CLASSIFICATION ASYMPTOTICS IN THE RANDOM MATRIX REGIME

Romain Couillet, Zhenyu Liao, Xiaoyi Mai

CentraleSupélec, University Paris-Saclay, France.
G-STATS Data Science Chair, GIPSA-lab, University Grenobles-Alpes, France.

ABSTRACT

This article discusses the asymptotic performance of classical machine learning classification methods (from discriminant analysis to neural networks) for simultaneously large and numerous Gaussian mixture modelled data. We first provide theoretical bounds on the minimally discriminable class means and covariances under an oracle setting, which are then compared to recent theoretical findings on the performance of machine learning. Non-obvious phenomena are discussed, among which surprising phase transitions in the optimal performance rates for specific hyperparameter settings.

Index Terms— Random matrix theory, classification, kernel methods, neural networks, LDA/QDA.

1. INTRODUCTION

The renewed interest for machine learning spurred by the big data movement leads statisticians to reconsider the asymptotic performance of statistical learning when both the number of data n and their dimension p grow simultaneously large, i.e., under a *random matrix regime*. In this regime, many conventional estimators, starting with the sample covariance matrix, are known to become inconsistent [1, 2, 3]. The field of random matrix theory has provided a deep understanding on the limitations and possible corrections of such "large p, large n" statistics, yet mostly for linear estimators (such as in array processing [4, 5], wireless communications [6, 7], detection and estimation [8], etc.). Dealing with non-linear operators, as in machine learning (kernel methods, neural nets), is more challenging and only very recent works have provided first steps into understanding and improving machine learning for large dimensional data [9, 10, 11].

The article aims to extract from these recent articles conclusions in terms of optimal discriminative performance rates in the classification of Gaussian mixture models and compare them to the optimal oracle performance of the Neyman–Pearson test for known class distributions.

This work is supported by the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

2. SYSTEM SETTING AND OBJECTIVES

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be a set of vectors to be classified, either in a supervised, semi-supervised, or unsupervised manner. For simplicity, we consider a binary setting: $x_1, \ldots, x_{n_1}$ belong to class $\mathcal{C}_1$ and $x_{n_1+1}, \ldots, x_n$ to class $\mathcal{C}_2$, with $n_i = |\mathcal{C}_i|$. We further restrict ourselves to the setting where the x_i's are independent and arise from a Gaussian mixture model,[1] i.e., for $\mu_1, \mu_2 \in \mathbb{R}^p$ and nonnegative definite $C_1, C_2 \in \mathbb{R}^{p \times p}$,

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\mu_a, C_a).$$

We will discuss the asymptotic performance of various classification methods as $n, p \to \infty$ with $p/n \to c_0 > 0$, our objective being twofold:

1. we first characterize the fundamental limits of Gaussian mixture classification under the oracle setting where μ_1, μ_2 and C_1, C_2 are known; these limits are expressed in terms of the minimal "distance scaling rate" between μ_1 and μ_2, and C_1 and C_1 with respect to n, p which ensures asymptotically non trivial classification (i.e., neither perfect nor impossible);

2. we compare the asymptotic oracle performance to those achieved by standard classification methods (discriminant analysis, kernel methods, random neural networks) with various degrees of available samples (from supervised classification to clustering).

The random matrix regime is interesting as it corresponds to the rate where C_1, C_2 can no longer be consistently estimated by sample covariances. As such, simple plug-in estimators (as we shall see with the popular QDA method) are bound to induce large performance losses with respect to oracle and thus possibly a loss in the minimally discriminable rates for $\|\mu_1 - \mu_2\|$ and $\|C_1 - C_2\|$. But random matrix theory has for long provided a series of consistent estimators for *functionals* of C_1, C_2 (i.e., mappings from $\mathbb{R}^{p \times p}$ to $\mathbb{R}$), some of which, as we shall see, are at the core of most well-performing classifiers; these classifiers often maintain close-to-optimal discriminative power.

[1] But many results to be introduced in the following are valid for a much larger scope of statistical distributions.

3. MAIN RESULTS

To avoid unnecessary complications, we make the following base hypotheses.

Assumption 1 (Growth Rate Control)**.** *The matrices C_1, C_2 are invertible and, as $p \to \infty$, for $\|\cdot\|$ the operator norm,*

$$\liminf_p \max\{\|C_a\|, \|C_a^{-1}\|\} < \infty \quad \textit{for } a \in \{1, 2\}.$$

When convenient, we further use the shortcut definitions

$$\mu \equiv \mu_1 - \mu_2, \quad \Delta C \equiv C_1 - C_2.$$

3.1. Oracle Classification

Under the setting of Section 2, let $x \in \mathbb{R}^p$ be a vector genuinely belonging to class $\mathcal{C}_1$, i.e., $x \sim \mathcal{N}(\mu_1, C_1)$. Further assume for simplicity that, when observed, x has prior probability $\frac{1}{2}$ to belong to class $\mathcal{C}_i$. Then, for perfectly known μ_1, μ_2 and C_1, C_2, the (decision optimal) Neyman–Pearson test for its belonging to $\mathcal{C}_1$ consists in the comparison test

$$(x-\mu_2)^\mathsf{T} C_2^{-1}(x-\mu_2) - (x-\mu_1)^\mathsf{T} C_1^{-1}(x-\mu_1) \lessgtr \log \frac{|C_1|}{|C_2|}. \tag{1}$$

Writing $x = \mu_1 + C_1^{\frac{1}{2}} w$, so that $w \sim \mathcal{N}(0, I_p)$, after elementary manipulation, the test is equivalent to the comparison

$$S(x) \equiv \frac{1}{p} w^\mathsf{T} (C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} - I_p) w + \frac{2}{p} \mu^\mathsf{T} C_2^{-1} C_1^{\frac{1}{2}} w + \frac{1}{p} \mu^\mathsf{T} C_2^{-1} \mu - \frac{1}{p} \log \frac{|C_1|}{|C_2|} \gtrless 0. \tag{2}$$

Since Uw for $U \in \mathbb{R}^{p \times p}$ an eigenvector basis for $C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} - I_p$ has the same distribution as w, the random variable $S(x)$ can be written as the sum of p independent random variables. A careful application of Lyapunov's central limit theorem [12], along with Assumption 1, reveals that, as $p \to \infty$,

$$V_S^{-\frac{1}{2}} (S(x) - \bar{S}) \to \mathcal{N}(0, 1)$$

in distribution, where

$$\bar{S} \equiv \frac{1}{p} \mathrm{tr}\, C_1 C_2^{-1} - 1 + \frac{1}{p} \mu^\mathsf{T} C_2^{-1} \mu - \frac{1}{p} \log \frac{|C_1|}{|C_2|} \tag{3}$$

$$V_S \equiv \frac{2}{p^2} \mathrm{tr}\, (C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} - I_p)^2 + \frac{4}{p^2} \mu^\mathsf{T} C_2^{-1} C_1 C_2^{-1} \mu. \tag{4}$$

The classification performance of x is thus only non-trivial (e.g., converging neither to 0 nor 1) if both $\bar{S}$ and $\sqrt{V_S}$ are of the same order of magnitude (with respect to p). Assume first that $C_1 = C_2 \equiv C$; then from Assumption 1,

$$\bar{S} = \frac{1}{p} \mu^\mathsf{T} C^{-1} \mu = O(\|\mu\|^2/p)$$
$$\sqrt{V_S} = \frac{2}{p} \sqrt{\mu^\mathsf{T} C^{-1} \mu} = O(\|\mu\|/p).$$

As a consequence, for a non-trivial asymptotic classification, we demand here that in the worst case, in order of magnitude, $\|\mu\| \sim_p \|\mu\|^2$, i.e., $\|\mu\| = O(1)$. The associated asymptotic correct classification probability $\mathbb{P}(S(x) > 0)$ thus satisfies

$$\mathbb{P}(S(x) > 0) - Q\left(-\frac{1}{2}\sqrt{\mu^\mathsf{T} C^{-1} \mu}\right) \to 0$$

with $Q(t) \equiv \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-\frac{1}{2}u^2} du$ the Gaussian tail function.

Under this minimal constraint $\|\mu\| = O(1)$, we now reincorporate $C_1 \neq C_2$ in such a way that $\|\Delta C\| = o(1)$. A Taylor expansion of (3)–(4) around $C = C_1$ leads to

$$\bar{S} = \frac{1}{p} \mu^\mathsf{T} C^{-1} \mu + \frac{1}{2p} \mathrm{tr}\, (C^{-1} \Delta C)^2 + o(p^{-1})$$
$$V_S = \frac{2}{p^2} \mathrm{tr}\, (C^{-1} \Delta C)^2 + \frac{4}{p^2} \mu^\mathsf{T} C^{-1} \mu + o(p^{-2})$$

from which it unfolds that $\|\Delta C\|$ must be at least of order $O(p^{-\frac{1}{2}})$ (so that $\mathrm{tr}\, (C^{-1}\Delta C)^2 = O(1)$) for the difference ΔC to have discriminative power. The associated asymptotic classification probability in this case satisfies

$$\mathbb{P}(S(x) > 0) \sim Q\left(-\frac{\sqrt{\mu^\mathsf{T} C_1^{-1} \mu + \frac{1}{2}\mathrm{tr}\, (C_1^{-1} C_2 - I_p)^2}}{2}\right).$$

Consequently, when μ_1, μ_2 and C_1, C_2 are a priori known, non-trivial classification is achieved for

$$\|\mu\| \geq O(1), \text{ or } \|\Delta C\| \geq O(p^{-\frac{1}{2}})$$

where $X_p \geq O(p^\alpha)$ stands for $\liminf_p p^{-\alpha} X_p > 0$. This forms an optimal asymptotic baseline for Gaussian mixture classification that statistical learning methods cannot outperform. In the following, we discuss the performances of a class of such methods, retrieved from a series of recent articles based on advanced random matrix considerations.

3.2. Asymptotic Performance of Machine Learning Tools

3.2.1. Supervised learning: LDA and QDA

In [13], the authors consider the popular linear and quadratic discriminant analysis (LDA and QDA) *supervised* classifiers. Those classifiers assume a two-class Gaussian mixture model for the data, as introduced presently. In its most general form (QDA), the classifier consists in estimating μ_a's and C_a's by sample means and covariances from two sets of data labelled in class $\mathcal{C}_1$ or $\mathcal{C}_2$. The estimators are then plugged in (1) in place of the genuine parameters. The respective numbers n_1 and n_2 of data from class $\mathcal{C}_1$ and $\mathcal{C}_2$ are such that $n_1, n_2 = O(p)$, with $n_1, n_2 > p$ (in order for sample covariance matrices to be invertible with probability one).

In [13], it is shown that the noise induced by the estimation of C_1 and C_2 has a debilitating effect on classification. Indeed, in order to perform the comparison (2),

$S(x)$ needs to be consistently estimated. If $x = \mu_1 + C_1^{\frac{1}{2}} w$ with $\|\mu_1\| = O(1)$, then $\frac{1}{p} x^\mathsf{T} C_2^{-1} x$ is a consistent estimator for $\frac{1}{p} w^\mathsf{T} C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} w$; but since C_2 is unknown and that $\hat{C}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\mathsf{T}$ ($\bar{x} = \frac{1}{n_2} \sum_{j=n_1+1}^{n} x_j$) is not a consistent estimator for C_2, $\frac{1}{p} x^\mathsf{T} \hat{C}_2^{-1} x$ is a biased estimator for $\frac{1}{p} w^\mathsf{T} C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} w$. Nonetheless, random matrix theory provides advanced results to correct these biases, and in particular $(1 - \frac{p}{n}) \frac{1}{p} x^\mathsf{T} \hat{C}_2^{-1} x$ consistently estimates $\frac{1}{p} w^\mathsf{T} C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} w$. The major issue though, is that $(1 - \frac{p}{n}) \frac{1}{p} x^\mathsf{T} \hat{C}_2^{-1} x = \frac{1}{p} w^\mathsf{T} C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} w + O(p^{-\frac{1}{2}})$. Similarly, $(1 - \frac{p}{n}) \frac{1}{p} x^\mathsf{T} \hat{C}_1^{-1} x = \frac{1}{p} w^\mathsf{T} I_p w + O(p^{-\frac{1}{2}})$. But under minimal discriminative oracle rates, $\frac{1}{p} w^\mathsf{T} (C_1^{\frac{1}{2}} C_2^{-1} C_1^{\frac{1}{2}} - I_p) w = O(p^{-1})$ so that the estimator $(1 - \frac{p}{n}) \frac{1}{p} x^\mathsf{T} (\hat{C}_2^{-1} - \hat{C}_1^{-1}) x$ induces a dominating $O(p^{-\frac{1}{2}})$ noise. This noise also dominates the term $\frac{1}{p} \mu^\mathsf{T} C_2^{-1} \mu$ in $S(x)$. As a consequence, to obtain a non-trivial classification rate, either constraint $\|\Delta C\| = O(p^{-\frac{1}{2}})$ or $\|\mu\|^2 = O(1)$ must be relaxed to at least an order of magnitude $\sqrt{p}$. The minimal distance scaling for non-trivial classification in this case becomes

$$\|\mu\| \geq O(p^{\frac{1}{4}}), \text{ or } \|\Delta C\| \geq O(1).$$

Remark 1 (Noise dominance when $\Delta C = 0$)**.** *Interestingly here, even when $\Delta C = 0$ (a property not known in advance), the minimal distance scaling of $\|\mu\|$ for non-trivial classification with QDA is still $\|\mu\| = O(p^{\frac{1}{4}})$, due to the non-discriminative, yet performed, estimation of C_1, C_2.*

When it comes to the simpler linear discriminant approach (LDA), meant to discriminate a Gaussian mixture with common covariances $C_1 = C_2 \equiv C$, a surprising phenomenon occurs. There, while C is still estimated by the sample covariance, this estimator *no longer* degrades the minimal distance rate $\|\mu\| = O(1)$. This result is due to the fact that, for LDA, only the sign of $(w - \frac{1}{2}\mu)^\mathsf{T} C^{-1} \mu = O(1)$ must be estimated. Substituting $(1 - \frac{p}{n}) \hat{C}^{-1}$ (with $\hat{C}$ the sample covariance of all n data) for C provides an estimator with fluctuations of order $O(1)$. Both estimates and fluctuations being of the same order, classification can be achieved at optimal order rate.

Even more surprisingly, the minimal discriminative distance $\|\mu\| = O(1)$ is also achieved when using LDA while C_1 and C_2 are *genuinely distinct* (where LDA supposes equal). A much unexpected practical consequence is that, at least in scenarios where $\|\Delta C\|$ does not largely overtake $\|\mu\|$, it is often more beneficial to use LDA rather than QDA classifiers even though one knows that $C_1 \neq C_2$.[2]

[2]This can be understood from a machine learning intuition as being related to the *overfitting* phenomenon by which, with too many unknown parameters to be estimated by a given algorithm, worse performance can be achieved than if less parameters were estimated in the first place.

3.2.2. Supervised learning: LS-SVM and kernel regression

Support vector machines (SVM) are probably the most popular supervised classification method. It consists in constructing a separating hyperplane between data issued from two similarity classes [14]. The asymptotic classification performance of SVM in the random matrix regime has been recently performed but offers so far little conclusive insight [15]. SVM is indeed the result of an optimization problem with no explicit solution, leading to non-trivial asymptotics. As an alternative, least square SVM (LS-SVM) [16] takes an explicit linear regression form by relaxing the SVM "hard" constraints.

In the Gaussian mixture setting, as in many classical application cases, data may not be directly linearly separated by an hyperplane and must then be "re-expressed" through a non-linear map prior to hyperplane separation. This leads to considering *kernel methods* [17] and in particular to use an affinity metric between x_i and x_j of the type, say, $K_{ij} \equiv f(\frac{1}{p}\|x_i - x_j\|^2)$, for some appropriately set function f. With the same notations as previously, LS-SVM (or similarly kernel regression) decides on the class of $x \in \mathbb{R}^p$ upon the test

$$S(x) = \sum_{i=1}^{n} \alpha_i f\left(\frac{1}{p}\|x_i - x\|^2\right) + b \gtrless 0 \tag{5}$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^\mathsf{T}$ and $b \in \mathbb{R}$ are given, for some hyperparameter $\gamma > 0$ and labels $y_i = (-1)^{1_{\{i > n_1\}}}$, by

$$\alpha = \left(K + \frac{n}{\gamma} I_n\right)^{-1} (y - b 1_n), \quad b = \frac{1_n^\mathsf{T} (K + \frac{n}{\gamma} I_n)^{-1} y}{1_n^\mathsf{T} (K + \frac{n}{\gamma} I_n)^{-1} 1_n}.$$

In [18], the asymptotic performance of LS-SVM is evaluated. To this end, the authors rely on a concentration approach allowing for a Taylor-expansion of the kernel matrix K [9], as $n, p \to \infty$. Precisely, under small enough rates for $\|\mu\|$ and $\|\Delta C\|$, $\frac{1}{p}\|x_i - x_j\|^2 \sim \frac{2}{p} \mathrm{tr}\,(\frac{n_1}{n} C_1 + \frac{n_2}{n} C_2) \equiv \tau$ so that $K_{ij} = f(\tau) + f'(\tau)(\frac{1}{p}\|x_i - x_j\|^2 - \tau) + \ldots$ ($i \neq j$). Developing carefully the expansion for K, it is proved that the minimal distance rates follow the rules below:

- for $f'(\tau) \neq 0$, then either (i) $\|\mu\| \geq O(1)$ and $\mathrm{tr}\,\Delta C \geq O(\sqrt{p})$ or (ii) $\mathrm{tr}\,([\Delta C]^2) \geq O(p)$;
- for $f'(\tau) = 0$ with $\mu = o(1)$ and $\mathrm{tr}\,\Delta C = o(\sqrt{p})$, then $\mathrm{tr}\,([\Delta C]^2) \geq O(p^{\frac{1}{2}})$.

This result is first surprising in that, for $f'(\tau) \neq 0$, it has a better minimal rate for $\|\mu\|$ than QDA, even though it is *not* designed dedicatedly for Gaussian models as in the QDA case. This gain is here due to the fact that no full covariance matrix needs be estimated. As a trade-off though, only the traces of ΔC and $[\Delta C]^2$ are used for discriminating covariances. Also, as $\|\Delta C\| = O(p^{-\frac{1}{2}}) \Rightarrow \mathrm{tr}\,([\Delta C]^2) = O(1)$, optimal covariance discriminative rates are not achieved. Besides, the best possible rate $\mathrm{tr}\,([\Delta C]^2) = O(p^{\frac{1}{2}})$ is only

achieved under the stringent condition that $\|\mu\| = o(1)$ and $\operatorname{tr}\Delta C = o(\sqrt{p})$ by selecting f such that $f'(\tau) = 0$.

The latter result is nonetheless quite interesting when it comes in practice to differentiating data with mostly different covariances but almost identical means. This is the case for EEG time series as well as multi-antenna wireless communication channels as shown in [19]. A deeper investigation of the scenario $f'(\tau) = 0$ is carried out in [20] with surprising findings when compared to the $f'(\tau) \neq 0$ case. Beyond the case $f'(\tau) = 0$, improved kernel refinements [21] have recently been proposed that ensure both the best rates $\|\mu\| \geq O(1)$ and $\operatorname{tr}([\Delta C]^2) \geq O(p^{\frac{1}{2}})$. These rely on choosing $f'(\tau) = O(p^{-\frac{1}{2}})$ rather than $f'(\tau) = 0$.

Remark 2 (Counter-intuitive behavior for large dimensional data)**.** *Interestingly, well-performing kernels with $f'(\tau) \simeq 0$ have a counter-intuitive behavior to the classical machine learning lore as a* "good" affinity function f is thus not a decreasing function of the distance between data vectors.

3.2.3. Supervised learning: Extreme learning machines

Extreme learning machines [22, 23] are practical single hidden-layer networks with a random (not optimized) first connection layer and a ridge-regression as the output layer. That is, the input-to-output relation for input data x reads

$$S(x) = \beta^{\mathsf{T}}\sigma(Wx)$$

where $\beta = \frac{1}{n}(\frac{1}{n}\sigma(WX)\sigma(WX)^{\mathsf{T}} + \gamma I_m)^{-1}\sigma(WX)y$, for training data $X = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_n] \in \{-1, 1\}^n$, $W \in \mathbb{R}^{m\times p}$ a random matrix dimensioned by the size-m neuron layer, $\sigma(\cdot)$ a non-linear activation function operating entry-wise, and some regularization parameter $\gamma \geq 0$. For supervised classification, the resulting test consists again here in deciding on the sign of $S(x)$.

In [11], the asymptotic performance of extreme learning machines is studied, as $n, m, p \to \infty$ at the same rate. It is shown that each triplet $(W, \sigma, \frac{m}{n})$ defines a kernel operator $(x, y) \mapsto f(x^{\mathsf{T}}y)$ which makes the extreme learning machine behave similar to LS-SVM. The family of kernels reached in this manner is however somewhat restricted and does not allow for easily setting constraints on the derivatives of f as previously. Yet the minimal discriminative rates for classification are preserved and the same as for LS-SVM.

3.2.4. Semi-supervised learning

The LS-SVM method has a natural extension to semi-supervised learning whereby a few samples are already labelled (their classes are known). A classical approach [24, 25], still relying on kernels, consists in computing "scores" $Y_{[u]} \in \mathbb{R}^{n_u\times 2}$ for n_u *unlabeled* data as a function of scores $Y_{[l]} \in \mathbb{R}^{n_l\times 2}$ for n_l *labelled* data, set as

Datasets	$\|\mu\|^2$	$\frac{1}{\sqrt{p}}\operatorname{tr}(\Delta C^2)$	Gauss	Quad
MNIST $(1, 7)$	613	1990	**98%**	**98%**
MNIST $(3, 6)$	441	1120	98%	**100%**
MNIST $(3, 8)$	212	658	84%	**95%**
EEG (sets A, E)	2.4	109	69%	**94%**

Table 1. Clustering performance for MNIST ($p = 784$) [31] and EEG ($p = 100$) [32] datasets, $n = 1024$, for $f(t) = \exp(-t^2/2)$ (Gauss) and for $f(t) = (t - \tau + \alpha p^{-\frac{1}{2}})^2$ (Quad).

$[Y_{[l]}]_{ij} = 1_{\{x_i\in\mathcal{C}_j\}}$. An optimization procedure leads to

$$Y_{[u]} = (I_{n_u} - K_{[uu]}D_{[u]}^{-1})^{-1}K_{[ul]}D_{[l]}^{-1}Y_{[l]}$$

where $D = \operatorname{diag}(K1_n)$ and both K and D were divided into matrix blocks as $K = \left[\begin{smallmatrix} K_{[uu]} & K_{[ul]} \\ K_{[lu]} & K_{[ll]} \end{smallmatrix}\right]$, $D = \operatorname{diag}(D_{[uu]}, D_{[ll]})$, with obvious notations. Here again, in the analysis performed in [26], under the assumption that $n_u, n_l = O(p)$, it is found that the optimal classification rate of a non-trivial number of the n_u unlabeled data is the same as in Section 3.2.2.

3.2.5. Unsupervised spectral classification (or clustering)

Spectral clustering consists in classifying n purely unlabelled data from the dominant eigenvectors of matrix K (upon which k-means is performed) [27]. In our present scenario, similar conclusions on optimal discriminable rates as in Section 3.2.2 still hold [10, 20]. The result is surprising as almost oracle rates are achieved despite the complete absence of labelled data. This suggests that the kernel spectral clustering method is capable on its own to isolate alike data as well as if in presence of labelled information (at least in terms of rate orders).

Performance figures borrowed from [21] are provided in Table 1, which confirm the large superiority of improved versus standard kernels, particularly for datasets only discriminable through their covariance structure (EEG case). This corroborates theoretical findings on *practical* datasets, so beyond the Gaussian assumption.

4. CONCLUSION

The article showed that, under a large p-dimensional binary Gaussian mixture model setting ($\mathcal{N}(\mu_a, C_a)$, $a \in \{1, 2\}$), elementary machine learning methods achieve non-trivial classification under almost Neyman–Pearson optimal discriminative rates for $\|\mu_1 - \mu_2\|$ and $\|C_2 - C_2\|$ as a function of p. But the picture is not that simple and many methods need properly set hyperparameters to achieve optimal discriminative power. The article summarizes some of these "proper choices". More importantly, the results presented here under a Gaussian setting are often shown to be applicable to real world (non Gaussian) data; as such, the article unveils a novel random matrix-improved paradigm in the understanding and optimization of machine learning methods for large dimensional data.

5. REFERENCES

[1] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, 1967.

[2] 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.

[3] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, 2009.

[4] X. Mestre, "On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices," *IEEE Transactions on Signal Processing*, vol. 56, no. 11, pp. 5353–5368, Nov. 2008.

[5] ——, "Improved estimation of eigenvalues of covariance matrices and their associated subspaces using their sample estimates," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5113–5129, Nov. 2008.

[6] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, 2004.

[7] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. NY, USA: Cambridge University Press, 2011.

[8] L. S. Cardoso, M. Debbah, P. Bianchi, and J. Najim, "Cooperative spectrum sensing using random matrix theory," in *IEEE Pervasive Computing (ISWPC'08)*, Santorini, Greece, May 2008, pp. 334–338.

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

[10] 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.

[11] 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.

[12] P. Billingsley, *Probability and Measure*, 3rd ed. Hoboken, NJ: John Wiley and Sons, Inc., 1995.

[13] K. Elkhalil, A. Kammoun, R. Couillet, T. Y. Al-Naffouri, and M.-S. Alouini, "A large dimensional analysis of regularized discriminant analysis classifiers," *arXiv preprint arXiv:1711.00382*, 2017.

[14] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[15] H. Huang, "Asymptotic behavior of support vector machine for spiked population model," *Journal of Machine Learning Research*, vol. 18, no. 45, pp. 1–21, 2017.

[16] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[17] A. J. Smola and B. Schölkopf, *Learning with kernels*. GMD-Forschungszentrum Informationstechnik, 1998.

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

[19] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *2016 Asilomar Conference on Signals, Systems, and Computers*, 2016.

[20] A. Kammoun and R. Couillet, "Subspace kernel clustering of large dimensional data," *(submitted to) Journal of Machine Learning Research*, 2017.

[21] H. Tiomoko Ali, A. Kammoun, and R. Couillet, "Random matrix-improved kernels for large dimensional spectral clustering," in *Statistical Signal Processing Workshop (SSP'18)*, Freiburg, Germany, 2018.

[22] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[23] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, pp. 513–529, 2012.

[24] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

[25] K. Avrachenkov, P. Gonçalves, A. Mishenin, and M. Sokol, "Generalized optimization framework for graph-based semi-supervised learning," in *Proceedings of SIAM Conference on Data Mining (SDM 2012)*, vol. 9. SIAM, 2012.

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

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

[28] M. Newman, "Spectral community detection in sparse networks," *arXiv preprint arXiv:1308.6494*, 2013.

[29] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

[30] H. Tiomoko Ali and R. Couillet, "Random matrix improved community detection in heterogeneous networks," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2016.

[31] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," 1998.

[32] 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.