Chapter 1

High-dimensional, or nonparametric statistics

1.1 Asymptotics in smooth, parametric models

In most introductory courses on mathematical statistics the focus is on classical statistical models in which the number of unknown parameters is small relative to the sample size. If such models depend smoothly on the parameter of interest, then under regularity conditions we typically have that as the sample size n tends to infinity, maximum likelihood or Bayesian estimators converge at the rate $\sqrt{n}$ to the parameter corresponding to the true distribution that generated the data. Moreover, asymptotically such estimators are normally distributed and efficient, for instance in the sense that they have minimal asymptotic variance.

Consider for example an i.i.d. sample $X_1, \ldots, X_n$ from a positive density p_θ depending smoothly on a real-valued parameter θ. Denote the true parameter by θ_0, i.e. the true marginal density of the sample is p_{θ_0}. Then the maximum likelihood estimator (MLE) $\hat{\theta}_n$ is the maximizer of the random function $\theta \mapsto n^{-1}\sum(\ell_\theta - \ell_{\theta_0})(X_i)$, where $\ell_\theta(x) = \log p_\theta(x)$. By the law of large numbers this function converges pointwise to the corresponding expectation under P_{θ_0}, which is minus the Kullback-Leibler (KL) divergence $-\mathrm{KL}(p_{\theta_0}, p_\theta)$. Here the KL-divergence between two probability measures with positive densities f and g with respect to a dominating measure μ is defined as

$$\mathrm{KL}(f, g) = \int f \log \frac{f}{g}\, d\mu.$$

If the model is identifiable the function $\theta \mapsto -\mathrm{KL}(p_{\theta_0}, p_\theta)$ has a unique maximum at θ_0 (see Exercise 1.1). Under regularity conditions it can be shown that the maximizer $\hat{\theta}_n$ of $\theta \mapsto n^{-1}\sum(\ell_\theta - \ell_{\theta_0})(X_i)$ converges in probability under P_{θ_0} to the maximizer θ_0 of the limiting function $\theta \mapsto -\mathrm{KL}(p_{\theta_0}, p_\theta)$, i.e. that the maximum likelihood estimator is consistent.

To derive further asympotic properties of the MLE we note that, again under some regularity conditions, the MLE $\hat{\theta}_n$ is also the (or a) zero of the score function s_n given by $s_n(\theta) = \sum \dot{\ell}_\theta(X_i)$, where $\dot{\ell}_\theta(x) = (\partial/\partial\theta)\ell_\theta(x)$. A Taylor expansion of the score function around the true parameter θ_0 then shows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Big(-\frac{1}{n}\dot{s}_n(\bar{\theta}_n)\Big)^{-1} \frac{1}{\sqrt{n}} s_n(\theta_0),$$

where $\bar{\theta}_n$ is a (random) point between θ_0 and $\hat{\theta}_n$. If we have consistency then $\bar{\theta}_n$ will converge to θ_0 in P_{θ_0}-probability, so we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \Big(-\frac{1}{n}\dot{s}_n(\theta_0)\Big)^{-1} \frac{1}{\sqrt{n}} s_n(\theta_0).$$

By the law of large numbers, the central limit theorem and Slutsky's lemma, we have under regularity conditions that the right-hand side converges in distribution to $i_{\theta_0}^{-1} N(0, i_{\theta_0}) = N(0, i_{\theta_0}^{-1})$ as $n \to \infty$, where $i_\theta = \mathrm{Var}_\theta \dot{\ell}_\theta(X_1) = -\mathrm{E}_{\theta_0}\ddot{\ell}_{\theta_0}(X_1)$ is the Fisher information in a single observation. In other words, for large n we have the approximation

$$\hat{\theta}_n \overset{d}{\approx} N(\theta_0, 1/(n i_{\theta_0})).$$

In particular, in these regular models the asymptotic variance of the MLE equals the Cramér-Rao lower bound $1/(n i_{\theta_0})$ for the variance of an unbiased estimator and the asymptotic bias is typically $o(1/\sqrt{n})$.

This sloppy asymptotic analysis of the MLE can be made precise and generalizes to the situation that the parameter θ belongs to $\mathbb{R}^k$ for some fixed $k > 1$. As a consequence of the Bernstein-von Mises theorem, Bayesian estimators typically have the same asymptotic behaviour as the MLE in regular models, under mild conditions. (Note that this means in particular that the influence of the prior disappears asymptotically.) Other estimators, such as moment estimators for instance, will usually converge at rate $\sqrt{n}$ and be asymptotically normal as well, but will typically have an asymptotic variance that is larger than the Cramér-Rao lower bound.

The rate $1/\sqrt{n}$ arises in the analysis of parametric testing problems as well. Suppose again that we have n i.i.d. observations from a density p_θ parametrized by a parameter $\theta \in \Theta \subset \mathbb{R}$. Then for two different, fixed points $\theta_0, \theta_1 \in \Theta$, the likelihood ratio test for the hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is typically consistent. Indeed, consider the test $\varphi_n = 1_{T_n \geq c}$ which rejects the null hypothesis if $T_n \geq c$, where

$$T_n = \sum \log \frac{p_{\theta_1}}{p_{\theta_0}}(X_i)$$

is the log-likelihood ratio statistic and $c \in \mathbb{R}$ is some threshold. By the law of large numbers we have the almost sure convergence

$$\frac{T_n}{n} \to \begin{cases} -\mathrm{KL}(p_{\theta_0}, p_{\theta_1}) & \text{under } H_0, \\ \mathrm{KL}(p_{\theta_1}, p_{\theta_0}) & \text{under } H_1. \end{cases}$$

It follows that for every value of $c \in \mathbb{R}$ the total error of the test satisfies $\mathrm{E}_{\theta_0}\varphi_n + \mathrm{E}_{\theta_1}(1-\varphi_n) \to 0$ as $n \to \infty$, provided that the model is identifiable (check!).

Now a natural question to ask is how fast we can let θ_1 tend to θ_0 as n grows, while still retaining consistency of the test, i.e. how far the hypotheses should be separated in order to be statistically distinguishable. Using the same notation as above, a Taylor expansion shows that

$$T_n = \sum(\ell_{\theta_1} - \ell_{\theta_0})(X_i) \approx (\theta_1 - \theta_0)\sum \dot{\ell}_{\theta_0}(X_i).$$

Hence if for some sequence $M_n \to \infty$ we have $\theta_1 - \theta_0 = M_n/\sqrt{n}$, then by the central limit theorem we have that T_n is approximately $N(0, M_n^2 i_{\theta_0})$-distributed under H_0. To understand the behaviour under H_1 we write

$$(\theta_1 - \theta_0)\sum \dot{\ell}_{\theta_0}(X_i) = (\theta_1 - \theta_0)\sum \dot{\ell}_{\theta_1}(X_i) + (\theta_1 - \theta_0)\sum(\dot{\ell}_{\theta_0} - \dot{\ell}_{\theta_1})(X_i).$$

Using another Taylor approximation we then get

$$T_n \approx (\theta_1 - \theta_0)\sum \dot{\ell}_{\theta_1}(X_i) - (\theta_1 - \theta_0)^2 \sum \ddot{\ell}_{\theta_1}(X_i).$$

So if $\theta_1 - \theta_0 = M_n/\sqrt{n}$ again, then by the central limit theorem and the law of large numbers, T_n is approximately $N(M_n^2 i_{\theta_1}, M_n^2 i_{\theta_1})$-distributed under H_1. By combining these facts we see that if in the definition of the test we take an appropriate level c, we still get a consistent test if the hypotheses are at distance of order larger than $1/\sqrt{n}$. Indeed, in view of the approximations for the distribution of T_n under the respective hypotheses, we have, for Z a standard normal variable,

$$\mathrm{E}_{\theta_0}\varphi_n = \mathrm{P}_{\theta_0}(T_n \geq c) \approx \mathrm{P}\Big(Z \geq \frac{c}{M_n\sqrt{i_{\theta_0}}}\Big) \to 0$$

if $c/M_n \to \infty$ and

$$\mathrm{E}_{\theta_1}(1-\varphi_n) = \mathrm{P}_{\theta_1}(T_n \leq c) \approx \mathrm{P}\Big(Z \leq \frac{c - M_n^2 i_{\theta_1}}{M_n\sqrt{i_{\theta_1}}}\Big) \to 0$$

if $c/M_n - M_n i_{\theta_1} \to -\infty$. Hence is we set for instance $c = M_n \log M_n$, we get a consistent test. If $\theta_1 - \theta_0 = o(1/\sqrt{n})$ however, the test statistic T_n converges to 0 at the same speed under both hypotheses and we loose consistency.

Precise results in the asymptotic theory of parametric estimators and tests, of various degrees of mathematical sophistication, can be found for instance in Ibragimov and Has'minskii (1981), Lehmann and Casella (1998), Van der Vaart (1998), and Lehmann and Romano (2005).

1.2 High-dimensional and nonparametric models

There are many situations in which it is natural, or desirable, to consider statistical models with an unknown parameter that is very high-dimensional compared to sample size, or even infinite-dimensional.

Example 1.2.1 (Nonparametric regression)**.** Suppose we observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and wish to model the relation between the x and y-variables, for instance with the goal to predict the y-values corresponding to new x-variables that we will get to observe later.

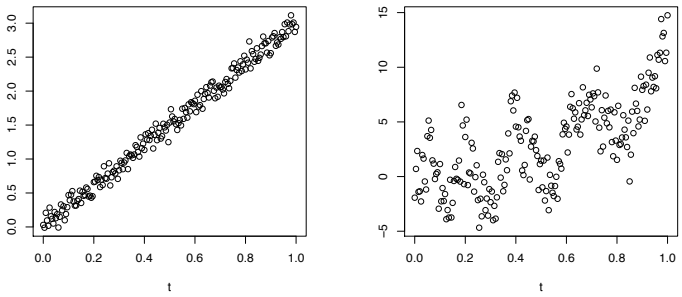

Figure 1.1: Parametric versus nonparametric regression.

If we plot the data points in a scatter plot and we see a picture like the left-hand panel of Figure 1.1, then it makes perfect sense to postulate a linear model with just a few parameters, maybe two to describe the linear relation and one more to describe the variance of the fluctuations around the line. The standard linear model would postulate in this case that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where β_0, β_1 are real parameters and the ε_i are independent, $N(0, \sigma^2)$-distributed variables. We can then use maximum likelihood, for instance, to estimate the parameter $\theta = (\beta_0, \beta_1, \sigma^2)$.

If the scatter plot looks like the right-hand panel however, it is less clear what a reasonable choice for a parametric model would be. Alternatively, we could choose to only assume some qualitative properties of the relation between the variables, for instance that the x and y-variables are, up to noise, related through a continuous, or differentiable function. We could postulate for instance that

$$Y_i = f(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where f is a differentiable function and the ε_i are independent $N(0, \sigma^2)$ variables. The full parameter of the model is then the pair $\theta = (f, \sigma)$, which lives in $\mathcal{F} \times (0, \infty)$, for some function space $\mathcal{F}$.

Since this model is not described in terms of a small number of real parameters, it is called *nonparametric*. A model like the one just postulated is called a *nonparametric regression model*. See for instance Wasserman (2006) for examples of statistical methods for nonparametric regression. ■

Example 1.2.2 (High-dimensional linear regression)**.** The basic Gaussian linear model considers observations $Y \in \mathbb{R}^n$ that satisfy the linear regression relation $Y = X\theta + \varepsilon$, where $\theta \in \mathbb{R}^p$ is an unknown parameter vector, X is a fixed, known $n \times p$ design matrix, and $\varepsilon \sim N_n(0, \sigma^2 I)$ for some $\sigma^2 > 0$. If p is small relative to n this is a classical parametric model as considered in Section 1.1. If p is fixed for instance and $n \to \infty$, then we have $\sqrt{n}$-consistency, asymptotic normality, and asymptotic efficiency of the maximum likelihood estimator (see for instance Chapter 8 of Davison (2003)).

There are however also situations in which p is not small compared to n. In certain applications in genomics for instance it is not uncommon that p is in the many thousands, while n is only in the hundreds. In such cases we speak of a *high-dimensional linear model*. A simple special case is the normal means model, which is the case $p = n$ and $X = I$, so that $Y = \theta + \varepsilon$. This arises for instance in image denoising, with θ representing a noiseless image and Y the noisy observation. Estimating θ then corresponds to removing the noise from the observed image.

Suppose for instance that we need to denoise a single noisy 200×200 pixel image. If we postulate the normal means model we then have $p = 40,000$ real parameters to estimate (plus one for the variance of the noise). Clearly this does not fall into the category of parametric problems as considered in Section 1.1. The maximum likelihood estimator for θ in the model $Y = \theta + \varepsilon$ is simply the data Y itself, which is of course rather useless. This indicates that in models of this type, we need to consider alternative approaches to making inference about unknown parameters. Although high-dimensional models have finitely many parameters, the number is so high that the problem has much more in common with a fully nonparametric problem as considered in the preceding example, than with a classical parametric problems with just a relatively small number of real parameters. For much more on this, see for instance Bühlmann and van de Geer (2011). ■

Example 1.2.3 (Nonparametric density estimation)**.** Suppose we observe a sample $X_1, \ldots, X_n$ from some unknown distribution. The classical approach to fitting a statistical model to such data, as considered in Section 1.1, is to postulate that the X_i form a sample from some parametric family of densities $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ and then to estimate the parameter θ from the data. Often it is however not clear what a suitable parametric family of densities is and we would prefer to avoid postulating a model that is potentially incorrect. As an example, see the data in Figure 1.2.

In such cases we might view the whole unkown density p of the X_i as parameter, and only assume that it belongs to some broad class $\mathcal{P}$ of densities, for instance the class of all continuous densities, or all densities with a certain degree of smoothness. In this case the whole function p is the parameter of the statistical model and the parameter space $\mathcal{P}$ is an infinite-dimensional function space. The problem of estimating p in such an infinite-dimensional space of densities is called *nonparametric density estimation*.

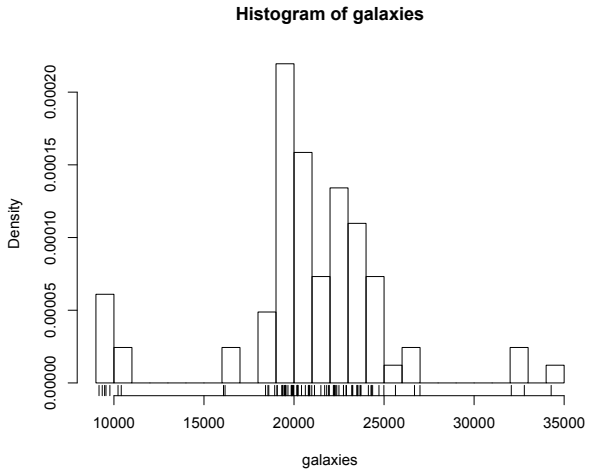

Figure 1.2: The galaxies dataset.

In Exercise 1.2 we consider some issues regarding parametric versus nonparametric maximum likelihood estimation for the density of the data in Figure 1.2. ■

Example 1.2.4 (High-dimensional classification)**.** Consider a binary classification problem in which we observe independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $X_i \in \mathcal{X} \subset \mathbb{R}^k$ and $Y_i \in \{0, 1\}$. Think for example of a situation in which the X_i represent images that can have two different labels, 0 or 1. We want to infer from the data the binary regression function $p : \mathcal{X} \to [0, 1]$ given by

$$p(x) = \mathrm{P}(Y_i = 1 \,|\, X_i = x).$$

An estimator $\hat{p}$ of this function can be used for instance to predict the label of new, unlabelled images. For a given $x \in \mathcal{X}$ the usual prediction rule is to predict the label to be 1 if $\hat{p}(x) > 1/2$ and 0 otherwise.

A popular parametric model for this setting is the logistic regression model, which postulates that the binary regression function is of the form

$$p_\theta(x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}}$$

for some $\theta \in \mathbb{R}^k$. Note that for this model we have $p_\theta(x) > 1/2$ if and only if $\theta^T x > 0$. Hence, label prediction becomes straightforward once the parameter θ has been estimated. If we have an estimator $\hat{\theta}$ for θ obtained from fitting the

logistic regression model to some training data, then a new, unlabelled $x \in \mathcal{X}$ can be classified by simply considering the sign of $x^T\hat{\theta}$.

If the X_i are deterministic, the log-likelihood is easily seen to be given by

$$\theta \mapsto \sum Y_i \theta^T X_i - \log(1 + e^{\theta^T X_i}).$$

There exists no closed-form expression for the maximizer of this function, but if k is small relative to n then the MLE can be efficiently computed numerically. In the high-dimensional situation that $k \geq n$ however, the MLE typically does not exist. Indeed, the jth partial derivative of the log-likelihood is given by

$$\sum Y_i X_{ij} - \frac{X_{ij} e^{\theta^T X_i}}{1 + e^{\theta^T X_i}} = \sum X_{ij} \frac{Y_i(1 + e^{\theta^T X_i}) - e^{\theta^T X_i}}{1 + e^{\theta^T X_i}} = (X^T v_\theta)_j,$$

where $X = (X_{ij})$ is an $n \times k$ matrix and v_θ is the n-dimensional vector with coordinates

$$v_{\theta,i} = \frac{Y_i(1 + e^{\theta^T X_i}) - e^{\theta^T X_i}}{1 + e^{\theta^T X_i}}.$$

If θ is a stationary point of the log-likelihood, then $X^T v_\theta = 0$. But if $k \geq n$, then X^T typically has full rank n. If that is the case, then for a stationary point θ it must hold that $v_\theta = 0$, i.e. that

$$Y_i(1 + e^{\theta^T X_i}) = e^{\theta^T X_i}$$

for every i. Clearly this is impossible.

So also in high-dimensional logistic regression the usual parametric approach breaks down and alternative ideas are necessary. See Exercise 1.3 for a concrete example. ■

The examples we have considered in this section have in common that they involve statistical models with parameters that are very high-dimensional compared to sample size, or even infinite-dimensional. Naively using methods like maximum likelihood is typically either impossible or useless in these cases, so we will have to develop alternative ideas to tackle such problems.

We will see that in high-dimensional or nonparametric models we usually have completely different behaviour of statistical procedures. Convergence rates are typically slower than $\sqrt{n}$, asymptotic normality is not guaranteed, and optimality of procedures can not be assessed in terms of minimal variance. Moreover, we will need to use other technical machinery for the mathematical analysis, since the "Taylor's formula + Central Limit Theorem" approach will not get us very far.

1.3 Exercises

Exercise 1.1 (Kullback-Leibler divergence)**.** Let f, g be positive probability densities with respect to some measure μ. Show that $\mathrm{KL}(f, g) \geq 0$ and that $\mathrm{KL}(f, g) = 0$ if and only if $f = g$, μ-almost everywhere. (Hint: lower bound the KL-divergence by the squared Hellinger distance, i.e. the squared L^2-distance between the square roots of the densities.)

Exercise 1.2 (Parametric versus nonparametric mixture models in density estimation)**.** A distribution on the line is called a (discrete) *mixture of normals* if its density f is of the form

$$f(x) = \sum_{k=1}^{K} p_k \frac{1}{\sigma_k} \varphi\Big(\frac{x - \mu_k}{\sigma_k}\Big), \tag{1.1}$$

where φ is the standard normal density, $K \in \mathbb{N}$ is the number of mixture components, $\mu_1, \ldots, \mu_K \in \mathbb{R}$ are the locations of the components, $\sigma_1, \ldots, \sigma_K > 0$ are the standard deviations, and $p_1, \ldots, p_K > 0$ are the component weights. Clearly, we should have $\sum p_k = 1$ to ensure that f is a probability density.

(i) If we fix the number of mixture components at some level K, a normal mixture model is just a parametric model with $3K - 1$ real parameters. Computing the maximum likelihood estimator for these parameters is not entirely straightforward however. Closed-form expressions can not be derived and one has to resort to numerical methods. An effective approach in this setting is to introduce, in addition to the datapoints, a layer of unobserved auxiliary variables that describe the mixture components to which the datapoints belong. The EM algorithm can then be used to iteratively compute the MLE for the parameters. See for instance Section 5.5.2. of Davison (2003) for details.

In R this estimation procedure has been implemented for instance in the function `normalmixEM` in the package `mixtools`. Use this function to fit normal mixture models with $K = 2, 3, 4$ and 5 components to the galaxies dataset from Figure 1.2, which can be found in the R package `MASS`. Report the estimates for the weights, the locations, and variances. Plot the weighted densities of the components and the estimated overall density over the histogram of Figure 1.2.

(ii) We can also leave the number of components unspecified, leaving it as a parameter that is to be estimated from the data. This turns the model into an infinite-dimensional model.

Prove that the MLE does not exist in this case. More precisely, let $\mathcal{F}$ be the class of all mixture densities of the form (1.1) and prove that if

$X_1, \ldots, X_n$ is a sample from any $f_0 \in \mathcal{F}$, then

$$\max_{f \in \mathcal{F}} \prod_{i=1}^{n} f(X_i) = \infty$$

almost surely.

Exercise 1.3 (High-dimensional versus low-dimensional classification)**.** The famous MNIST dataset contains a large number of 28×28 pixel, grayscale images of handwritten digits. These images are labeled, i.e. the true number that they depict is in the dataset as well. We have extracted from the MNIST database a file containing 2215 labeled images of 0's and 1's. From this file, 100 labeled images were selected at random, see Figure 1.3. In this exercise we explore the possibility of fitting a logistic regression model to this training dataset, with the goal of predicting the labels of the other 2115 images.

(i) The training data are available in the file `traindata01.csv` on the webpage. This file contains 100 lines, each corresponding to a labeled image. A line contains numeric values separated by commas. The first value on a line is the label of the image, which is a 0 or a 1. The other $28 \times 28 = 784$ values are integers between 0 and 255, which are the gray-scale values of the pixels in the image. The value 0 corresponds to white, the value 255 corresponds to black.

We can treat the data as independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ as in Example 1.2.4, with $n = 100$ and for every i, X_i a 784-dimensional vector.

Show that the matrix $X = (X_{ij})_{i=1,\ldots,100;j=1,\ldots,784}$, as considered in Example 1.2.4, has rank 100 in this case. Hence, we are in the high-dimensional setting in which the MLE does not exist.

Nevertheless, the R function `glm` can still be used to fit a logistic regression model to the data. This does not produce an error message that the rank of X is too low. Instead, R basically removes columns from X that are linear combinations of other columns. For those columns, it produces an `NA` as estimate of the corresponding coordinate of the parameter vector θ.

Fit the logistic regression model using `glm`. Determine which columns of X were retained by R, i.e. which coordinates of the estimate for θ contain non-`NA`'s. How many such columns are there?

Let $\tilde{X}$ and $\tilde{\theta}$ be the matrix X and the estimated vector θ from which the columns and coordinates corresponding to the `NA`'s have been removed. Show that by considering the sign of $\tilde{X}\tilde{\theta}$, we obtain predictions for the labels of the training images that are correct in all 100 cases.

Figure 1.3: The 100 training images from the MNIST dataset.

(ii) Instead of using the complete images as explanatory vectors in the logistic regression model we can try to come up with reasonable low-dimensional *feature* vectors that also have good predictive power, so that we can can fit an ordinary, low-dimensional logistic regression model and avoid the issues arising in the high-dimensional setting.

Show that if instead of with the complete training images we just work with the 2×2 pixel sub-images at the center of each image, we can fit a logistic regression model, do not obtain `NA`'s, and also predict all 100 training image labels correctly.

(iii) Parts (i) and (ii) give rise to two different predictors for labels of images which take a 28×28 image as input and produce a label in $\{0, 1\}$ as output.

Compare the prediction performance of these two methods on the whole collection of 2215 images, available in the file `alldata01.csv`.

Can you explain why there is a difference in the *generalization* performance, i.e. in the quality of the label predictions for images of which the label has not been observed?