Cambridge University Press
978-1-108-49802-9 — High-Dimensional Statistics
Martin J. Wainwright
Excerpt
More Information

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

Cambridge University Press
978-1-108-49802-9 — High-Dimensional Statistics
Martin J. Wainwright
Excerpt
More Information

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 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 $\{X_i\}_{i=1}^{n_1+n_2}$.

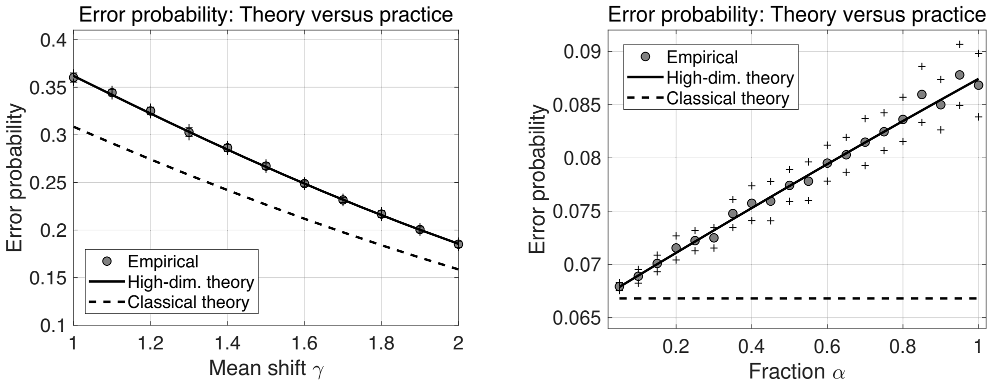
In the 1960s, Kolmogorov analyzed a simple version of the Fisher linear discriminant, in which the covariance matrix $\mathbf{\Sigma}$ is known *a priori* to be the identity, so that the linear statistic (1.4) simplifies to

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

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 $\Phi(t) := \mathbb{P}[Z \le t]$ 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.



(a)

(b)

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 $\mathbf{\Sigma} = \mathbf{I}_d$.

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 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.

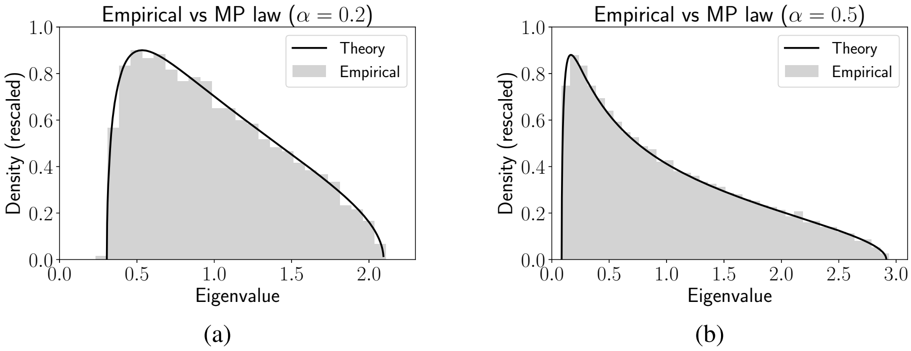


Figure 1.2 Empirical distribution of the eigenvalues of a sample covariance matrix $\widehat{\mathbf{\Sigma}}$ versus the asymptotic prediction of the Marčenko–Pastur law. It is specified by a density of the form $f_{\mathrm{MP}}(\gamma) \propto \sqrt{\frac{(t_{\max}(\alpha)-\gamma)\,(\gamma - t_{\min}(\alpha))}{\gamma}}$, 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

$d/n \to \alpha \in (0,1)$—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

$X_1, \ldots, X_n$, 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.

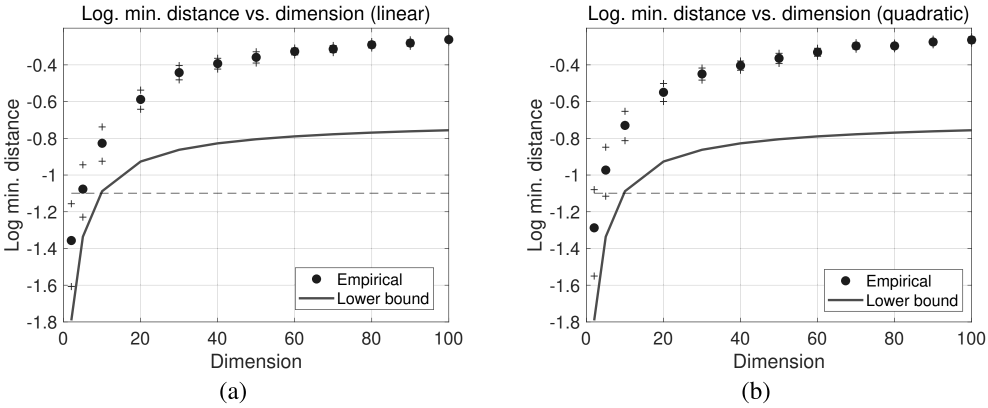


Figure 1.3 Behavior of the quantity $\rho_\infty(n,d)$ versus the dimension d, for different scalings of the pair (n,d). Full circles correspond to the average over 20 trials, with confidence bands shown with plus signs, whereas the solid curve provides the theoretical lower bound (1.14). (a) Behavior 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.

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

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

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.

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 nonzero, 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

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

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 $\text{Err}(\widetilde{\Psi})$