1

Introduction

The focus of this book is non-asymptotic theory in high-dimensional statistics. As an area of intellectual inquiry, high-dimensional statistics is not new: it has roots going back to the seminal work of Rao, Wigner, Kolmogorov, Huber and others, from the 1950s onwards. What is new—and very exciting—is the dramatic surge of interest and activity in high-dimensional analysis over the past two decades. The impetus for this research is the nature of data sets arising in modern science and engineering: many of them are extremely large, often with the dimension of the same order as, or possibly even larger than, the sample size. In such regimes, classical asymptotic theory often fails to provide useful predictions, and standard methods may break down in dramatic ways. These phenomena call for the development of new theory as well as new methods. Developments in high-dimensional statistics have connections with many areas of applied mathematics—among them machine learning, optimization, numerical analysis, functional and geometric analysis, information theory, approximation theory and probability theory. The goal of this book is to provide a coherent introduction to this body of work.

1.1 Classical versus high-dimensional theory

What is meant by the term "high-dimensional", and why is it important and interesting to study high-dimensional problems? In order to answer these questions, we first need to understand the distinction between classical as opposed to high-dimensional theory.

Classical theory in probability and statistics provides statements that apply to a fixed class of models, parameterized by an index n that is allowed to increase. In statistical settings, this integer-valued index has an interpretation as a sample size. The canonical instance of such a theoretical statement is the *law of large numbers*. In its simplest instantiation, it concerns the limiting behavior of the sample mean of n independent and identically distributed d-dimensional random vectors $\{X_i\}_{i=1}^n$, say, with mean $\mu = \mathbb{E}[X_1]$ and a finite variance. The law of large numbers guarantees that the sample mean $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to μ. Consequently, the sample mean $\hat{\mu}_n$ is a consistent estimator of the unknown population mean. A more refined statement is provided by the *central limit theorem*, which guarantees that the rescaled deviation $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(X_1)$. These two theoretical statements underlie the analysis of a wide range of classical statistical estimators—in particular, ensuring their consistency and asymptotic normality, respectively.

In a classical theoretical framework, the ambient dimension d of the data space is typically

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

viewed as fixed. In order to appreciate the motivation for high-dimensional statistics, it is worthwhile considering the following:

Question Suppose that we are given $n = 1000$ samples from a statistical model in $d = 500$ dimensions. Will theory that requires $n \to +\infty$ with the dimension d remaining fixed provide useful predictions?

Of course, this question cannot be answered definitively without further details on the model under consideration. Some essential facts that motivate our discussion in this book are the following:

1. The data sets arising in many parts of modern science and engineering have a "high-dimensional flavor", with d on the same order as, or possibly larger than, the sample size n.
2. For many of these applications, classical "large n, fixed d" theory fails to provide useful predictions.
3. Classical methods can break down dramatically in high-dimensional regimes.

These facts motivate the study of high-dimensional statistical models, as well as the associated methodology and theory for estimation, testing and inference in such models.

1.2 What can go wrong in high dimensions?

In order to appreciate the challenges associated with high-dimensional problems, it is worthwhile considering some simple problems in which classical results break down. Accordingly, this section is devoted to three brief forays into some examples of high-dimensional phenomena.

1.2.1 Linear discriminant analysis

In the problem of binary hypothesis testing, the goal is to determine whether an observed vector $x \in \mathbb{R}^d$ has been drawn from one of two possible distributions, say $\mathbb{P}_1$ versus $\mathbb{P}_2$. When these two distributions are known, then a natural decision rule is based on thresholding the log-likelihood ratio $\log \frac{\mathbb{P}_2[x]}{\mathbb{P}_1[x]}$; varying the setting of the threshold allows for a principled trade-off between the two types of errors—namely, deciding $\mathbb{P}_1$ when the true distribution is $\mathbb{P}_2$, and vice versa. The celebrated Neyman–Pearson lemma guarantees that this family of decision rules, possibly with randomization, are optimal in the sense that they trace out the curve giving the best possible trade-off between the two error types.

As a special case, suppose that the two classes are distributed as multivariate Gaussians, say $N(\mu_1, \mathbf{\Sigma})$ and $N(\mu_2, \mathbf{\Sigma})$, respectively, differing only in their mean vectors. In this case, the log-likelihood ratio reduces to the linear statistic

$$\Psi(x) := \left\langle \mu_1 - \mu_2, \, \mathbf{\Sigma}^{-1}\left(x - \frac{\mu_1 + \mu_2}{2}\right)\right\rangle, \tag{1.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^d$. The optimal decision rule is based on thresholding this statistic. We can evaluate the quality of this decision rule by computing the

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

probability of incorrect classification. Concretely, if the two classes are equally likely, this probability is given by

$$\mathrm{Err}(\Psi) := \tfrac{1}{2}\mathbb{P}_1[\Psi(X') \leq 0] + \tfrac{1}{2}\mathbb{P}_2[\Psi(X'') > 0],$$

where X' and X'' are random vectors drawn from the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. Given our Gaussian assumptions, some algebra shows that the error probability can be written in terms of the Gaussian cumulative distribution function Φ as

$$\mathrm{Err}(\Psi) = \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\gamma/2} e^{-t^2/2}\, dt}_{\Phi(-\gamma/2)}, \qquad \text{where } \gamma = \sqrt{(\mu_1 - \mu_2)^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2)}. \tag{1.2}$$

In practice, the class conditional distributions are not known, but instead one observes a collection of labeled samples, say $\{x_1, \ldots, x_{n_1}\}$ drawn independently from $\mathbb{P}_1$, and $\{x_{n_1+1}, \ldots, x_{n_1+n_2}\}$ drawn independently from $\mathbb{P}_2$. A natural approach is to use these samples in order to estimate the class conditional distributions, and then "plug" these estimates into the log-likelihood ratio. In the Gaussian case, estimating the distributions is equivalent to estimating the mean vectors μ_1 and μ_2, as well as the covariance matrix $\mathbf{\Sigma}$, and standard estimates are the samples means

$$\hat{\mu}_1 := \frac{1}{n_1} \sum_{i=1}^{n_1} x_i \quad \text{and} \quad \hat{\mu}_2 := \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} x_i, \tag{1.3a}$$

as well as the pooled sample covariance matrix

$$\widehat{\mathbf{\Sigma}} := \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^{\mathrm{T}} + \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^{\mathrm{T}}. \tag{1.3b}$$

Substituting these estimates into the log-likelihood ratio (1.1) yields the *Fisher linear discriminant function*

$$\widehat{\Psi}(x) = \left\langle \hat{\mu}_1 - \hat{\mu}_2,\, \widehat{\mathbf{\Sigma}}^{-1} \left(x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right) \right\rangle. \tag{1.4}$$

Here we have assumed that the sample covariance is invertible, and hence are assuming implicitly that $n_i > d$.

Let us assume that the two classes are equally likely *a priori*. In this case, the error probability obtained by using a zero threshold is given by

$$\mathrm{Err}(\widehat{\Psi}) := \tfrac{1}{2}\mathbb{P}_1[\widehat{\Psi}(X') \leq 0] + \tfrac{1}{2}\mathbb{P}_2[\widehat{\Psi}(X'') > 0],$$

where X' and X'' are samples drawn independently from the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. Note that the error probability is itself a random variable, since the discriminant function $\widehat{\Psi}$ is a function of the samples $\{X_i\}_{i=1}^{n_1+n_2}$.

In the 1960s, Kolmogorov analyzed a simple version of the Fisher linear discriminant, in which the covariance matrix $\mathbf{\Sigma}$ is known *a priori* to be the identity, so that the linear statistic (1.4) simplifies to

$$\widehat{\Psi}_{\mathrm{id}}(x) = \left\langle \hat{\mu}_1 - \hat{\mu}_2,\, x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right\rangle. \tag{1.5}$$

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

Working under an assumption of Gaussian data, he analyzed the behavior of this method under a form of high-dimensional asymptotics, in which the triple (n_1, n_2, d) all tend to infinity, with the ratios d/n_i, for $i = 1, 2$, converging to some non-negative fraction $\alpha > 0$, and the Euclidean[1] distance $\|\mu_1 - \mu_2\|_2$ converging to a constant $\gamma > 0$. Under this type of high-dimensional scaling, he showed that the error $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ converges in probability to a fixed number—in particular,

$$\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}}) \xrightarrow{\text{prob.}} \Phi\Big(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\Big), \tag{1.6}$$

where $\Phi(t) := \mathbb{P}[Z \leq t]$ is the cumulative distribution function of a standard normal variable. Thus, if $d/n_i \to 0$, then the asymptotic error probability is simply $\Phi(-\gamma/2)$, as is predicted by classical scaling (1.2). However, when the ratios d/n_i converge to a strictly positive number $\alpha > 0$, then the asymptotic error probability is strictly larger than the classical prediction, since the quantity $\frac{\gamma^2}{2\sqrt{\gamma^2+2\alpha}}$ is shifted towards zero.

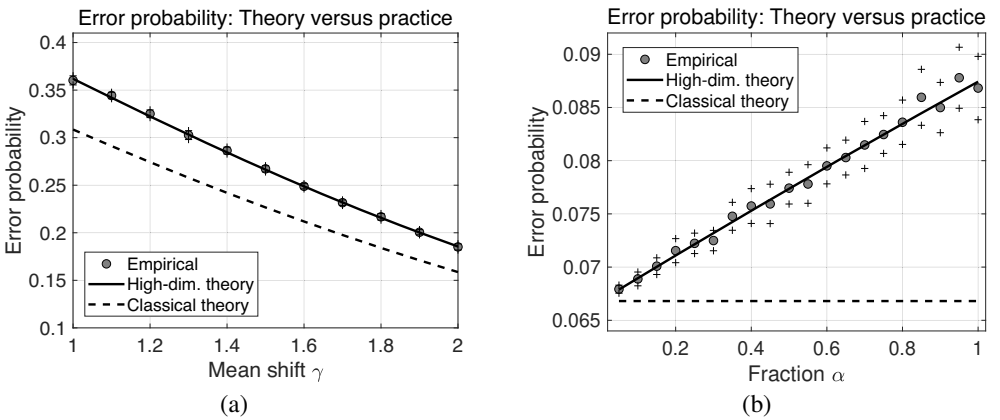


Figure 1.1 (a) Plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the mean shift parameter $\gamma \in [1, 2]$ for $d = 400$ and fraction $\alpha = 0.5$, so that $n_1 = n_2 = 800$. Gray circles correspond to the empirical error probabilities, averaged over 50 trials and confidence bands shown with plus signs, as defined by three times the standard error. The solid curve gives the high-dimensional prediction (1.6), whereas the dashed curve gives the classical prediction (1.2). (b) Plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the fraction $\alpha \in [0, 1]$ for $d = 400$ and $\gamma = 2$. In this case, the classical prediction $\Phi(-\gamma/2)$ plotted as a dashed line remains flat, since it is independent of α.

Recalling our original motivating question from Section 1.1, it is natural to ask whether the error probability of the test $\widehat{\Psi}_{\mathrm{id}}$, for some finite triple (d, n_1, n_2), is better described by the classical prediction (1.2), or the high-dimensional analog (1.6). In Figure 1.1, we plot comparisons between the empirical behavior and theoretical predictions for different choices of the mean shift parameter γ and limiting fraction α. Figure 1.1(a) shows plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the mean shift parameter γ for dimension $d = 400$ and fraction $\alpha = 0.5$, meaning that $n_1 = n_2 = 800$. Gray circles correspond to the empirical

[1] We note that the Mahalanobis distance from equation (1.2) reduces to the Euclidean distance when $\boldsymbol{\Sigma} = \mathbf{I}_d$.

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

performance averaged over 50 trials, whereas the solid and dashed lines correspond to the high-dimensional and classical predictions, respectively. Note that the high-dimensional prediction (1.6) with $\alpha = 0.5$ shows excellent agreement with the behavior in practice, whereas the classical prediction $\Phi(-\gamma)$ drastically underestimates the error rate. Figure 1.1(b) shows a similar plot, again with dimension $d = 400$ but with $\gamma = 2$ and the fraction α ranging in the interval $[0.05, 1]$. In this case, the classical prediction is flat, since it has no dependence on α. Once again, the empirical behavior shows good agreement with the high-dimensional prediction.

A failure to take into account high-dimensional effects can also lead to sub-optimality. A simple instance of this phenomenon arises when the two fractions d/n_i, $i = 1, 2$, converge to possibly different quantities $\alpha_i \geq 0$ for $i = 1, 2$. For reasons to become clear shortly, it is natural to consider the behavior of the discriminant function $\widehat{\Psi}_{\mathrm{id}}$ for a general choice of threshold $t \in \mathbb{R}$, in which case the associated error probability takes the form

$$\mathrm{Err}_t(\widehat{\Psi}_{\mathrm{id}}) = \tfrac{1}{2}\mathbb{P}_1[\widehat{\Psi}_{\mathrm{id}}(X') \leq t] + \tfrac{1}{2}\mathbb{P}_2[\widehat{\Psi}_{\mathrm{id}}(X'') > t], \tag{1.7}$$

where X' and X'' are again independent samples from $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively. For this set-up, it can be shown that

$$\mathrm{Err}_t(\widehat{\Psi}_{\mathrm{id}}) \xrightarrow{\text{prob.}} \frac{1}{2}\Phi\left(-\frac{\gamma^2 + 2t + (\alpha_1 - \alpha_2)}{2\sqrt{\gamma^2 + \alpha_1 + \alpha_2}}\right) + \frac{1}{2}\Phi\left(-\frac{\gamma^2 - 2t - (\alpha_1 - \alpha_2)}{2\sqrt{\gamma^2 + \alpha_1 + \alpha_2}}\right),$$

a formula which reduces to the earlier expression (1.6) in the special case when $\alpha_1 = \alpha_2 = \alpha$ and $t = 0$. Due to the additional term $\alpha_1 - \alpha_2$, whose sign differs between the two terms, the choice $t = 0$ is no longer asymptotically optimal, even though we have assumed that the two classes are equally likely *a priori*. Instead, the optimal choice of the threshold is $t = \frac{\alpha_2 - \alpha_1}{2}$, a choice that takes into account the different sample sizes between the two classes.

1.2.2 Covariance estimation

We now turn to an exploration of high-dimensional effects for the problem of covariance estimation. In concrete terms, suppose that we are given a collection of random vectors $\{x_1, \ldots, x_n\}$, where each x_i is drawn in an independent and identically distributed (i.i.d.) manner from some zero-mean distribution in $\mathbb{R}^d$, and our goal is to estimate the unknown covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(X)$. A natural estimator is the *sample covariance matrix*

$$\widehat{\mathbf{\Sigma}} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}, \tag{1.8}$$

a $d \times d$ random matrix corresponding to the sample average of the outer products $x_i x_i^T \in \mathbb{R}^{d \times d}$. By construction, the sample covariance $\widehat{\mathbf{\Sigma}}$ is an unbiased estimate, meaning that $\mathbb{E}[\widehat{\mathbf{\Sigma}}] = \mathbf{\Sigma}$.

A classical analysis considers the behavior of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ as the sample size n increases while the ambient dimension d stays fixed. There are different ways in which to measure the distance between the random matrix $\widehat{\mathbf{\Sigma}}$ and the population covariance matrix $\mathbf{\Sigma}$, but, regardless of which norm is used, the sample covariance is a consistent

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

estimate. One useful matrix norm is the ℓ_2-operator norm, given by

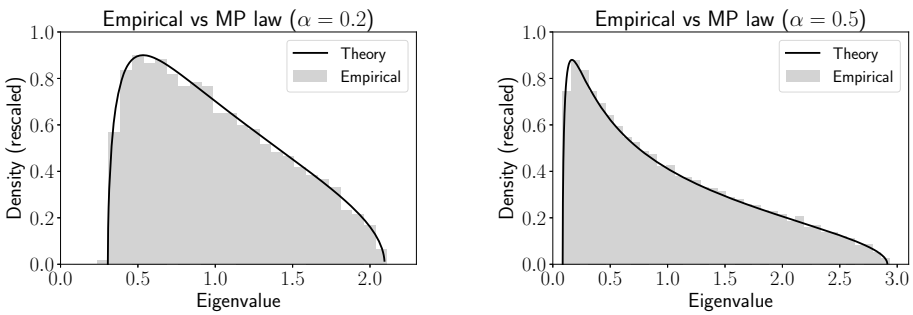
$$|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2 := \sup_{u \neq 0} \frac{\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})u\|_2}{\|u\|_2}. \tag{1.9}$$

Under mild moment conditions, an argument based on the classical law of large numbers can be used to show that the difference $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$ converges to zero almost surely as $n \to \infty$. Consequently, the sample covariance is a strongly consistent estimate of the population covariance in the classical setting.

Is this type of consistency preserved if we also allow the dimension d to tend to infinity? In order to pose the question more crisply, let us consider sequences of problems $(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma})$ indexed by the pair (n, d), and suppose that we allow both n and d to increase with their ratio remaining fixed—in particular, say $d/n = \alpha \in (0, 1)$. In Figure 1.2, we plot the results of simulations for a random ensemble $\mathbf{\Sigma} = \mathbf{I}_d$, with each $X_i \sim N(0, \mathbf{I}_d)$ for $i = 1, \ldots, n$. Using these n samples, we generated the sample covariance matrix (1.8), and then computed its vector of eigenvalues $\gamma(\widehat{\mathbf{\Sigma}}) \in \mathbb{R}^d$, say arranged in non-increasing order as

$$\gamma_{\max}(\widehat{\mathbf{\Sigma}}) = \gamma_1(\widehat{\mathbf{\Sigma}}) \geq \gamma_2(\widehat{\mathbf{\Sigma}}) \geq \cdots \geq \gamma_d(\widehat{\mathbf{\Sigma}}) = \gamma_{\min}(\widehat{\mathbf{\Sigma}}) \geq 0.$$

Each plot shows a histogram of the vector $\gamma(\widehat{\mathbf{\Sigma}}) \in \mathbb{R}^d$ of eigenvalues: Figure 1.2(a) corresponds to the case $(n, d) = (4000, 800)$ or $\alpha = 0.2$, whereas Figure 1.2(b) shows the pair $(n, d) = (4000, 2000)$ or $\alpha = 0.5$. If the sample covariance matrix were converging to the identity matrix, then the vector of eigenvalues $\gamma(\widehat{\mathbf{\Sigma}})$ should converge to the all-ones vector, and the corresponding histograms should concentrate around 1. Instead, the histograms in both plots are highly dispersed around 1, with differing shapes depending on the aspect ratios.



(a) (b)

Figure 1.2 Empirical distribution of the eigenvalues of a sample covariance matrix $\widehat{\mathbf{\Sigma}}$ versus the asymptotic prediction of the Marčenko–Pastur law. It is specified by a density of the form $f_{\mathrm{MP}}(\gamma) \propto \sqrt{\frac{(t_{\max}(\alpha) - \gamma)(\gamma - t_{\min}(\alpha))}{\gamma}}$, supported on the interval $[t_{\min}(\alpha), t_{\max}(\alpha)] = [(1 - \sqrt{\alpha})^2, (1 + \sqrt{\alpha})^2]$. (a) Aspect ratio $\alpha = 0.2$ and $(n, d) = (4000, 800)$. (b) Aspect ratio $\alpha = 0.5$ and $(n, d) = (4000, 2000)$. In both cases, the maximum eigenvalue $\gamma_{\max}(\mathbf{\Sigma})$ is very close to $(1 + \sqrt{\alpha})^2$, consistent with theory.

These shapes—if we let both the sample size and dimension increase in such a way that

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

$d/n \to \alpha \in (0, 1)$—are characterized by an asymptotic distribution known as the Marčenko–Pastur law. Under some mild moment conditions, this theory predicts convergence to a strictly positive density supported on the interval $[t_{\min}(\alpha), t_{\max}(\alpha)]$, where

$$t_{\min}(\alpha) := (1 - \sqrt{\alpha})^2 \quad \text{and} \quad t_{\max}(\alpha) := (1 + \sqrt{\alpha})^2. \tag{1.10}$$

See the caption of Figure 1.2 for more details.

The Marčenko–Pastur law is an asymptotic statement, albeit of a non-classical flavor since it allows both the sample size and dimension to diverge. By contrast, the primary focus of this book are results that are non-asymptotic in nature—that is, in the current context, we seek results that hold for *all* choices of the pair (n, d), and that provide explicit bounds on the events of interest. For example, as we discuss at more length in Chapter 6, in the setting of Figure 1.2, it can be shown that the maximum eigenvalue $\gamma_{\max}(\widehat{\mathbf{\Sigma}})$ satisfies the upper deviation inequality

$$\mathbb{P}[\gamma_{\max}(\widehat{\mathbf{\Sigma}}) \geq (1 + \sqrt{d/n} + \delta)^2] \leq e^{-n\delta^2/2} \qquad \text{for all } \delta \geq 0, \tag{1.11}$$

with an analogous lower deviation inequality for the minimum eigenvalue $\gamma_{\min}(\widehat{\mathbf{\Sigma}})$ in the regime $n \geq d$. This result gives us more refined information about the maximum eigenvalue, showing that the probability that it deviates above $(1 + \sqrt{d/n})^2$ is exponentially small in the sample size n. In addition, this inequality (and related results) can be used to show that the sample covariance matrix $\widehat{\mathbf{\Sigma}}$ is an operator-norm-consistent estimate of the population covariance matrix $\mathbf{\Sigma}$ as long as $d/n \to 0$.

1.2.3 Nonparametric regression

The effects of high dimensions on regression problems can be even more dramatic. In one instance of the problem known as *nonparametric regression*, we are interested in estimating a function from the unit hypercube $[0, 1]^d$ to the real line $\mathbb{R}$; this function can be viewed as mapping a vector $x \in [0, 1]^d$ of predictors or covariates to a scalar response variable $y \in \mathbb{R}$. If we view the pair (X, Y) as random variables, then we can ask for the function f that minimizes the least-squares prediction error $\mathbb{E}[(Y - f(X))^2]$. An easy calculation shows that the optimal such function is defined by the conditional expectation $f(x) = \mathbb{E}[Y \mid x]$, and it is known as the regression function.

In practice, the joint distribution $\mathbb{P}_{X,Y}$ of (X, Y) is unknown, so that computing f directly is not possible. Instead, we are given samples (X_i, Y_i) for $i = 1, \ldots, n$, drawn in an i.i.d. manner from $\mathbb{P}_{X,Y}$, and our goal is to find a function $\widehat{f}$ for which the mean-squared error (MSE)

$$\|\widehat{f} - f\|_{L^2}^2 := \mathbb{E}_X[(\widehat{f}(X) - f(X))^2] \tag{1.12}$$

is as small as possible.

It turns out that this problem becomes extremely difficult in high dimensions, a manifestation of what is known as the *curse of dimensionality*. This notion will be made precise in our discussion of nonparametric regression in Chapter 13. Here, let us do some simple simulations to address the following question: How many samples n should be required as a function of the problem dimension d? For concreteness, let us suppose that the covariate vector X is uniformly distributed over $[0, 1]^d$, so that $\mathbb{P}_X$ is the uniform distribution, denoted by $\text{Uni}([0, 1]^d)$. If we are able to generate a good estimate of $\widehat{f}$ based on the samples

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

$X_1, \ldots, X_n$, then it should be the case that a typical vector $X' \in [0,1]^d$ is relatively close to at least one of our samples. To formalize this notation, we might study the quantity

$$\rho_\infty(n,d) := \mathbb{E}_{X',X}\Big[\min_{i=1,\ldots,n} \|X' - X_i\|_\infty\Big], \tag{1.13}$$

which measures the average distance between an independently drawn sample X', again from the uniform distribution $\mathrm{Uni}([0,1]^d)$, and our original data set $\{X_1, \ldots, X_n\}$.

How many samples n do we need to collect as a function of the dimension d so as to ensure that $\rho_\infty(n,d)$ falls below some threshold δ? For illustrative purposes, we use $\delta = 1/3$ in the simulations to follow. As in the previous sections, let us first consider a scaling in which the ratio d/n converges to some constant $\alpha > 0$, say $\alpha = 0.5$ for concreteness, so that $n = 2d$. Figure 1.3(a) shows the results of estimating the quantity $\rho_\infty(2d,d)$ on the basis of 20 trials. As shown by the gray circles, in practice, the closest point (on average) to a data set based on $n = 2d$ samples tends to increase with dimension, and certainly stays bounded above $1/3$. What happens if we try a more aggressive scaling of the sample size? Figure 1.3(b) shows the results of the same experiments with $n = d^2$ samples; again, the minimum distance tends to increase as the dimension increases, and stays bounded well above $1/3$.

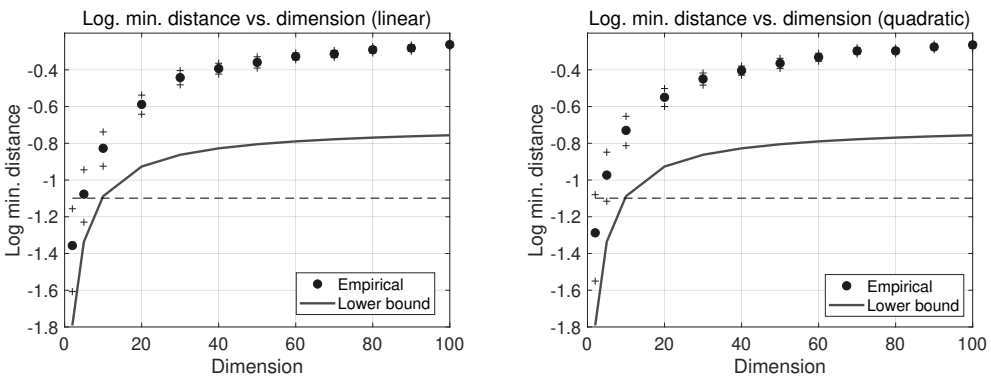


Figure 1.3 Behavior of the quantity $\rho_\infty(n,d)$ versus the dimension d, for different scalings of the pair (n,d). Full circles correspond to the average over 20 trials, with confidence bands shown with plus signs, whereas the solid curve provides the theoretical lower bound (1.14). (a) Behavior of the variable $\rho_\infty(2d,d)$. (b) Behavior of the variable $\rho_\infty(d^2,d)$. In both cases, the expected minimum distance remains bounded above $1/3$, corresponding to $\log(1/3) \approx -1.1$ (horizontal dashed line) on this logarithmic scale.

In fact, we would need to take an *exponentially large* sample size in order to ensure that $\rho_\infty(n,d)$ remained below δ as the dimension increased. This fact can be confirmed by proving the lower bound

$$\log \rho_\infty(n,d) \geq \log \frac{d}{2(d+1)} - \frac{\log n}{d}, \tag{1.14}$$

which implies that a sample size $n > (1/\delta)^d$ is required to ensure that the upper bound $\rho_\infty(n,d) \leq \delta$ holds. We leave the proof of the bound (1.14) as an exercise for the reader.

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

We have chosen to illustrate this exponential explosion in a randomized setting, where the covariates X are drawn uniformly from the hypercube $[0, 1]^d$. But the curse of dimensionality manifests itself with equal ferocity in the deterministic setting, where we are given the freedom of choosing some collection $\{x_i\}_{i=1}^n$ of vectors in the hypercube $[0, 1]^d$. Let us investigate the minimal number n required to ensure that any vector $x' \in [0, 1]^d$ is at most distance δ in the ℓ_∞-norm to some vector in our collection—that is, such that

$$\sup_{x' \in [0,1]^d} \min_{i=1,\ldots,n} \|x' - x_i\|_\infty \leq \delta. \tag{1.15}$$

The most straightforward way of ensuring this approximation quality is by a uniform gridding of the unit hypercube: in particular, suppose that we divide each of the d sides of the cube into $\lceil 1/(2\delta) \rceil$ sub-intervals,[2] each of length 2δ. Taking the Cartesian products of these sub-intervals yields a total of $\lceil 1/(2\delta) \rceil^d$ boxes. Placing one of our points x_i at the center of each of these boxes yields the desired approximation (1.15).

This construction provides an instance of what is known as a δ-covering of the unit hypercube in the ℓ_∞-norm, and we see that its size must grow exponentially in the dimension. By studying a related quantity known as a δ-packing, this exponential scaling can be shown to be inescapable—that is, there is not a covering set with substantially fewer elements. See Chapter 5 for a much more detailed treatment of the notions of packing and covering.

1.3 What can help us in high dimensions?

An important fact is that the high-dimensional phenomena described in the previous sections are *all unavoidable*. Concretely, for the classification problem described in Section 1.2.1, if the ratio d/n stays bounded strictly above zero, then it is not possible to achieve the optimal classification rate (1.2). For the covariance estimation problem described in Section 1.2.2, there is no consistent estimator of the covariance matrix in ℓ_2-operator norm when d/n remains bounded away from zero. Finally, for the nonparametric regression problem in Section 1.2.3, given the goal of estimating a differentiable regression function f, no consistent procedure is possible unless the sample size n grows exponentially in the dimension d. All of these statements can be made rigorous via the notions of metric entropy and minimax lower bounds, to be developed in Chapters 5 and 15, respectively.

Given these "no free lunch" guarantees, what can help us in the high-dimensional setting? Essentially, our only hope is that the data is endowed with some form of *low-dimensional structure,* one which makes it simpler than the high-dimensional view might suggest. Much of high-dimensional statistics involves constructing models of high-dimensional phenomena that involve some implicit form of low-dimensional structure, and then studying the statistical and computational gains afforded by exploiting this structure. In order to illustrate, let us revisit our earlier three vignettes, and show how the behavior can change dramatically when low-dimensional structure is present.

[2] Here $\lceil a \rceil$ denotes the ceiling of a, or the smallest integer greater than or equal to a.

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

1.3.1 Sparsity in vectors

Recall the simple classification problem described in Section 1.2.1, in which, for $j = 1, 2$, we observe n_j samples of a multivariate Gaussian with mean $\mu_j \in \mathbb{R}^d$ and identity covariance matrix $\mathbf{I}_d$. Setting $n = n_1 = n_2$, let us recall the scaling in which the ratios d/n_j are fixed to some number $\alpha \in (0, \infty)$. What is the underlying cause of the inaccuracy of the classical prediction shown in Figure 1.1? Recalling that $\hat{\mu}_j$ denotes the sample mean of the n_j samples, the squared Euclidean error $\|\hat{\mu}_j - \mu_j\|_2^2$ turns out to concentrate sharply around $\frac{d}{n_j} = \alpha$. This fact is a straightforward consequence of the chi-squared (χ^2) tail bounds to be developed in Chapter 2—in particular, see Example 2.11. When $\alpha > 0$, there is a constant level of error, for which reason the classical prediction (1.2) of the error rate is overly optimistic.

But the sample mean is not the only possible estimate of the true mean: when the true mean vector is equipped with some type of low-dimensional structure, there can be much better estimators. Perhaps the simplest form of structure is sparsity: suppose that we knew that each mean vector μ_j were relatively sparse, with only s of its d entries being non-zero, for some sparsity parameter $s \ll d$. In this case, we can obtain a substantially better estimator by applying some form of thresholding to the sample means. As an example, for a given threshold level $\lambda > 0$, the hard-thresholding estimator is given by

$$H_\lambda(x) = x \mathbb{I}[|x| > \lambda] = \begin{cases} x & \text{if } |x| > \lambda, \\ 0 & \text{otherwise,} \end{cases} \tag{1.16}$$

where $\mathbb{I}[|x| > \lambda]$ is a 0–1 indicator for the event $\{|x| > \lambda\}$. As shown by the solid curve in Figure 1.4(a), it is a "keep-or-kill" function that zeroes out x whenever its absolute value falls below the threshold λ, and does nothing otherwise. A closely related function is the soft-thresholding operator

$$T_\lambda(x) = \mathbb{I}[|x| > \lambda](x - \lambda \operatorname{sign}(x)) = \begin{cases} x - \lambda \operatorname{sign}(x) & \text{if } |x| > \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{1.17}$$

As shown by the dashed line in Figure 1.4(a), it has been shifted so as to be continuous, in contrast to the hard-thresholding function.

In the context of our classification problem, instead of using the sample means $\hat{\mu}_j$ in the plug-in classification rule (1.5), suppose that we used hard-thresholded versions of the sample means—namely

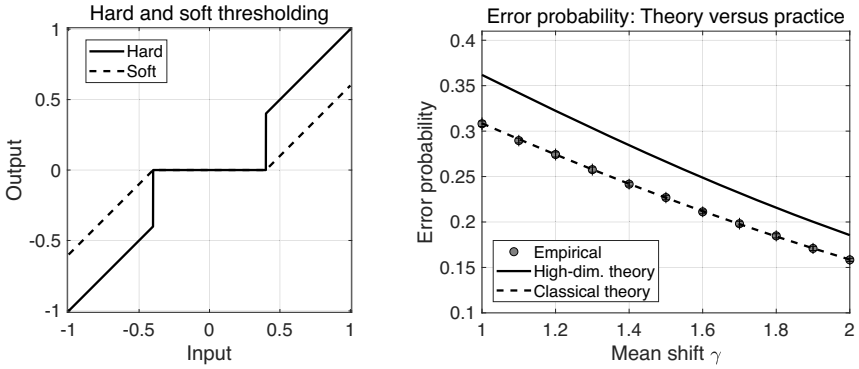
$$\widetilde{\mu}_j = H_{\lambda_n}(\hat{\mu}_j) \quad \text{for } j = 1, 2 \qquad \text{where } \lambda_n := \sqrt{\frac{2 \log d}{n}}. \tag{1.18}$$

Standard tail bounds to be developed in Chapter 2—see Exercise 2.12 in particular—will illuminate why this particular choice of threshold λ_n is a good one. Using these thresholded estimates, we can then implement a classifier based on the linear discriminant

$$\widetilde{\Psi}(x) := \left\langle \widetilde{\mu}_1 - \widetilde{\mu}_2, \, x - \frac{\widetilde{\mu}_1 + \widetilde{\mu}_2}{2} \right\rangle. \tag{1.19}$$

In order to explore the performance of this classifier, we performed simulations using the same parameters as those in Figure 1.1(a); Figure 1.4(b) gives a plot of the error $\operatorname{Err}(\widetilde{\Psi})$

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press



(a) (b)

Figure 1.4 (a) Plots of the hard-thresholding and soft-thresholding functions at some level $\lambda > 0$. (b) Plots of the error probability $\mathrm{Err}(\widehat{\Psi}_{\mathrm{id}})$ versus the mean shift parameter $\gamma \in [1, 2]$ with the same set-up as the simulations in Figure 1.1: dimension $d = 400$, and sample sizes $n = n_1 = n_2 = 800$. In this case, the mean vectors μ_1 and μ_2 each had $s = 5$ non-zero entries, and the classification was based on hard-thresholded versions of the sample means at the level $\lambda_n = \sqrt{\frac{2 \log d}{n}}$. Gray circles correspond to the empirical error probabilities, averaged over 50 trials and confidence intervals defined by three times the standard error. The solid curve gives the high-dimensional prediction (1.6), whereas the dashed curve gives the classical prediction (1.2). In contrast to Figure 1.1(a), the classical prediction is now accurate.

versus the mean shift γ. Overlaid for comparison are both the classical (1.2) and high-dimensional (1.6) predictions. In contrast to Figure 1.1(a), the classical prediction now gives an excellent fit to the observed behavior. In fact, the classical limit prediction is exact whenever the ratio $\log \binom{d}{s}/n$ approaches zero. Our theory on sparse vector estimation in Chapter 7 can be used to provide a rigorous justification of this claim.

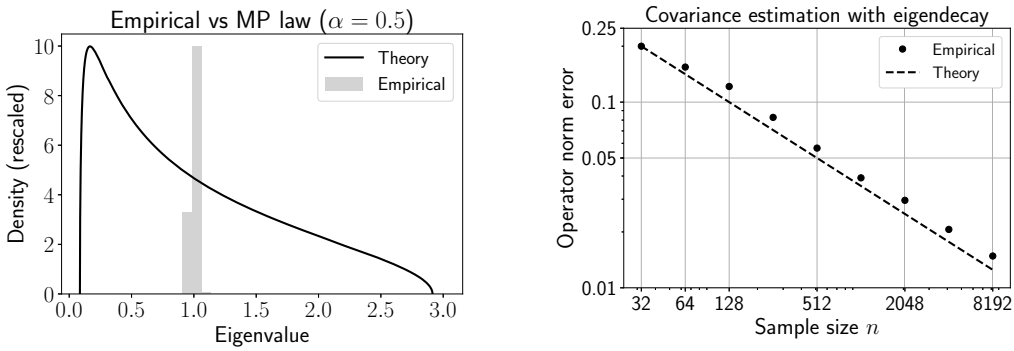
1.3.2 Structure in covariance matrices

In Section 1.2.2, we analyzed the behavior of the eigenvalues of a sample covariance matrix $\widehat{\mathbf{\Sigma}}$ based on n samples of a d-dimensional random vector with the identity matrix as its covariance. As shown in Figure 1.2, when the ratio d/n remains bounded away from zero, the sample eigenspectrum $\gamma(\widehat{\mathbf{\Sigma}})$ remains highly dispersed around 1, showing that $\widehat{\mathbf{\Sigma}}$ is not a good estimate of the population covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$. Again, we can ask the questions: What types of low-dimensional structure might be appropriate for modeling covariance matrices? And how can they can be exploited to construct better estimators?

As a very simple example, suppose that our goal is to estimate a covariance matrix known to be diagonal. It is then intuitively clear that the sample covariance matrix can be improved by zeroing out its non-diagonal entries, leading to the diagonal covariance estimate $\widehat{\mathbf{D}}$. A little more realistically, if the covariance matrix $\mathbf{\Sigma}$ were assumed to be sparse but the positions were unknown, then a reasonable estimator would be the hard-thresholded version $\widetilde{\mathbf{\Sigma}} := T_{\lambda_n}(\widehat{\mathbf{\Sigma}})$ of the sample covariance, say with $\lambda_n = \sqrt{\frac{2 \log d}{n}}$ as before. Figure 1.5(a)

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

shows the resulting eigenspectrum $\gamma(\widetilde{\mathbf{\Sigma}})$ of this estimator with aspect ratio $\alpha = 0.2$ and $(n, d) = (4000, 800)$—that is, the same settings as Figure 1.2(a). In contrast to the Marčenko–Pastur behavior shown in the former figure, we now see that the eigenspectrum $\gamma(\widetilde{\mathbf{\Sigma}})$ is sharply concentrated around the point mass at 1. Tail bounds and theory from Chapters 2 and 6 can be used to show that $|||\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2 \precsim \sqrt{\frac{\log d}{n}}$ with high probability.



(a) (b)

Figure 1.5 (a) Behavior of the eigenspectrum $\gamma(\widetilde{\mathbf{\Sigma}})$ for a hard-thresholded version of the sample covariance matrix. Unlike the sample covariance matrix itself, it can be a consistent estimator of a sparse covariance matrix even for scalings such that $d/n = \alpha > 0$. (b) Behavior of the sample covariance matrix for estimating sequences of covariance matrices of increasing dimension but all satisfying the constraint $\text{trace}(\mathbf{\Sigma}) \leq 20$. Consistent with theoretical predictions, the operator norm error $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$ for this sequence decays at the rate $1/\sqrt{n}$, as shown by the solid line on the log–log plot.

An alternative form of low-dimensional structure for symmetric matrices is that of fast decay in their eigenspectra. If we again consider sequences of problems indexed by (n, d), suppose that our sequence of covariance matrices have a bounded trace—that is, $\text{trace}(\mathbf{\Sigma}) \leq R$, independent of the dimension d. This requirement means that the ordered eigenvalues $\gamma_j(\mathbf{\Sigma})$ must decay a little more quickly than j^{-1}. As we discuss in Chapter 10, these types of eigendecay conditions hold in a variety of applications. Figure 1.5(b) shows a log–log plot of the operator norm error $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$ over a range of pairs (n, d), all with the fixed ratio $d/n = 0.2$, for a sequence of covariance matrices that all satisfy the constraint $\text{trace}(\mathbf{\Sigma}) \leq 20$. Theoretical results to be developed in Chapter 6 predict that, for such a sequence of covariance matrices, the error $|||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|||_2$ should decay as $n^{-1/2}$, even if the dimension d grows in proportion to the sample size n. See also Chapters 8 and 10 for discussion of other forms of matrix estimation in which these types of rank or eigendecay constraints play a role.

1.3.3 Structured forms of regression

As discussed in Section 1.2.3, a generic regression problem in high dimensions suffers from a severe curse of dimensionality. What type of structure can alleviate this curse? There are

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

various forms of low-dimensional structure that have been studied in past and on-going work on high-dimensional regression.

One form of structure is that of an *additive decomposition* in the regression function—say of the form

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} g_j(x_j), \tag{1.20}$$

where each univariate function $g_j : \mathbb{R} \to \mathbb{R}$ is chosen from some base class. For such functions, the problem of regression is reduced to estimating a collection of d separate univariate functions. The general theory developed in Chapters 13 and 14 can be used to show how the additive assumption (1.20) largely circumvents[3] the curse of dimensionality. A very special case of the additive decomposition (1.20) is the classical linear model, in which, for each $j = 1, \ldots, d$, the univariate function takes the form $g_j(x_j) = \theta_j x_j$ for some coefficients $\theta_j \in \mathbb{R}$. More generally, we might assume that each g_j belongs to a reproducing kernel Hilbert space, a class of function spaces studied at length in Chapter 12.

Assumptions of sparsity also play an important role in the regression setting. The *sparse additive model* (SPAM) is based on positing the existence of some subset $S \subset \{1, 2, \ldots, d\}$ of cardinality $s = |S|$ such that the regression function can be decomposed as

$$f(x_1, \ldots, x_d) = \sum_{j \in S} g_j(x_j). \tag{1.21}$$

In this model, there are two different classes of objects to be estimated: (i) the unknown subset S that ranges over all $\binom{d}{s}$ possible subsets of size s; and (ii) the univariate functions $\{g_j, j \in S\}$ associated with this subset. A special case of the SPAM decomposition (1.21) is the *sparse linear model*, in which $f(x) = \sum_{j=1}^{d} \theta_j x_j$ for some vector $\theta \in \mathbb{R}^d$ that is s-sparse. See Chapter 7 for a detailed discussion of this class of models, and the conditions under which accurate estimation is possible even when $d \gg n$.

There are a variety of other types of structured regression models to which the methods and theory developed in this book can be applied. Examples include the *multiple-index model*, in which the regression function takes the form

$$f(x_1, \ldots, x_d) = h(\mathbf{A}x), \tag{1.22}$$

for some matrix $\mathbf{A} \in \mathbb{R}^{s \times d}$, and function $h : \mathbb{R}^s \to \mathbb{R}$. The single-index model is the special case of this model with $s = 1$, so that $f(x) = h(\langle a, x \rangle)$ for some vector $a \in \mathbb{R}^d$. Another special case of this more general family is the SPAM class (1.21): it can be obtained by letting the rows of $\mathbf{A}$ be the standard basis vectors $\{e_j, j \in S\}$, and letting the function h belong to the additive class (1.20).

Taking sums of single-index models leads to a method known as *projection pursuit regression*, involving functions of the form

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{M} g_j(\langle a_j, x \rangle), \tag{1.23}$$

for some collection of univariate functions $\{g_j\}_{j=1}^{M}$, and a collection of d vectors $\{a_j\}_{j=1}^{M}$. Such

[3] In particular, see Exercise 13.9, as well as Examples 14.11 and 14.14.

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

models can also help alleviate the curse of dimensionality, as long as the number of terms M can be kept relatively small while retaining a good fit to the regression function.

1.4 What is the non-asymptotic viewpoint?

As indicated by its title, this book emphasizes non-asymptotic results in high-dimensional statistics. In order to put this emphasis in context, we can distinguish between at least three types of statistical analysis, depending on how the sample size behaves relative to the dimension and other problem parameters:

- *Classical asymptotics.* The sample size n is taken to infinity, with the dimension d and all other problem parameters remaining fixed. The standard laws of large numbers and central limit theorem are examples of this type of theory.
- *High-dimensional asymptotics.* The pair (n, d) is taken to infinity simultaneously, while enforcing that, for some scaling function Ψ, the sequence $\Psi(n, d)$ remains fixed, or converges to some value $\alpha \in [0, \infty]$. For example, in our discussions of linear discriminant analysis (Section 1.2.1) and covariance estimation (Section 1.2.2), we considered such scalings with the function $\Psi(n, d) = d/n$. More generally, the scaling function might depend on other problem parameters in addition to (n, d). For example, in studying vector estimation problems involving a sparsity parameter s, the scaling function $\Psi(n, d, s) = \log \binom{d}{s} / n$ might be used. Here the numerator reflects that there are $\binom{d}{s}$ possible subsets of cardinality s contained in the set of all possible indices $\{1, 2, \ldots, d\}$.
- *Non-asymptotic bounds.* The pair (n, d), as well as other problem parameters, are viewed as fixed, and high-probability statements are made as a function of them. The previously stated bound (1.11) on the maximum eigenvalue of a sample covariance matrix is a standard example of such a result. Results of this type—that is, tail bounds and concentration inequalities on the performance of statistical estimators—are the primary focus of this book.

To be clear, these modes of analysis are closely related. Tail bounds and concentration inequalities typically underlie the proofs of classical asymptotic theorems, such as almost sure convergence of a sequence of random variables. Non-asymptotic theory can be used to predict some aspects of high-dimensional asymptotic phenomena—for instance, it can be used to derive the limiting forms of the error probabilities (1.6) for linear discriminant analysis. In random matrix theory, it can be used to establish that the sample eigenspectrum of a sample covariance matrix with $d/n = \alpha$ lies within[4] the interval $[(1 - \sqrt{\alpha})^2, (1 + \sqrt{\alpha})^2]$ with probability one as (n, d) grow—cf. Figure 1.2. Finally, the functions that arise in a non-asymptotic analysis can suggest appropriate forms of scaling functions Ψ suitable for performing a high-dimensional asymptotic analysis so as to unveil limiting distributional behavior.

One topic *not* covered in this book—due to space constraints—is an evolving line of work that seeks to characterize the asymptotic behavior of low-dimensional functions of a given high-dimensional estimator; see the bibliography in Section 1.6 for some references.

[4] To be clear, it does not predict the precise shape of the distribution on this interval, as given by the Marčenko–Pastur law.

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

For instance, in sparse vector estimation, one natural goal is to seek a confidence interval for a given coordinate of the d-dimensional vector. At the heart of such analyses are non-asymptotic tail bounds, which allow for control of residuals within the asymptotics. Consequently, the reader who has mastered the techniques laid out in this book will be well equipped to follow these types of derivations.

1.5 Overview of the book

With this motivation in hand, let us now turn to a broad overview of the structure of this book, as well as some suggestions regarding its potential use in a teaching context.

1.5.1 Chapter structure and synopses

The chapters follow a rough division into two types: material on *Tools and techniques* (TT), and material on *Models and estimators* (ME). Chapters of the TT type are foundational in nature, meant to develop techniques and derive theory that is broadly applicable in high-dimensional statistics. The ME chapters are meant to be complementary in nature: each such chapter focuses on a particular class of statistical estimation problems, and brings to bear the methods developed in the foundational chapters.

Tools and techniques

- Chapter 2: This chapter provides an introduction to standard techniques in deriving tail bounds and concentration inequalities. It is required reading for all other chapters in the book.
- Chapter 3: Following directly from Chapter 2, this chapter is devoted to more advanced material on concentration of measure, including the entropic method, log-Sobolev inequalities, and transportation cost inequalities. It is meant for the reader interested in a deeper understanding of the concentration phenomenon, but is not required reading for the remaining chapters. The concentration inequalities in Section 3.4 for empirical processes are used in later analysis of nonparametric models.
- Chapter 4: This chapter is again required reading for most other chapters, as it introduces the foundational ideas of uniform laws of large numbers, along with techniques such as symmetrization, which leads naturally to the Rademacher complexity of a set. It also covers the notion of Vapnik–Chervonenkis (VC) dimension as a particular way of bounding the Rademacher complexity.
- Chapter 5: This chapter introduces the geometric notions of covering and packing in metric spaces, along with the associated discretization and chaining arguments that underlie proofs of uniform laws via entropic arguments. These arguments, including Dudley's entropy integral, are required for later study of nonparametric models in Chapters 13 and 14. Also covered in this chapter are various connections to Gaussian processes, including the Sudakov–Fernique and Gordon–Slepian bounds, as well as Sudakov's lower bound.
- Chapter 12: This chapter provides a self-contained introduction to reproducing kernel Hilbert spaces, including material on kernel functions, Mercer's theorem and eigenvalues, the representer theorem, and applications to function interpolation and estimation via kernel ridge regression. This material is not a prerequisite for reading Chapters 13 and 14,

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

but is required for understanding the kernel-based examples covered in these chapters on nonparametric problems.

- Chapter 14: This chapter follows the material from Chapters 4 and 13, and is devoted to more advanced material on uniform laws, including an in-depth analysis of two-sided and one-sided uniform laws for the population and empirical L^2-norms. It also includes some extensions to certain Lipschitz cost functions, along with applications to nonparametric density estimation.
- Chapter 15: This chapter provides a self-contained introduction to techniques for proving minimax lower bounds, including in-depth discussions of Le Cam's method in both its naive and general forms, the local and Yang–Barron versions of the Fano method, along with various examples. It can be read independently of any other chapter, but does make reference (for comparison) to upper bounds proved in other chapters.

Models and estimators

- Chapter 6: This chapter is devoted to the problem of covariance estimation. It develops various non-asymptotic bounds for the singular values and operator norms of random matrices, using methods based on comparison inequalities for Gaussian matrices, discretization methods for sub-Gaussian and sub-exponential variables, as well as tail bounds of the Ahlswede–Winter type. It also covers the estimation of sparse and structured covariance matrices via thresholding and related techniques. Material from Chapters 2, 4 and 5 is needed for a full understanding of the proofs in this chapter.
- Chapter 7: The sparse linear model is possibly the most widely studied instance of a high-dimensional statistical model, and arises in various applications. This chapter is devoted to theoretical results on the behavior of ℓ_1-relaxations for estimating sparse vectors, including results on exact recovery for noiseless models, estimation in ℓ_2-norm and prediction semi-norms for noisy models, as well as results on variable selection. It makes substantial use of various tail bounds from Chapter 2.
- Chapter 8: Principal component analysis is a standard method in multivariate data analysis, and exhibits a number of interesting phenomena in the high-dimensional setting. This chapter is devoted to a non-asymptotic study of its properties, in both its unstructured and sparse versions. The underlying analysis makes use of techniques from Chapters 2 and 6.
- Chapter 9: This chapter develops general techniques for analyzing estimators that are based on decomposable regularizers, including the ℓ_1-norm and nuclear norm as special cases. It builds on the material on sparse linear regression from Chapter 7, and makes uses of techniques from Chapters 2 and 4.
- Chapter 10: There are various applications that involve the estimation of low-rank matrices in high dimensions, and this chapter is devoted to estimators based on replacing the rank constraint with a nuclear norm penalty. It makes direct use of the framework from Chapter 9, as well as tail bounds and random matrix theory from Chapters 2 and 6.
- Chapter 11: Graphical models combine ideas from probability theory and graph theory, and are widely used in modeling high-dimensional data. This chapter addresses various types of estimation and model selection problems that arise in graphical models. It requires background from Chapters 2 and 7.
- Chapter 13: This chapter is devoted to an in-depth analysis of least-squares estimation

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

in the general nonparametric setting, with a broad range of examples. It exploits techniques from Chapters 2, 4 and 5, along with some concentration inequalities for empirical processes from Chapter 3.

1.5.2 Recommended background

This book is targeted at graduate students with an interest in applied mathematics broadly defined, including mathematically oriented branches of statistics, computer science, electrical engineering and econometrics. As such, it assumes a strong undergraduate background in basic aspects of mathematics, including the following:

- A course in linear algebra, including material on matrices, eigenvalues and eigendecompositions, singular values, and so on.
- A course in basic real analysis, at the level of Rudin's elementary book (Rudin, 1964), covering convergence of sequences and series, metric spaces and abstract integration.
- A course in probability theory, including both discrete and continuous variables, laws of large numbers, as well as central limit theory. A measure-theoretic version is not required, but the ability to deal with the abstraction of this type is useful. Some useful books include Breiman (1992), Chung (1991), Durrett (2010) and Williams (1991).
- A course in classical mathematical statistics, including some background on decision theory, basics of estimation and testing, maximum likelihood estimation and some asymptotic theory. Some standard books at the appropriate level include Keener (2010), Bickel and Doksum (2015) and Shao (2007).

Probably the most subtle requirement is a certain degree of mathematical maturity on the part of the reader. This book is meant for the person who is interested in gaining a deep understanding of the core issues in high-dimensional statistics. As with anything worthwhile in life, doing so requires effort. This basic fact should be kept in mind while working through the proofs, examples and exercises in this book.

At the same time, this book has been written with self-study and/or teaching in mind. To wit, we have often sacrificed generality or sharpness in theorem statements for the sake of proof clarity. In lieu of an exhaustive treatment, our primary emphasis is on developing techniques that can be used to analyze many different problems. To this end, each chapter is seeded with a large number of examples, in which we derive specific consequences of more abstract statements. Working through these examples in detail, as well as through some of the many exercises at the end of each chapter, is the best way to gain a robust grasp of the material. As a warning to the reader: the exercises range in difficulty from relatively straightforward to extremely challenging. *Don't be discouraged* if you find an exercise to be challenging; some of them are meant to be!

1.5.3 Teaching possibilities and a flow diagram

This book has been used for teaching one-semester graduate courses on high-dimensional statistics at various universities, including the University of California Berkeley, Carnegie

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

Mellon University, Massachusetts Institute of Technology and Yale University. The book has far too much material for a one-semester class, but there are various ways of working through different subsets of chapters over time periods ranging from five to 15 weeks. See Figure 1.6 for a flow diagram that illustrates some of these different pathways through the book.

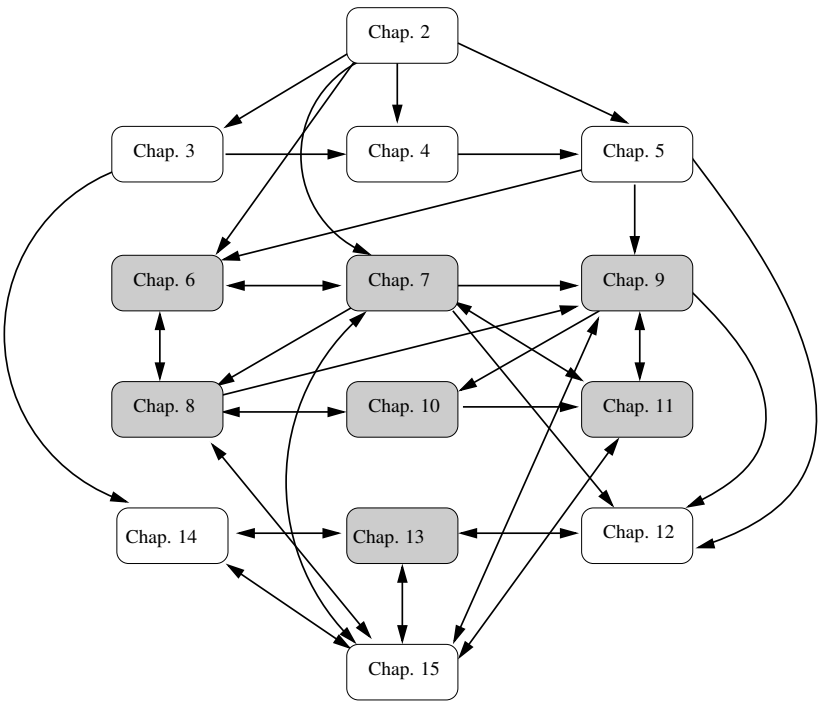


Figure 1.6 A flow diagram of Chapters 2–15 and some of their dependence structure. Various tours of subsets of chapters are possible; see the text for more details.

A short introduction. Given a shorter period of a few weeks, it would be reasonable to cover Chapter 2 followed by Chapter 7 on sparse linear regression, followed by parts of Chapter 6 on covariance estimation. Other brief tours beginning with Chapter 2 are also possible.

A longer look. Given a few more weeks, a longer look could be obtained by supplementing the short introduction with some material from Chapter 5 on metric entropy and Dudley's entropy integral, followed by Chapter 13 on nonparametric least squares. This supplement would give a taste of the nonparametric material in the book. Alternative additions are possible, depending on interests.

A full semester course. A semester-length tour through the book could include Chapter 2 on tail bounds, Chapter 4 on uniform laws, the material in Sections 5.1 through 5.3.3 on metric entropy through to Dudley's entropy integral, followed by parts of Chapter 6 on covariance

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

estimation, Chapter 7 on sparse linear regression, and Chapter 8 on principal component analysis. A second component of the course could consist of Chapter 12 on reproducing kernel Hilbert spaces, followed by Chapter 13 on nonparametric least squares. Depending on the semester length, it could also be possible to cover some material on minimax lower bounds from Chapter 15.

1.6 Bibliographic details and background

Rao (1949) was one of the first authors to consider high-dimensional effects in two-sample testing problems. The high-dimensional linear discriminant problem discussed in Section 1.2.1 was first proposed and analyzed by Kolmogorov in the 1960s. Deev, working in the group of Kolmogorov, analyzed the high-dimensional asymptotics of the general Fisher linear discriminant for fractions $\alpha_i \in [0, 1)$. See the book by Serdobolskii (2000) and the survey paper by Raudys and Young (2004) for further detail on this early line of Russian research in high-dimensional classification.

The study of high-dimensional random matrices, as treated briefly in Section 1.2.2, also has deep roots, dating back to the seminal work from the 1950s onwards (e.g., Wigner, 1955, 1958; Marčenko and Pastur, 1967; Pastur, 1972; Wachter, 1978; Geman, 1980). The high-dimensional asymptotic law for the eigenvalues of a sample covariance matrix illustrated in Figure 1.2 is due to Marčenko and Pastur (1967); this asymptotic prediction has been shown to be a remarkably robust phenomenon, requiring only mild moment conditions (e.g., Silverstein, 1995; Bai and Silverstein, 2010). See also the paper by Götze and Tikhomirov (2004) for quantitative bounds on the distance to this limiting distribution.

In his Wald Memorial Lecture, Huber (1973) studied the asymptotics of robust regression under a high-dimensional scaling with d/n constant. Portnoy (1984; 1985) studied M-estimators for high-dimensional linear regression models, proving consistency when the ratio $\frac{d \log d}{n}$ goes to zero, and asymptotic normality under somewhat more stringent conditions. See also Portnoy (1988) for extensions to more general exponential family models. The high-dimensional asymptotics of various forms of robust regression estimators have been studied in recent work by El Karoui and co-authors (e.g., Bean et al., 2013; El Karoui, 2013; El Karoui et al., 2013), as well as by Donoho and Montanari (2013).

Thresholding estimators are widely used in statistical problems in which the estimand is expected to be sparse. See the book by Johnstone (2015) for an extensive discussion of thresholding estimators in the context of the normal sequence model, with various applications in nonparametric estimation and density estimation. See also Chapters 6 and 7 for some discussion and analysis of thresholding estimators. Soft thresholding is very closely related to ℓ_1-regularization, a method with a lengthy history (e.g., Levy and Fullagar, 1981; Santosa and Symes, 1986; Tibshirani, 1996; Chen et al., 1998; Juditsky and Nemirovski, 2000; Donoho and Huo, 2001; Elad and Bruckstein, 2002; Candès and Tao, 2005; Donoho, 2006b; Bickel et al., 2009); see Chapter 7 for an in-depth discussion.

Stone (1985) introduced the class of additive models (1.20) for nonparametric regression; see the book by Hastie and Tibshirani (1990) for more details. The SPAM class (1.21) has been studied by many researchers (e.g., Meier et al., 2009; Ravikumar et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012). The single-index model (1.22), as a particular instance of a semiparametric model, has also been widely studied; for instance, see the var-

https://doi.org/10.1017/9781108627771.001 Published online by Cambridge University Press

ious papers (Härdle and Stoker, 1989; Härdle et al., 1993; Ichimura, 1993; Hristache et al., 2001) and references therein for further details. Friedman and Stuetzle (1981) introduced the idea of projection pursuit regression (1.23). In broad terms, projection pursuit methods are based on seeking "interesting" projections of high-dimensional data (Kruskal, 1969; Huber, 1985; Friedman and Tukey, 1994), and projection pursuit regression is based on this idea in the context of regression.