RANDOM MATRICES AND MACHINE LEARNING

ENS DIPLOMA THESIS
IN
MATHEMATICS

Submitted by

GREGOIRE MACQUERON

Under the supervision of

GIAMBATTISTA GIACOMIN

DEPARTMENT OF APPLIED MATHEMATICS (DMA)
ECOLE NORMALE SUPERIEURE
45 rue d'Ulm, FRANCE PARIS 75005

MAY, 2022

Abstract

The big data revolution comes along with the challenging need to parse, mine, compress large amount of large dimensional and possibly heterogeneous data. In many applications, the dimension p of the observations is as large as – if not much larger than – their number n. In this context, Random Matrix Theory (RMT) - that deals with the eigenvalue distribution and the eigenvectors statistical behavior of large dimensional random matrices - becomes a really powerful tool for the performance analysis of numerous learning methods. Indeed, many Machine Learning methods such as Principal Component Analysis (PCA), Spectral Clustering and some Semi-Supervised learning techniques are built directly upon the eigenspace spanned by the several top eigenvectors. However, lots of algorithms were originally developed under the inappropriate finite-dimensional intuition assumption ($n \gg p$) which is radically different from the large dimensional one. Hence those algorithms are likely to fail or at least to perform inefficiently, yet RMT is able to analyze, improve, and evoke a whole new paradigm for large dimensional learning. This Diploma Thesis takes inspiration from the MVA course Notes on the subject [1] and the associated draft book [2] of Romain Couillet.

Contents

Curse of Dimensionality

We will see along this chapter that finite-dimensional intuitions which are at the core of many Machine Learning algorithms (starting with spectral clustering [3],[4]) may strikingly fail when applied in a simultaneously large n, p setting.

1.1 Sample Covariance Matrices

Let's consider $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ with $\mathbf{x}_i \underset{iid}{\sim} \mathcal{N}(0, \mathbf{C})$, n being the number of observations and p the number of features per observation. The Maximum Likelyhood Estimator for $\mathbf{C}$ is the sample covariance matrix (SCM):

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T, \text{ where } \mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$$

Finite-Dimensional Intuition:

For a fixed $p \in \mathbb{N}^*$ and $n \to \infty$:

$$\hat{\mathbf{C}} \xrightarrow{a.s.} \mathbf{C} \iff \|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \xrightarrow{a.s.} 0 \iff \|\hat{\mathbf{C}} - \mathbf{C}\| \underset{n\to\infty}{\longrightarrow} 0$$

from the fact that $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$, where $\|.\|$ denotes the operator norm.

Large-Dimensional Intuition:

This " $\iff$ " no longer holds if $p, n \to \infty$ with $\frac{p}{n} \to c \in]0, \infty[$. Let's consider the simple case $\mathbf{x}_i \underset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_p)$, with $p = cn$ and $c > 1$. Then we still have point-wise convergence and by a concentration inequality argument we even have that:

$$\max_{1\leq i,j\leq p} |[\hat{\mathbf{C}} - \mathbf{I}_p]_{i,j}| = \max_{1\leq i,j\leq p} |[\frac{1}{n}\sum_{k=1}^{n} [\mathbf{x}_k]_i [\mathbf{x}_k]_j - \delta_{ij}| \xrightarrow{a.s.} 0 \iff \|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \xrightarrow{a.s.} 0$$

However, there is no convergence in spectral norm as $\hat{\mathbf{C}}$ is always singular ($\forall n \in \mathbb{N}^*, p > n$) and $\mathbf{I}_p$ is not, hence:

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\| \not\to 0$$

More generally for $p, n \to \infty$ with $\frac{p}{n} \to c \in]0, \infty[$, the Empirical Spectral Measure converges in law to the deterministic smooth limit called the Marcenko Pastur [5] Law:

$$\mu_p = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\hat{C})} \underset{n,p\to\infty}{\longrightarrow} \mu(dx) = \left(\frac{1-c}{c}\right)^+ \delta_0 + \frac{1}{2\pi c x}\sqrt{(x-(1-\sqrt{c})^2)^+(x-(1+\sqrt{c})^2)^+}dx$$

In particular, for $n = 100p$, where one would expect a sufficiently large number of samples for $\hat{\mathbf{C}}$ to properly estimate $\mathbf{C} = \mathbf{I}_p$, the eigenvalues span on a range of

$$(1+\sqrt{c})^2) - (1-\sqrt{c})^2 = 4\sqrt{c} = 0.4$$

which is a large spread around the mean (and true) eigenvalue 1 see 4.3.

1.2 Euclidean Distances Concentration Phenomena

Another important example is the loss of relevance of the notion of Euclidean distance between large dimensional data vectors. To be more precise, we will see in this section, that under **asymptotically non-trivial** classification settings, data vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ extracted from a multi-class Gaussian mixture model ($\mathbf{x}_i \in \mathcal{C}_j \iff \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{C}_j)$) tend to be asymptotically at equal distance from one another, independently of their mixture class:

$$\max_{1 \leq i \neq j \leq p} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \underset{n,p \to \infty}{\longrightarrow} 0 \tag{1.1}$$

with $\tau = \frac{2}{p} \mathrm{tr} \mathbf{C}^o, \boldsymbol{\mu}^o = \sum_{i=1}^k \frac{n_i}{n} \boldsymbol{\mu}_i$, and $\mathbf{C}^o = \sum_{i=1}^k \frac{n_i}{n} \mathbf{C}_i$

This asymptotic behavior is extremely counterintuitive and drastically opposed to the finite-dimensional intuition - that forged many of the leading machine learning algorithms such as Spectral Clustering [3],[4] - in which two data points belong to the same class if they are "close" in Euclidean distance.

Proposition 1.2.1 (Asymptotically non-Trivial Classification Regime)**.** *We give ourselves a training set $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ of n samples independently drawn from the k-class $(\mathcal{C}_1, \dots, \mathcal{C}_k)$ Gaussian mixture i.e. $\mathbf{x}_i \in \mathcal{C}_j \iff \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{C}_j)$. As $n, p \to \infty$ we need to be sure that asymptotic classification is neither too simple nor too hard.*

To do so we will find the limit-case scenario when everything is perfectly known. This study leads to the following control growth rates:

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}^o\| = O(1), \ tr(\mathbf{C}_i - \mathbf{C}^o) = O(\sqrt{p}) \ and \ \|\mathbf{C}_i\| = O(1) \tag{1.2}$$

Proof Idea. Let's consider the case $k = 2$ with equally draw class:

- Suppose $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$, the (decision optimal) Neyman-Pearson test gives : $\mathbb{P}(\mathbf{x} \to \mathcal{C}_1 \mid \mathbf{x} \in \mathcal{C}_2) = \|\boldsymbol{\mu}\| \int_{\|\boldsymbol{\mu}\|}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$ which is only non-trivial for $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 2\|\boldsymbol{\mu}\| = O(1)$ (w.r.t. p).
- Suppose $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ and $C_1 = \mathbf{I}_p$ and $\mathbf{C}_2 = (1+\epsilon)\mathbf{I}_p$ (such that $\mathrm{tr}(\mathbf{C}_1 - \mathbf{C}_2) = \epsilon p$), following the same procedure, we find that the problem is non-trivial only if $\epsilon \to 0$, in which case: $\mathbb{P}(\mathbf{x} \to \mathcal{C}_1 \mid \mathbf{x} \in \mathcal{C}_2) \approx \mathbb{P}_{\mathcal{N}(0,1)}\left(w > \sqrt{\frac{p}{3\epsilon}}\right)$, which is only non-trivial for $\epsilon = O\left(\frac{1}{\sqrt{p}}\right)$.

Now suppose $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $\mathbf{C}_1 = \mathbf{I}_p$ and $\mathbf{C}_2 = (1+\epsilon)\mathbf{I}_p$ and that we respect the control growth rates above:

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\| = \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\| + O\left(\frac{1}{\sqrt{p}}\right) \text{ with } \mathbf{z}_i \underset{iid}{\sim} \mathcal{N}(0, \mathbf{I}_p)$$

hence, $\max_{1 \leq i \neq j \leq p} \left\{ \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - 2 \right\} \underset{n,p \to \infty}{\longrightarrow} 0$, ($\|\mathbf{z}_i - \mathbf{z}_j\| \sim \chi^2$ with p degree of freedom). $\square$

Basics of Random Matrix Theory

In this section, I will present key concepts used in RMT in order to access a random matrix spectral measure, the location of its isolated eigenvalues and the statistical behavior of their associated eigenvectors. Then I will state two major Theorems that I will will constitute the cornerstone of the whole Diploma Thesis.

2.1 Key Tools

Definition 2.1.1 (Resolvent). *For a symmetric matrix* $\mathbf{M} \in \mathbb{R}^{n\times n}$*, the resolvent* $\mathbf{Q_M}(z)$ *of* $\mathbf{M}$ *is defined, for* $z \in \mathbb{C}$ *not eigenvalue of* $\mathbf{M}$*, as*

$$\mathbf{Q_M}(z) = (\mathbf{M} - z\mathbf{I}_n)^{-1} \tag{2.1}$$

Definition 2.1.2 (Empirical Spectral Measure). *For a symmetric matrix* $\mathbf{M} \in \mathbb{R}^{n\times n}$*, the spectral measure or empirical spectral measure or empirical spectral distribution (e.s.d.)* $\mu_{\mathbf{M}}(z)$ *of* $\mathbf{M}$ *is defined as the normalized counting measure of the eigenvalues* $\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M})$ *of* $\mathbf{M}$*,*

$$\mu_{\mathbf{M}}(z) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})} \tag{2.2}$$

Definition 2.1.3 (Stieltjes Transform). *For a real probability measure* μ *with support* $supp(\mu)$*, the Stieltjes transform* $m_\mu(z)$ *is defined, for all* $z \in \mathbb{C}\backslash supp(\mu)$*, as*

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{t-z}\mu(dt) \tag{2.3}$$

Proposition 2.1.4. *The important relation between the empirical spectral measure* $\mu_{\mathbf{M}}(z)$ *of* $\mathbf{M} \in \mathbb{R}^{n\times\ n}$*, the Stieltjes transform* $m_{\mu_{\mathbf{M}}}(z)$ *and the resolvent* $\mathbf{Q_M}$ *lies in the fact that:*

$$m_\mu(z) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t-z}\mu(dt) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lambda_i(\mathbf{M})-z}\mu(dt) = \frac{1}{n}tr(\mathbf{Q_M}) \tag{2.4}$$

2.2 Marcenko-Pastur

Theorem 2.2.1 (Marcenko Pastur [5]). *Let's consider* $\mathbf{X} \in \mathbb{R}^{p\times n}$ *having i.i.d zero mean* $(\mathbb{E}(\mathbf{X_{ij}}) = \mathbf{0})$*, unit variance* $(\mathbb{E} \mid \mathbf{X_{ij}} \mid^{\mathbf{2}} = \mathbf{0})$ *and smooth tail condition* $(\mathbb{E} \mid \mathbf{X_{ij}} \mid^{\mathbf{4}} < \infty)$*. Then as* $p, n \to \infty$ *with* $\frac{p}{n} \to c \in]0, \infty[$*, the spectral distribution* $\mu_{\mathbf{Z}_n}$ *of* $\frac{1}{n}\mathbf{X}\mathbf{X}^T$*:*

$$\mu_{\mathbf{Z}_n} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(\mathbf{Z}_n)}, \text{ with } \mathbf{Z}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

converges in law to the Marcenko-Pastur Law:

$$\mu_{MP}(dx) = \left(\frac{1-c}{c}\right)^{+}\delta_0(x) + \frac{1}{2\pi cx}\sqrt{(x-(1-\sqrt{c})^2)^{+}(x-(1+\sqrt{c})^2)^{+}}dx \tag{2.5}$$

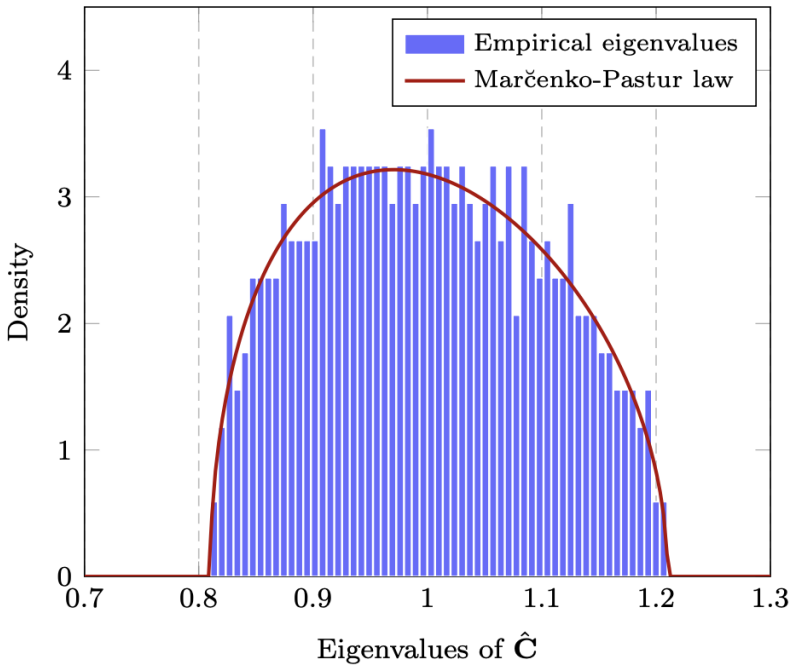

Figure 2.1: Histogram of the empirical spectral distribution of $\hat{\mathbf{C}}$ for a sample covariance matrix studied in 1.1 with for $p = 500$ and $n = 50000$ v.s. the Marcenko-Pastur law

2.3 Spiked Models

In Machine Learning, where one is interested in the correlation matrix $\mathbf{X}\mathbf{X}^T$, the *i.i.d.* assumption is often too limiting. It is expected that the columns $\mathbf{x}_i \in \mathbb{R}^p$ of $\mathbf{X}$ exhibit a correlation structure and be non-necessarily independent. We will introduces a very special, yet practically far reaching, case of sample covariance matrix models for which the limiting spectral measure coincides with the Marcenko Pastur [5] Law + some spikes, by letting the covariance matrix $\mathbb{C}$ be a low rank perturbation of the identity matrix.

Theorem 2.3.1 (Spiked Model [6])**.** *Let's consider* $\mathbf{X} \in \mathbb{R}^{p\times n}$ *as in 2.2.1 and the covariance matrix* $\mathbf{C}$ *of the form* $\mathbf{C} = \mathbf{I}_p + \mathbf{P}$ *with:*

$$\mathbf{P} = \sum_{i=1}^{k} l_i \mathbf{u}_i \mathbf{u}_i^T, \text{ where } k \text{ and } l_1 \geq \cdots \geq l_k > 0 \text{ are fixed w.r.t. } p,n$$

We pose $\mathbf{Y} = \mathbf{C}^{\frac{1}{2}}\mathbf{X}$*, then, by denoting* $\lambda_1 \geq \cdots \geq \lambda_p$ *the eigenvalues of* $\frac{1}{n}\mathbf{Y}\mathbf{Y}^T$ *as* $p, n \to \infty$ *with* $\frac{p}{n} \to c \in]0, \infty[$*:*

$$\lambda_i \xrightarrow{a.s.} \begin{cases} 1 + l_i + c\frac{1+l_i}{l_i} > (1+\sqrt{c})^2 & , l_i > \sqrt{c} \\ (1+\sqrt{c})^2 & , l_i \leq \sqrt{c} \end{cases} \tag{2.6}$$

Actually, the "spiked model" terminology goes beyond sample covariance matrix models with $\mathbf{C} = \mathbf{I}_p + \mathbf{P}$, for $\mathbf{P}$ a low rank matrix. In the literature, spiked models loosely refer to as "low rank perturbative" models in the following sense: there exists an underlying random matrix model $\mathbf{X}$, the spectral measure of which converges to a well-defined measure with compact support and having eigenvalues converging to the support which is then modified in some way by a low rank perturbation matrix $\mathbf{P}$; the resulting matrix has the same limiting spectral measure as that of $\mathbf{X}$ but with possibly some spurious (isolated) eigenvalues.

Proof Idea.

$$\begin{aligned}0 =& \det\left(\frac{1}{n}\mathbf{Y}\mathbf{Y}^T - \lambda\mathbf{I}_p\right) \iff 0 = \det(\mathbf{C})\det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{C}^{-1}\right)\\ \iff& 0 = \det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}_p + \lambda\left(\mathbf{I}_p - \mathbf{C}^{-1}\right)\right)\\ \iff& 0 = \det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}_p\right)\det\left(\mathbf{I}_p + \lambda\left(\mathbf{I}_p - \mathbf{C}^{-1}\right)\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}_p\right)^{-1}\right)\end{aligned}$$

Using the low rank property: $\mathbf{C} = \mathbf{I}_p + \mathbf{P}$ where $\mathbf{P} = \mathbf{U}\mathbf{\Omega}\mathbf{U}^T$ with $\mathbf{\Omega} = \mathrm{diag}(\lambda_1, \dots, \lambda_k)$, to obtain that $\mathbf{I}_p - \mathbf{C}^{-1} = \mathbf{U}\left(\mathbf{I}_k - \mathbf{\Omega}^{-1}\right)^{-1}\mathbf{U}^T$ and Sylvester's identity ($\det(\mathbf{I} + \mathbf{A}\mathbf{B}) = \det(\mathbf{I} + \mathbf{B}\mathbf{A})$) we have:

$$0 = \underbrace{\det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}_p\right)}_{\mu_{MP}(\lambda)}\det\left(\mathbf{I}_k + \lambda\left(\mathbf{I}_k - \mathbf{\Omega}^{-1}\right)^{-1}\underbrace{\mathbf{U}^T\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - \lambda\mathbf{I}_p\right)^{-1}\mathbf{U}}_{\xrightarrow[a.s.]{}\mu_{MP}(\lambda)\mathbf{I}_k \text{ by } \star}\right)$$

*** Isotropic Marcenko-Pastur**: $\forall z \in \mathbb{C}\backslash\mathrm{supp}(\mu_{MP})$,

$$\mathbf{U}^T\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T - z\mathbf{I}_p\right)^{-1}\mathbf{U} \xrightarrow[a.s.]{} \mu_{MP}(z)\mathbf{I}_k$$

($\mathbf{X}$ being "almost-unitary invariant" $\mathbf{U}$ made of "i.i.d.-like" random vector.)
Hence,

$$0 \approx \prod_{i=1}^{k}\left(1 + \frac{l_k}{1+l_k}\lambda\mu_{MP}(\lambda)\right) \iff \lambda\mu_{MP}(\lambda) = -\frac{1+l_k}{l_k}$$

But $\lambda\mu_{MP}(\lambda)$ maps $](1+\sqrt{c})^2, +\infty[$ to $]-\frac{1+\sqrt{c}}{\sqrt{c}}, 0[$. Therefore, we have a solution only when $l_i > \sqrt{c}$ which is:

$$1 + l_i + c\frac{1+l_i}{l_i}$$

□

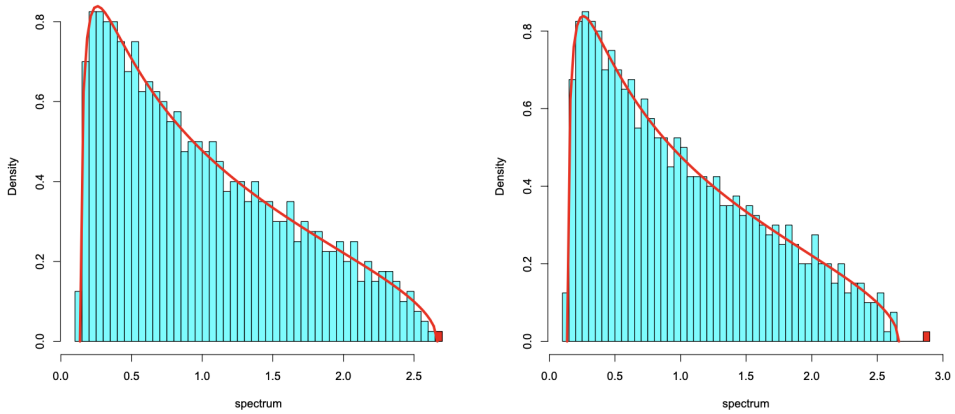

Figure 2.2: 1^{st} Order Perturbation: Influence of the largest eigenvalue on the spectral distribution. (**Left**) $l_1 = 0.5 < \sqrt{c}$, (**Right**) $l_1 = 2 > \sqrt{c}$, where $c = \frac{p}{n} = \frac{800}{2000} = 0.63$.

Spectral Clustering

3.1 Defintion

We would like to classify n data vectors $\mathbf{x}_1, \dots, \mathbf{x}_n \mathbb{R}^p$ into k-classes $\mathcal{C}_1, \dots, \mathcal{C}_k$:

$$\begin{cases} \mathbf{x}_1, \dots, \mathbf{x}_{n_1} \in \mathcal{C}_1 \\ \qquad \vdots \\ \mathbf{x}_{n-n_k+1}, \dots, \mathbf{x}_n \in \mathcal{C}_k \end{cases}$$

To do so we introduce an **affinity metric**: for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$\kappa(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^+, \text{ e.g. } \kappa(\mathbf{x}, \mathbf{y}) = f\left(\frac{1}{p}\|\mathbf{x} - \mathbf{y}\|_2^2\right)$$

Remark 3.1.1 (Finite-Dimensional Intuition)**.** *For an affinity function as above, it is natural to assume that f be a non- increasing function, as close-by data $\mathbf{x}_i, \mathbf{x}_j$ should have a stronger affinity than distant $\mathbf{x}_i, \mathbf{x}_j$. The popular choice $f(t) = \exp(-\frac{t}{2})$ (Gaussian or Heat kernel) is particularly appealing as it brings arbitrarily close data to a unit affinity and far data to a null affinity.*

We will subsequently show in this section that this natural reasoning often collapses when dealing with realistic large dimensional data, leading to erroneous intuitions and disrupting many conventional ideas behind kernel-based machine learning.

In order to perform the unsupervised classification, we will try to minimize a clustering metric:

$$\begin{aligned}
&\min_{\substack{\hat{\mathcal{C}}_1,\dots,\hat{\mathcal{C}}_k \\ \sqcup_{i=1}^k \hat{\mathcal{C}}_i = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}}} \sum_{i=1}^k \sum_{\substack{\mathbf{x}_j \in \hat{\mathcal{C}}_i \\ \mathbf{x}_l \notin \hat{\mathcal{C}}_i}} \frac{\kappa(\mathbf{x}_j, \mathbf{x}_l)}{|\,\hat{\mathcal{C}}_i\,|} \iff \min_{\substack{\hat{\mathcal{C}}_1,\dots,\hat{\mathcal{C}}_k \\ \sqcup_{i=1}^k \hat{\mathcal{C}}_i = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}}} \sum_{i=1}^k \left[\sum_{\substack{\mathbf{x}_j \in \hat{\mathcal{C}}_i \\ \mathbf{x}_l}} \frac{\kappa(\mathbf{x}_j, \mathbf{x}_l)}{|\,\hat{\mathcal{C}}_i\,|} - \sum_{\substack{\mathbf{x}_j \in \hat{\mathcal{C}}_i \\ \mathbf{x}_l \in \hat{\mathcal{C}}_i}} \frac{\kappa(\mathbf{x}_j, \mathbf{x}_l)}{|\,\hat{\mathcal{C}}_i\,|} \right] \\
&\iff \min_{\substack{\hat{\mathcal{C}}_1,\dots,\hat{\mathcal{C}}_k \\ \sqcup_{i=1}^k \hat{\mathcal{C}}_i = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}}} \sum_{i=1}^k \left[\sum_{\mathbf{x}_j} \delta^2_{\mathbf{x}_j \in \hat{\mathcal{C}}_i} \frac{\sum_{\mathbf{x}_l} \kappa(\mathbf{x}_j, \mathbf{x}_l)}{|\,\hat{\mathcal{C}}_i\,|} - \sum_{\mathbf{x}_j, \mathbf{x}_l} \delta_{\mathbf{x}_j \in \hat{\mathcal{C}}_i} \frac{\kappa(\mathbf{x}_j, \mathbf{x}_l)}{|\,\hat{\mathcal{C}}_i\,|} \delta_{\mathbf{x}_l \in \hat{\mathcal{C}}_i} \right] \\
&\iff \min_{\mathbf{F} \in \mathcal{F}} \sum_{i=1}^k \left[\sum_j \mathbf{F}^2_{ji} \mathbf{D}_j - \sum_{j,l} \mathbf{F}_{ji} \mathbf{K}_{j,l} \mathbf{F}_{li} \right] \\
&\iff \min_{\mathbf{F} \in \mathcal{F}} \sum_{i=1}^k \mathbf{F}^T_{\cdot i} (\mathbf{D} - \mathbf{k}) \mathbf{F}_{\cdot i} \iff \min_{\mathbf{F} \in \mathcal{F}} \operatorname{tr}\left(\mathbf{F}^{\mathbf{T}} (\mathbf{D} - \mathbf{K}) \mathbf{F}\right)
\end{aligned}$$

with $\mathcal{F} \subset \mathbb{R}^{n,k}$ the discrete set of valid $\mathbf{F}$'s, where:

$$\mathbf{F} = \left\{ \frac{1}{\sqrt{|\,\hat{\mathcal{C}}_j\,|}} \delta_{\mathbf{x}_i \in \hat{\mathcal{C}}_j} \right\}_{i,j=1}^{n,k}, \mathbf{K} = \left\{ \kappa(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1}^n, \mathbf{D} = \operatorname{diag}\left(\left\{ \sum_{j=1}^n \kappa(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i=1}^n \right)$$

Now the idea is to relax the problem, and transform $\mathcal{F}$ into $\left\{\mathbf{F} \in \mathbb{R}^{n,k}, \mathbf{F}^T\mathbf{F} = \mathbf{I}_k\right\}$:

$$\min_{\mathbf{F}, \mathbf{F}^T\mathbf{F}=\mathbf{I}_k} \operatorname{tr}\left(\mathbf{F}^\mathbf{T}(\mathbf{D}-\mathbf{K})\mathbf{F}\right) \tag{3.1}$$

The solution of this problem is well-known, it is the k **smallest** eigenvectors of $\mathbf{D}-\mathbf{K}$.

3.2 Finite Dimension Intuition

If we use the famous Gaussian Kernel: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2})$, the general intuition would be:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \begin{cases} \gg 1, & \text{if } \mathbf{x}_i, \mathbf{x}_j \text{ in different classes} \\ \ll 1, & \text{if } \mathbf{x}_i, \mathbf{x}_j \text{ in same class} \end{cases}$$

Hence,

$$\mathbf{K} \approx \begin{pmatrix} \mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & [\boldsymbol{\epsilon}]_{n_1 \times n_2} & \dots \\ [\boldsymbol{\epsilon}]_{n_2 \times n_1} & \mathbf{1}_{n_2}\mathbf{1}_{n_2}^T & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \text{ with } \boldsymbol{\epsilon}_{ij} \approx 0$$

in particular, $\forall i \in \{1, \dots, k\}$:

$$\mathbf{K}\mathbf{u}_i \approx \mathbf{D}\mathbf{u}_i, \text{ where } \mathbf{u}_i = [\underbrace{0, \dots, 0}_{n_1+\dots+n_{i-1}}, \underbrace{1, \dots, 1}_{n_i}, \underbrace{0, \dots, 0}_{n_{i+1}+\dots+n_k}]^T$$

So u_i is the canonical vector of $\mathcal{C}_i$ and eigenvector of $\mathbf{1D} - \mathbf{K}$.

Remark 3.2.1. *Notice that we also have :* $\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}^{\frac{1}{2}}\mathbf{u}_i) = \mathbf{D}^{\frac{1}{2}}\mathbf{u}_i$. *In practice it is more stable to study* $\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} - \mathbf{I}_p$ *then* $\mathbf{D} - \mathbf{K}$. *We will do so in the continuing, we are looking for the* ***largest*** *eigenvectors of* $\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} - \mathbf{I}_p$.

However, we do not obtain those results in large p, n regime. Here an example with equally draw 2-class Gaussian mixture data (with asymptotically non-trivial classification condition): $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$ with $\boldsymbol{\mu} = [2, \mathbf{0}_{p-1}]$.

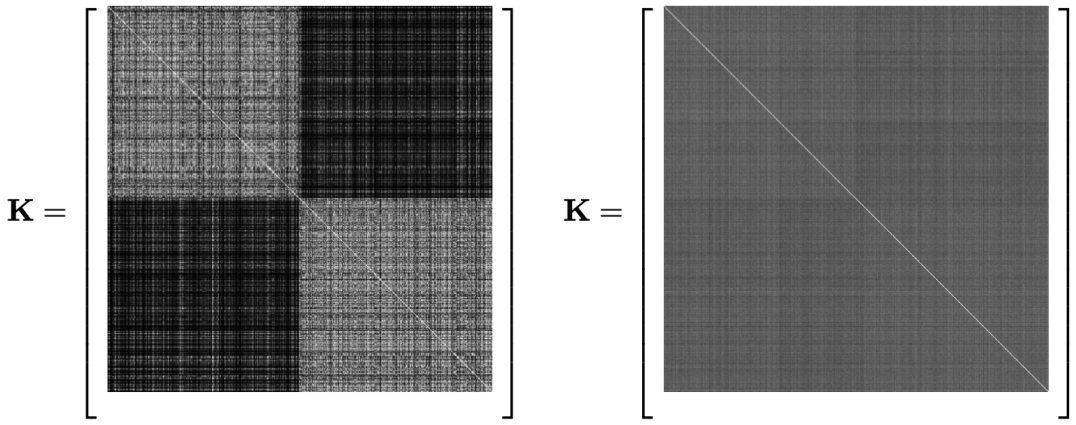

Figure 3.1: Kernel matrices $\mathbf{K}$ for small ($p = 5, n = 500$) and large ($p = 250, n = 500$) dimensional data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with $\mathbf{x}_1, \dots, \mathbf{x}_{\frac{n}{2}} \in \mathcal{C}_1$ and $\mathbf{x}_{\frac{n}{2}+1}, \dots, \mathbf{x}_n \in \mathcal{C}_2$.

3.3 Large Dimensional Intuition

We have already seen in 1.2 that this intuition is wrong! Indeed, in the case k-class Gaussian mixture data, in asymptotically non-trivial classification regime we have:

$$\underset{1\leq i\neq j\leq p}{max}\left\{\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|^2-\tau\right\}\underset{n,p\to\infty}{\longrightarrow}0 \tag{3.2}$$

Hence,

$$\mathbf{K}\approx\tau\begin{pmatrix}1&\dots&1\\ \vdots&\ddots&\vdots\\ 1&\dots&1\end{pmatrix}\neq\begin{pmatrix}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T&[\boldsymbol{\epsilon}]_{n_1\times n_2}&\dots\\ [\boldsymbol{\epsilon}]_{n_2\times n_1}&\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T&\dots\\ \vdots&\vdots&\ddots\end{pmatrix}$$

Question: Does this mean that the eigenvalues of $\mathbf{K}$ tends to 0 or n? No, to think that way would be to make the same mistake we made in the Sample Covariance Example in the first section.

Theorem 3.3.1 ([7])**.** *In the k-class Gaussian mixture case with asymptotically non-trivial classification conditions 1.2.1 and a kernel of the form $\mathbf{K}_{ij}=f\left(\frac{1}{p}\|\mathbf{x}_i-\mathbf{x}_j\|_2^2\right)$, we have an spectral asymptotic equivalent for the renormalized Laplacian matrix $\mathbf{L}$:*

$$\begin{aligned}\mathbf{L}&=n\mathbf{D}^{-\frac{1}{2}}\left(\mathbf{K}-\frac{\mathbf{d}\mathbf{d}^T}{\mathbf{d}^T\mathbf{1}_n}\right)\mathbf{D}^{-\frac{1}{2}}\\ \hat{\mathbf{L}}&=-2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}\underbrace{\mathbf{P}\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{P}}_{\approx MP}+\frac{1}{p}\underbrace{\mathbf{J}\mathbf{B}\mathbf{J}^{\mathbf{T}}}_{Low\ rank}+\underbrace{\star}_{Negligible}\end{aligned} \tag{3.3}$$

$$\textit{where}\begin{cases}\mathbf{P}=\mathbf{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T,\mathbf{W}=[\mathbf{w}_1,\dots,\mathbf{w}_n]\in\mathbb{R}^{p\times n},\ \textit{with}\ \mathbf{x}_i=\boldsymbol{\mu}_a+\mathbf{w}_i\\ \mathbf{J}=[\mathbf{j}_1,\dots,\mathbf{j}_n],\mathbf{d}=[n_1,\dots,n_k]\\ \mathbf{B}=-2\frac{f'(\tau)}{f(\tau)}\mathbf{M}^{\mathbf{T}}\mathbf{M}+\left(\frac{f''(\tau)}{f(\tau)}-\frac{5f'(\tau)^2}{4f(\tau)^2}\right)\mathbf{t}\mathbf{t}^T+2\frac{f''(\tau)}{f(\tau)}\mathbf{T}+\star\\ \mathbf{M}=[\boldsymbol{\mu}_1^o,\dots,\boldsymbol{\mu}_n],t=\frac{1}{\sqrt{p}}[tr\mathbf{C}_1^o,\dots,tr\mathbf{C}_k^o],\mathbf{T}=\left\{\frac{1}{p}tr\mathbf{C}_a^o\mathbf{C}_b^o\right\}_{a,b=1}^k\end{cases} \tag{3.4}$$

Main Intuition. The key idea is to do a Matrix Taylor Expansion and not term by term expansion. It is not because in the scalar expansion, the order decrease term after term, that it will be the case in terms of spectral energy in the matrix expansion. For simplicity, take $\mathbf{K}=\left\{f\left(\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j\right)\right\}_{i,j=1}^n$ (we forget the terms $\|\mathbf{x}_i\|_2^2$ and $\|\mathbf{x}_i\|_2^2$), and we perform a Taylor expansion around $f(0)$:

$$f\left(\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j\right)=\underbrace{f(0)}_{O(1)}+\underbrace{f'(0)\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j}_{O(\frac{1}{\sqrt{p}})}+\underbrace{\frac{1}{2}f''(0)\left(\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j\right)^2}_{O(\frac{1}{p})}+\star$$

to matrix expansion (neglect diagonal effect):

$$\mathbf{K}=f(0)\mathbf{1}_n\mathbf{1}_n^T+f'(0)\frac{1}{p}\mathbf{X}^T\mathbf{X}+\frac{1}{2}f''(0)\left\{\left(\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j\right)^2\right\}_{i,j=1}^n+\star$$

Now let's evaluate the spectral energy of each term, with $\mathbf{x}_i \in \mathcal{C}_a$, we can write $\mathbf{x}_i = \boldsymbol{\mu}_a + \mathbf{w}_i$ with $\|\boldsymbol{\mu}_a\|_2 = O(1)$ and $\mathbf{w}_i = O(\sqrt{p})$ then $\mathbf{X} = \mathbf{W} + \mathbf{M}\mathbf{J}^\mathbf{T}$:

$$\frac{1}{p}\mathbf{X}^T\mathbf{X} = O_{\|.\|}(1) \text{ because } \begin{cases} \|\frac{1}{p}\mathbf{W}^T\mathbf{W}\| = O(1)1.1 \\ \|\frac{1}{p}(\mathbf{M}\mathbf{J}^\mathbf{T})^\mathbf{T}(\mathbf{M}\mathbf{J}^\mathbf{T})\| = \|\frac{1}{p}\mathbf{J}\mathbf{M}^\mathbf{T}\mathbf{M}\mathbf{J}^\mathbf{T}\| = O(1) \end{cases}$$

in addition, writing $(\mathbf{x}_i^T\mathbf{x}_j)^2 = \mathbb{E}[(\mathbf{x}_i^T\mathbf{x}_j)^2] + \mathbf{Z}_{ij}$, and with $\mathrm{tr}(\mathbf{C}_a\mathbf{C}_b) = O(p)$:

$$\left\{\left(\frac{1}{p}\mathbf{x}_i^T\mathbf{x}_j\right)^2\right\}_{i,j=1}^n = \underbrace{\frac{1}{p^2}\begin{pmatrix} \mathrm{tr}\mathbf{C}_1^2\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & \mathrm{tr}\mathbf{C}_1\mathbf{C}_2\mathbf{1}_{n_1}\mathbf{1}_{n_2}^T & \dots \\ \mathrm{tr}\mathbf{C}_2\mathbf{C}_1^2\mathbf{1}_{n_2}\mathbf{1}_{n_1}^T & \mathrm{tr}\mathbf{C}_2^2\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}}_{O_{\|.\|}(1)} + \underbrace{\mathbf{Z}}_{\substack{\mathbb{E}[\mathbf{Z}_{ij}]=0 \\ Var(\mathbf{Z}_{ij})=O(\frac{1}{p^2})}} + \underbrace{\star}_{o_{\|.\|}(1)}$$

So we finally obtain:

$$K = f(0)\underbrace{\mathbf{1}_n\mathbf{1}_n^T}_{O_{\|.\|}(n)} + \underbrace{f'(0)\frac{1}{p}(\mathbf{W} + \mathbf{M}\mathbf{J}^\mathbf{T})^\mathbf{T}(\mathbf{W} + \mathbf{M}\mathbf{J}^\mathbf{T})}_{O_{\|.\|}(1)} + \underbrace{\frac{f''(0)}{2}\frac{1}{p}\mathbf{J}\mathbf{T}\mathbf{J}^\mathbf{T}}_{O_{\|.\|}(1)} + \underbrace{\star}_{o_{\|.\|}(1)}$$

The first term will have for eigenvalues $\{0, n\}$ that will blow up, the second term can disappear if we take $f'(0) = 0$ and the third one is deterministic, and contains classes information. □

Hence, with this matrix Taylor expansion, we obtained a spiked model that we perfectly know how to exploit:

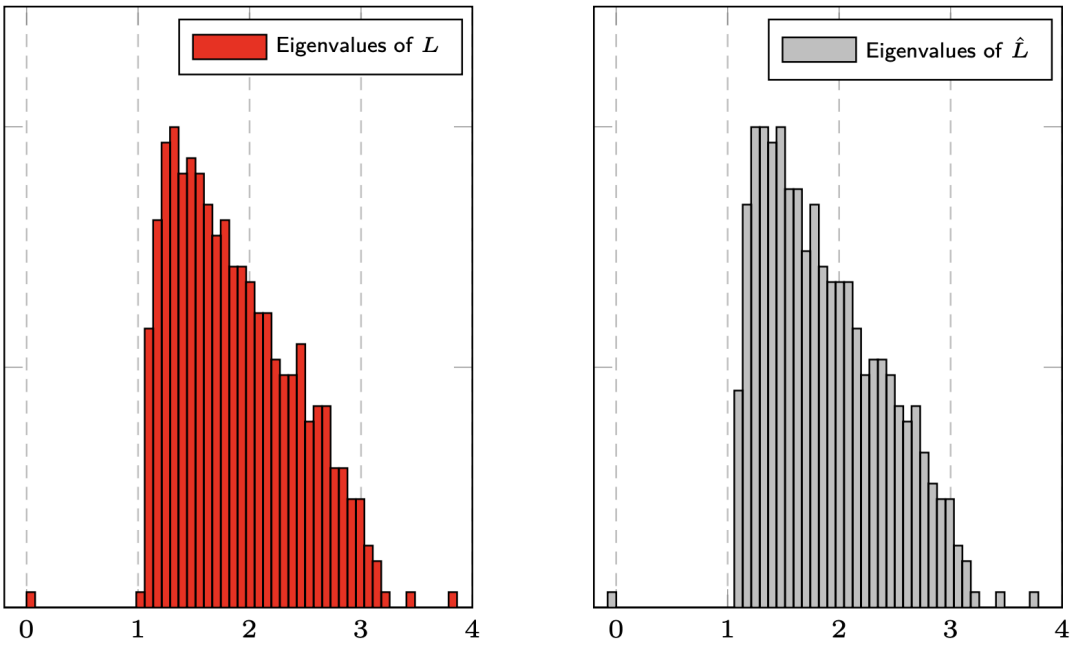

Figure 3.2: Eigenvalues of $\mathbf{L}$ and $\hat{\mathbf{L}}$, $k = 3, p = 2048, n = 512, 2c_1 = 2c_2 = c_3 = 1/2, [\boldsymbol{\mu}_{aj}] = 4\delta_{aj}, \mathbf{C}_a = (1 + 2(a-1)/\sqrt{p})I_p, f(x) = exp(-x/2)$.

3.4 Heat Kernel

A really striking consequence of 3.3.1 if that for a choice of f such as $f(\tau) = 0$ we kill the noise term ($\mathbf{PW^TWP}$) and are left with a deterministic term ($\mathbf{JBJ^T}$) that contain classes information [8].

$$\hat{\mathbf{L}} = -2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}\underbrace{\mathbf{PW^TWP}}_{Noise} + \frac{1}{p}\underbrace{\mathbf{JBJ^T}}_{\text{Information}} + \underbrace{\star}_{Negligible}$$

It is important to highlight that by setting $f(\tau) = 0$ we also kill the information concerning the means ($\mathbf{M^TM}$) in $\mathbf{B}$. So we will not be able to separate classes with different means but same covariance. However we can perfectly distinguish classes with different covariance matrices.

$$\mathbf{B} = -2\frac{f'(\tau)}{f(\tau)}\mathbf{M^TM} + \left(\frac{f''(\tau)}{f(\tau)} - \frac{5f'(\tau)^2}{4f(\tau)^2}\right)\mathbf{tt^T} + 2\frac{f''(\tau)}{f(\tau)}\mathbf{T} + \star$$

Hence, we see that the finite dimension intuition concerning the Gaussian Kernel has been proved wrong once again in large dimension ($f'(\mathbf{x}) > 0$). Here is an example (**Top**) for two equally drawn classes $\mathcal{N}(0, \mathbf{C}_1)$ and $\mathcal{N}(0, \mathbf{C}_2)$ with $\mathbf{C}_1 = \mathbf{I}_p$ and $\mathbf{C}_2 = \mathbf{I}_p + 2\mathbf{ZZ}^T/p$ where $\mathbf{Z} \in \mathbb{R}^{p \times \frac{p}{2}}$ has independent standard Gaussian entries. This is also verified with real life Data, here is an example **Bottom** of electrocardiogram data of sane v.s. epileptic patients:

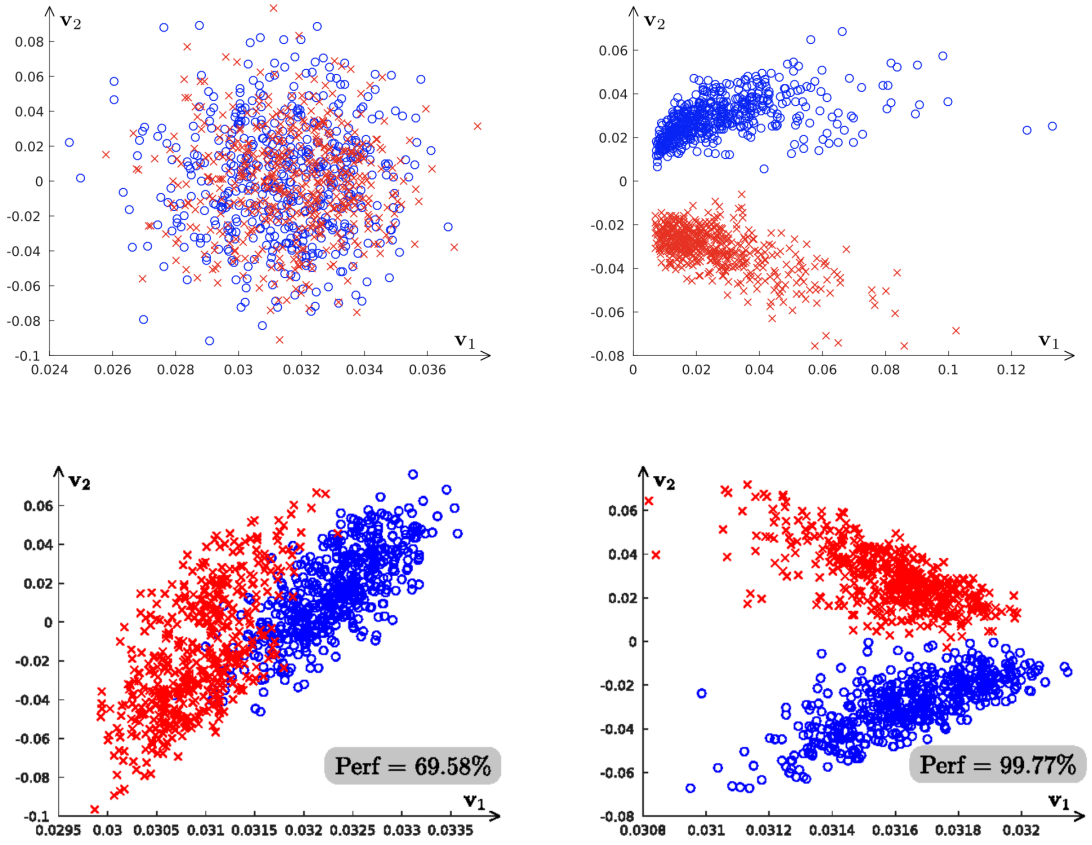

Figure 3.3: (**Left**) $\mathbf{K}_{ij} = exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2p}\right)$, (**Right**) $\mathbf{K}_{ij} = \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{p} - \tau\right)^2$

Real Data

In the previous chapters, we have repeatedly worked under the assumption that data arise from a Gaussian mixture model or is represented by vectors with independent entries to elaborate asymptotic performance analyses. Those assumptions primarily arises for mathematical convenience but these vector distributions are naturally deemed unrealistic models for realistic data. Yet, we do claim that scalar observations obtained from large data tend to behave as if the data were Gaussian mixtures in the first place. We justify this claim below with two strong arguments: one theoretical and one empirical.

4.1 Concentrated Random Vectors (CRV)

The modelling assumption that data $\mathbf{x}_i$ are affine maps $\mathbf{x}_i = \mathbf{A}\mathbf{z}_i + b$ of vectors $\mathbf{z}_i$ constituted of i.i.d. entries is a severe practical limitation as few real datasets are of this simplistic form. In [9], El Karoui provided a first means for RMT to go beyond the "vector of independent entries" assumption. There, relying on concentration of measure theory - extensively developed by Ledoux in [10] - El Karoui essentially showed that some of the early random matrix results from Pastur (2.2.1), Bai, and Silverstein (2.3.1), remain valid under the assumption that the $\mathbf{x}_i$'s are concentrated random vectors (CRV).

Intuitively, a CRV is a (random) point in high dimensional space having "predictable observations" $f(\mathbf{x})$, in the sense that, with (exponentially) high probability, $f(\mathbf{x})$ takes values very close to the deterministic M_f .

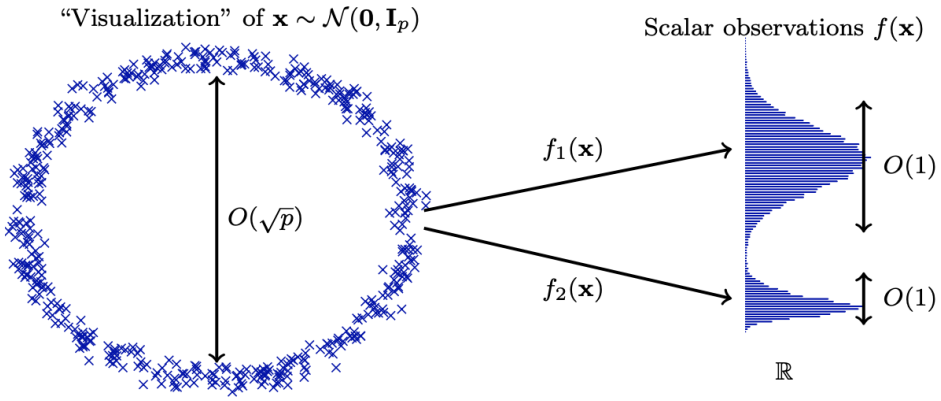

Figure 4.1: (**Left**) A visual "interpretation" of 500 independent drawings of $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_n)$.(**Right**) Concentration of the observation for linear ($f_1(\mathbf{x}) = \mathbf{x}^T\mathbf{1}_p/\sqrt{p}$) and Lipschitz ($f_2(\mathbf{x}) = \|\mathbf{x}\|_\infty$) maps.

Definition 4.1.1. *A random vector $\mathbf{x} \in \mathbb{R}^p$ is concentrated id, for a certain family of functions $f : \mathbb{R}^p \to \mathbb{R}$, there exists a deterministic scalar $M_f \in \mathbb{R}$ such that:*

$$\mathbb{P}(|\ f(\mathbf{x}) - M_f\ | \geq t) \leq Ce^{-ct^q},\ \textit{for } q > 0, C, c > 0$$

Thus, in the "observable world", the observation $f(\mathbf{x})$ appears to be "stable" for any CRV $\mathbf{x}$.

Paradoxically, very few "classical" multivariate distributions are known to produce CRV, and yet, this is enough to bring an outstanding practical competitive advantage against vectors with independent entries, when it comes to modelling real data for machine learning.

4.2 GAN: Generative Adversarial Networks

Indeed, Gaussian random vectors are *Lipschitz-concentrated*, hence they are stable by any lipschitz mapping from $\mathbb{R}^P \to \mathbb{R}^q$. But today, there exist Machine Learning techniques that learn to produce artificial but extremely realistic data, exclusively based on Lipschitz maps. The most popular of these methods are the Generative Adversarial Networks (GANs) initially proposed by Goodfellow et al. [11]. Those are feedforward neural networks which, after training, generate highly realistic data $f(\mathbf{x})$ from a standard Gaussian vector input $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_p)$). Since a feedforward neural network is a sequence of linear operators (interlayer connections and convolution operators) and Lipschitz nonlinear activation functions (sigmoidal, ReLU, etc.), f is indeed Lipschitz.

Figure 4.2: Image samples generated by BigGAN in [12]

Finally, an important finding in [13] showed that for most conventional Machine Learning methods, the asymptotic performances achieved on CRV mixtures coincide with those obtained on Gaussian mixtures - having the same means and covariances per class.

Those results combined - (i) GANs produced CRV so realistic that even human beings cannot distinguish synthetic data from real ones (ii) asymptotic performances on CRV can be obtain using adequate Gaussian mixtures - strongly suggests that (i) CRV are accurate models of real-world data, (ii) performances obtained from large data tend to behave as if the data were Gaussian mixtures.

4.3 MNIST Dataset

Figure 4.3: $28 \times 28 = 784$ pixels image of 0, 1, 2 from the Database MNIST

Now we will show this claim using Real Data. To do so, we will work with the famous MNIST Dataset [14]. We select three classes ($k = 3$), with 64 elements per class ($n_1 = n_2 = n_3 = 64$): 0, 1, 2. Then:

- We compute the Gaussain Kernel matrix $\mathbf{K}$ and compute the 4 largest eigenvectors of the normalized Laplacian Matrix $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} - \mathbf{I}_p$.
- We compute spectral asymptotic equivalent of $\mathbf{L}$, $\hat{\mathbf{L}} = -2\frac{f'(\tau)}{f(\tau)}\frac{1}{p}\mathbf{P}\mathbf{W}^{\mathbf{T}}\mathbf{W}\mathbf{P} + \frac{1}{p}\mathbf{J}\mathbf{B}\mathbf{J}^{\mathbf{T}}$ as if the Data was a Gaussian mixture (same mean and covariance).

And compare them:

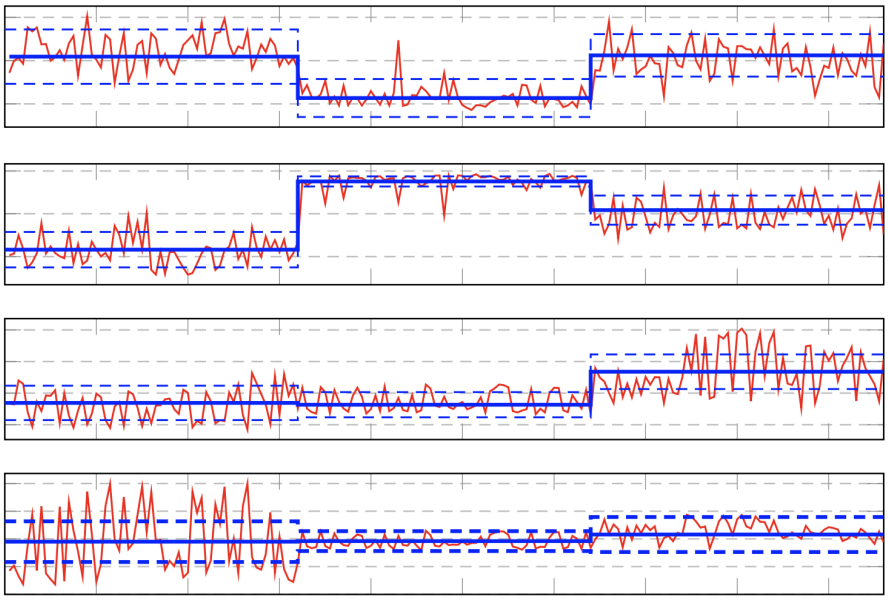

Figure 4.4: (**Red**) Eigenvalues of $\mathbf{L}$, (**Blue**) Eigenvalues of $\hat{\mathbf{L}}$ as if Gaussian mixture model. Top to Bottom: largest to smallest eigenvector

We see that we obtain extremely close results using a Gaussian mixture model. This is of a major importance: it means that we can predict the results and performance of a machine learning method beforehand using a Gaussian method. And use Gaussian Data to analyse the results, fine-tune the kernel, improve the pre-processing to finally - at the last moment - run everything on real Data.

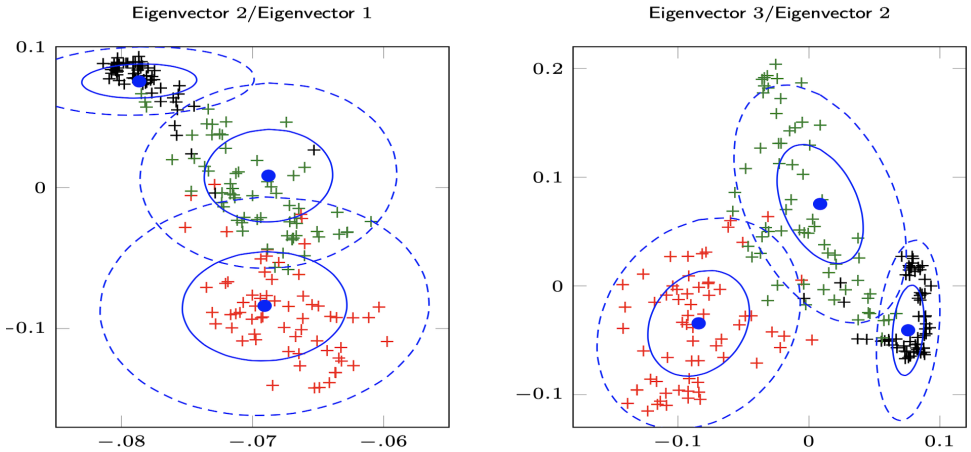

Figure 4.5: 2D representation of eigenvectors of $\mathbf{L}$, for the MNIST dataset. Theoretical means and 1- and 2-std in blue.

Conclusion

The main objective of this Diploma Thesis was to cast a light on the resulting biases of finite-dimensional intuition and their consequences on Machine Learning methods but also to provide a systematic random matrix framework to improve these algorithms: (i) one first needs to conceive the limitations of low dimensional intuitions and understand the reach of the large dimensional intuitions, (ii) capture the effect of the main mathematical objects at play in machine learning on large data models, (iii) include these objects in a mathematical framework of performance analysis, and (iv) foresee improvement methods based on the newly acquired large dimensional intuitions and mathematical understanding.

Bibliography

[1] R. C. Jamal Najim, *Introduction a la theorie des grandes matrices aleatoires*. Course Notes: http://polaris.imag.fr/romain.couillet/docs/courses/rmt/poly.pdf, 2021.

[2] Z. L. Romain Couillet, *Random Matrix Theory for Machine Learning*. Draft Book: http://polaris.imag.fr/romain.couillet/docs/RMT_ML_Book.pdf, 2022.

[3] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.

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

[5] L. A. Pastur and M. Shcherbina, *Eigenvalue distribution of large random matrices*. American Mathematical Soc., 2011, no. 171.

[6] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. Springer, 2010, vol. 20.

[7] 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.

[8] H. T. 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.

[9] N. El Karoui, "Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond," *The Annals of Applied Probability*, vol. 19, no. 6, pp. 2362–2405, 2009.

[10] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[12] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[13] C. Louart and R. Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," *arXiv preprint arXiv:1805.08295*, 2018.

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