CAMBRIDGE

Cambridge University Press & Assessment
978-1-009-12323-5 — Random Matrix Methods for Machine Learning
Romain Couillet , Zhenyu Liao
Excerpt
More Information

1 Introduction

This chapter discusses fundamentally different mental images of large- versus small-dimensional machine learning through examples of sample covariance and kernel matrices, on both synthetic and real data. Random matrix theory is presented as a flexible and powerful tool to assess, understand, and improve classical machine learning methods in this modern large-dimensional setting.

1.1 Motivation: The Pitfalls of Large-Dimensional Statistics

1.1.1 The Big Data Era: When n Is No Longer Much Larger than p

The big data revolution comes along with the challenging needs to parse, mine, and compress a 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 array processing and wireless communications, the number of antennas required for fine localization resolution or increased communication throughput may be as large (today in the order of hundreds) as the number of available independent signal observations [Li and Stoica, 2007, Lu et al., 2014]. In genomics, the identification of correlations among hundreds of thousands of genes based on a limited number of independent (and expensive) samples induces an even larger ratio p/n [Arnold et al., 1994]. In statistical finance, portfolio optimization relies on the need to invest on a large number p of assets to reduce volatility but at the same time to estimate the current (rather than past) asset statistics from a relatively small number n of asset return records [Laloux et al., 2000].

As we shall demonstrate in the following section, the fact that in these problems n is not *much larger* than p annihilates most of the results from standard asymptotic statistics that assume n alone is large [Vaart, 2000]. As a rule of thumb, by "much larger" we mean here that n must be at least 100 times larger than p for standard asymptotic statistics to be of practical convenience (see our argument in Section 1.1.2). Many algorithms in statistics, signal processing, and machine learning are precisely derived from this $n \gg p$ assumption that is no longer appropriate today. A major objective of this book is to cast some light on the resulting biases and problems incurred and to provide a systematic random matrix framework to improve these algorithms.

Possibly more importantly, we will see in this book that (small p) small-dimensional intuitions at the core of many machine learning algorithms (starting with spectral clustering [Ng et al., 2002, Luxburg, 2007]) may strikingly fail when applied in a simultaneously large n,p setting. A compelling example lies in the notion of "distance" between vectors. Most classification methods in machine learning are rooted in the observation that random data vectors arising from a mixture distribution (say Gaussian) gather in "groups" of close-by vectors in the Euclidean norm. When dealing with large-dimensional data, however, concentration phenomena arise that make Euclidean distances useless, if not counterproductive: Vectors from the *same* mixture class may be further away in Euclidean distance than vectors arising from *different* classes. While classification may still be doable, it works in a rather different way from our small-dimensional intuition. The book intends to prepare the reader for the multiple traps caused by this "curse of dimensionality."

1.1.2 Sample Covariance Matrices in the Large *n,p* Regime

Let us consider the following example that illustrates a first elementary, yet counterintuitive, result: For simultaneously large n,p, the sample covariance matrix $\hat{\mathbf{C}} \in \mathbb{R}^{p\times p}$ based on n samples $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{C})$ is an *entry-wise* consistent estimator of the population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ (i.e., $\|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \to 0$ as $p,n \to \infty$ for $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$) while overall being an extremely poor estimator in a (more practical) operator norm sense (i.e., $\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0$, with $\|\cdot\|$ being the operator norm here). Matrix norms are, in particular, *not* equivalent in the large n,p scenario.

Let us detail this claim, in the simplest case where $\mathbf{C} = \mathbf{I}_p$. Consider a dataset $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n] \in \mathbb{R}^{p\times n}$ of n independent and identically distributed (i.i.d.) observations from a p-dimensional standard Gaussian distribution, that is, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}_p)$ for $i \in \{1,\ldots,n\}$. We wish to estimate the population covariance matrix $\mathbf{C} = \mathbf{I}_p$ from the n available samples. The maximum likelihood estimator in this zero-mean Gaussian setting is the sample covariance matrix $\hat{\mathbf{C}}$ defined by

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}. \tag{1.1}$$

By the strong law of large numbers, for fixed p, $\hat{\mathbf{C}} \to \mathbf{I}_p$ almost surely as $n \to \infty$, so that $\|\hat{\mathbf{C}} - \mathbf{I}_p\| \xrightarrow{\text{a.s.}} 0$ holds for any standard matrix norm and in particular for the operator norm.

One must be more careful when dealing with the case $n,p \to \infty$ with the ratio $p/n \to c \in (0,\infty)$ (or, from a practical standpoint, n is *not much larger* than p). First, note that the entry-wise convergence still holds since, invoking the law of large numbers again,

$$[\hat{\mathbf{C}}]_{ij} = \frac{1}{n}\sum_{l=1}^{n} [\mathbf{X}]_{il}[\mathbf{X}]_{jl} \xrightarrow{\text{a.s.}} \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

Besides, by a concentration inequality argument, it can even be shown that

$$\max_{1\le i,j\le p} \left|[\hat{\mathbf{C}} - \mathbf{I}_p]_{ij}\right| \xrightarrow{\text{a.s.}} 0,$$

which holds as long as p is no larger than a polynomial function of n, and thus:

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Consider now the case $p > n$. Since $\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$ is the sum of n rank-one matrices, the rank of $\hat{\mathbf{C}}$ is *at most* equal to n and thus, being a $p \times p$ matrix with $p > n$, the sample covariance matrix $\hat{\mathbf{C}}$ must be a *singular* matrix having at least $p - n > 0$ null eigenvalues. As a consequence,

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

for $\|\cdot\|$ the matrix operator (or spectral) norm. This last result actually extends to the general case where $p/n \to c \in (0,\infty)$. As such, matrix norms cannot be considered equivalent in the regime where p is not negligible compared to n. This follows from the fact that the coefficients involved in the *equivalence of norm* relation between the infinity and operator norm *depend on* p; here, for instance, we have that for symmetric matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\|\mathbf{A}\|_\infty \le \|\mathbf{A}\| \le p\|\mathbf{A}\|_\infty$.

Unfortunately, in practice, the (nonconverging) operator norm is of more practical interest than the (converging) infinity norm.

Remark 1.1 (On the importance of operator norm)**.** *For practical purposes, this "loss" of norm equivalence for large p raises the question of the relevant matrix norm to consider for a given application. For the purpose of the present book, and for most applications in machine learning, the operator (or spectral) norm is the most relevant. First, the operator norm is the matrix norm induced by the Euclidean norm of vectors. Thus, the study of regression vectors or label/score vectors in classification is naturally attached to the spectral study of matrices. Besides, we will often be interested in the* asymptotic equivalence *of families of large-dimensional symmetric matrices. If* $\|\mathbf{A}_p - \mathbf{B}_p\| \to 0$ *for matrix sequences* $\{\mathbf{A}_p\}$ *and* $\{\mathbf{B}_p\}$*, indexed by their dimension p, then according to Weyl's inequality (see, e.g., Lemma 2.10 in Section 2.2.1),*

$$\max_i \left|\lambda_i(\mathbf{A}_p) - \lambda_i(\mathbf{B}_p)\right| \to 0$$

for $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \cdots$*, the eigenvalues of* $\mathbf{A}$ *in a decreasing order. Besides, for* $\mathbf{u}_i(\mathbf{A}_p)$*, an eigenvector of* $\mathbf{A}_p$ *associated with an* isolated eigenvalue $\lambda_i(\mathbf{A}_p)$ *(i.e., such that* $\min\{|\lambda_{i+1}(\mathbf{A}_p) - \lambda_i(\mathbf{A}_p)|, |\lambda_i(\mathbf{A}_p) - \lambda_{i-1}(\mathbf{A}_p)|\} > \varepsilon$ *for some* $\varepsilon > 0$ *uniformly on p),*

$$\left\|\mathbf{u}_i(\mathbf{A}_p) - \mathbf{u}_i(\mathbf{B}_p)\right\| \to 0.$$

These results ensure that, as far as spectral properties are concerned, $\mathbf{A}_p$ *can be studied* equivalently *through* $\mathbf{B}_p$*. We will often use this argument to investigate intractable random matrices* $\mathbf{A}_p$ *by means of a more tractable "proxy"* $\mathbf{B}_p$*.*

The pitfall that consists in assuming that $\hat{\mathbf{C}}$ *is a valid estimator of* $\mathbf{C}$ *since* $\|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \xrightarrow{\text{a.s.}} 0$ *may thus have deleterious practical consequences when n is not* significantly larger *than p.*

Resuming our discussion of norm convergence, it is now natural to ask whether $\hat{\mathbf{C}}$, which badly estimates $\mathbf{C}$, has a controlled asymptotic behavior. There precisely lay the first theoretical interests of random matrix theory. While $\hat{\mathbf{C}}$ itself does not converge in

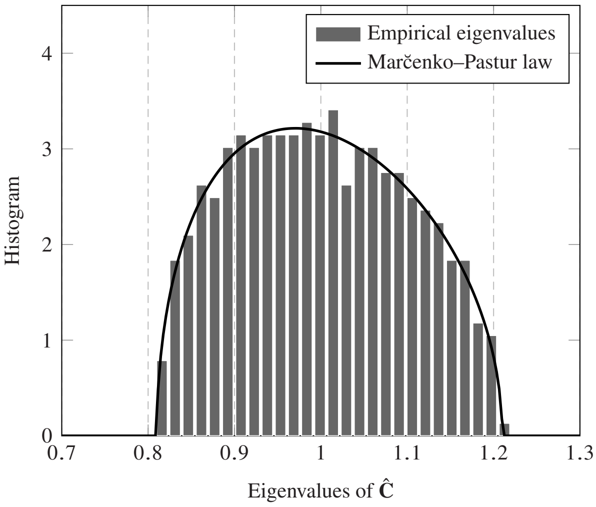


Figure 1.1 Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the Marčenko–Pastur law, for $\mathbf{X}$ having standard Gaussian entries, $p = 500$ and $n = 50\,000$. Code on web: MATLAB and Python.

any useful way, its eigenvalue distribution does exhibit a traceable limiting behavior [Marčenko and Pastur, 1967, Silverstein and Bai, 1995, Bai and Silverstein, 2010]. The seminal result in this direction, due to Marčenko and Pastur, states that, for $\mathbf{C} = \mathbf{I}_p$, as $n,p \to \infty$, with $p/n \to c \in (0,\infty)$, it holds with probability 1 that the random *discrete eigenvalue/empirical spectral distribution*

$$\mu_p \equiv \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\hat{\mathbf{C}})}$$

converges in law to a nonrandom *smooth* limit, today referred to as the "Marčenko–Pastur law" [Marčenko and Pastur, 1967],

$$\mu(dx) = (1-c^{-1})^+\delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x-E_-)^+(E_+-x)^+}\,dx, \tag{1.2}$$

where $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ \equiv \max(x,0)$.

Figure 1.1 compares the empirical spectral distribution of $\hat{\mathbf{C}}$ to the limiting Marčenko–Pastur law given in (1.2), for $p = 500$ and $n = 50\,000$.

The elementary Marčenko–Pastur result is already quite instructive and insightful.

Remark 1.2 (When is one under the random matrix regime?). *Equation* (1.2) *reveals that the eigenvalues of* $\hat{\mathbf{C}}$*, instead of concentrating at* $x = 1$ *as a large-n alone analysis would suggest, are spread from* $(1-\sqrt{c})^2$ *to* $(1+\sqrt{c})^2$*. As such, the eigenvalues span on a range*

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

This is a slow decaying behavior with respect to $c = \lim p/n$*. In particular, for* $n = 100p$*, in which case, one would expect a sufficiently large number of samples for* $\hat{\mathbf{C}}$ *to properly estimate* $\mathbf{C} = \mathbf{I}_p$*, one has* $4\sqrt{c} = 0.4$*, which is a large spread around* *the mean (and true) eigenvalue* 1*. This is visually confirmed by Figure 1.1 for* $p = 500$ *and* $n = 50\,000$*, where the histogram of the eigenvalues is nowhere near concentrated at* $x = 1$*. Therefore, random matrix results will be much more accurate than classical asymptotic statistics even when* $n \sim 100p$*. As a telling example, estimating the covariance matrix of each digit from the popular Modified National Institute of Standards and Technology (MNIST) dataset [LeCun et al., 1998], made of no more than* 60 000 *training samples (and thus about* $n = 6\,000$ *samples per digit) of size* $p = 784$*, is likely a hazardous undertaking.*

Remark 1.3 (On universality)**.** *Although introduced here in the context of a Gaussian distribution for* $\mathbf{x}_i$*, the Marčenko–Pastur law applies to much more general cases. Indeed, the result remains valid as long as the* $\mathbf{x}_i$*s have independent normalized entries of zero mean and unit variance (and even beyond this setting, see El Karoui [2009] and Louart and Couillet [2018]). Similar to the law of large numbers in standard asymptotic statistics, this* universality *phenomenon commonly arises in random matrix theory and large-dimensional statistics. We will exploit this phenomenon in the book to justify the wide applicability of the presented results, even to real datasets. See Chapter 8 for more detail.*

1.1.3 Kernel Matrices of Large-Dimensional Data

Another less-known but equally important example of the curse of dimensionality in machine learning involves the loss of relevance of (the notion of) Euclidean distance between large-dimensional data vectors. To be more precise, we will see in the sequel that, *in an asymptotically nontrivial classification setting* (i.e., ensuring that asymptotic classification is neither trivially easy nor impossible), large and numerous data vectors $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^p$ extracted from a few-class (say two-class) mixture model tend to be asymptotically at *equal* (Euclidean) distance from one another, irrespective of their corresponding class. Roughly speaking, in this nontrivial setting and under some reasonable statistical assumptions on the $\mathbf{x}_i$s, we have

$$\max_{1\le i\neq j\le n} \left\{ \frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \to 0 \tag{1.3}$$

for some constant $\tau > 0$ as $n,p \to \infty$, *independently of the classes (same or different) of* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ (here the normalization by p is used for compliance with the notations in the remainder of this book and has no particular importance).

This asymptotic behavior is extremely counterintuitive and conveys the idea that classification by standard methods ought *not* to be doable in this large-dimensional regime. Indeed, in the conventional small-dimensional intuition that forged many of the leading machine learning algorithms of everyday use (such as spectral clustering [Ng et al., 2002, Luxburg, 2007]), two data points are assigned to the same class if they are “close” in Euclidean distance. Here we claim that, when p is large, *data pairs are neither close nor far* from each other, regardless of their belonging to the same class or not. Despite this troubling loss of *individual* discriminative power between data pairs, we subsequently show that, thanks to a *collective* behavior of all data belonging to the same (few and thus large) classes, data classification or clustering is still achievable. Better, we shall see that, while many conventional methods devised from small-dimensional intuitions do fail in this large-dimensional regime, some popular approaches, such as the Ng–Jordan–Weiss spectral clustering method [Ng et al., 2002] or the PageRank semisupervised learning approach [Avrachenkov et al., 2012], still function. But the core reasons for their functioning are strikingly different from the reasons of their initial designs, and they often operate far from optimally.

The Nontrivial Classification Regime

To get a clear picture of the source of Equation (1.3), we first need to clarify what we refer to as the "asymptotically nontrivial" classification setting. Consider the simplest scenario of a binary Gaussian mixture classification: Given a training set $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^p$ of n samples independently drawn from the two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) Gaussian mixture,

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu},\mathbf{I}_p), \quad \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu},\mathbf{I}_p+\mathbf{E}), \tag{1.4}$$

each drawn with probability 1/2, for some deterministic $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric $\mathbf{E} \in \mathbb{R}^{p\times p}$, both possibly depending on p. In the ideal case where $\boldsymbol{\mu}$ and $\mathbf{E}$ are perfectly known, one can devise a (decision optimal) Neyman–Pearson test. For an unknown $\mathbf{x}$, genuinely belonging to $\mathcal{C}_1$, the Neyman–Pearson test to decide on the class of $\mathbf{x}$ reads

$$(\mathbf{x}+\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{I}_p+\mathbf{E})^{-1}(\mathbf{x}+\boldsymbol{\mu}) - (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x}-\boldsymbol{\mu}) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} -\log\det(\mathbf{I}_p+\mathbf{E}). \tag{1.5}$$

Writing $\mathbf{x} = \boldsymbol{\mu}+\mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_p)$, the above test is equivalent to

$$\begin{aligned} T(\mathbf{x}) \equiv{}& 4\boldsymbol{\mu}^{\mathsf{T}}(\mathbf{I}_p+\mathbf{E})^{-1}\boldsymbol{\mu} + 4\boldsymbol{\mu}^{\mathsf{T}}(\mathbf{I}_p+\mathbf{E})^{-1}\mathbf{z} + \mathbf{z}^{\mathsf{T}}\left((\mathbf{I}_p+\mathbf{E})^{-1}-\mathbf{I}_p\right)\mathbf{z} \\ &+\log\det(\mathbf{I}_p+\mathbf{E}) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} 0. \end{aligned} \tag{1.6}$$

Since $\mathbf{U}\mathbf{z}$ for $\mathbf{U} \in \mathbb{R}^{p\times p}$, an eigenvector basis of $(\mathbf{I}_p+\mathbf{E})^{-1}$ (and thus of $(\mathbf{I}_p+\mathbf{E})^{-1} - \mathbf{I}_p$), follows the same distribution as $\mathbf{z}$, the random variable $T(\mathbf{x})$ can be written as the sum of p independent random variables. Further assuming that $\|\boldsymbol{\mu}\| = O(1)$ with respect to p, by Lyapunov's central limit theorem (e.g., [Billingsley, 2012, Theorem 27.3]) and the fact that $\operatorname{Var}[\mathbf{z}^{\mathsf{T}}\mathbf{A}\mathbf{z}] = 2\operatorname{tr}(\mathbf{A}^2)$ for symmetric $\mathbf{A} \in \mathbb{R}^{p\times p}$ and Gaussian $\mathbf{z}$, we have, as $p \to \infty$,

$$V_T^{-1/2}(T(\mathbf{x})-\bar{T}) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned} \bar{T} &\equiv 4\boldsymbol{\mu}^{\mathsf{T}}(\mathbf{I}_p+\mathbf{E})^{-1}\boldsymbol{\mu} + \operatorname{tr}(\mathbf{I}_p+\mathbf{E})^{-1} - p + \log\det(\mathbf{I}_p+\mathbf{E}), \\ V_T &\equiv 16\boldsymbol{\mu}^{\mathsf{T}}(\mathbf{I}_p+\mathbf{E})^{-2}\boldsymbol{\mu} + 2\operatorname{tr}\left((\mathbf{I}_p+\mathbf{E})^{-1}-\mathbf{I}_p\right)^2. \end{aligned}$$

CAMBRIDGE

Cambridge University Press & Assessment
978-1-009-12323-5 — Random Matrix Methods for Machine Learning
Romain Couillet , Zhenyu Liao
Excerpt
More Information

As a consequence, the classification of $\mathbf{x} \in \mathcal{C}_1$ is asymptotically nontrivial (i.e., the classification error neither goes to 0 nor 1 as $p \to \infty$) if and only if $\bar{T}$ is of the same order as $\sqrt{V_T}$. Considering the (worst-case) scenario where $\mathbf{E} = \mathbf{0}$, we must have $\|\boldsymbol{\mu}\| \geq O(1)$ with respect to p (indeed, if instead $\|\boldsymbol{\mu}\| = o(1)$, the classification of $\mathbf{x}$ is asymptotically impossible).

Under the constraint $\|\boldsymbol{\mu}\| = O(1)$, we move on to consider the case $\mathbf{E} \neq \mathbf{0}$ with the spectral norm constraint $\|\mathbf{E}\| = o(1)$. By a Taylor expansion of both $(\mathbf{I}_p + \mathbf{E})^{-1}$ and $\log\det(\mathbf{I}_p + \mathbf{E})$ around $\mathbf{I}_p$, we obtain

$$\bar{T} = 4\|\boldsymbol{\mu}\|^2 - \frac{1}{2}\operatorname{tr}(\mathbf{E}^2) + o(1);$$
$$V_T = 16\|\boldsymbol{\mu}\|^2 + 2\operatorname{tr}(\mathbf{E}^2) + o(1),$$

which demands $\operatorname{tr}(\mathbf{E}^2)$ to be of order $O(1)$ (same as $\|\boldsymbol{\mu}\|$) so as to have discriminative power. Since $\operatorname{tr}(\mathbf{E}^2) \leq p\|\mathbf{E}\|^2$, with equality if and only if $\mathbf{E}$ is proportional to the identity, that is, $\mathbf{E} = \epsilon\mathbf{I}_p$, one must have $\|\mathbf{E}\| \geq O(p^{-1/2})$. Also, since $O(1) = \operatorname{tr}(\mathbf{E}^2) \leq (\operatorname{tr}\mathbf{E})^2$, we must have $|\operatorname{tr}\mathbf{E}| \geq O(1)$. This allows us to conclude on the following nontrivial classification conditions:

$$\|\boldsymbol{\mu}\| \geq O(1), \quad \|\mathbf{E}\| \geq O(p^{-1/2}), \quad |\operatorname{tr}(\mathbf{E})| \geq O(1), \quad \operatorname{tr}(\mathbf{E}^2) \geq O(1). \tag{1.7}$$

These are the *minimal* conditions for classification in the case of perfectly known means and covariances in the following sense: (i) if none of the inequalities hold (i.e., if the means and covariances from both classes are too close), asymptotic classification must fail and (ii) if at least one of the inequalities is not tight (say if $\|\boldsymbol{\mu}\| \geq O(\sqrt{p})$), asymptotic classification becomes trivial.[1]

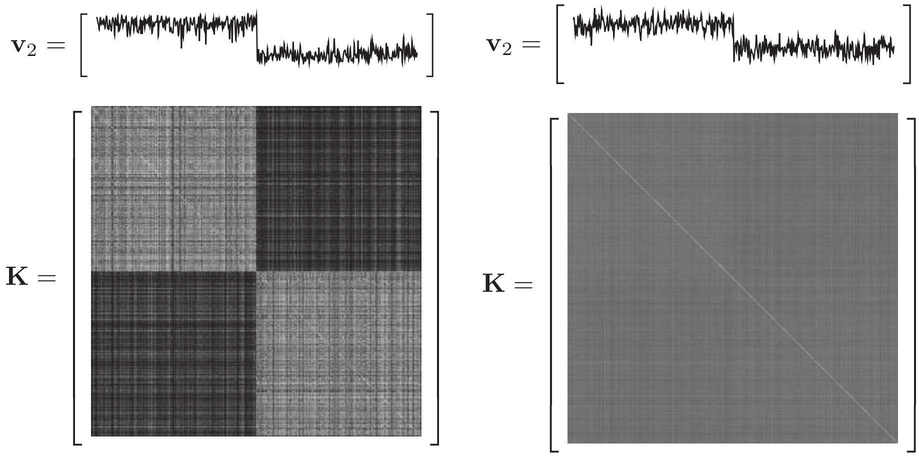
We shall subsequently see that (1.7) precisely induces the asymptotic loss of distance discrimination raised in (1.3) but that standard spectral clustering methods based on $n \sim p$ data remain valid.

Asymptotic Loss of Pairwise Distance Discrimination

Under the equality case for the conditions in (1.7), consider the (normalized) Euclidean distance between two distinct data vectors $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b, i \neq j$, given by

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \begin{cases} \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + Ap^{-1}, & \text{for } a = b = 2 \\ \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + Bp^{-1}, & \text{for } a = 1, b = 2, \end{cases} \tag{1.8}$$

[1] It should be noted here that, unlike in computer science, we will stick in this book with the notation $O(\cdot)$ indifferently from the complexity notations $\Omega(\cdot)$, $O(\cdot)$, and $\Theta(\cdot)$. The exact meaning of $O(\cdot)$ will be clear in context. For instance, under computer science notations, Equation (1.7) would be $\|\boldsymbol{\mu}\| \geq \Theta(1)$, $\|\mathbf{E}\| \geq \Theta(p^{-1/2})$, $|\operatorname{tr}(\mathbf{E})| \geq \Theta(1)$, and $\operatorname{tr}(\mathbf{E}^2) \geq \Theta(1)$.



(a) $p = 5$ (b) $p = 250$

Figure 1.2 Gaussian kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for **(a)** small- and **(b)** large-dimensional data $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n] \in \mathbb{R}^{p\times n}$, with $\mathbf{x}_1,\ldots,\mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1},\ldots,\mathbf{x}_n \in \mathcal{C}_2$ for $n = 5000$. Code on web: MATLAB and Python.

where

$$A = \mathbf{z}_i^\mathsf{T}\mathbf{E}\mathbf{z}_i + \mathbf{z}_j^\mathsf{T}\mathbf{E}\mathbf{z}_j - 2\mathbf{z}_i^\mathsf{T}\mathbf{E}\mathbf{z}_j \text{ and}$$

$$B = \mathbf{z}_j^\mathsf{T}(\mathbf{E}+\mathbf{E}^2/4)\mathbf{z}_j - \mathbf{z}_i^\mathsf{T}\mathbf{E}\mathbf{z}_j + 4\|\boldsymbol{\mu}\|^2 + 4\boldsymbol{\mu}^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j) + o(1)$$

are both of order $O(1)$ $\big($and thus both Ap^{-1} and Bp^{-1} are of order $O(p^{-1})\big)$, while the leading term $\frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2$ of (1.8) is of order $O(1)$. As such,

$$\max_{1\le i\neq j\le n}\left\{\frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 - 2\right\} \to 0$$

almost surely as $n,p \to \infty$ (this follows by exploiting the fact that $\|\mathbf{z}_i - \mathbf{z}_j\|^2$ is a chi-square random variable with p degrees of freedom). As a consequence, as previously claimed in (1.3),

$$\max_{1\le i\neq j\le n}\left\{\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau\right\} \to 0$$

for $\tau = 2$ here. Besides, on a closer inspection of (1.8), we find that, beyond this common value τ of order $O(1)$, the discriminative class information in means $4\|\boldsymbol{\mu}\|^2/p$ and that in covariances $\mathbf{z}_j^\mathsf{T}(\mathbf{E}+\mathbf{E}^2/4)\mathbf{z}_j/p \simeq \operatorname{tr}(\mathbf{E}+\mathbf{E}^2/4)/p$ are both of order $O(p^{-1})$, while by the central limit theorem, $\|\mathbf{z}_i - \mathbf{z}_j\|^2/p = 2 + O(p^{-1/2})$. The class information is thus largely overtaken by the random fluctuations. As a consequence, asymptotically, the pairwise distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ contains *no* exploitable statistical information (about $\boldsymbol{\mu}$ or $\mathbf{E}$) to distinguish if the $\mathbf{x}_i$ and $\mathbf{x}_j$ vectors belong to the same or different classes.

To visually confirm this joint convergence of the data distances, in Figure 1.2, we display the content of the Gaussian (heat) kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, with $[\mathbf{K}]_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2p)\right)$, and the associated second dominant eigenvector $\mathbf{v}_2$ for a two-class Gaussian mixture $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu},\mathbf{I}_p)$, with $\boldsymbol{\mu} = [2;\ \mathbf{0}_{p-1}]$. For a constant $n = 500$, we take $p = 5$ in Figure 1.2(a) and $p = 250$ in Figure 1.2(b).

While the "block-structure" in the case of $p = 5$ of Figure 1.2(a) does agree with the small-dimensional intuition – data vectors from the same class are "closer" to one another in diagonal blocks with larger values (since $\exp(-x/2)$ decreases with x) than in nondiagonal blocks – this intuition collapses when large-dimensional data vectors are considered. Indeed, in the large data setting of Figure 1.2(b), all entries (except obviously on the diagonal) of $\mathbf{K}$ have approximately the same value, which, we now know from (1.3), is $\exp(-1)$.

This is no longer surprising to us. However, what remains surprising in Figure 1.2 at this stage of our analysis is that the eigenvector $\mathbf{v}_2$ of $\mathbf{K}$ seems *not* affected by this (asymptotic) loss of class-wise discrimination of individual distances. And spectral clustering seems to work equally well for $p = 5$ and for $p = 250$, despite the radical and intuitively destructive change in the behavior of $\mathbf{K}$ for $p = 250$.

Explaining Kernel Methods with Random Matrix Theory

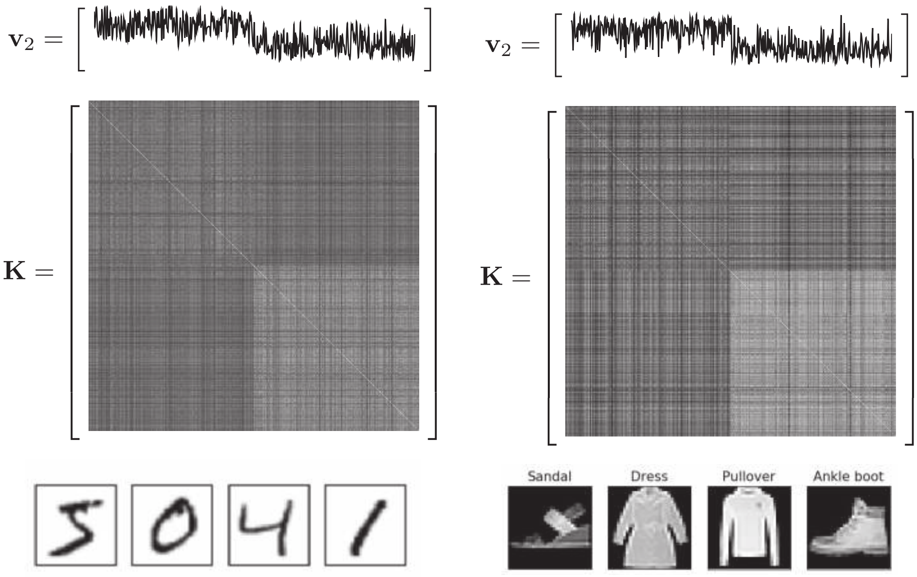
The fundamental reason behind this surprising behavior lies in the *accumulated* effect of the $n/2$ small "hidden" informative terms $\|\boldsymbol{\mu}\|^2$, $\operatorname{tr}\mathbf{E}$ and $\operatorname{tr}(\mathbf{E}^2)$ in each class, which collectively "steer" the several top eigenvectors of $\mathbf{K}$. More explicitly, we shall see in the course of this book that the Gaussian kernel matrix $\mathbf{K}$ can be asymptotically expanded as

$$\mathbf{K} = \exp(-1)\left(\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}} + \frac{1}{p}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right) + f(\boldsymbol{\mu},\mathbf{E})\cdot\frac{1}{p}\mathbf{j}\mathbf{j}^{\mathsf{T}} + * + o_{\|\cdot\|}(1), \tag{1.9}$$

where $\mathbf{Z} = [\mathbf{z}_1,\ldots,\mathbf{z}_n] \in \mathbb{R}^{p\times n}$ is a Gaussian noise matrix, $f(\boldsymbol{\mu},\mathbf{E}) = O(1)$, and $\mathbf{j} = [\mathbf{1}_{n/2};\ \ -\mathbf{1}_{n/2}]$ is the class-information "label" vector (as in the setting of Figure 1.2). Here "*" symbolizes extra terms of marginal importance to the present discussion, and $o_{\|\cdot\|}(1)$ represents terms of asymptotically vanishing *operator* norm as $n,p \to \infty$. The important remark to be made here is that

(i) Under this description, $[\mathbf{K}]_{ij} = \exp(-1)(1 + \mathbf{z}_i^{\mathsf{T}}\mathbf{z}_j/p) \pm f(\boldsymbol{\mu},\mathbf{E})/p + *$, with $f(\boldsymbol{\mu},\mathbf{E})/p \ll \mathbf{z}_i^{\mathsf{T}}\mathbf{z}_j/p = O(p^{-1/2})$; this is consistent with our previous discussion: The statistical information is *entry-wise* dominated by noise.

(ii) From a *spectral* viewpoint, $\|\mathbf{Z}^{\mathsf{T}}\mathbf{Z}/p\| = O(1)$, as per the Marčenko–Pastur theorem [Marčenko and Pastur, 1967] discussed in Section 1.1.2 and visually confirmed in Figure 1.1, while $\|f(\boldsymbol{\mu},\mathbf{E})\cdot\mathbf{j}\mathbf{j}^{\mathsf{T}}/p\| = O(1)$: Thus, *spectrum-wise*, the information stands on even ground with noise.



(a) MNIST data

(b) Fashion-MNIST data

Figure 1.3 Gaussian kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for **(a)** MNIST [LeCun et al., 1998] (class 8 versus 9) and **(b)** Fashion-MNIST [Xiao et al., 2017] data (class 5 versus 7), with $\mathbf{x}_1,\ldots,\mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1},\ldots,\mathbf{x}_n \in \mathcal{C}_2$ for $n = 5\,000$. Code on web: MATLAB and Python.

The mathematical magic at play here lies in $f(\boldsymbol{\mu},\mathbf{E}) \cdot \mathbf{j}\mathbf{j}^\mathsf{T}/p$ having entries of order $O(p^{-1})$ while being a low-rank (here unit-rank) matrix: All its "energy" *concentrates* in a single nonzero eigenvalue. As for $\mathbf{Z}^\mathsf{T}\mathbf{Z}/p$, with larger $O(p^{-1/2})$ amplitude entries, it is composed of "essentially independent" zero-mean random variables and tends to be of full rank and *spreads* its energy over its n eigenvalues. Spectrum-wise, both $f(\boldsymbol{\mu},\mathbf{E}) \cdot \mathbf{j}\mathbf{j}^\mathsf{T}/p$ and $\mathbf{Z}^\mathsf{T}\mathbf{Z}/p$ meet on even ground under the nontrivial classification setting of (1.7).

We shall see in Section 4 that things are actually not as clear-cut and, in particular, that not all choices of kernel functions can achieve the same nontrivial classification rates. In particular, the popular Gaussian (radial basis function [RBF]) kernel will be shown to be largely suboptimal in this respect.

Do Real Data Follow Small- or Large-Dimensional Intuitions?

A first glimpse into this riddle, fundamental for the practical design of machine learning algorithms, is provided in Figure 1.3. Similar to Figure 1.2 for synthetic Gaussian data, Figure 1.3 depicts the content of kernel matrices built from the MNIST [LeCun et al., 1998] and Fashion-MNIST data [Xiao et al., 2017], with $p = 28 \times 28 = 784$ and $n = 5\,000$ in both cases. In Figure 1.4, instead of raw data, we display the *features* extracted from popular deep neural networks, such as VGG-16 [Simonyan and Zisserman, 2014] of the more complex CIFAR-10 images (with