HIGH DIMENSIONAL ROBUST CLASSIFICATION: A RANDOM MATRIX ANALYSIS

Romain Couillet

GIPSA-lab, University Grenoble-Alpes
CentraleSupélec, University ParisSaclay.

ABSTRACT

This article proposes a random matrix analysis of the spectral properties of a new robust kernel matrix model adapted to elliptically distributed large dimensional data mixtures. It is shown that these kernel matrices, based on robust estimators of scatter, when finely tuned, can perform asymptotic non-trivial (unsupervised) classification while sample covariance matrices are ineffective. Unlike in conventional robust statistics wisdom though, the "maximally robust" estimators (such as Tyler's estimator of scatter) also break asymptotic classification feasibility. This entails the existence of an optimal robustness-classification trade-off which we discuss.

Index Terms— Random matrix theory; robust statistics; classification.

1. INTRODUCTION

Machine learning algorithms for (supervised or unsupervised) classification draw their strengths from the information redundancy contained in the numerous data from each class. Spectral clustering methods [1] precisely rely on this idea: the induced redundancy of the data "pulls" the dominant eigenvectors of the affinity (or kernel) matrix $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ (pairwise comparing through $\kappa(\cdot,\cdot)$ the n data vectors $x_1, \ldots, x_n \in \mathbb{R}^p$) to be strongly aligned to the canonical vectors ($j_1, \ldots, j_k \in \mathbb{R}^n$ with $[j_a]_i = \delta_{x_i \in \mathcal{C}_a}$) of the classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ [2].

Until recently though, the non-linearity of κ and the non-trivial dependence in the entries of K have prevented statisticians to fully comprehend the statistical behavior of the eigenvectors of such a matrix K, even in simple model settings. New techniques in random matrix theory [3] have lately emerged that now allow for a deep understanding of spectral clustering methods (and related kernel classification and learning [4, 5, 6]) under the regime where both data size p and number n are simultaneously large.

These recent findings, mostly envisioned under a Gaussian mixture model for the data so far, now open the door to more elaborate data modelling considerations. The question raised in this article concerns classification of a mixture model under an *elliptical* noise setting. That is, we assume that data vectors belonging to class $\mathcal{C}_a$ are of the type $x_i = \mu_a + \sqrt{\tau_i} C^{\frac{1}{2}} w_i$ for $\mu_a \in \mathbb{R}^p$ the class mean, $C \in \mathbb{R}^{p \times p}$ a common "covariance" (or scatter) matrix, $w_i \in \mathbb{R}^n$ random of fixed norm and $\tau_i > 0$ modelling noise impulses of arbitrary amplitude. These models are standard in array processing, and notably in SAR and hyperspectral imaging [7]. However, most conventional detection and estimation algorithms suffer to cope with heavy-tailed distributions, which have instead been treated in the literature by means of robust statistical methods [8]. It is in particular known that the family of M-estimators of scatter, originally devised by Huber, Maronna and Tyler [9, 10, 11], are appropriate "robust" estimators for the covariance matrix C, with Tyler's estimator claimed as the "most robust" of the family.

The objective of this article is to demonstrate that these M-estimators of scatter can be effectively used to perform unsupervised classification for the aforementioned data model, where conventional kernel methods provably fail. The fundamental intuition behind this claim lies in a finding from [12]: when the covariance matrix C of a data model $x_i = \sqrt{\tau_i} C^{\frac{1}{2}} w_i$ assumes the form $C = I_p + \lambda v v^\mathsf{T}$ with $v \in \mathbb{R}^p$ of unit norm, then the asymptotic spectrum of the robust estimator of scatter exhibits an isolated largest eigenvalue when λ exceeds a threshold and the associated eigenvector is aligned to u; this does not hold for the sample covariance matrix from which it is impossible to estimate v by principle component analysis. Our main contribution is to rigorously demonstrate that a similar phenomenon arises in the present classification setting. The results are however not completely straightforward: while our proposed kernel approach based on M-estimators of scatter is shown to allow for non-trivial class recovery, a robustness-classification trade-off must be found. Indeed, unlike in the $C = I_p + \lambda v v^\mathsf{T}$ model, maximally robust estimators, such as Tyler's estimator, tend to "break" class structures, while non-robust estimators fail to bring the aforementioned data redundancy into the dominant kernel matrix eigenvectors. The three theorems presented in the article provide an accurate evaluation of the large n, p spectral behavior of the "robust kernel" model under study, making it possible to determine the robust estimator achieving optimal asymptotic classification rate.

The proofs of all results are deferred to an extended version of the article.

2. MODEL AND PRELIMINARY RESULTS

Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ arising from a k-class mixture model, with $X_1 = [x_1, \ldots, x_{n_1}]$ in class $\mathcal{C}_1$, ..., $X_k = [x_{n-n_k+1}, \ldots, x_n]$ in class $\mathcal{C}_k$, where

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i = \mu_a + \sqrt{\tau_i} C^{\frac{1}{2}} w_i$$

for $w_i \in \mathbb{R}^p$ distributed uniformly at random on the $\sqrt{p}$-radius sphere $\sqrt{p}\mathcal{S}^{p-1}$ of $\mathbb{R}^p$, $\mu_a \in \mathbb{R}^p$ and $C \in \mathbb{R}^{p \times p}$ deterministic, and $\tau_i > 0$ deterministic (or random independent of w_i) and independent of p.

The additional parameters τ_i provide a natural extension of Gaussian mixture models with possibly heavy noise tails. The objective is to classify the data in an unsupervised manner, in the

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

regime where both p and n are large. To this end, as in [3], we request that $p, n \to \infty$ in such a way that $p/n \to c \in (0,1)$, $\|C\| = O(1)$ and, for each a, $n_a/n = O(1)$ and $\|\mu_a\| = O(1)$. The requirement $\|\mu_a\| = O(1)$ ensures that classification does not become asymptotically impossible nor too easy, as $p \to \infty$.

Our proposed object of interest is the inner-product kernel

$$K = \frac{1}{n} D^{\frac{1}{2}} X^\mathsf{T} X D^{\frac{1}{2}}$$

where D is diagonal and defined through the fixed-point system

$$D = \mathrm{diag}(\{d_i\}_{i=1}^n),\ d_i = u\left(\frac{1}{p} x_i^\mathsf{T} \hat{C}^{-1} x_i\right),\ \hat{C} = \frac{1}{n} X D X^\mathsf{T}$$

with $u : \mathbb{R}_+ \to \mathbb{R}_+$ nonincreasing and such that $\varphi(t) = tu(t)$ is increasing and bounded as $|\varphi|_\infty < \frac{1}{c}$. Under these conditions, it is proved that the d_i's are well defined and unique; the matrix $\hat{C}$ is known as Maronna's estimator of scatter, extensively studied in [10, 13].

Our first result shows that, as $n, p \to \infty$, the scalars d_i concentrate *independently of the μ_a's*.

Theorem 1 (Limit of d_i) *As $n, p \to \infty$, we have*

$$\max_{1 \le i \le n} |d_i - v(\tau_i \gamma)| = o_p(1)$$

with γ defined as the unique positive solution to

$$1 = \frac{1}{n} \sum_{i=1}^n \frac{\psi(\tau_i \gamma)}{1 + c\psi(\tau_i \gamma)}$$

with $\psi(t) = tv(t)$ and $v = u \circ g^{-1}$ for $g : \mathbb{R}_+ \to \mathbb{R}_+$, $t \mapsto \frac{t}{1 - c\varphi(t)}$ and g^{-1} its functional inverse.

Despite the presence of the μ_a's, this is the same result as in [13] where $\mu_1 = \ldots = \mu_k = 0$.

In plain words, Theorem 1 states that the d_i's, and thus $\hat{C}$ and K, have a controllable asymptotic behavior as $n, p \to \infty$. In particular, assume now for simplicity that the τ_i's are i.i.d. with law $\mathcal{T}$. Then Theorem 1 may be restated as

$$\max_{1 \le i \le n} |d_i - v(\tau_i \gamma)| = o_p(1)$$

where γ is the unique positive solution to

$$1 = \int \frac{\psi(t\gamma)}{1 + c\psi(t\gamma)} \mathcal{T}(dt).$$

Under these conditions, the spectral measure $\frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(K)}$ of the eigenvalues of K is known to have an almost sure weak limit $\mathcal{K}$ [14]. The latter is quantified through its Stieltjes transform $m(z) = \int (t - z)^{-1} \mathcal{K}(dt)$ for $z \in \mathbb{C} \setminus \mathrm{supp}(\mathcal{K})$, defined as the unique solution of the equation

$$m(z) = \left(-z + \frac{1}{\gamma} \int \frac{\psi(t\gamma) \mathcal{T}(dt)}{1 + c\frac{m(z)}{\gamma} \psi(t\gamma)} \right)^{-1}. \tag{1}$$

The value $x_+ \equiv \sup(\mathrm{supp}\mathcal{K})$ can be determined as $x_+ = x(m_+)$ for $x(\cdot) : \mathbb{R} \setminus \{m | 0 \in 1 + c\frac{m}{\gamma} \psi(\gamma \mathcal{T})\} \to \mathbb{R}$ the explicit function

$$x(m) = -\frac{1}{m} + \frac{1}{\gamma} \int \frac{\psi(t\gamma) \mathcal{T}(dt)}{1 + c\frac{m}{\gamma} \psi(t\gamma)}$$

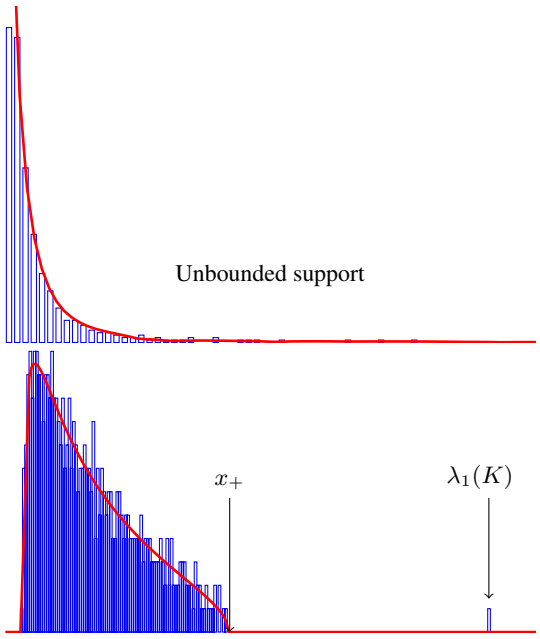

Fig. 1. Empirical eigenvalue distribution of K (histrogram) versus limiting spectral measure $\mathcal{K}$ (red) for $u(t) = 1$ (non robust, **top**) versus $u(t) = 2/(1+t)$ (robust, **bottom**). Here for τ_i Student distributed with unit degree of freedom, $k = 3$, $n = 2000$, $p = 500$, $n_1 = n_2 = n_3/2$, $[\mu_a]_i = 4\delta_{ai}$.

and $m_+ = \sup\{m \mid x'(m) \le 0\}$.

Figure 1 depicts the histogram of the eigenvalue distribution of K (or equivalently of $\hat{C}$), respectively for $u(t) = 1$ (i.e., for the non-robust $K = \frac{1}{n} X^\mathsf{T} X$) and for $u(t) = 2/(1+t)$, versus the limiting spectral distribution $\mathcal{K}$ obtained from Equation (1) and from the inverse Stieltjes transform formula

$$\mathcal{K}([a,b]) = \lim_{\varepsilon \to 0} \frac{1}{\pi} \int_a^b \Im[m(x + \imath\varepsilon)] dx$$

for all continuity points $a < b$ of $\mathcal{K}$. It is observed, as claimed, that the support of $\mathcal{K}$ is unbounded for $u(t) = 1$ while it remains compact for robust choices of $u(\cdot)$.

Interestingly, the setting of Figure 1 assumes the presence of three classes and the spectrum of K is seen to exhibit at least one isolated eigenvalue, clearly beyond x_+, when $u(\cdot)$ is chosen in the family of robust estimators. From our previous discussion, this eigenvalue, invisible in the spectrum of $\frac{1}{n} X^\mathsf{T} X$, is necessarily induced by the existence of classes. This is the main motivation of the present article. We will detail next when this phenomenon is observed and we will show that the eigenvector associated with this (hypothetical) isolated eigenvalue has a non-trivial correlation to the canonical vectors of classes $j_1, \ldots, j_k \in \mathbb{R}^n$, where $[j_a]_i = \delta_{x_i \in \mathcal{C}_a}$.

3. MAIN RESULTS

Our main objective is to determine the conditions under which isolated eigenvalues in the spectrum of K are found and to demonstrate that their associated eigenvectors have an asymptotic non-trivial alignment to the canonical class vectors $j_a \in \mathbb{R}^n$, $a = 1, \ldots, k$,

defined by $[j_a]_i = \delta_{x_i \in \mathcal{C}_a}$; that is, these eigenvectors can be used to achieve non-trivial spectral clustering of the data.

Our first result concerns the eigenvalues of K. Under the notations above, we have the following result that determines the presence and locate isolated dominant eigenvalues in the spectrum of K.

Theorem 2 (Isolated Eigenvalues and Clustering Phase Transition) *The number of eigenvalues of K found beyond $x \geq x_+$ (i.e., at macroscopic distance) asymptotically equals the number of eigenvalues ℓ of $D_{\frac{n_\cdot}{n}} M^\mathsf{T} M$ such that, for all large n, p,*

$$\int \frac{m(x)v(t\gamma)\mathcal{T}(dt)}{1 + cm(x)tv(t\gamma)} < -\frac{1}{\ell}$$

where $M = [\mu_1, \ldots, \mu_k]$ and we denoted $D_{Z_\cdot} = \mathrm{diag}\{Z_a\}_{a=1}^k$. In particular, spectral clustering asymptotically leads to non-trivial classification so long that the inequality is met for $\ell = \|D_{\frac{n_\cdot}{n}} M^\mathsf{T} M\|$ and $x = x_+$. These largest eigenvalues $\lambda_i(K)$ asymptotically satisfy

$$\int \frac{m(\lambda_i(K))v(t\gamma)\mathcal{T}(dt)}{1 + cm(\lambda_i(K))tv(t\gamma)} = -\frac{1}{\lambda_i(D_{\frac{n_\cdot}{n}} M^\mathsf{T} M)} + o(1).$$

It is not easy to intuitively assess from Theorem 2 the dependence in the function $u(\cdot)$ of the existence and expected positions of isolated eigenvalues in the spectrum of K. Numerically, it is observed that $u(t) \sim 1/t$ bring the earliest phase transition in the sense that isolated eigenvalues of K are found for small values of $\lambda_i(D_{\frac{n_\cdot}{n}} M^\mathsf{T} M)$. On the opposite, for $u(t) \sim 1$, the effect is opposite: only large values of $\lambda_i(D_{\frac{n_\cdot}{n}} M^\mathsf{T} M)$ induce isolated eigenvalues in the spectrum of K. At this point, it thus seems of utmost interest to chose $u(\cdot)$ to be "maximally" robust, i.e., close to Tyler's estimator.

We will now show that this reasoning does not propagate to the eigenvectors: maximally robust $u(\cdot)$ functions have the undesirable property to "break" the class information contained in the dominant eigenvectors of K. To see this, we now evaluate the alignment between these dominant eigenvectors and the canonical class vectors $j_1, \ldots, j_k$.

Theorem 3 (Informative Eigenvectors) *Let $(\lambda, \mathbf{u}) \equiv (\lambda(K), u(K))$ be an eigenpair of K linked to the eigenpair of $D_{\sqrt{\frac{n_\cdot}{n}}} M^\mathsf{T} M D_{\sqrt{\frac{n_\cdot}{n}}}$ $(\ell, \mathbf{v}) \equiv (\lambda(D_{\frac{n_\cdot}{n}} M^\mathsf{T} M), u(D_{\sqrt{\frac{n_\cdot}{n}}} M^\mathsf{T} M D_{\sqrt{\frac{n_\cdot}{n}}}))$ as defined in the previous theorem. Then, for unit-norm deterministic vectors $a, b \in \mathbb{R}^n$,*

$$a^\mathsf{T} \mathbf{u}\mathbf{u}^\mathsf{T} b = \frac{1}{n} \sum_{i,j} a_i b_j \frac{\sqrt{v(\gamma\tau_i)}}{1 + cm(\lambda)\tau_i v(\gamma\tau_i)} \frac{\sqrt{v(\gamma\tau_j)}}{1 + cm(\lambda)\tau_j v(\gamma\tau_j)} \times \frac{e_i^\mathsf{T} J D_{\sqrt{\frac{n}{n_\cdot}}} \mathbf{v}\mathbf{v}^\mathsf{T} D_{\sqrt{\frac{n}{n_\cdot}}} J e_j}{\int \frac{v(\gamma t)\mathcal{T}(dt)}{(1+cm(\lambda)tv(\gamma t))^2}} \frac{-m(\lambda)}{\lambda m'(\lambda)} + o_p(1).$$

In particular, letting $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$,

$$D_{\frac{1}{\sqrt{n_\cdot}}} J^\mathsf{T} \mathbf{u}\mathbf{u}^\mathsf{T} J D_{\frac{1}{\sqrt{n_\cdot}}} = \frac{-m(\lambda)}{\lambda m'(\lambda)} \frac{\left(\int \frac{\sqrt{v(\gamma t)}\mathcal{T}(dt)}{1+cm(\lambda)tv(\gamma t)} \right)^2}{\int \frac{v(\gamma t)\mathcal{T}(dt)}{(1+cm(\lambda)tv(\gamma t))^2}} \mathbf{v}\mathbf{v}^\mathsf{T} + o_p(1).$$

A few remarks are in order to best understand and interpret the statement of Theorem 3.

Remark 1 (Comments on the eigenvector behavior) *By Cauchy–Schwarz's inequality and basic algebraic manipulations, it is easily shown from the second result that*

$$D_{\frac{1}{\sqrt{n_\cdot}}} J^\mathsf{T} \mathbf{u}\mathbf{u}^\mathsf{T} J D_{\frac{1}{\sqrt{n_\cdot}}} \preceq \mathbf{v}\mathbf{v}^\mathsf{T} + o_p(1)$$

in the order of nonnegative definite matrices and, taking the trace on left and right hand sides, we naturally have

$$\mathbf{u}^\mathsf{T} J D_{\frac{1}{n_\cdot}} J^\mathsf{T} \mathbf{u} \leq 1 + o_p(1).$$

The left-hand quantity above measures the effective alignment of $\mathbf{u}$ to the subspace spanned by the canonical class vectors j_a: the closer to 1 the less "noisy" the eigenvector. Note here that the case of (asymptotic) equality is only met as $\lambda \to \infty$ (that is for easy clustering) and *for $v(\cdot) = 1$ (i.e., $u(\cdot) = 1$). Indeed, for arbitrary $u(\cdot)$, as $\lambda \to \infty$, $m(\lambda) \sim \frac{1}{\lambda}$, $m'(\lambda) \sim -\frac{1}{\lambda^2}$, and thus*

$$\mathbf{u}^\mathsf{T} J D_{\frac{1}{n_\cdot}} J^\mathsf{T} \mathbf{u} \xrightarrow{\lambda \to \infty} \frac{\left(\int \sqrt{v(\gamma t)}\mathcal{T}(dt) \right)^2}{\int v(\gamma t)\mathcal{T}(dt)} \tag{2}$$

which in general is strictly less than 1, unless $\mathcal{T} = \delta_1$ or $v(t) = 1$ for all t. As such, from a "global" standpoint, no robust $u(\cdot)$ allows for asymptotically perfect eigenvector alignment to the canonical class-vectors $j_1, \ldots, j_k$. In fact, the more different $u(\cdot)$ from 1 (so in particular for $u(t) = 1/t$), the smaller the right-hand side limit: maximally robust estimators thus induce minimal alignment of the dominant eigenvectors to $j_1, \ldots, j_k$ when $\lambda \to \infty$.

But let us fall back on a more standard non-trivial scenario where λ is not extremely large, for which we know that $u(\cdot) = 1$, i.e., $K = \frac{1}{n} X^\mathsf{T} X$, is not appropriate (since no eigenvector can be trusted). For the sake of argumentation, consider the setting where $k = 2$, $n_1 = n_2 = n/2$ and $\mu_1 = -\mu_2$, so that $\mathbf{v} = [1, -1]^\mathsf{T}/\sqrt{2}$ is the only informative eigenvector of $D_{\sqrt{\frac{n_\cdot}{n}}} M^\mathsf{T} M D_{\sqrt{\frac{n_\cdot}{n}}}$. In this setting, correctly clustering vector x_i depends on the sign and amplitude of $[\mathbf{u}]_i$ for $\mathbf{u}$ the dominant eigenvector of K. Note that $u(0) = v(0)$ and that both $u(t)$ and $v(t)$ behave as $1/t$ for large t. From the first equation in the theorem statement with $a = e_i$ and $b = e_j$, it appears that,

- *if $v(0)$ is large (in the extreme case of Tyler's function $u(t) = 1/t$, $v(0) = \infty$), the small τ_i's will induce spurious large values in $\mathbf{u}$ which, as a result, reduces the amplitudes of the majority of the $[\mathbf{u}]_j$'s associated to non-small τ_j's. This has the effect of having the two classes collapse onto one another.*
- *yet, for $tv(t)$ almost constant (which is the case for Tyler), the denominators $1+cm(\lambda)v(\gamma\tau_i)$ do not change sign. And thus, in this particular "two class of even sizes" scenario, almost not a single $[\mathbf{u}]_i$ can cross zero and change sign. This may be seen as very advantageous. However, from the previous item, the consequence is that most of the entries of $\mathbf{u}$, except a few ones (those associated to the smallest τ_i's), collapse, thus preventing the last classification step of the spectral clustering method (via k-means for instance) to perform any useful clustering. Also, this reasoning only holds for $k = 2$ classes and cannot extend beyond: for $k > 2$, only the cluster collapsing effect remains which is quite detrimental to clustering.*

The convenient trade-off then consists in (i) avoiding a large spread of the values of $\tau_i v(\gamma\tau_i)$ to prevent changes in sign in $[\mathbf{u}]_i$ (especially for small λ's) and (ii) avoiding too large values for $v(0)$ to prevent the collapse effect.

Figure 2 extends this basic interpretation of Theorem 3 to the case of three classes. Remark that, without any robust processing (top display), the dominant eigenvectors are meaningless and do not carry any information about the classes, as predicted. At the other extreme, for $u(t) = 1/t$ (i.e., for the "maximally robust" Tyler estimator, central display), the main distribution of the data points collapses at the $(0, 0)$ coordinate but are still "visually" somewhat distinguishable. For an optimal choice of α in the class of functions $u(t) = (1+\alpha)/(t+\alpha)$ (bottom display), $\mathbf{u}_1$ alone is quite discriminative and $(\mathbf{u}_1, \mathbf{u}_2)$ has a 2D-Gaussian mixture like shape.

Figure 3 complements Figure 2 by providing the limiting theoretical alignment of $\mathbf{u}_1$ to the subspace $\mathrm{span}(j_1, \ldots, j_k)$, as per Theorem 3, for $u(t) = (1 + \alpha)/(\alpha + t)$ with different values of α. Note that, as the dominant eigenvalue $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M)$ increases, the alignment improves but saturates, thereby confirming Equation (2). For decreasing values of α (thus for more robust u), non-trivial alignment emerges for smaller values of $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M)$ but with a lower saturation. As specified by the green marker, the choice $\alpha = 1$ is optimal for the value $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M) = 5.56$ corresponding to the setting of Figure 2. In this setting, note for instance that the eigenvector $\mathbf{u}_1$ carries no class information when $\alpha \geq 2.5$ (i.e., towards less robust estimators) and has very weak correlation for $\alpha \leq 0.1$ (i.e., towards more robust estimators), which imposes a very careful choice for α.

4. CONCLUDING REMARKS

The main message of the article is that, while standard kernels are inapt to retrieve class information under an impulsive data noise setting, the proposed new class of robust kernels can retrieve the class information by an appropriate robustness-classification trade-off in the function $u(\cdot)$. The next natural question is whether $u(\cdot)$ can be a priori found, given $x_1, \ldots, x_n$. It can in fact be shown that, under the setting of Theorem 1, $\frac{1}{p} x_i^\mathsf{T} \hat{C}_{-i}^{-1} x_i$ with $\hat{C}_i = \hat{C} - \frac{1}{n} u(\frac{1}{p} x_i^\mathsf{T} \hat{C} x_i) x_i x_i^\mathsf{T}$ is a consistent estimator for τ_i; from this, γ in Theorem 1 and $m(z)$, $\mathcal{T}$ in (1) can be estimated. This further implies that all quantities of Theorem 3 are known, except for the spectrum of $D_{\frac{n}{n}} M^\mathsf{T} M$. One may then produce an empirical version of the graph of Figure 3, however without being capable of a priori estimating $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M)$. To cope with this problem, a two-step approach can be imagined whereby a first matrix K is produced for some $u(\cdot)$ function, the dominant eigenvalue of which is used to estimate $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M)$ (by Theorem 2), thus allowing to decide on an improved choice for $u(\cdot)$.

It remains to know whether the proposed approach, which after all is quite heuristic, is the best heavy-tailed noise processing procedure. It would also be desirable to consider the possibly simpler setting where the τ_i's follow a mixture "Gaussian versus outlier" distribution with a vanishing proportion of outliers. In this case, most of the results above translate with $\mathcal{T} = \delta_1$ except that a few entries of $\mathbf{u}_1, \mathbf{u}_2, \ldots$ have a spurious behavior, which needs be properly analyzed. This setting would allow for a "robustification" of classification algorithms against rare outliers.

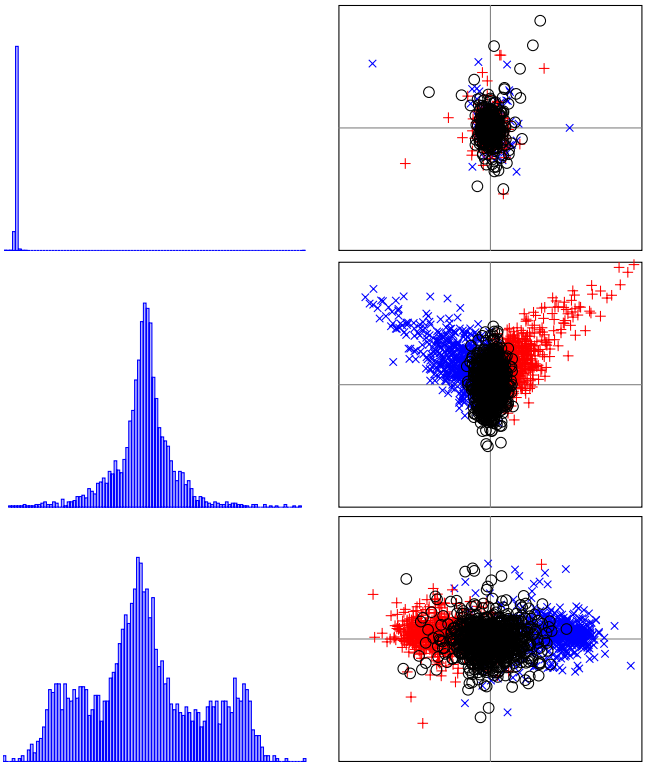

Fig. 2. Histogram of the dominant eigenvector $\mathbf{u}_1$ (left-hand side) and a 2D-scatter plot of the two dominant eigenvectors $\mathbf{u}_1$ vs. $\mathbf{u}_2$ (right-hand side) for $K = \frac{1}{n} X^\mathsf{T} X$ (top) and $K = \frac{1}{n} D^{\frac{1}{2}} X^\mathsf{T} X D^{\frac{1}{2}}$ for $u(t) = 1/t$ (center) or $u(t) = \frac{\alpha+1}{\alpha+t}$ for optimized oracle $\alpha = 1$ (bottom). Three classes with $[\mu_a]_i = 4\delta_{ia}$, $n_1 = n_2 = n_3/2$, $n = 2000$, $p = 500$, τ_i i.i.d. Student with one degree of freedom.

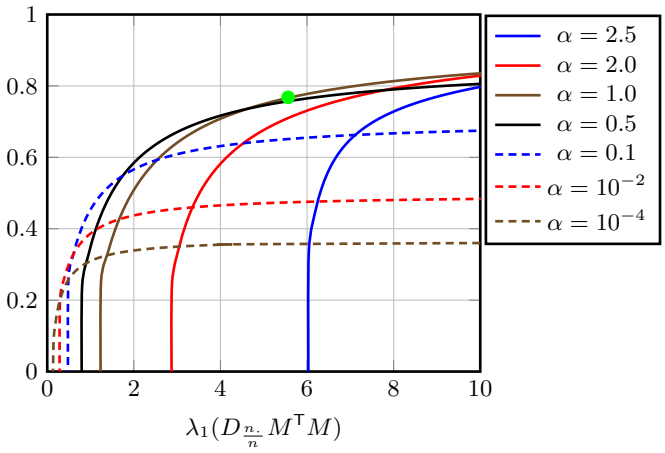

Fig. 3. Alignment $\mathbf{u}_1^\mathsf{T} J D_{\frac{1}{n}} D^\mathsf{T} \mathbf{u}_1$ between $\mathbf{u}_1$ and $\mathrm{span}(j_1, \ldots, j_k)$ for $u(t) = (1 + \alpha)/(t + \alpha)$ in the setting of Figure 2, as a function of $\lambda_1(D_{\frac{n}{n}} M^\mathsf{T} M)$, for different values of α. Marker indicates setting of Figure 2 (bottom display).

5. REFERENCES

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

[2] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press*, vol. 14, pp. 849–856, 2001.

[3] 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.

[4] Z. Liao and R. Couillet, "A large dimensional analysis of least squares support vector machines," *(in Press) IEEE Transactions on Signal Processing, arXiv preprint arXiv:1701.02967*, 2018.

[5] X. Mai and R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *(in Press) Journal of Machine Learning Research, arXiv preprint arXiv:1711.03404*, 2017.

[6] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *Ann. Appl. Probab.*, vol. 28, no. 2, pp. 1190–1248, 2018.

[7] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to stap detection problem," *Signal Processing, IEEE Transactions on*, vol. 62, no. 21, pp. 5640–5651, 2014.

[8] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust statistics*. J. Wiley, 2006.

[9] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[10] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *The Annals of Statistics*, vol. 4, pp. 51–67, 1976.

[11] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[12] R. Couillet, "Robust spiked random matrices and a robust G-MUSIC estimator," *J. Multivar. Anal.*, vol. 140, pp. 139–151, 2015.

[13] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of Maronna's M-estimator with elliptically distributed samples," *Journal of Multivariate Analysis*, vol. 139, pp. 56–78, 2015.

[14] J. W. Silverstein and Z. D. 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.

[15] Z. D. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of large dimensional sample covariance matrices," *The Annals of Probability*, vol. 26, no. 1, pp. 316–345, 1998.