ASYMPTOTIC SPECTRAL BEHAVIOR OF KERNEL MATRICES IN COMPLEX VALUED OBSERVATIONS

Xavier Mestre, Roberto Pereira, David Gregoratti

Centre Tecnològic de Telecomunicacions de Catalunya
Av. Carl Friedrich Gauss 7, 08860 Castelldefels (Spain)

ABSTRACT

The spectral behavior of kernel matrices built from complex multivariate data is established in the asymptotic regime where both the number of observations and their dimensionality increase without bound at the same rate. The result is an extension of currently available results for inner product based kernel matrices formed from real valued observations to the case where the input data is complex valued. In particular, assuming complex independent standardized Gaussian inputs and imposing certain conditions on the kernel function, it is shown that the empirical distribution of eigenvalues of this type of matrices converges almost surely to a probability measure in this asymptotic domain. Furthermore, the asymptotic spectral density can be obtained by solving a quartic polynomial equation involving its Stieltjes transform and some coefficients depending on the Hermite-like expansion of the kernel function. This is in stark contrast with the equivalent result for real valued observations, in which the underlying polynomial equation is cubic.

Index Terms— Kernel matrices, complex observations, limiting spectrum, Hermite polynomials.

1. INTRODUCTION

Practical problems in modern data science cannot possibly be described using only linear interactions among the data, and this fact that has traditionally motivated the need for non-linear data analytical procedures. The introduction of kernel methods has spurred the development of a number of non-linear extensions to classical linear algorithms, which have been usually referred to as kernel methods [1]. Kernel methods have been shown to outperform linear techniques in a number of classification, regression and structure extraction procedures inherent to modern machine learning. Relevant examples include Spectral Clustering [2, 3, 4, 5], Principal Component Analysis [6, 7, 8, 9], Support Vector Machines [10], Discriminant Analysis [11, 12, 13, 14] or Canonical Correlation Analysis [15, 16, 17, 18].

Kernel-based learning methods conventionally work on the kernel matrix $\mathbf{K}$, which is built from the observations by applying a non-linear function to some distance measure between pairs of observations. More specifically, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a collection of p-dimensional observations and $f(\cdot)$ is a certain non-linear function, the kernel matrix $\mathbf{K}$ is defined as a $n \times n$ random matrix with entries $K_{i,j} = f(d(\mathbf{x}_i, \mathbf{x}_j))$ where $d(\mathbf{x}_i, \mathbf{x}_j)$ is a certain distance measure between the observations $\mathbf{x}_i$ and $\mathbf{x}_j$, usually chosen for real valued data as either $d(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ (inner product kernel matrices) or $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|^2$ (Euclidean distance kernel matrices).

This work has been partially funded by the European Commission under the Windmill project (contract 813999) and the Spanish government under the Aristides project (RTI2018-099722-B-I00).

It turns out that the spectral behavior of the kernel matrix $\mathbf{K}$ (i.e. the behavior of its eigenvalues and eigenvectors) is of fundamental importance in order to determine the performance of the corresponding learning algorithms. Unfortunately, given the inherent non-linearity of the kernel function $f(\cdot)$ and the statistical nature of the observed data, it becomes very difficult to infer statistical properties of $\mathbf{K}$ for fixed p, n. This has motivated a number of studies on the asymptotic behavior of the spectrum of $\mathbf{K}$.

Traditional studies have mainly focused on the behavior of these kernel matrices in the limit as $n \to \infty$ for a fixed observation dimension p. For instance, [19] and [20] studied the relationship between the eigenvalues of $\mathbf{K}$ and the eigenfunctions of the associated integral operator as $n \to \infty$. It was argued in these papers that this connection can be exploited in order to design kernel functions that are specifically tuned for classification problems. Unfortunately, the asymptotic regime in which the sample size $n \to \infty$ for fixed p is not representative in many machine learning applications, especially in high dimensional data problems whereby p may have the same magnitude (or be even larger) than n. This has recently motivated the study of the behavior of the kernel matrix $\mathbf{K}$ under more realistic asymptotic conditions where both the observation dimension p and the number of observations n converge to infinity at the same rate (i.e. $n, p \to \infty$ but $n/p \to \gamma$, where $0 < \gamma < \infty$). These asymptotic approximations are much more realistic in order to study learning mechanisms operating on data sets with the number of observations n commensurable with their dimension p.

Asymptotic studies so far have focused on both inner product and Euclidean distance kernel matrices, although the former seem to have received more attention in the literature. One of the first results for this type of matrix was obtained in [21], which established the convergence of the empirical eigenvalue distribution of the random kernel matrix $\mathbf{K}$ with entries

$$K_{ij} = f\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{p}\right) \delta_{i \neq j} \quad (1)$$

where the observations $\mathbf{x}_j$ are independent zero mean p-dimensional Gaussian random vectors with covariance matrix $\mathbf{C}$ and where $f(\cdot)$ is a sufficiently smooth function. Later, this result was generalized in [22] to the more relevant setting where the observations follow a (non necessarily zero mean) Gaussian mixture model. One of the main concerns with the above model is the fact that, according to the law of large numbers, the entries of the above matrix converge almost surely to 0, and as a consequence the asymptotic eigenvalue distribution of $\mathbf{K}$ only depends on the kernel function $f(\cdot)$ through its local behavior at zero (namely $f(0)$ plus the two derivatives $f'(0)$ and $f''(0)$).

An alternative approach was followed in [23], which established the almost sure convergence of the empirical eigenvalue distribution of the kernel matrix with entries

$$K_{ij} = \frac{1}{\sqrt{p}} f\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\sqrt{p}}\right) \delta_{i \neq j} \tag{2}$$

where the observations $\mathbf{x}_j$ are independent standardized p-dimensional Gaussian random vectors and where $f(\cdot)$ follows some regularity conditions but is not necessarily continuous. One of the main advantages of the above model with respect to (1) is the fact that now the arguments of the kernel function $f(\cdot)$ associated to off-diagonal elements of $\mathbf{K}$ converge in law to a standard Gaussian random variable instead of zero, so that the *global* behavior of $f(\cdot)$ comes into play in the asymptotic spectrum of $\mathbf{K}$. In particular, it was established in [23] that the asymptotic eigenvalue distribution of the kernel matrix can be obtained by solving a third order polynomial equation whose coefficients depend on $f(\cdot)$ via two quantities related to the expansion of $f(\cdot)$ in the Hermite orthogonal polynomial system.

This celebrated result for the model in (2) has recently been extended along multiple directions. For example, in [24] the above result was generalized to kernel matrices built from non-necessarily Gaussian observations. Furthermore, in [25] the authors studied the behavior of the largest eigenvalue of $\mathbf{K}$ and established that it converges to the largest point in the support of the asymptotic eigenvalue distribution. Recently, the analysis of the model in (2) has been extended in [26] to multi-class mixture model for the observations.

All the results reviewed above focused on the spectral analysis of the kernel matrix $\mathbf{K}$ for real valued $f(\cdot)$ and real valued observations $\mathbf{x}_i$. Unfortunately, there exist a number of applications in signal processing where the data to be classified belongs to the complex field, a situation that cannot be studied using the above results. One relevant example is the clustering of wireless communication channels in multi-antenna settings, where n users need to be clustered according to their associated channel as measured in a multi-antenna setting composed of p elements. It has recently been recognized that learning separately from the real and imaginary parts of the observation is clearly suboptimal, and kernels operating on the actual complex data are much more efficient. This has motivated a number of studies on complex valued kernels in a number of applications, including filtering [27, 28], principal component analysis [29] or regression [30, 31].

Quite unexpectedly, it will be shown below that when the observations are drawn from a multivariate circularly symmetric complex distribution, the empirical eigenvalue distribution of $\mathbf{K}$ converges almost surely to a limit that is different from the one obtained with real valued inputs. This is in contrast with conventional random matrix model results, which typically behave in the same way regardless of whether their entries are complex or real valued (as long as they have equivalent moments).

2. PROBLEM STATEMENT AND DEFINITIONS

Our objective is to analyze the asymptotic behavior of the eigenvalue distribution of inner product kernel matrices built from complex observations. In this work, we focus on the case where observations are complex circularly symmetric standardized Gaussian random variables. This can be seen as a first step in order to obtain wider results for the more realistic scenarios, in which observations follow a multi-class mixture model. See for instance [26], where the "uninformative" result (equivalent to the one derived here but for real valued observations) is modified in order to account for different means/covariances in the input data.

More specifically, we consider here the $n \times n$ kernel matrix $\mathbf{K}$ with entries[1]

$$K_{ij} = \frac{1}{\sqrt{p}} k\left(\frac{\mathbf{x}_i^H \mathbf{x}_j}{\sqrt{p}}\right) \delta_{i \neq j} \tag{3}$$

where $\mathbf{x}_i$ are independent circularly symmetric standard complex Gaussian random vectors (i.e. they have independent real and imaginary parts with zero mean and variance $0.5\mathbf{I}_p$) and where $k(z)$ is a certain (generally complex valued) kernel function of complex variable. The main difference between the above model and (2) is the fact that we consider complex observations, so that the argument of $k(z)$ is a complex number. We will generally assume that the complex function may take complex values and is such that $k(z^*) = k^*(z)$. This implies that the kernel matrix $\mathbf{K}$ is Hermitian, so that it has real valued eigenvalues.

It is important to point out that the complex model in (3) cannot be obtained as a particular instance of (2) because there is no direct way of expressing $\mathbf{x}_i^H \mathbf{x}_j$ in terms of the scalar product of the corresponding real valued components. Indeed, if we write $\mathbf{x}_i = \mathbf{u}_i + \mathrm{i}\mathbf{v}_i$ (where $\mathbf{u}_i$ and $\mathbf{v}_i$ contain the real and imaginary parts of $\mathbf{x}_i$), we have

$$\mathbf{x}_i^H \mathbf{x}_j = \left(\mathbf{u}_i^T \mathbf{u}_j + \mathbf{v}_i^T \mathbf{v}_j\right) + \mathrm{i}\left(\mathbf{u}_i^T \mathbf{v}_j - \mathbf{v}_i^T \mathbf{u}_j\right) \tag{4}$$

and we see here that the real valued model is only valid to describe the complex one when $k(z)$ is a real valued function of either $\mathrm{Re}(z)$ or $\mathrm{Im}(z)$.

In some occasions, it will be effective to consider $k(z)$ as a function of the real and imaginary parts of z. Throughout the paper, x and y will denote the real and imaginary parts of the complex variable z, so that $z = x + \mathrm{i}y$. Furthermore, we will identify

$$k(z) = k(x + \mathrm{i}y) = f(x, y)$$

where $f(x, y)$ is now a bivariate function of real valued variables.

2.1. Orthogonal polynomials

The spectral convergence result presented in this paper holds for a wide family of kernel functions $k(z)$, which do not even need to be continuous. However, as it will be made explicit below, we need $k(z)$ to have some regularity conditions with respect to the probability measures of the real and imaginary parts of $\mathbf{x}_i^H \mathbf{x}_j / \sqrt{p}$, which will be denoted as μ_p and ν_p respectively. We will denote by $\{P_{l,p}, l \geq 0\}$ and $\{Q_{l,p}, l \geq 0\}$ the orthonormal polynomials associated to these two probability measures, which can be obtained via the conventional Gram-Schmidt orthogonalization procedure.

By the central limit theorem, as $p \to \infty$ the random variable $\mathbf{x}_i^H \mathbf{x}_j / \sqrt{p}$ converges in law to a complex circularly symmetric standardized Gaussian random variable, which has probability measure $\pi^{-1} \exp\left(-|z|^2\right) dxdy$. This means that as $p \to \infty$ the coefficients of the polynomials $P_{l,p}(x)$ and $Q_{l,p}(x)$ will both converge to those of the orthonormal polynomials associated with the probability measure $\pi^{-1/2} \exp\left(-x^2\right) dx$, see further [23, Lemma 4.1]. We will denote as $h_m(x)$ the m-th normalized orthogonal polynomial associated to this measure (sometimes this polynomial is referred to as the "physicist" Hermite polynomial of order m), which is such that

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} h_m(x) h_n(x) \exp\left(-x^2\right) dx = \delta_{m-n}.$$

[1] Here and throughout the paper z^* indicates the complex conjugate of z and $(\cdot)^H$ complex conjugate transpose.

We can easily find a closed form expression for the coefficients of these polynomials as

$$h_m(x) = \sqrt{\frac{m!}{2^m}} \sum_{l=0}^{\lfloor m/2 \rfloor} \frac{(-1)^l}{l!\,(m-2l)!} (2x)^{m-2l} . \tag{5}$$

In particular, we have $h_0(x) = 1$, $h_1(x) = \sqrt{2}x$ and $h_2(x) = \sqrt{2}(x^2 - 1/2)$.

3. MAIN RESULT

Having introduced the main definitions that are relevant to our problem, we are now in a position to formulate the main assumptions regarding the observations as well as the kernel function $k(z)$ (equivalently $f(x, y)$).

(As1) The observations $\mathbf{x}_i$ are modeled as independent circularly symmetric standard Gaussian complex random vectors, so that they have independent real and imaginary parts with zero mean and variance $0.5\mathbf{I}_p$.

(As2) The kernel function $f(x, y)$ is square integrable with respect to the probability measure $\mu_p \times \nu_p$, so that we can define (by Fubini)

$$a_{k,l}^{(p)} = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) P_{k,p}(x) P_{l,p}(y) d\mu_p(x) d\nu_p(y)$$

for $k, l \geq 0$. Then, the double series $\sum_{k,l\geq 0} a_{k,l}^{(p)} P_{k,p}(x) Q_{l,p}(y)$ converges uniformly in p in the family of square integrable functions with respect to the measure $\mu_p \times \nu_p$. In other words, for any $\epsilon > 0$ there exist p_0 and L such that

$$\sum_{k,l\geq L+1} \left| a_{k,l}^{(p)} \right|^2 < \epsilon$$

for any $p > p_0$.

(As3) The coefficients $a_{k,l}^{(p)}$ in the above expansion converge as $p \to \infty$ for $k, l \in \{0, 1\}$. We will denote as $a_{k,l}$ the corresponding limits, which can also be expressed as

$$a_{k,l} = \frac{1}{\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) h_k(x) h_l(y) \exp\left(-|z|^2\right) dxdy$$

with $h_k(x)$ as defined above. Furthermore, we will assume that the following double series also converges to a non-negative real number C as $p \to \infty$

$$\sum_{k,l\geq 0} \left| a_{k,l}^{(p)} \right|^2 \to C > 0$$

and we will consider $a_{0,0}^{(p)} = 0$.

The last assumption implies that $k(z)$ has zero mean with respect to the complex circularly symmetric Gaussian measure, that is

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k(z) \exp\left(-|z|^2\right) dxdy = 0.$$

Contrary to what it may seem, this assumption is irrelevant for the purposes of establishing the asymptotic eigenvalue distribution, since if $a_{0,0}^{(p)} \neq 0$ we can always rewrite our kernel matrix as

$$\mathbf{K} = \mathbf{K}_0 + \frac{a_{0,0}^{(p)}}{\sqrt{p}} \left(1_n 1_n^T - \mathbf{I}_n\right) \tag{6}$$

where $\mathbf{K}_0$ is a kernel matrix built with the function $k(z) - a_{0,0}^{(p)}$ instead of $k(z)$. Assuming that $a_{0,0}^{(p)}$ is bounded, the two matrices $\mathbf{K}$ and $\mathbf{K}_0$ have the same asymptotic eigenvalue distribution, since they are equivalent up to a negative displacement $a_{0,0}^{(p)}/\sqrt{p} \to 0$ (note that rank one perturbations do not alter the asymptotic eigenvalue distribution).

Theorem 1. *Let $p, n \to \infty$ with $n/p \to \gamma$, $0 < \gamma < \infty$ and assume that $(\mathbf{As1})$-$(\mathbf{As3})$ hold. With probability one, the empirical eigenvalue distribution of $\mathbf{K}$ converges weakly to a probability measure ξ, uniquely determined by its Stieltjes transform $m(z) = \int_{\mathbb{R}^+} (t - z)^{-1} d\xi(t)$ for $z \in \mathbb{C}^+$ (the upper complex semi-plane) as the unique solution in $\mathbb{C}^+$ to the following quartic equation*

$$\frac{-1}{m(z)} = z + \gamma m(z)\omega + \left(\frac{|\tilde{\alpha}|^2}{1 + \tilde{\alpha}\gamma m(z)} + \frac{|\alpha|^2}{1 + \alpha\gamma m(z)} \right) \gamma m(z) \tag{7}$$

where

$$\omega = \sum_{q,r=0}^{\infty} |a_{q,r}|^2 - |a_{1,0}|^2 - |a_{0,1}|^2$$

and where we have introduced the two complex coefficients

$$\alpha = \frac{1}{\sqrt{2}} (a_{1,0} + \mathrm{i} a_{0,1}) \text{ and } \tilde{\alpha} = \frac{1}{\sqrt{2}} (a_{1,0} - \mathrm{i} a_{0,1}) .$$

Proof. The proof essentially follows from the proof in [23]. More details are provided in the supplementary material. □

Theorem 1 offers a simple method to retrieve the asymptotic eigenvalue distribution of the kernel matrix $\mathbf{K}$. Indeed, one only needs to solve the equation in (7) by a polynomial rooting procedure to obtain $m(z)$ and use the inverse Stieltjes method to obtain the asymptotic density of eigenvalues as

$$\frac{d\xi(\lambda)}{d\lambda} = \lim_{y\to 0^+} \frac{1}{\pi} \operatorname{Im}\left[m(\lambda + \mathrm{i}y)\right] .$$

In this sense, it is somewhat surprising to see that the Stietjes transform $m(z)$ is obtained as a solution to a quartic polynomial equation, which is in contrast with the cubic equation that fully describes the real valued case (see [23]). This result is quite exceptional if we compare it to more common random matrix theory models, where the asymptotic eigenvalue density is typically the same regardless of whether the associated entries are real or complex valued (up to moment equivalence). Here, this is clearly not the case, and we see that the asymptotic eigenvalue distribution formed with complex entries leads to a completely different spectral behavior.

The asymptotic eigenvalue distribution of the kernel matrix $\mathbf{K}$ depends on the kernel function through three different parameters, namely α, $\tilde{\alpha}$ and ω. If $\zeta \sim \mathcal{CN}(0, 1)$ denotes a circularly symmetric complex standard Gaussian random variable, these parameters can be expressed explicitly with respect to the kernel function as

$$\alpha = \mathbb{E}\left[\zeta k(\zeta)\right] \quad \tilde{\alpha} = \mathbb{E}\left[\zeta^* k(\zeta)\right]$$

and

$$\omega = \mathbb{E}\left[k^2(\zeta)\right] - \frac{|\alpha + \tilde{\alpha}|^2}{2} - \frac{|\alpha - \tilde{\alpha}|^2}{2}.$$

It is interesting to particularize the above equation to the case where the kernel function $k(z)$ is real valued (still of complex variable). Whenever this is the case, the two coefficients $\alpha, \tilde{\alpha}$ become the complex conjugate of one another, so that ω takes the simpler form

$$\omega = \operatorname{Var}\left(k(\zeta)\right) - 2\left|\alpha\right|^2 .$$

Note, however, that even when the kernel function is real valued, the polynomial equation (7) is still quartic (rather than cubic, as in the real valued case). We will see below that it is possible to retrieve the real valued result in [23] as a particular instance of Theorem 1.

3.1. Recovering the conventional real valued result

We can try to recover the original result from this, by observing that, according to (4) we are able to write

$$2\,\mathrm{Re}\left(\mathbf{x}_i^H\mathbf{x}_j\right) = \mathbf{w}_i^T\mathbf{w}_j$$

where $\mathbf{w}_i = \sqrt{2}[\mathbf{u}_i^T, \mathbf{v}_i^T]^T$ is a $2p$-dimensional real vector of independent and identically distributed standardized Gaussian entries. Now, we can particularize Theorem 1 to the specific choice

$$k(z) = \frac{1}{\sqrt{2}} g\left(\sqrt{2}\,\mathrm{Re}(z)\right)$$

where g is now a real valued function, so that

$$K_{ij} = \frac{1}{\sqrt{p}} k\left(\frac{\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{p}}\right)\delta_{i\neq j} = \frac{1}{\sqrt{2p}} g\left(\frac{\mathbf{w}_i^T\mathbf{w}_j}{\sqrt{2p}}\right)\delta_{i\neq j}.$$

With this choice, the kernel matrix $\mathbf{K}$ fits into the conventional real valued model in (2) but with observation dimensionality equal to $2p$ instead of p. As $p \to \infty$, the coefficients $a_{k,l}^{(p)}$ will converge to $a_{k,l} = \frac{1}{\sqrt{2}} a_k \delta_l$ with a_k defined as

$$a_k = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g\left(x\right) h_k\left(\frac{x}{\sqrt{2}}\right) \exp\left(-\frac{x^2}{2}\right) dx$$

as $p \to \infty$, which coincides with the corresponding coefficient for the real valued series expansion in [23]. With these definitions, we readily see from Theorem 1 that the empirical eigenvalue distribution of $\mathbf{K}$ converges towards a measure with Stieltjes transform as the unique solution in $\mathbb{C}^+$ to the following equation

$$\frac{-1}{m(z)} = z + \frac{\gamma}{2} m(z) \sum_{k\geq 1} a_k^2 + \frac{\gamma}{2} \frac{a_1^2 m(z)}{1 + a_1 \frac{\gamma}{2} m(z)}$$

which is the same cubic equation that was originally obtained in [23] (with the equivalent of γ replaced by $\gamma/2$).

4. PARTICULARIZATION TO SOME SPECIFIC KERNEL FUNCTIONS

4.1. Gaussian radial basis function

Consider the complex kernel function

$$k(z) = \exp\left(-\lambda\left|z\right|^2\right)$$

where λ is a certain positive parameter. We observe here that for this particular kernel choice, it is not true that $a_{0,0}^{(p)} = 0$. However, as explained above, this does not really affect the asymptotic eigenvalue distribution. Using the definition of the polynomials in (5) we can readily establish that $a_{1,0} = a_{0,1} = 0$ and also

$$\omega = \frac{1}{1+2\lambda} \frac{\lambda^2}{(1+\lambda)^2}.$$

This implies that the empirical density of eigenvalues converges to a probability measure with Stieltjes transform

$$\frac{1}{m(z)} = -z - \gamma m(z) \frac{1}{1+2\lambda} \frac{\lambda^2}{(1+\lambda)^2}$$

which is a semicircle law with density

$$\lim_{y\to 0^+} \frac{1}{\pi}\,\mathrm{Im}\left[m(x+\mathrm{i}y)\right] = \frac{1}{2\pi\gamma\omega}\sqrt{4\gamma\omega - x^2}$$

on the support $\left(-2\sqrt{\gamma\omega}, 2\sqrt{\gamma\omega}\right)$. For the real valued case, where $f(x) = \exp\left(-\lambda x^2\right)$ it can readily be shown that the eigenvalue density also converges to a semi-circle law with the same density, but now with

$$\omega = \frac{1}{\sqrt{1+4\lambda}} - \frac{1}{1+2\lambda}.$$

Hence, even if we obtain a semi-circular law in both cases, we see that the dependence of the support of the asymptotic eigenvalue distribution on the kernel parameter λ is significantly different.

4.2. Sign function

To illustrate a situation in which the coefficients α, $\tilde{\alpha}$ are different from zero, we consider here a direct extension of the sign kernel to the complex domain by choosing the complex kernel matrix as $k(z) = k(x+\mathrm{i}y) = \mathrm{sign}(x) + \mathrm{i}{\cdot}\mathrm{sign}(y)$. In this case, we can readily see that

$$a_{1,0} = \sqrt{\frac{2}{\pi}}, \quad a_{0,1} = \mathrm{i}\sqrt{\frac{2}{\pi}}.$$

On the other hand, it can easily be reasoned that $\sum_{m,n\geq 0}\left|a_{m,n}\right|^2 = 2$. This means that the Stieltjes transform $m(z)$ can be obtained by finding the only root of the following cubic equation on the upper complex plane:

$$\frac{-1}{m(z)} = z + 2\left(1 - \frac{2}{\pi}\right)\gamma m(z) + \frac{4}{\pi}\frac{\gamma m(z)}{1 + \frac{2}{\sqrt{\pi}}\gamma m(z)}.$$

The corresponding equation for real valued observations can be obtained by choosing $k(z) = \frac{1}{\sqrt{2}}\mathrm{sign}\left(\sqrt{2}\,\mathrm{Re}(z)\right)$ and noting that, in this situation, $a_{1,0} = \frac{1}{\sqrt{\pi}}$, $a_{0,1} = 0$ and $\sum_{m,n\geq 0}\left|a_{m,n}\right|^2 = \frac{1}{2}$. This means that $m(z)$ is a solution to the equation

$$\frac{-1}{m(z)} = z + \left(1 - \frac{2}{\pi}\right)\gamma m(z) + \frac{2}{\pi}\frac{\gamma m(z)}{1 + \sqrt{\frac{2}{\pi}}\gamma m(z)}$$

which is slightly different from the equation obtained in the complex domain. In the numerical evaluation section below, we will compare the two corresponding eigenvalue densities and point out the differences.

4.3. Quadratic polynomial

We finally particularize the results in Theorem 1 to the choice $k(z) = \lambda\left|z\right|^2$. In this particular case, one can readily establish that

$$a_{m,n} = \lambda\left(\frac{1}{2}\delta_m + \frac{1}{\sqrt{2}}\delta_m\right)\left(\frac{1}{2}\delta_n + \frac{1}{\sqrt{2}}\delta_n\right)$$

which directly implies that we have $a_{0,1} = a_{1,0} = 0$, while $\omega = \frac{\lambda^2}{2}$. This means that $m(z)$ is obtained from the equation

$$\frac{1}{m(z)} = -z - \gamma m(z) \left(\frac{\lambda^2}{2} \right)$$

which is, again, a semicircle law. A similar conclusion can be obtained for the case where the observations are real valued [32].

5. NUMERICAL VALIDATION

In order to illustrate the result in Theorem 1 we considered some examples of kernel matrices generated with the kernel functions discussed above. For complex valued observations, real and imaginary parts of the observations are independently generated with zero mean and variance $1/2$, whereas real valued observations were generated also with zero mean but with variance 1. Both the Gaussian radial basis function (with $\lambda = 2$) and the sign kernels presented above were considered in the simulations. Figure 1 compares the histogram of the eigenvalues obtained for some specific choices of n and p against the asymptotic eigenvalue density as established in Theorem 1 (solid red lines).

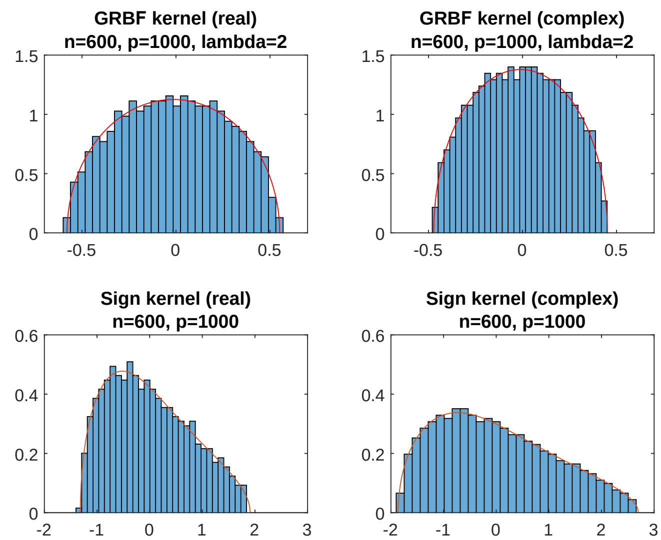
Observe that there is a very good match between the asymptotic and the empirical eigenvalue distribution regardless of whether $p > n$ or $p < n$. Furthermore, these plots also exemplify the fact that the eigenvalue density behaves essentially differently in the real and complex value cases, especially with regards to the support of the asymptotic eigenvalue distribution. The determination of this support will be crucial in order to extend the present work to the case where the observations follow a multi-class mixture model.

An interesting issue related to the kernel matrix formed with the Gaussian radial basis function is the fact that an isolated eigenvalue appears outside the support of the asymptotic eigenvalue distribution. This eigenvalue has not been represented in the histograms of Figure 1. The presence of this eigenvalue, which of course does not contribute to the asymptotic eigenvalue distribution, is caused by the fact that $a_{00} \neq 0$ in this particular choice of kernel function. According to the reasoning in (6), this fact originates an additional eigenvalue at $\sqrt{p}a_{0,0}$, which will be arbitrarily large as $p \to \infty$. As it can be seen from (6), this eigenvalue is associated to an all-ones eigenvector, which can be easily canceled out by subtracting the empirical mean along rows and columns of $\mathbf{K}$.

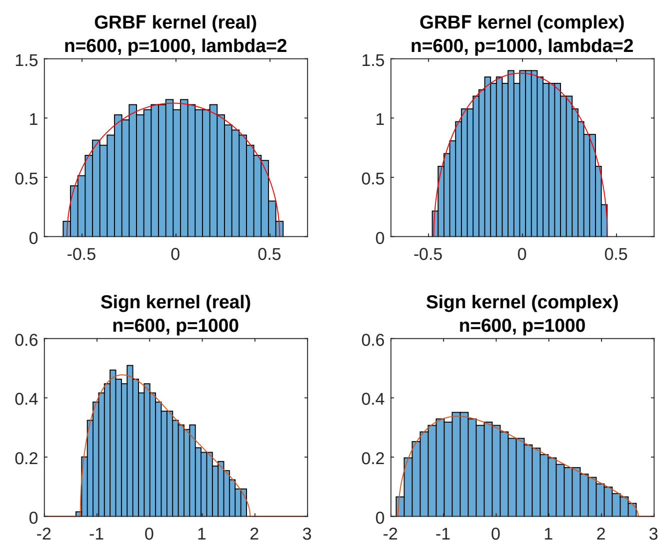
6. CONCLUSIONS

The asymptotic eigenvalue distribution of inner product based kernel matrices formed from complex observations has been studied in the asymptotic regime whereby both the observation dimension p and the number of input data n increase without bound at the same rate. It has been shown that the empirical distribution of eigenvalues converges almost surely to a probability measure with Stieltjes transform that can be obtained from a quartic polynomial equation. This contrasts with the real valued observation case, where the Stieltjes transform is a root of a cubic polynomial instead.

The effect of the kernel function on the asymptotic eigenvalue distribution is established through three different parameters (α, $\tilde{\alpha}$ and ω) which are directly related to its bivariate expansion in terms of Hermite polynomials. The result generalizes the real valued observation case, which can be retrieved as a particular case of the presented result, by restricting the kernel function to be real valued of a real variable. The result has been particularized to three different choices of kernel function, namely Gaussian radial basis function, sign function and quadratic polynomial respectively. Numerical simulations confirm the validity of the presented results and illustrate the different behavior of real valued and complex valued kernel matrices. Results will be useful towards the characterization of kernel-based machine learning algorithms working on complex signals, for which conventional results based on real valued observations are not applicable.



(a) Case $n = 1000$ and $p = 600$.



(b) Case $n = 600$ and $p = 1000$.

Fig. 1. Comparison between eigenvalue histograms and asymptotic eigenvalue distribution for different kernel functions and values of p and n.

7. REFERENCES

[1] B. Scholkopf and A. Smola, *Learning with kernels: support vector machines, regularization, optimization and beyond.* United States of America: MIT press, 2001.

[2] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proceedings of the Neural Information Processing Systems Conference*, pp. 849–856, 2002.

[3] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nystrom method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 214–225, Feb 2004.

[4] C. Alzate and M. Sinn, "Improved electricity load forecasting via kernel spectral clustering of smart meters," in *2013 IEEE 13th Int. Conference on Data Mining*, pp. 943–948, Dec 2013.

[5] D. S. Satish and C. C. Sekhar, "Kernel based clustering and vector quantization for speech recognition," in *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pp. 315–324, Sept 2004.

[6] D. Widjaja, C. Varon, A. Dorado, J. A. K. Suykens, and S. V. Huffel, "Application of kernel principal component analysis for single-lead-ECG-derived respiration," *IEEE Trans. on Biomedical Engin.*, vol. 59, pp. 1169–1176, April 2012.

[7] Q. Wang, "Kernel principal component analysis and its applications in face recognition and active shape models," *arXiv*, no. arXiv:1207.3538, 2012. https://arxiv.org/pdf/1207.3538.pdf.

[8] S. Choi, C. Lee, J. Lee, J. Park, and I. Lee, "Fault detection and identification of nonlinear processes based on kernel PCA," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 1, pp. 55–67, 2005.

[9] Y. Gu, Y. Liu, and Y. Zhang, "A selective kernel PCA algorithm for anomaly detection in hyperspectral imagery," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2, pp. II–II, May 2006.

[10] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[11] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Transactions on Signal Processing*, vol. 51, pp. 1702–1715, 2003.

[12] J. Yang, A. F. Frangi, J.-Y. Yang, D. Zhang, and Z. Jin, "Kpca plus lda: a complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 230–244, Feb 2005.

[13] L. Wei, Y. Yang, R. M. Nishikawa, and Y. Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 371–380, March 2005.

[14] M. A. Tahir, E. Khan, and A. A. Salem, "Medical text categorization using sebla and kernel discriminant analysis," in *2015 2nd World Symposium on Web Applications and Networking (WSWAN)*, pp. 1–6, March 2015.

[15] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, 2012.

[16] Y. Yamanishi, J. Vert, and M. Kanehisa, *Kernel Methods in Computational Biology*, ch. Heterogeneous Data Comparison and Gene Selection with Kernel Canonical Correlation Analysis, p. 209–229. MIT Press, 2004.

[17] Y. Li and J. Shawe-Taylor, "Using KCCA for japanese-english cross-language information retrieval and classification," *Journal of Intelligent Information Systems*, vol. 27, no. 2, p. 117–133, 2005.

[18] F. Bach and M. Jordan, "Kernel independent component analysis," *Jour. Machine Learn. Res.*, vol. 3, pp. 1–48, 2002.

[19] C. Williams and M. Seeger, "The effect of the input density distribution on kernel-based classifiers," in *International Conference on Machine Learning*, vol. 17, pp. 1159–1166, 2000.

[20] V. Koltchinskii and E. Gine, "Random matrix approximation of spectra of integral operators," *Bernouilli*, no. 6, pp. 113–167, 2000.

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

[22] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, pp. 1593–1454, 2016.

[23] X. Cheng and A. Singer, "The spectrum of random inner-produuct kernel matrices," *Random Matrices: Theory and Applications*, vol. 02, no. 04, 2013.

[24] 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. 3, 2013.

[25] Z. Fan and A. Montanari, "The spectral norm of random inner-product kernel matrices," *Probability Theory and Related Fields*, 2018.

[26] Z. Liao and R. Couillet, "On inner-product kernels of high dimensional data," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 579–583, 2019.

[27] P. Bouboulis and S. Theodoridis, "Extension of Wirtinger's calculus to reproducing kernel Hilbert spaces and the complex kernel LMS," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 964–978, 2011.

[28] T. Ogunfunmi and T. Paul, "On the complex kernel-based adaptive filter," in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, pp. 1263–1266, 2011.

[29] A. Papaioannou and S. Zafeiriou, "Principal component analysis with complex kernel: The widely linear model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1719–1726, 2014.

[30] F. A. Tobar, A. Kuh, and D. P. Mandic, "A novel augmented complex valued kernel LMS," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 473–476, 2012.

[31] R. Boloix-Tortosa, J. J. Murillo-Fuentes, I. Santos, and F. Perez-Cruz, "Widely linear complex-valued kernel methods for regression," *IEEE Transactions on Signal Processing*, vol. 65, no. 19, pp. 5240–5248, 2017.

[32] A. Kammoun and R. Couillet, "Subspace kernel spectral clustering of large dimensional data," *Submitted to the Annals of Applied Statistics*, 2020.