Berkeley
UNIVERSITY OF CALIFORNIA

High-Dimensional Statistics

Notes By: Yuhang Cai

Summer 2023

CONTENTS

CHAPTER 1

INTRODUCTION

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 highdimensional analysis over the past two decades.

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.

1.1 Classical versus 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.

We have two major theorems:

- Law of large numbers,
- Central limit theorem.

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

Example 1.1.
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 are the following:

- The data sets arising in many parts of modern science and engineering have a 'highdimensional flavor', with d on the same order as, or possibly larger than, the sample size n.
- For many of these applications, classical "large n, fixed d" theory fails to provide useful predictions.
- 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 HD?

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

We consider the binary hypothesis testing problem. Given a vector $x \in \mathbb{R}^d$, the goal is to determine whether it's drawn from $\mathbb{P}_1$ or $\mathbb{P}_2$.

As for the classical method, if we assume two distributions are known:

- A natural decision rule is $\log \frac{P_2[x]}{P_1[x]}$,
- Varying the threshold leads to the trade-off between two types of errors,
- Neyman-Pearson lemma guarantees this family of decision rules works.

A special case is to consider two multivariate Gaussian, $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$. Then the log-likelihood ratio reduces to the linear statistics:

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

If $\mathbb{P}(x \text{ from } \mathbb{P}_1) = \mathbb{P}(x \text{ from } \mathbb{P}_2) = \frac{1}{2}$, then

$$\operatorname{Err}(\Psi) := \frac{1}{2}\mathbb{P}_1[\psi(X') \leqslant 0] + \frac{1}{2}\mathbb{P}_2[\Psi(X'') > 0],$$

where $X' \sim \mathbb{P}_1$ and $X'' \sim \mathbb{P}_2$. We can show that

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

In practice, we don't know the distributions, but observes a collection if 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 $\mathbb{P}_2$. Then, we will estimate the means and covariance:

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

and

$$\hat{\Sigma} := \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^{\mathrm{T}} + \frac{1}{n_2 - 1} \sum_{i=n_1+1}^{n_1+n_2} (x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)^{\mathrm{T}}. \tag{1.4}$$

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

$$\hat{\Psi}(x) = \left\langle \hat{\mu}_1 - \hat{\mu}_2, \hat{\Sigma}^{-1} \left(x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right) \right\rangle. \tag{1.5}$$

Note that the sample covariance is invertible, and hence we need $n_i > d$. 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

$$\operatorname{Err}(\widehat{\Psi}) := \frac{1}{2}\mathbb{P}_1\left[\widehat{\Psi}(X') \leqslant 0\right] + \frac{1}{2}\mathbb{P}_2\left[\widehat{\Psi}(X'') > 0\right],$$

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 Σ is known a priori to be the identity, so that the linear statistic (1.1) simplifies to

$$\widehat{\Psi}_{\mathrm{id}}(x) = \left\langle \hat{\mu}_1 - \hat{\mu}_2, x - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} \right\rangle. \tag{1.6}$$

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,

$$\mathrm{Err}\left(\widehat{\Psi}_{\mathrm{id}}\right) \xrightarrow{\text{prob.}} \Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}}\right). \tag{1.7}$$

If $\frac{d}{n_i} \to 0$, the error is simply $\Phi(-\gamma/2)$, as is predicted by classical scaling (1.2). If $\alpha > 0$, the error will be larger.