigni and S. Péché, "Eigenvalue distribution of nonlinear models of random matrices," *arXiv preprint arXiv:1904.03090*, 2019.