Johannes Schmidt-Hieber

Lecture notes

Lecture Notes

Foreword

In statistics, we estimate/reconstruct objects from data. Given a noisy image for example, the aim is to reconstruct the underlying true image.

A first course in statistics typically deals with reconstruction of finite dimensional parameters, such as the mean or the standard deviation of a distribution. For many interesting applications, however, we want to assume as little as possible about the true underlying objects. Taking a fixed number of parameters is then not appropriate. Instead this should be modelled by assuming a high-dimensional or even infinite dimensional parameter space. To reconstruct an image, for instance, we can think of it as a two-dimensional function and take as a parameter space a function class.

A statistical model with complex data structure refers to settings where the classical parametric theory fails. It comprises two subfields, nonparametric statistics where the parameter space is infinite dimensional (e.g. a function class) and high-dimensional statistics with parameter spaces that are typically vector spaces with dimensions that increase with the sample size. We choose a mathematical approach that allows to treat these two cases simultaneously.

The mathematical theory of complex statistical models has been developed largely during the past years but remains a topic of active research with many challenging open problems. One of the nice features is that there is a notion of optimality and estimators (reconstruction methods) can be constructed that (nearly) achieve this optimal behaviour.

Nonparametric and high-dimensional statistics is certainly among the most active current research areas in applied mathematics. This is due to the countless number of applications across all sciences. Mathematics has always evolved along its applications. Newton's discovery that mechanical systems are based on differential equations initiated the development of a sound theory of differential equations. Today's applications are driven by the enormous amounts of data that we collect. Often, these questions are of the following kind: Given data reconstruct some complex latent object. For example, reconstruct the true image from a noisy version. The text aims to summarize mathematical concepts that underly statistical methods for these type of questions.

The lecture notes are structured as follows. In a first part, we introduce and motivate different statistical models. Based on that, we identify the most commonly used nonparametric and high-dimensional models. We then discuss different estimation principles for nonparametric problems, based on smoothing and series expansions and study their estimation error under different loss functions. It turns out that these methods always rely on a good choice of a bandwidth/smoothness/truncation parameter which needs to be carefully selected and we discuss different strategies how to do that. One of the most common strategies in nonparametric estimation is series estimation with shrinkage of the series coefficients. This is closely related to Stein's phenomenon which shows that minimax estimators can be improved, in particular in high dimensional settings. Based on that, we discuss high-dimensional models under constraints such as sparsity. In Chapter 9, we then ask whether these estimators are optimal. Here, optimality is studied for large sample size and we are mainly interested in the optimal rate of convergence. For that, we use a general reduction scheme that relates lower bounds to control with respects to information distances. We also study another approach which links minimax lower bounds to the Bayes risk. This allows to obtain very precise non-asymptotic lower bounds for some simple

problems.

We assume that the reader has knowledge of basic parametric statistical theory as commonly covered by an introductory course in statistics.

There are many excellent books on nonparametric statistics available. A very good survey of the field is given in Wasserman [30]. The book covers the most relevant estimation strategies with as little details as possible. Proofs are mainly omitted which allows to get a quick overview of the theory. The standard reference for the nonparametric minimax theory is Tsybakov [28] containing in particular the most complete description of proving strategies for deriving lower bounds on the minimax risk. This book is based on the older text Korostelev and Tsybakov [16] which was more specifically written on image recovery and therefore contains additional material on change-point and edge-reconstruction problems. The Saint-Flour lecture notes by Nemirovski [8] are a bit of the same flavor discussing estimation of functionals in great detail.

Additionally to these more general treatments there are several books devoted to specific models or estimation techniques. Estimation in the Gaussian sequence model might be viewed as the fundamental problem in nonparametric function estimation and is extensively studied in the excellent draft Johnstone [15] (available from `http://statweb.stanford.edu/~imj/GE_08_09_17.pdf`). This book is particularly recommended as additional reading. Wand and Jones [29] is a very detailed and easy to read book on kernel estimation. There are many books on wavelets methods. A good introduction to the statistical theory of wavelets is Haerdle et *al.* [13]. Giné and Nickl [11] approach nonparametric statistics from empirical process theory. This book contains also a lot on function spaces. A more statistical learning point of view on nonparametric statistics was adopted in Györfi et *al.* [12]. One of the achievements of this book is the summary of the mathematical convergence theory of neural networks. Estimation and inference in high-dimensional statistics are summarized in the books Bühlmann and van de Geer [5] and Hastie et *al.* [14].

We are grateful to Gino Kpogbezan, Tyron Lardy and Lincen Yang for valuable comments and suggestion.

If you find typos or you think that certain parts are unclear, please feel free to mention that to me. Suggestions and comments are always welcome and can be send to schmidthieberaj@math.leidenuniv.nl

Notation: We use mainly standard notation which is recalled here for convenience.

Sets: For a finite set A, $\#A$ denotes the number of elements in A. For an arbitrary set B, $\mathbf{1}(B)$ denotes the indicator function on B.

Functions: To define a function we often use the $\cdot$ operator. For example, instead of $x \mapsto g(x) = \sin(2x)$ we write $g = \sin(2\cdot)$. As usual we write $L^p(\mathbb{R})$ for the Lebesgue function spaces on $\mathbb{R}$.

Vectors: We often denote vectors by bold letters. For the transpose, we write $^\top$.

Sequences and order symbols: For two sequences of positive real numbers $(a_n)_n$ and $(b_n)_n$ we write $a_n \ll b_n$ or $a_n = o(b_n)$ if $\limsup_n a_n/b_n = 0$; the notation $a_n \lesssim b_n$ or $a_n = O(b_n)$ is used if $\limsup_n a_n/b_n < \infty$; and $a_n \asymp b_n$ if $a_n \lesssim b_n \lesssim a_n$. In particular it is convenient to write $o(1)$ for a sequence tending to zero.

Random variables: We abbreviate the terms cumulative distribution function and probability distribution function by c.d.f. and p.d.f. The abbreviation i.i.d. stands for independent and identically distributed. If X is a random variable drawn from a distribution P, we write $X \sim P$. We also sometimes write $X \sim F$ or $X \sim f$ for a c.d.f. F or a p.d.f. f to indicate that X follows the distribution induced by F or f, respectively. The expectation taken with respect to the probability measures P_θ or P_f is denoted by E_θ or E_f, respectively. As usual, we write $\mathrm{Cov}(X, Y)$ for the covariance of the random variables X and Y and $\mathrm{Var}(Y)$ for the variance of Y. If X follows a normal distribution with mean μ and variance σ^2, we write $X \sim \mathcal{N}(\mu, \sigma^2)$. If X is multivariate normal with mean vector μ and covariance matrix Σ, we write $X \sim \mathcal{N}(\mu, \Sigma)$. In particular, I_n denotes the $n \times n$ identity matrix. The distribution of a Poisson random variable with intensity parameter λ is denoted by $\mathrm{Poisson}(\lambda)$.

Measures: For two probability measures P, Q defined on the same measurable space, we write $P \ll Q$ if P is dominated by Q, that is, $P(A) = 0$ whenever $Q(A) = 0$.

Miscellaneous: In an equation, $:=$ means that the left hand (l.h.s.) is defined as the r.h.s. In heuristic arguments, we use $x \approx y$ to indicate that x and y are close. We define $(x)_+ = \max(x, 0)$ and write $x \vee y$, $x \wedge y$ for the maximum and minimum of x and y, respectively. As usual, we write $\mathrm{argmin}_{\theta \in \Theta} F(\theta)$ for the set of all elements in the closure of Θ that minimize the expression $F(\theta)$.

Contents

Chapter 1

Introduction

1.1 From parametric to nonparametric statistics

Figure 1 shows a scatterplot of a dataset which comes with the r-package ”np”. It plots the age against the logarithm of the wage for a group of Canadians with similar educational background, cf. [17]. A classical idea is to explain this dataset by a linear relationship between the two observed variables. This amounts to fitting a straight line through the data. In the least squares sense, the ”optimal line” is the one which minimizes the residual sum of squares. This means that if we ”subtract” the line from the data points, that is, if we look at the residual, then the least squares fit returns the line minimizing the sum of squared residuals.

It is not hard to see that using a linear model for this dataset results in a very poor fit. As statisticians, we have many options how to improve the model. One possibility would be instead of considering log(wage) on the y-axis, to take a different transformation, which gives a more linear relationship. Alternatively, there are plenty of other models that we could try and which allow to incorporate the nonlinearity of the drift. For instance we could fit a parabola to the data.

All such models, can be stated in the following form. We observe n pairs

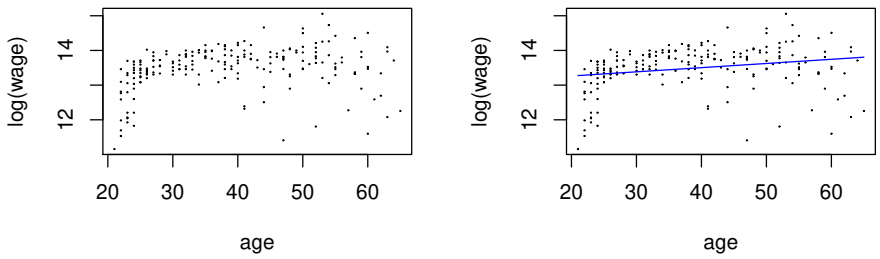


Figure 1: Data example with linear least squares fit

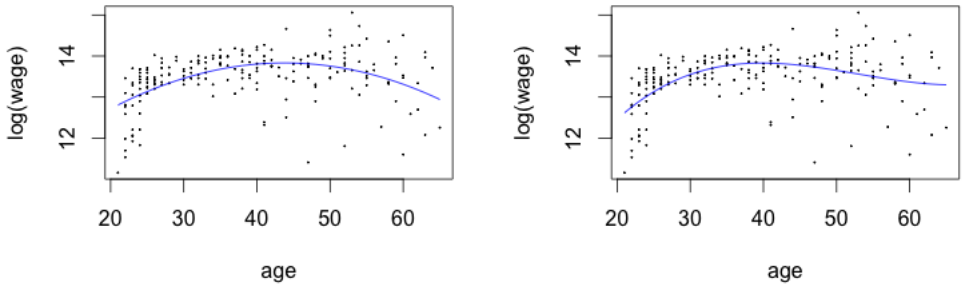


Figure 2: Least squares fit with second order and third order polynomials

$(X_1, Y_1), \ldots, (X_n Y_n)$, with

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1.1}$$

and f in some class of functions. Here and in the following, f is called the *regression function*. The random variables ε_i model the *(measurement) noise*. We do not observe these noise variables directly but typically have some knowledge about the distribution of ε_i.

In the example above, n is the sample size, X_i is the age of the i-th observed person and Y_i is the corresponding $\log$(wage). If we believe in a linear relationship, we assume that $f(x) = ax + b$ with a, b real numbers. We may equivalently write then, that $f \in \mathcal{F}$ with

$$\mathcal{F} = \{f : f(x) = ax + b,\ a, b \in \mathbb{R}\}. \tag{1.1.2}$$

The class of target functions $\mathcal{F}$ is called the *parameter space*.

If instead, we want to model the non-linearity, we might add a squared term and would then take instead the function class

$$\mathcal{F} = \{f : f(x) = ax^2 + bx + c,\ a, b, c \in \mathbb{R}\}$$

as parameter space. Fitting a quadratic function seems to lead to a better fit of the data than the linear model. Nevertheless, there would still be a discrepancy between the model and the data and this would become visible if we would add more data for instance. In Figure 2, the least squares fit for polynomials of degree two and three are displayed. We see that even then, there is quite a big difference between the two reconstructions. Also the interpretation of the estimators would be different. The reconstruction for quadratic polynomials suggests a much faster decay in the $\log$(wage) than the estimator based on the third order polynomials. Moreover, the maximum in the first plot seems to be attained for the age between 40 and 50 and in the second plot for the age below 40. This difference is an artifact induced by the characteristic shape of 2nd and 3rd order polynomials.

We might argue that all models are wrong. Any parameter space is therefore a subjective selection of candidate functions by the statistician. We can overcome this problem to a large extent by allowing for a much bigger parameter space, for instance the parameter space of all Lipschitz functions. Recall that a function f is Lipschitz on the domain D, if $|f(x) - f(y)| \leq L|x - y|$, for all $x, y \in D$. The smallest constant L for which this holds is called the Lipschitz constant. We might take this space to be the parameter space, that is,

$$\mathcal{F} = \{f : \ f \text{ is a Lipschitz function}\}.$$

In particular, this space contains all polynomials on a bounded interval. In contrast to the parameter spaces considered above, this space cannot be parametrized by finitely many real parameters.

1.2 Nonparametric and high-dimensional statistical models

A statistical decision problem is called *parametric* if the parameter space is parametrizable by finitely many *real* parameters and the number of real parameters does not grow with the sample size. The parameter space of linear functions (1.1.2) can be for instance parametrized by two real parameters and gives henceforth rise to a parametric model. Nonparametric just means that the problem is not parametric. Differently speaking, a statistical decision problem is called *nonparametric* if the parameter space *cannot* be parametrized by finitely many real parameters such as in the example with the space of Lipschitz functions above or if the number of parameters grows with the sample size.

A dependence of the sample size on the number of parameters might seem at first sight unnatural. It will turn out later that an infinite dimensional problem has typically many irrelevant dimensions and can be reduced to a subspace whose dimension grows with the sample size.

The term nonparametric is misleading as it suggests that there is no parameter anymore. If the parameter space is the class of Lipschitz functions, we would still refer to $\mathcal{F}$ as the parameter space and to a Lipschitz function f as a parameter. This means that a nonparametric statistical model still has parameters and a parameter space.

In all the applications of nonparametric models that we discuss, the parameter space is a function class. One can therefore also think of a nonparametric statistical problem as reconstruction of functions from data.

A statistical model is called high-dimensional if the number of real parameters needed to parametrize the parameter space exceeds the sample size. We will see later that this is the correct framework for many real-world applications.

> A statistical decision problem is called *high-dimensional* if the number of real parameters exceeds the number of observations. A statistical decision problem is called *nonparametric* if the number of real parameters grows with the sample size or if the parameter space is infinite dimensional.

Notice that this is not a precise mathematical definition since not every space has a well-defined notion of dimension and the notion might change for instance by changing the underlying topology.

In statistical theory, high-dimensional is typically understood as having more parameters than data. In applications the term high-dimensional is rather used to denote statistical models with many parameters. A model with sample size 1000 and 100 parameters would often be called high-dimensional, for instance.

One of the distinctive features of a parametric model is that if we have a sample of size n we can in many cases reconstruct the true parameter up to error $n^{-1/2}$. The term nonparametric is occasionally used to denote statistical decision problems with convergence rate slower than the parametric rate $n^{-1/2}$. Large parameter space often lead to slower convergence rates, but this is not always the case (a standard example is estimation of the c.d.f. discussed in Section 1.3) and there are also many estimation problems with finite dimensional parameter space which have convergence rate slower than $n^{-1/2}$. The advantage of this definition is that it can be made mathematical precise, but it is not very useful and we therefore avoid it.

1.3 Some basic nonparametric and high-dimensional statistical models

There is a huge number and variation of proposed nonparametric models. Although the models might be quite different, most of them can be treated with similar methods. To develop nonparametric statistic theory, it will be enough to consider few basic types of models, which we present below.

Nonparametric regression

Nonparametric regression is also often called the signal plus noise model. In this model, we observe a deterministic function subject to some additive noise, which typically models measurement errors. In its simplest form, we observe a vector of independent random variables $\mathbf{Y} = (Y_{1,n}, \dots, Y_{n,n})$ with

$$Y_{i,n} = f\left(\tfrac{i}{n}\right) + \varepsilon_{i,n}, \quad i = 1, \dots, n. \tag{1.3.1}$$

Here, $\varepsilon_{i,n} \sim \mathcal{N}(0,1)$, $i = 1, \dots, n$, are *independent and identically distributed* (i.i.d.) random variables. Equivalently, we can write this model as

$$Y_{in} \sim \mathcal{N}\left(f\left(\tfrac{i}{n}\right), 1\right), \quad i = 1, \dots, n.$$

The regression function $f \in \mathcal{F}$ is unknown and the choice of the parameter space $\mathcal{F}$ depends on the specific application.

This is the simplest model within a whole class of regression models. In order to model noisy images, we can extend the model to regression functions on the square $[0,1]^2$. Then, we observe a matrix $\mathbf{Y} = (Y_{ijn})_{i,j=1,\dots,n}$ with

$$Y_{ijn} = f\left(\tfrac{i}{n}, \tfrac{j}{n}\right) + \varepsilon_{ijn}, \quad i, j = 1, \dots, n.$$

Figure 3 displays a grayscale image. On each pixel, the grayscale is converted into numeric values of $f(i/n, j/n)$ and Y_{ijn}.

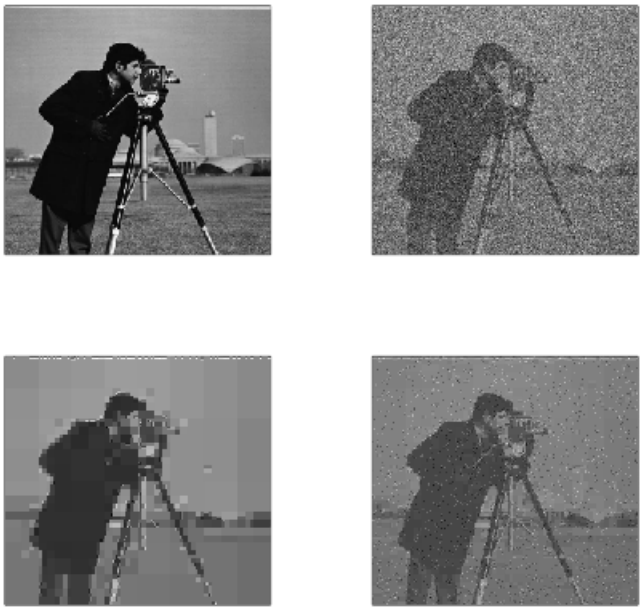


Figure 3: *Top:* True image (left) and observed noisy image (right). *Bottom:* Two reconstructions, cf. Section 5.6.

The assumption in model (1.3.1) that we observe f at time points i/n, $i = 1, \dots, n$ is not always met in applications and many extensions to other designs exist. We can distinguish two cases, namely *deterministic design* and *random design.* In deterministic design, we have numbers $x_1, \dots, x_n$ and observe the random vector $\mathbf{Y} = (Y_{1,n}, \dots, Y_{n,n})$ with $Y_{i,n} = f(x_i) + \varepsilon_{i,n}, \quad i = 1, \dots, n$. In random design, we assume that we observe n i.i.d. pairs (X_i, Y_i) with

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n. \tag{1.3.2}$$

This is the model which we already discussed in (1.1.1).

Another generalization of model (1.3.1) is to work under a more general noise assumption. In order to account for heteroscedasticity, we might introduce a variance function σ^2 and replace (1.3.1) by

$$Y_{i,n} = f\left(\tfrac{i}{n}\right) + \sigma\left(\tfrac{i}{n}\right)\varepsilon_{i,n}, \quad i = 1, \dots, n. \tag{1.3.3}$$

Depending on the application, σ is assumed to be known or unknown. For many applications, the Gaussian assumption on the noise is violated and we should impose other noise distributions instead, for instance one which allows for extreme outliers. It might also be enough to assume that ε_{in} are i.i.d., centered (that is, $E[\varepsilon_{in}] = 0$), and that the first r absolute moments are finite for some $r \geq 1$. If we do not make the assumption that ε_{in} is centered, there is no way to tell from the data whether the regression function is f or $f+$constant.

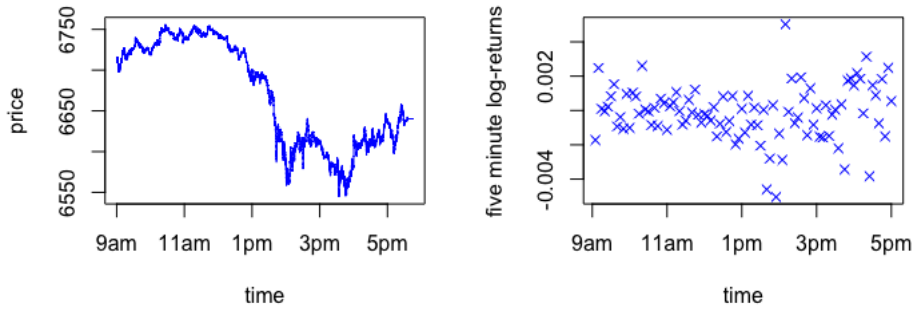


Figure 4: Left: Prices of a German Bund future over one trading day. Right: Five minute log-returns.

To understand how estimation for nonparametric regression works, it is enough to study the toy model (1.3.1). Theoretical results for the generalized models, can then be obtained using similar strategies and often only lead to additional technicalities in the proofs.

Volatility estimation

An important problem in finance is to determine the volatility from historic data. The volatility can be used as risk measure or for pricing of financial derivatives. Suppose, we have observations from a financial asset over a finite time interval say $[0, T]$ such as a day or a month. For convenience set $T = 1$. Divide the interval $[0, 1]$ in n smaller time windows of equal length, such as five-minute intervals. On each of these smaller intervals, we compute the change in the logarithm of the price. A very simple model for these so-called log-returns is

$$Y_{i,n} = \frac{\sigma(i/n)}{\sqrt{n}}\varepsilon_{i,n}, \quad i = 1, \dots, n.$$

Figure 4 displays a real data example of a FGBL future on one trading day together with the corresponding five minute log-returns. The unknown function $t \mapsto \sigma(t)$ is called the spot volatility, that is, the volatility as a function over time. The statistical task is to estimate the spot volatility from the data. The model is essentially the same as (1.3.3) with regression function $f = 0$.

One might question the use of the asymptotic statement $n \to \infty$ in this setup. Indeed if n gets large this means that we take log-returns on smaller intervals. An alternative would be to consider a longer observation period of the underlying asset, that is, $T \to \infty$. Both asymptotics are of interest and are known as high-frequency and low-frequency limit.

Gaussian white noise model

Working with the nonparametric regression model (1.3.1), it turns out that the discreteness of the design does not play a role, but makes the proofs a bit technical. The idea is to replace (1.3.1) by

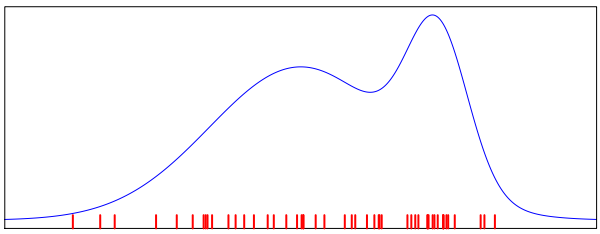


Figure 5: Simulated sample (red) drawn from a density (blue).

a continuous version.

Recall that a Brownian motion $(W_t)_{t>0}$ is a centered Gaussian process with covariance structure $\mathrm{Cov}(W_s, W_t) = s \wedge t$. It is the limit of suitably scaled partial sum processes. Indeed if we sum the first k observations in (1.3.1), we obtain

$$\frac{1}{n}\sum_{i=1}^{k} Y_{i,n} = \frac{1}{n}\sum_{i=1}^{k} f\left(\tfrac{i}{n}\right) + \frac{1}{n}\sum_{i=1}^{k} \varepsilon_{i,n} \approx \int_0^{k/n} f(u)\,du + n^{-1/2} W_{k/n}$$

(cf. Exercise 1.1). This is close to the model, where we observe the path of the process $(Z_t)_{t\in[0,1]}$ with

$$Z_t = \int_0^t f(u)\,du + n^{-1/2} W_t, \quad t \in [0,1].$$

This model is called the *Gaussian white noise model*. This is a continuous model, that is, we observe a stochastic process on an interval. The Brownian motion models the noise and is scaled by the sample size n. Obviously, if n becomes large the noise term vanishes.

Estimation of c.d.f. and p.d.f.

Given a random variable X, recall that $t \mapsto F(t) = P(X \le t)$ is the *cumulative distribution function* (c.d.f.). If X has a density f with respect to the Lebesgue measure on $\mathbb{R}$, then $f(t) = F'(t)$ is the *probability density function* (p.d.f.). Consider the statistical model, where the sample consists of n i.i.d. copies of X, that is, we observe $X_1, \dots, X_n \sim F$, i.i.d. The unknown parameters are the c.d.f. F and its density f.

To estimate F, we can use the estimator

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le t). \tag{1.3.4}$$

By the strong law of large numbers, $\widehat{F}_n(t) \to F(t)$ a.s. Application of the central limit theorem yields moreover

$$\sqrt{n}\big(\widehat{F}_n(t) - F(t)\big) \to \mathcal{N}\big(0, F(t)(1-F(t))\big) \tag{1.3.5}$$

Figure 6: Artificial images of birds that are black with green and have a very short beak. Citation needed ???

implying the parametric $n^{-1/2}$-rate of convergence.

Estimation of the density f is more involved. The statistical problem is displayed in Figure 5 based on simulated data. The red tick marks are located at the values $X_1, \ldots, X_n$. The true (unknown) density f is plotted from which the data were generated. The statistical problem is to reconstruct the function f from the observations.

Since $f = F'$, one could try to use the derivative of the estimator $\widehat{F}_n$ to estimate the density f. Since $\widehat{F}_n$ is a step function, the estimator $\widehat{F}_n'$ is a weighted sum of delta functions which does not converge to f pointwise. Thus, this estimator does not work. A very common technique to estimate the density f is to use a histogram. In Chapter 3, we consider so called kernel density estimators which allow to incorporate underlying smoothness of the signal.

The input X does not need to be a random variable but could also be a more general random element. A very recent applications are so called generative adversarial networks (GANs). The idea is to take as X_i images from the same object, say birds. Once we estimated the density we can sample from it new (artificial) images of birds. We can also specify an x and obtain the corresponding bird image $\widehat{f}(x)$. This means we can generate artificial images of birds with specific properties such as their color. An illustration is given in Figure 6.

Deconvolution

In this model, we observe n i.i.d. copies of the sum of two independent random variables. More precisely, we observe

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \ldots, n, \quad X_i \overset{i.i.d.}{\sim} F, \ \ \varepsilon_i \overset{i.i.d.}{\sim} F_\varepsilon. \tag{1.3.6}$$

It is common to assume that the random vectors $(X_1, \ldots, X_n)$ and $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent implying that the Y_i are i.i.d. Suppose that bounded Lebesgue densities $f = F'$ and $f_\varepsilon = F_\varepsilon'$ exist. If G denotes the distribution function of Y_1 then

$$\begin{aligned} G(t) = P(Y_1 \le t) = P(X_1 + \varepsilon_1 \le t) &= \int_{-\infty}^{\infty} P(X_1 \le t - x \,|\, \varepsilon_1 = x) f_\varepsilon(x) dx \\ &= \int_{-\infty}^{\infty} F(t-x) f_\varepsilon(x) dx =: (F \star f_\varepsilon)(t), \end{aligned}$$

Figure 7: *Top:* True image (left) and convolved image (right). *Bottom:* True image with noise and convolved image with noise.

where $\star$ denotes the so called *convolution product* $h_1 \star h_2(t) = \int_{-\infty}^{\infty} h_1(t-x)h_2(x)dx$. By Lemma 13,

$$g(t) = \int_{-\infty}^{\infty} f(t-x)f_\varepsilon(x)dx =: (f \star f_\varepsilon)(t).$$

The statistical task is now to reconstruct the density f from the observations $(Y_1, \ldots, Y_n)$ assuming that f_ε is known. Thus, we have to invert the convolution step. These models are therefore called deconvolution models. The model combines density estimation with additive errors. It occurs in many applications where a sample from a distribution is obtained that is corrupted by measurement noise.

There are many applications where we observe a convolved image possibly also perturbed by additional measurement noise. Figure 7 provides an example, where we observe the cameraman image subject to convolution and measurement noise. The data were generated in this case from the model

$$Y_{ijn} = (f \star f_\varepsilon)\left(\tfrac{i}{n}, \tfrac{j}{n}\right) + \varepsilon_{ijn}, \quad i, j = 1, \ldots, n,$$

with f_ε a known density, f the true image and $\varepsilon_{ijn} \sim \mathcal{N}(0,1)$, i.i.d. Figure 7 also shows the data without the additive errors. In this case one can see clearly that the convolution has a blurring

effect. Reconstruction of the true function in a convolution model, is therefore also often called deblurring.

High-dimensional linear regression

In the high-dimensional regression model, we observe a vector $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and a $n \times p$ design matrix X with

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.3.7}$$

Here, $\boldsymbol{\beta}$ is an unobserved p-dimensional coefficient vector. Moreover, I_n denotes the $n \times n$ identity matrix and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ is a centered random vector from a multivariate normal distribution with covariance matrix I_n. The design matrix X is known and the goal is to estimate the unknown coefficient vector $\boldsymbol{\beta}$.

This model is high-dimensional if $p > n$, since in this case we have more parameters than observations. Typically we think of p as a sequence of n, for instance $p = n^2$ or $p = e^n$. Compared to the nonparametric models that we have seen so far, the high-dimensional regression model is philosophically a bit different as we think of the parameter as a quantity that depends on the sample size n, whereas before the parameter was considered to be fixed. In particular, the parameter space depends on n through p but we typically omit the dependence in the notation.

To indicate the dimensions of the objects in the high-dimensional regression model, we can also write the model more explicitly as

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} X_{11} & \dots & \dots & \dots & \dots & X_{1p} \\ \vdots & & & & & \vdots \\ X_{n1} & \dots & \dots & \dots & \dots & X_{np} \end{pmatrix}}_{n\times p} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix}}_{p\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{n\times 1}.$$

High-dimensional regression has various applications. In genomics we want to discover which combination of genes is responsible for a certain phenotype. This can be modelled by (7.2.1), with $\mathbf{Y}$ a vector of observed phenotypes and the regression vector $\boldsymbol{\beta}$ containing the effect of each gene. The human genome consists of about 30.000 genes and typical studies have several hundred patients. The number of unknown parameters is therefore much bigger than the sample size and we are in a high-dimensional regime.

Even if we neglect the noise in the high-dimensional regression model and assume that $\mathbf{Y} = X\boldsymbol{\beta}$ is observed directly, the linear system is underdetermined since there are more unknowns than equations. In such a situation the model is also said to be *not identifiable*. In order to ensure identifiability, additional strutural assumptions on the regression vector $\boldsymbol{\beta}$ need to be made.

In genetics we indeed know that only few genes are responsible for a specific phenotype. Nevertheless we neither know the exact number nor the location of the active genes. The regression

vector has consequently only few non-zero entries and most components are zero. Such vectors are called *sparse*. Sparsity also occurs in astronomy where the sky is scanned for specific objects. The non-zero components in the vector correspond then to the light emmitted from these objects. Since most measured positions do not contain any signal the regression vector is sparse. Because of this application, sparse vectors are also sometimes referred to as *nearly black*.

Given the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, we can define *the support of* $\boldsymbol{\beta}$ or *active set* as the non-zero components $S_{\boldsymbol{\beta}} = \{j : \boldsymbol{\beta}_j \neq 0\}$ and the *sparsity (index)* of $\boldsymbol{\beta}$ as $s_{\boldsymbol{\beta}} = |S_{\boldsymbol{\beta}}| =$ the cardinality of $S_{\boldsymbol{\beta}}$. Because the parameter changes with n, the active set and the sparsity are also n dependent but again this dependence is omitted in the notation.

We say that *a model is sparse if the number of non-zero parameters is of smaller order than the number of observations* which is equivalent to $s_{\boldsymbol{\beta}} \ll n$.

Matrix completion

In the matrix completion model with measurement errors, we observe few, noisy entries of a matrix and want to recover the full matrix under a low-rank constraint.

Define $[p] := \{1, \ldots, p\}$ and $[q] := \{1, \ldots, q\}$. Let $A = (a_{\ell,k})_{\ell \in [p], k \in [q]}$ be a $p \times q$ matrix with real valued entries. Suppose, we observe n independent variables (Y_i, L_i, K_i), where the indices L_i and K_i are independently drawn uniformly at random from $[p]$ and $[q]$, respectively and

$$Y_i = a_{L_i, K_i} + \varepsilon_i,$$

with $\varepsilon_i \sim \mathcal{N}(0, 1)$ independent of (L_i, K_i). In the matrix completion model, we thus see n entries which are perturbed by measurement noise. To "visualize" the model, one can think of a sample as a matrix

$$\begin{pmatrix} & & & * \\ * & * & & \\ * & * & * & \\ & * & & * \end{pmatrix}$$

where $*$ stands for an observed noisy entry.

High-dimensional principal component analysis

Suppose we observe a sample of n independent p-dimensional vectors

$$\mathbf{X}_i \sim \mathcal{N}(0, \Sigma), \quad i = 1, \ldots, n.$$

Here $\mathcal{N}(0, \Sigma)$ denotes the multivariate normal distribution with $p \times p$ covariance matrix Σ. One can equivalently write this model as

$$\mathbf{X}_i = \sum_{j=1}^{p} \sqrt{\lambda_j} \mathbf{v}_j \varepsilon_{ij},$$

with $\varepsilon_{ij} \sim \mathcal{N}(0,1)$, i.i.d., $\mathbf{v}_j$ the eigenvectors of Σ and λ_j the corresponding eigenvalues. Written in this form, we see some resemblance with the linear regression model, but the signal is now in the variation and not in the mean anymore. The aim in principal component analysis is to recover the eigenvalues λ_j and the eigenvectors $\mathbf{v}_j$ from the data.

If we observe many variables simultaneously, the correct way to model this is to allow p to grow with the sample size n. Since Σ has p^2 many parameters, this is henceforth a truly high-dimensional problem if $p \geq \sqrt{n}$.

The empirical covariance is given by

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top.$$

Since $\widehat{\Sigma} \approx \Sigma$ one can study the eigenvectors and eigenvalues of $\widehat{\Sigma}$. In the high-dimensional setup these estimated eigenvectors and eigenvalues are not close anymore to the true eigenvalues and eigenvectors.

High-dimensional sparse additive models

There are several popular models that combine sparsity in a high-dimensional setting with function estimation. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote n i.i.d. copies from a d-dimensional distribution on the hypercube $[0,1]^d$. In the nonparametric regression model with random design we observe the n pairs $(\mathbf{X}_i, Y_i)$, with

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \text{ i.i.d.}, \quad i = 1, \ldots, n$$

and $f : [0,1]^d \to \mathbb{R}$ the unknown regression function. We will see later that if the dimension d is large no good reconstruction can be obtained unless we have an extremely large sample size. One way around this so called curse of dimensionality is to impose additional structure on the regression function f. In additive regression models, we assume that the regression function is of the form

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} f_j(x_j).$$

For $d = 2$, $f(x_1, x_2) = x_1^2 - 2x_2$ is for instance of this form, while $f(x_1, x_2) = x_1 x_2$ cannot be written in an additive form. Additive models form a rich class of functions. In the high-dimensional additive regression model, we assume that the dimension d grows with n. On top of this, we assume that there is sparsity in the sense that only few functions f_j are not identically zero. This means that the true function does not depend on most of the x_i. As in the high-dimensional regression model, we do not know which f_j are active.

1.4 Exercises

Ex. 1.1 — Let $\varepsilon_i \sim \mathcal{N}(0,1)$, i.i.d. $i = 1, \dots, n$ and denote by $(W_t)_{t\geq 0}$ a Brownian motion. Show that

$$\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{k} \varepsilon_i\Big)_{k=1,\dots,n} \overset{\mathcal{D}}{=} (W_{k/n})_{k=1,\dots,n},$$

where $\overset{\mathcal{D}}{=}$ denotes equality in distribution.

Ex. 1.2 — Prove (1.3.5).

Chapter 2

Statistical estimation theory

2.1 Statistical models and estimators

A statistical model describes the data generating process. It is either given to the statistician or chosen. As an example for a simple data generating process, assume that we observe a random variable Y with

$$Y = \mu + \varepsilon, \tag{2.1.1}$$

$\varepsilon \sim \mathcal{N}(0,1)$ a standard normal random variable and $\mu \in \mathbb{R}$ the unknown mean. We might equivalently write this as $Y \sim P_\mu$, where $P_\mu = \mathcal{N}(\mu, 1)$ denotes the distribution of Y. This means that we observe a realization from one of the probability distributions $(P_\mu : \mu \in \mathbb{R})$. We do, however, not know which one since the parameter μ is unknown. The statistical challenge consists in reconstructing μ from observing P_μ. Suppose now that we know in advance that the parameter μ is not an arbitrary real number but lies in some subset $\Theta \subset \mathbb{R}$. Then, the set of candidate probability measures can be restricted to $(P_\mu : \mu \in \Theta)$.

Any data generating mechanism such as (2.1.1) can be rewritten as a collection of probability measures $(P_\theta : \theta \in \Theta)$ with θ the unknown parameter and Θ the parameter space and we call such a collection a *(statistical) model*.

We need to be a bit careful with this definition of a statistical model since it requires an underlying measurable space, say $(\Omega, \mathcal{A})$, on which the probability measures P_θ are defined. For statistical estimation this measurable space is almost irrelevant and we therefore omit it most of the time. A more precise notion is the term *statistical experiment* which stands for the triple $(\Omega, \mathcal{A}, (P_\theta : \theta \in \Theta))$.

For many statistical concepts it is convenient to work with the data generating process instead of the collection of probability measures. As usual, we apply the term (statistical) model also to the data generating process. For instance, we say that (2.1.1) is a statistical model. It is important that one always specifies the observed/unobserved quantities in the data generating mechanism. In addition to (2.1.1) for instance, we need to mention that Y is observed and that μ is unknown.

But there are also statistical problems for which the interpretation of a statistical model as a collection of probability measures is more convenient to work with. This is for instance true for the lower bounds derived in Chapter 9.

If we speak about a generic parameter, we call it θ and the corresponding parameter space Θ. For specific estimation problems, we use the standard letters, for instance for the mean and the standard deviation of a normal random variable we write μ and σ, respectively. Whenever we study estimation of a function, we call this function f and the corresponding parameter space $\mathcal{F}$.

Given data from a model, the statistical task is to make statements about the true underlying parameter. Any measurable function that maps the data to a parameter value is called a *(point) estimator*. We do not require that the estimator takes values in the parameter space Θ. Instead it is enough if it maps to some larger space $\Theta' \supseteq \Theta$ called the *action space*. This slight relaxation will be useful later if the parameter space consists of non-negative functions. The estimator may still attain negative values simplifying the analysis.

2.2 Loss and risk of an estimator

To infer the quality of an estimator, we need to measure the distance between the estimator and the true parameter. Recall that the estimator maps into the action space Θ'. Evaluating the performance of an estimator, we need therefore a function on $\Theta' \times \Theta$ with values in $[0, \infty)$. Such a function ℓ is called a *loss function* if $\ell(\theta, \theta) = 0$ for all $(\theta, \theta) \in \Theta' \times \Theta$, which means that we suffer no loss if we correctly estimate the parameter θ.

In applications, the loss has typically additional structure. In the cases that we consider, ℓ will be a (transformed) semi-metric. Working with metric structure is very natural, since estimators that are in some sense "further away" from the truth should result in larger loss.

The loss quantifies the quality of an estimator for a specific realization of the data. From a statistical point of view, it is more convenient to work with the average loss or *risk*, that is,

$$R(\widehat{\theta}, \theta) := E_\theta\big[\ell(\widehat{\theta}, \theta)\big],$$

where E_θ denotes the expectation with respect to P_θ. Notice that changing the loss function might result in completely different risk.

The statistical risk averages the loss over the data distribution. One can construct statistical models, where specific realizations of the data are much more informative about the underlying parameter. An example is given in Exercise ???. In such a case the statistical risk is the wrong concept.

For many nonparametric estimators, the exact risk is unknown and only upper bounds of the risk can be derived. Therefore, we are mainly interested in risk bounds depending on the sample size n. Consider a sequence of models $(P_\theta^n : \theta \in \Theta)$ indexed in $n = 1, 2, \ldots$ In many cases n is the sample size as for instance in the nonparametric regression model. Now, consider an estimator $\widehat{\theta}_n$ in the model $(P_\theta^n : \theta \in \Theta)$. If there exists a finite constant C which is independent of n and a sequence $(\psi_n)_n$ such that

$$\sup_{\theta \in \Theta} R(\widehat{\theta}_n, \theta) \leq C\psi_n, \quad \text{for all } n,$$

then we say that the estimator $\widehat{\theta}_n$ converges with the *rate of convergence* ψ_n.

The rate of convergence might be sometimes misleading. Because of huge constants an estimator can have a fast convergence rate and still perform badly. An example is given in Exercise **??**.

A somewhat dual concept is the concept of *sample complexity*. It is very natural to ask for the minimal required sample size such that the risk of an estimator is smaller than, say δ. This might be the correct formulation if one wants to conduct an experience and aims to achieve a certain precision. As we can typically only get convergence rates, we can also infer only the order of the sample size in terms of δ. From that we can derive an understanding how much more data we need to improve the risk by a given factor but it is not possible to compute the exact sample size.

2.3 Loss functions on function classes

Below we collect some examples of loss functions on function classes of univariate real-valued functions on $\mathcal{X} \subset \mathbb{R}$. One should think of $\mathcal{X}$ as an interval or $\mathcal{X} = \mathbb{R}$.

Squared pointwise loss: $\ell(f, g) = (f(x) - g(x))^2$ for fixed $x \in \mathcal{X}$. This loss measures how close the estimator comes to the true function at the point x. The corresponding risk is called *mean squared error* (MSE)

$$\mathrm{MSE} = \mathrm{MSE}\left(\widehat{f}_n(x)\right) = E_f\left[(\widehat{f}_n(x) - f(x))^2\right].$$

One of the attractive properties of the MSE is the following decomposition

$$\mathrm{MSE}\left(\widehat{f}_n(x)\right) = E_f\left[(\widehat{f}_n(x) - f(x))^2\right] = \mathrm{Bias}^2\left(\widehat{f}_n(x)\right) + \mathrm{Var}\left(\widehat{f}_n(x)\right) \tag{2.3.1}$$

with

$$\begin{aligned}\mathrm{Bias}\left(\widehat{f}_n(x)\right) &:= E_f\left[\widehat{f}_n(x)\right] - f(x)\\ \mathrm{Var}\left(\widehat{f}_n(x)\right) &:= E_f\left[\left(\widehat{f}_n(x) - E_f\widehat{f}_n(x)\right)^2\right].\end{aligned}$$

The identity should be checked as an exercise. It separates the deterministic/systematic error (also called *bias*) from the stochastic error or variance. Below we bound the two terms separately. For the variance, we have to control the stochastic fluctuations of the estimator around its mean and for the bias, we need to apply approximation theory.

The disadvantage of the pointwise loss is that it controls the loss at one point x only. Even if we can derive a uniform risk bound of the form $\sup_{x\in\mathbb{R}} \mathrm{MSE}(\widehat{f}_n(x)) \le \delta$ this does of course not imply that the estimator is $\sqrt{\delta}$-close to the truth for all x simultaneously. In order to make such a statement we would need to take the supremum inside the expectation. For nonparametric estimation problems, $\sup_x E_f[(\widehat{f}_n(x) - f(x))^2]$ is typically of a smaller order in n compared with $E_f[\sup_x(\widehat{f}_n(x) - f(x))^2]$. These risks are therefore not comparable and this motivates to consider the following loss.

Supremum-norm or sup-norm loss: For functions $f, g : \mathcal{X} \to \mathbb{R}$, this loss is defined as

$$\ell(f, g) = \sup_{x\in\mathcal{X}} |f(x) - g(x)| = \|f - g\|_{L^\infty(\mathcal{X})}.$$

In most applications, the space $\mathcal{X}$ is either $[0,1]$ or $\mathbb{R}$. If it is clear which $\mathcal{X}$ is meant, we also write $\|f-g\|_\infty := \|f-g\|_{L^\infty(\mathcal{X})}$. This loss function controls the worst case pointwise distance between the truth and the estimator. The risk belonging to the sup-norm loss is

$$E_f\big[\|\widehat{f}_n - f\|_{L^\infty(\mathcal{X})}\big].$$

Squared L^2-loss: If $f, g : \mathcal{X} \to \mathbb{R}$,

$$\ell(f,g) = \int_{\mathcal{X}} (f(x)-g(x))^2\,dx = \|f-g\|^2_{L^2(\mathcal{X})}.$$

Working with L^2-loss often leads to a particularly simple analysis thanks to the Hilbert space structure of L^2. Compared to the other two loss function, a disadvantage is that a bound with respect to L^2-loss has a less obvious interpretation. The corresponding risk is called *mean integrated squared error*

$$\mathrm{MISE} = E_f\big[\|\widehat{f}_n - f\|^2_{L^2(\mathcal{X})}\big].$$

2.4 0-1 loss

Let n be the sample size and consider a sequence of estimators $\widehat{\theta}_n$. Given a loss function ℓ and a positive sequence $(\delta_n)_n$ we can then define the corresponding 0-1 loss function

$$\widetilde{\ell}(\widehat{\theta}_n, \theta) = \mathbf{1}\big(\ell(\widehat{\theta}_n,\theta) \geq \delta_n\big).$$

Obviously, the loss function can only attain the values zero and one, depending whether the estimator is δ_n-close to θ. This explains the name. If written in this form, the loss function is n dependent. For the risk

$$E_\theta\big[\widetilde{\ell}(\widehat{\theta}_n,\theta)\big] = P_\theta\big(\ell(\widehat{\theta}_n,\theta) \geq \delta_n\big).$$

By Markov's inequality, $E_\theta\big[\widetilde{\ell}(\widehat{\theta}_n,\theta)\big] \leq \delta_n^{-1} E_\theta\big[\ell(\widehat{\theta}_n,\theta)\big]$. This means that the risk with respect to loss $\widetilde{\ell}$ is bounded by $\delta_n^{-1}\times$ the risk with respect to ℓ. Differently speaking, if $E_\theta\big[\ell(\widehat{\theta}_n,\theta)\big]$ tends to zero with a faster rate than δ_n, also $E_\theta\big[\widetilde{\ell}(\widehat{\theta}_n,\theta)\big]$ converges to zero.

The previous inequality shows that for the 0-1 loss function the object of primary interest is to determine δ_n such that the risk converges to zero. In nonparametric problems, we often encounter the phenomenon that there is a phase transition in the sense that there exists constants $0 < c < C$ and a rate r_n such that $P_\theta(\ell(\widehat{\theta}_n,\theta) \geq cr_n) \to 1$ and $P_\theta(\ell(\widehat{\theta}_n,\theta) \geq Cr_n)$ converges extremely fast to zero. It is conceivable but remains open whether this even holds for some c and C that are arbitrary close.

If the probability $P_\theta(\ell(\widehat{\theta}_n,\theta) \geq Cr_n)$ is very small, it means that it is very unlikely to overshoot the loss by more than Cr_n. The speed also allows us to combine different estimators. In practice it is common to try many different methods and tuning parameters and select the estimator that we like best. Suppose that the risk converges with a rate $o(N_n^{-1})$ and assume we

try N_n different estimators $\widehat{\theta}_i$ $i = 1, \dots, N_n$ of which we pick $\widehat{\theta}_{\widehat{i}}$ where $\widehat{i}$ is a random variable in $\{1, \dots, N_n\}$. Then, by the union bound

$$P_{\theta_0}\big(\ell(\widehat{\theta}_{\widehat{i}}, \theta_0) \geq Cr_n\big) \leq \sum_{i=1}^{N_n} P_{\theta_0}\big(\ell(\widehat{\theta}_i, \theta_0) \geq Cr_n\big) = o(1).$$

2.5 * Loss via linear functionals

Suppose that $\Theta \subseteq V$ with V a vector space. Let $\langle \cdot, \cdot \rangle$ be a scalar product on V and $M \subseteq V$ a symmetric set, that is, $M = -M := \{-v : v \in M\}$. This induces the loss function $\ell : V \times \Theta \to [0, \infty)$,

$$\ell(\theta', \theta) = \sup_{a \in M} \langle a, \theta' - \theta \rangle.$$

The symmetry of M ensures that the loss is indeed non-negative. M is also called the set of discriminators.

One advantage is that we can compare two loss functions via their sets of discriminators. If ℓ and $\widetilde{\ell}$ are loss functions of the type above with sets of discriminators M and $\widetilde{M}$, then ℓ is dominated by $\widetilde{\ell}$ if $M \subseteq \widetilde{M}$.

If$M = V$, then $\ell(\theta', \theta) = \infty$ whenever $\theta \neq \theta'$. This means that if the set of discriminators is too large, the loss becomes too strong for statistical inference.

Let $(V, \| \cdot \|)$ be a normed vector space. The dual norm $\| \cdot \|_*$ is defined as $\|v\|_* = \sup_{w:\|w\| \leq 1} \langle w, v \rangle$. The loss above contains therefore as a special case all dual norms

$$\|\theta' - \theta\|_* = \sup_{a:\|a\| \leq 1} \langle a, \theta' - \theta \rangle.$$

2.6 * Model induced loss functions

Suppose there is a function a that assigns to any two probability measures on the same probability space a non-negative value. This induces then a loss function via $\ell(\theta, \theta') := a(P_\theta, P_{\theta'})$. Notice that the loss depends on the model in the sense that for two statistical models and the same function a the loss ℓ might be very different.

These loss functions play an important role and turn out to be natural choices of loss functions.

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and P, Q two probability measures on $(\mathcal{X}, \mathcal{A})$. Suppose that ν is a probability measure dominating P and Q, that is, $\nu(A) = 0$ for $A \in \mathcal{A}$ implies $P(A) = Q(A) = 0$ (as usual, we write this as $P \ll \nu$ and $Q \ll \nu$). Such a ν can always be found, take for instance $\nu = \frac{1}{2}(P + Q)$. Then, the ν-densities $p = \frac{dP}{d\nu}$ and $q = \frac{dQ}{d\nu}$ exist.

Definition 1 (Total variation distance)**.** *The total variation distance between P and Q is*

$$V(P, Q) = \sup_{A \in \mathcal{A}} \big|P(A) - Q(A)\big| = \sup_{A \in \mathcal{A}} \Big| \int_A (p - q)\, d\nu \Big|.$$

Definition 2 (Kullback-Leibler divergence)**.** *The Kullback-Leibler divergence between P and Q is defined as*

$$K(P,Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP, & \text{if } P \ll Q, \\ \infty, & \text{otherwise.} \end{cases}$$

The total variation should not be confused with the variance of a random variable. We make frequently use of the following properties of the total variation distance. Firstly, the total variation takes values in the interval $[0,1]$, that is $0 \leq V(P,Q) \leq 1$. By Scheffé's theorem (cf. Lemma 2.1 in [28]),

$$V(P,Q) = \frac{1}{2}\int |p-q|\, d\nu = 1 - \int \min(p,q)\, d\nu.$$

A consequence is that the total variation is a distance on the probability measures over a measurable space.

The Kullback-Leibler divergence is not a distance. In particular, it is not symmetric. But it is always non-negative. This follows from applying Jensen's inequality to $K(P,Q) = -\int \log(\frac{dQ}{dP})\, dP \geq -\log \int \frac{dQ}{dP}\, dP = 0$. For Gaussian measures, the Kullback-Leibler divergence leads typically to explicit and relatively simple expressions. Moreover, it is additive under independence in the sense that if $(\mathcal{X},\mathcal{A}) = (\mathcal{X}_1,\mathcal{A}_1)\times(\mathcal{X}_2,\mathcal{A}_2)$ and $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, with P_i, Q_i defined on $(\mathcal{X}_i,\mathcal{A}_i)$, then also $K(P,Q) = K(P_1,Q_1) + K(P_2,Q_2)$.

The total variation distance and the Kullback-Leibler are related through Pinsker's inequality,

$$V(P,Q) \leq \sqrt{\frac{K(P,Q)}{2}} \tag{2.6.1}$$

cf. Lemma 2.5 in [28].

2.7 * Which loss function should one pick?

An estimator that gives fast convergence rates with respect to one loss function might still perform badly with respect to another loss. Suppose that we consider a function estimation problem. An estimator achieving fast rates with respect to the sup-norm loss will lie in a small band around the true function. But the reconstruction can be much rougher than the true function, see Figure 8 for an example. Even though such a reconstruction has small risk it might still lead to an incorrect interpretation of the data. If we also consider as a second loss function the sup norm of the derivatives, an estimator that is much rougher than the truth will be ruled out.

Studying an estimation problem under several loss functions provides us in general with a better understanding of the statistical problem and leads to more robust statistical procedures.

Normally the loss is chosen depending on the underlying parameter space. If the parameter space is a function space of locally regular functions, then we should also select a loss functions that controls the pointwise approximation error. We will see some examples in the subsequent chapters.

2.8 Exercises

Ex. 2.1 — Let $\mathcal{F}$ denote the parameter space in (1.3.1). According to Section 2.1, the nonparametric regression model can be written as collection of probability measures $(P_f : f \in \mathcal{F})$. Determine P_f.

Ex. 2.2 — Suppose we observe $X_1, \dots, X_n \sim F$ with $F \in \mathcal{F}$ an unknown c.d.f. Rewrite this as statistical model of the form $(P_F : F \in \mathcal{F})$ and determine P_F.

Ex. 2.3 — Prove (2.3.1).

Ex. 2.4 — [difficult !!!] Let P_θ be the distribution with $P_\theta(X = \theta+1) = P_\theta(X = \theta-1) = 1/2$. Suppose we observe two independent copies $X_1, X_2 \sim P_\theta$. Show that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} E_\theta \left[|\widehat{\theta} - \theta| \right] \geq \frac{1}{2},$$

where the infimum is taken over all estimators. Show that there is an estimator $\widehat{\theta}$ such that $P_\theta(\widehat{\theta} = \theta) = 3/4$.
This shows a drawback of the concept of estimation risk. The risk is the expectation over the dataset but this is misleading as in practice, we only observe one dataset. With probability 75% the dataset in the setting above will allow us to reconstruct the true parameter perfectly.

Ex. 2.5 — Suppose we observe a sample of n i.i.d. binomial random variables,

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} \mathrm{Bin}(m, p).$$

Both parameters m and p are unknown and we are interested in estimating m. A possible estimator for m is the sample maximum $\widehat{m} = \max_{i=1,\dots,n} X_i$. If m and p are fixed, this estimator converges exponentially fast to m as $n \to \infty$ since

$$\mathbb{P}\big(\widehat{m} = m\big) = 1 - \mathbb{P}\big(\widehat{m} < m\big) = 1 - \mathbb{P}(X_1 < m)^n = 1 - (1 - \mathbb{P}(X_1 = m))^n = 1 - (1 - p^m)^n.$$

Exponential convergence indicates that the estimator should approach the true value extremely fast. Although the convergence rate is exponential, the sample maximum estimator $\widehat{m}$ does not perform well in practice if p is small. Consider for instance the setting $p = 0.1$ and $m = 10$. The probability that for sample size $n = 10^9$ (one billion !) the sample maximum coincides with the true value is still less than 10%.

Part I

Fundamental principles of estimation

Chapter 3

Kernel density estimation

3.1 Introduction

Recall that in the density estimation model, we observe $X_1, \dots, X_n \sim f$ i.i.d. for some unknown Lebesgue density $f \in \mathcal{F}$. In this chapter, kernel density estimators are introduced and discussed.

Figure 5 is a good start to understand the difficulty of the problem. The statistical task is to reconstruct the density in the right panel from the observed sample displayed in the left panel. Clearly, the relative frequency of observations within a small interval corresponds to the height of the density in this interval. Differently speaking, if U is a neighborhood of a point x_0, we can approximate

$$f(x_0) \approx \frac{1}{|U|} \int_U f(u)du \approx \frac{\#\{i : X_i \in U\}}{n|U|}, \tag{3.1.1}$$

where $|U|$ denotes the Lebesgue measure of the set U. The right hand side can be computed from the data. All approaches to density estimation are variations of this very elementary fact.

3.2 Histogram estimator

For a real number a and a binwidth $h > 0$ consider the number of observations in the interval $(a + kh, a + (k + 1)h]$ of length h, that is,

$$N_{kha} := \#\big\{i : X_i \in (a + kh, a + (k + 1)h]\big\}, \quad k \in \mathbb{Z},$$

where $\#$ denotes the cardinality (or number of elements) in the set. Notice that at most n of the N_{kha}'s are non-zero. The histogram estimator $\widehat{f}^{\text{hist}}_{nha}$ is then defined as

$$\widehat{f}^{\text{hist}}_{nha} = \frac{1}{nh} \sum_{k=-\infty}^{\infty} N_{kha} \mathbf{1}\big(\cdot \in (a + kh, a + (k + 1)h]\big). \tag{3.2.1}$$

This estimator coincides with (3.1.1) for U the interval $(a + kh, a + (k + 1)h]$ containing x_0. Convergence rates for the histogram estimator are derived in Exercise 3.4 and Exercise 3.5. The

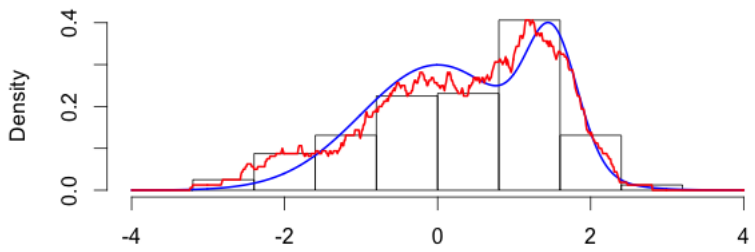


Figure 8: True density (blue), histogram with parameters $a = 0, h = 0.8$ (black), and kernel density estimator with rectangular kernel and $h = 0.4$ (red).

histogram can also be viewed as the MLE with parameter space consisting of piecewise constant densities on the intervals $(a + kh, a + (k + 1)h]$, see Exercise 3.6.

The parameter a in the histogram seems to be very arbitrary and there is often no natural way where to include the jumps. Notice that if we add an integer multiple of h to a, the estimator remains the same. To avoid choosing $a \in [0, h)$, we can consider the average $\widehat{f}_{nh} = h^{-1} \int_0^h \widehat{f}^{\text{hist}}_{nha}\, da$. It can be shown that

$$\widehat{f}_{nh} = \frac{1}{nh} \sum_{i=1}^{n} \Big(1 - \frac{|X_i - x|}{h}\Big)_+ \tag{3.2.2}$$

where $(y)_+ := \max(y, 0)$. This motivates kernel density estimators.

3.3 Smoothing and the kernel density estimator

Using $U = [x - h, x + h]$ as a neighborhood of x in (3.1.1), we obtain

$$f(x) \approx \frac{\#\{i : X_i \in (x - h, x + h]\}}{2nh} = \frac{1}{nh} \sum_{i=1}^{n} K_0\Big(\frac{X_i - x}{h}\Big) \tag{3.3.1}$$

with $K_0 = \frac{1}{2}\mathbf{1}(\cdot \in (-1, 1])$. The expression on the r.h.s. is an estimator of $f(x)$ since it only depends on the data and not on the unknown density f anymore. To introduce a generic function K_0 has teh advantage that we can write several estimators in this form such as for instance (3.2.2) with $K(x) = (1 - |x|)_+$.

Recall that for a density $\int f = 1$ and this should (approximately) also hold for the estimated density. If we replace K_0 by any other function $K : \mathbb{R} \to \mathbb{R}$ integrating to one, that is, $\int_{-\infty}^{\infty} K(u)\, du = 1$, also the right hand side in (3.3.1) integrates to one. This motivates the following definition.

Definition 3. *A kernel is a Lebesgue integrable function $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{-\infty}^{\infty} K(u)\,du = 1$. Moreover, for any $h > 0$,*

$$\widehat{f}_n(x) := \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) \tag{3.3.2}$$

is called kernel density estimator with kernel K and bandwidth h.

The kernel density estimator depends therefore on the choice of the kernel K and the bandwidth h. The bandwidth choice determines how local the estimator is. Taking the bandwidth to zero, we recover the bad estimator $\widehat{F}_n'$. For h large we start to average everything out and the first approximation in (3.3.1) becomes very imprecise.

Although we can take of course any function K that integrates to one, there are kernels that are very popular or which have specific properties. Here are some examples. We refer to the kernel $K_0 = \frac{1}{2}\mathbf{1}(\cdot \in (-1,1])$ considered above as *rectangular kernel* and to the kernel $K(u) = (1 - |u|)_+$ as *triangular kernel.* The kernel $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}(\cdot \in (-1,1])$ is called *parabolic or Epanechnikov kernel* and $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$ is the *Gaussian kernel.*

The kernel density estimator with rectangular kernel is not the same as the histogram. The difference is plotted in Figure 8. The number of jumps in the histogram is determined by the binwidth, whereas the kernel density estmator with rectangular kernel has almost surely $2n$ jumps with jump size $1/(2nh)$ (cf. Exercise 3.9).

Kernel density estimators still have a connection to the maximum likelihood principle, see Exercise 3.14.

3.4 Reduction of the MSE to bias and variance

In this section we study risk bound for the kernel density estimator $\widehat{f}_n$ defined in (3.3.2). We study the loss under the mean squared error and use the decomposition (2.3.1).

Lemma 1 (Variance bound)**.** *For any $h > 0$,*

$$\operatorname{Var}\big(\widehat{f}_n(x)\big) \le \frac{\|f\|_{L^\infty(\mathbb{R})} \|K\|^2_{L^2(\mathbb{R})}}{nh}.$$

Proof. Using that the X_i are independent, $\operatorname{Var}(Y) = E[Y^2] - E^2[Y] \le E[Y^2]$ for any random variable Y, and substituting $v = (u-x)/h$ gives

$$\begin{aligned}
\operatorname{Var}\big(\widehat{f}_n(x)\big) &= \frac{1}{(nh)^2} \sum_{i=1}^{n} \operatorname{Var}\Big(K\Big(\frac{X_i - x}{h}\Big)\Big) \le \frac{1}{nh^2} E\Big[K^2\Big(\frac{X_i - x}{h}\Big)\Big] \\
&= \frac{1}{nh^2} \int_{-\infty}^{\infty} K^2\Big(\frac{u-x}{h}\Big) f(u)\,du = \frac{1}{nh} \int_{-\infty}^{\infty} K^2(v) f(vh + x)\,dv \\
&\le \frac{\|f\|_{L^\infty(\mathbb{R})} \|K\|^2_{L^2(\mathbb{R})}}{nh}.
\end{aligned}$$

□

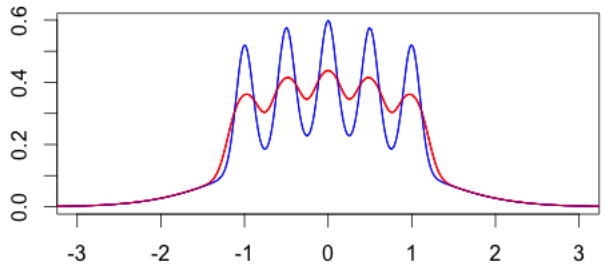


Figure 9: The plot displays the density f (blue) and the approximation $h^{-1}\int_{-\infty}^{\infty} K((u-\cdot)/h)f(u)\,du$ for K the rectangular kernel and $h = 0.2$ (red). The bias of the kernel density estimator is the difference between the two curves. The magnitude of teh bias heavily depends on the local fluctuation behavior of the density. In particular, the bias is small in regions where the density is smooth.

After having derived a bound on the variance, we will now study the bias term in the MSE. Using $\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^n K(\frac{X_i - x}{h})$, the bias has the more explicit representation

$$\text{Bias}\left(\widehat{f}_n(x)\right) = E_f\left[\widehat{f}_n(x)\right] - f(x) = \frac{1}{h}\int_{-\infty}^{\infty} K\Big(\frac{u-x}{h}\Big) f(u)\,du - f(x) =: \Delta_{h,K}(f). \tag{3.4.1}$$

The first term on the r.h.s. is a locally smoothed version of f. The size of the bias is therefore highly dependent on the allowed amount of local fluctuation of f. This, however, depends on the underlying parameter space. If $\mathcal{F}$ consists of very smooth functions only, the bias will be small. On the contrary, if $\mathcal{F}$ contains rough functions, the locally smoothed version of f can be far away from f, cf. the illustration in Figure 9.

3.5 Hölder spaces

The bound on (3.4.1) depends therefore on the underlying parameter space. One of the parameter spaces is the space of Lipschitz functions that we already used as a motivation in Chapter 1. All differentiable functions with bounded derivative are Lipschitz, but the opposite is not true. Indeed the space of Lipschitz functions is slightly larger than the space of differentiable function. For an example consider the function $f(x) = |x|$ which is Lipschitz but not differentiable. Whereas the notion of Lipschitz function is closely tied to one derivative, we now introduce a family of function spaces, which can be viewed as the space of β-times differentiable functions. Here, $\beta > 0$ does not necessarily need to be an integer.

We write $\lfloor \beta \rfloor$ for the largest integer that is strictly smaller than β. For instance, $\lfloor 3 \rfloor = 2$ and $\lfloor 3/2 \rfloor = 1$.

Definition 4 (Hölder space)**.** *Let $\beta > 0$ and $D \subset \mathbb{R}$. For $f : D \to \mathbb{R}$, define the Hölder seminorm*

$$|f|_{\mathcal{C}^\beta} := \sup_{x \neq y \; x,y \in D} \frac{|f^{\lfloor \beta \rfloor}(x) - f^{\lfloor \beta \rfloor}(y)|}{|x - y|^{\beta - \lfloor \beta \rfloor}}$$

and the Hölder norm

$$\|f\|_{\mathcal{C}^\beta} := \sum_{r=0}^{\lfloor \beta \rfloor} \|f^{(r)}\|_{L^\infty(D)} + |f|_{\mathcal{C}^\beta}.$$

The Hölder space with index β is then

$$\mathcal{C}^\beta := \big\{ f : D \to \mathbb{R} \text{ with } \|f\|_{\mathcal{C}^\beta} < \infty \big\}.$$

For each index $\beta > 0$, we obtain therefore a space of "β-smooth" functions on the domain D. The index β is also called *smoothness index*. It is not difficult to check that $|\cdot|_{\mathcal{C}^\beta}$ and $\|\cdot\|_{\mathcal{C}^\beta}$ are indeed (semi)norms on $\mathcal{C}^\beta$. Observe that for $\beta = 1$, the Hölder seminorm coincides with the Lipschitz constant.

Hölder spaces appear in different disciplines in mathematics and there are variations in the definition/notation. Our definition coincides with the space $\mathcal{C}^{\beta - \lfloor \beta \rfloor, \lfloor \beta \rfloor}(D)$ from PDE theory, cf. [9]. In statistics, the Hölder norm is often defined as $\|f\|_{L^\infty(D)} + \|f^{(\lfloor \beta \rfloor)}\|_{L^\infty(D)} + |f|_{\mathcal{C}^\beta}$, that is, only the sup-norm of f and $f^{\lfloor \beta \rfloor}$ are taken into account, see for instance [30], p.213. For convenience we omit the dependence on the domain D in the notation.

Hölder spaces control the local fluctuation of the function over different scales and are therefore well-suited as parameter spaces for risk bounds with respect to pointwise loss. For the theory, it would be enough to impose the Hölder property in a neighborhood of x. For statistics it is convenient to work with the *Hölder ball*, which for $L > 0$ is defined as

$$\mathcal{C}^\beta(L) = \big\{ f : D \to \mathbb{R} \text{ with } \|f\|_{\mathcal{C}^\beta} \leq L \big\}.$$

The word 'ball' comes from the geometric interpretation of a ball with radius R as set of all elements with norm bounded by R. As an analogy, the constant L in the definition of Hölder ball is also called the radius. Below we consider Hölder balls as parameter spaces. First we collect some useful facts about Hölder spaces. The proof is left as an exercise.

Lemma 2.

(i) If $f \in \mathcal{C}^\beta(L)$ and $\tau \in \mathbb{R}$, then $f(\cdot - \tau) \in \mathcal{C}^\beta(L)$.
(ii) $K \in \mathcal{C}^\beta(L)$ and $h \leq 1$ imply $h^\beta K(\cdot/h) \in \mathcal{C}^\beta(L)$,
(iii) If $f \in \mathcal{C}^\beta(L)$ with $\beta > 1$ then f is continuously differentiable and $f' \in \mathcal{C}^{\beta-1}(L)$.

For bounds on the bias, we use the following property of $\mathcal{C}^\beta$ functions: For each x in the domain D, the $\lfloor \beta \rfloor$-th order Taylor approximation around x approximates f at any point $y \in D$ with remainder term bounded by $|f|_{\mathcal{C}^\beta} |y - x|^\beta / \lfloor \beta \rfloor !$. Vice versa, if a function f can be approximated

at any point x by a $\lfloor\beta\rfloor$-th order polynomial with remainder term scaling with (distance to $x)^\beta$, then, f is Hölder with smoothness index β. A more intuitive way to interpret Hölder functions is thus to think of them as functions that can be locally well-approximated by polynomials. This characterization is also useful to introduce the Hölder property on abstract metric spaces in order to carry out a similar statistical analysis.

To kill the lower order terms in the Taylor approximation we need the following refinement of kernels.

Definition 5. *Let ℓ be a positive integer. A kernel $K : \mathbb{R} \to \mathbb{R}$ is said to be of order ℓ, if $\int |u^{\ell+1}K(u)|\,du < \infty$ and $\int_{-\infty}^{\infty} u^j K(u)\,du = 0$, $j = 1, \dots, \ell$.*

Because of $|u^j K(u)| \le (1 + |u|^{\ell+1})|K(u)|$ all integrals $\int_{-\infty}^{\infty} u^j K(u)\,du$ in the previous definition exist. The key property are the vanishing moments. This condition states that the kernel K should be orthogonal in $L^2(\mathbb{R})$ to the subspace spanned by $U = \{u, u^2, \dots, u^\ell\}$. For explicit constructions, see Exercise 3.11 and Exercise 3.12.

Now, we have the tools to bound the right hand side in (3.4.1). Notice that by definition of an $\lfloor\beta\rfloor$-th order kernel, $\int_{-\infty}^{\infty} |K(v)v^\beta|\,dv$ is always finite.

Lemma 3. *Given $f \in \mathcal{C}^\beta$ and K a kernel of order $\ell = \lfloor\beta\rfloor$. Then, for any $h > 0$,*

$$\big|\Delta_{h,K}(f)\big| \le h^\beta \frac{|f|_{\mathcal{C}^\beta}}{\ell!} \int_{-\infty}^{\infty} |K(v)v^\beta|\,dv.$$

Proof. Use the notation Substituting $v = (u - x)/h$ in (3.4.1) yields $\Delta_{h,K}(f) = \int_{-\infty}^{\infty} K(v)\big[f(x + vh) - f(x)\big]\,dv$. Since $f \in \mathcal{C}^\beta(L)$, we find by $\ell = \lfloor\beta\rfloor$-th order Taylor expansion,

$$f(x + vh) = f(x) + f'(x)vh + \ldots + \frac{f^{(\ell-1)}(x)}{(\ell-1)!}(vh)^{\ell-1} + \frac{f^{(\ell)}(x + \tau vh)}{\ell!}(vh)^\ell$$

for some $\tau \in [-1, 1]$. If we plug this expansion into the integral above and use that since K is a kernel of order ℓ, $\int u^j K(u)\,du = 0$ for $j = 1, \dots, \ell$, and $|\tau| \le 1$,

$$\begin{aligned}\Big|\frac{1}{h}\int_{-\infty}^{\infty} K\Big(\frac{u-x}{h}\Big) f(u)\,du - f(x)\Big| &= \Big|\int_{-\infty}^{\infty} K(v)\frac{(vh)^\ell}{\ell!}\big[f^{(\ell)}(x + \tau vh) - f^{(\ell)}(x)\big]\,dv\Big| \\ &\le h^\ell \frac{1}{\ell!}\int_{-\infty}^{\infty} \big|K(v)v^\ell\big||\tau vh|^{\beta-\ell}|f|_{\mathcal{C}^\beta}\,dv \\ &\le h^\beta \frac{|f|_{\mathcal{C}^\beta}}{\ell!}\int_{-\infty}^{\infty} \big|K(v)v^\beta\big|\,dv.\end{aligned}$$

□

3.6 The MSE for kernel density estimators

For density estimation, the true function f is a density, that is, f is non-negative and integrates to one. The parameter space is therefore the intersection of $\mathcal{C}^\beta(L)$-functions and densities

$$\mathcal{F}^\beta(L) := \{f : f \geq 0, \int_{-\infty}^{\infty} f(u)\,du = 1\} \cap \mathcal{C}^\beta(L). \tag{3.6.1}$$

Theorem 1. *Work in the nonparametric density estimation model. Let $\beta > 0$ and L a positive constant. Consider the kernel density estimator $\widehat{f}_n(x)$ for a kernel K of order $\lfloor\beta\rfloor$ that satisfies $\|K\|_2 < \infty$. If $h = \alpha n^{-\frac{1}{2\beta+1}}$, then,*

$$\sup_{f\in\mathcal{F}^\beta(L)} \sup_{x\in\mathbb{R}} \mathrm{MSE}\left(\widehat{f}_n(x)\right) \leq C n^{-\frac{2\beta}{2\beta+1}},$$

for some constant C which does not depend on n.

Proof. Combining (2.3.1), Lemma 1 and Lemma 3, we see that there is a constant C which does not depend on n, such that for any $h > 0$,

$$\sup_{f\in\mathcal{F}^\beta(L)} \sup_{x\in\mathbb{R}} \mathrm{MSE}\left(\widehat{f}_n(x)\right) \leq \sup_{f\in\mathcal{F}^\beta(L)} \sup_{x\in\mathbb{R}} \mathrm{Bias}^2\left(\widehat{f}_n(x)\right) + \mathrm{Var}\left(\widehat{f}_n(x)\right) \leq C\Big(h^{2\beta} + \frac{1}{nh}\Big).$$

The specific choice $h = \alpha n^{-\frac{1}{2\beta+1}}$ yields the result. □

Kernel density estimators can be easily implemented and do not involve difficult optimization steps. The main challenge is to find a good value of the bandwidth h. The theoretical bandwidth choice $h = \alpha n^{-\frac{1}{2\beta+1}}$ in the previous theorem balances the squared bias and the variance in the MSE in an optimal way but depends on the underlying smoothness of the true density which often is unknown in applications. We address this issue later.

The MSE of the estimator $\widehat{f}_n(x)$ has therefore the rate of convergence $n^{-\frac{2\beta}{2\beta+1}}$. It describes the approximation quality in dependence on the sample size n. Obviously, the MSE converges to zero for increasing sample size whenever $\beta > 0$. This matches our intuition that for large dataset we should be able to recover the true function value $f(x)$ with arbitrary precision.

The rate of convergence $n^{-\frac{2\beta}{2\beta+1}}$ at which the MSE decays to zero becomes faster for large smoothness index β. This does not come as a surprise since a smooth density fluctuates less and estimation should be easier. In the extreme case $\beta \to \infty$ we approach the rate n^{-1}. Because of the squared loss this is the same rate that a parametric estimator would obtain that is at a distance of order $n^{-1/2}$ away from the truth.

We can use the rate of convergence to compare estimators. Estimators with fast convergence rate are of course preferable. Later we prove that the rate of convergence $n^{-\frac{2\beta}{2\beta+1}}$ cannot be improved by any estimator for squared pointwise loss. The kernel density estimator with properly chosen bandwidth achieves therefore the optimal rate of convergence.

3.7 * Estimation of derivatives

It is of interest to study pointwise estimation of derivatives of the density f, that is , of $f'(x), f''(x), \ldots$ (provided the derivatives exist). A natural question is whether the derivatives of the kernel density estimator converge to the derivatives of f. We will show that under weak assumptions on the kernel K, the answer to this question is positive.

Let r be a positive integer and assume that the kernel $K : \mathbb{R} \to \mathbb{R}$ is r-times continuously differentiable. The r-th derivative of the kernel density estimator is then

$$\widehat{f}_n^{(r)}(x) = \frac{(-1)^r}{nh^{r+1}} \sum_{i=1}^{n} K^{(r)}\Big(\frac{X_i - x}{h}\Big).$$

Decomposing the MSE as in (2.3.1), we have the following bound on the variance.

Lemma 4 (Bound of the variance of $\widehat{f}_n^{(r)}$)**.** *Suppose that $\|f\|_{L^\infty(\mathbb{R})} < \infty$. If $\|K^{(r)}\|_{L^2(\mathbb{R})} < \infty$, then for any $h > 0$,*

$$\operatorname{Var}\big(\widehat{f}_n^{(r)}(x)\big) \le \frac{\|f\|_{L^\infty(\mathbb{R})} \|K^{(r)}\|_{L^2(\mathbb{R})}^2}{nh^{2r+1}}.$$

The proof is the same as for Lemma 1. The bound for the bias is similar to Lemma 3.

Lemma 5 (Bias estimate)**.** *Suppose that $f \in \mathcal{F}^\beta(L)$ with $\beta > r$ and K is a kernel of order $\ell = \lfloor \beta - r \rfloor$ with compact support. Then, for any $h > 0$,*

$$\big|\operatorname{Bias}\big(\widehat{f}_n^{(r)}(x)\big)\big| \le h^{\beta-r} \frac{L}{\ell!} \int_{-\infty}^{\infty} |K(v) v^{\beta-r}|\, dv.$$

Proof. Since K is r-times continuously differentiable with compact support, we can apply r-times the integration by parts formula and obtain for the bias,

$$\begin{aligned}\operatorname{Bias}\big(\widehat{f}_n^{(r)}(x)\big) &= \frac{(-1)^r}{h^{r+1}} \int K^{(r)}\Big(\frac{u-x}{h}\Big) f(u)\, du - f^{(r)}(x) \\ &= \frac{1}{h} \int K\Big(\frac{u-x}{h}\Big) f^{(r)}(u)\, du - f^{(r)}(x).\end{aligned}$$

The remaining proof follows by the same arguments as the proof of Lemma 3 since $f^{(r)} \in \mathcal{C}^{\beta-r}(L)$ by Lemma 2 (iii). □

Thus, under the joint assumptions of Lemma 4 and Lemma 5, there is a constant not depending on n or h, such that

$$\operatorname{MSE}\big(\widehat{f}_n^{(r)}(x)\big) \le Ch^{2\beta-2r} + \frac{C}{nh^{2r+1}}.$$

The bandwidth choice $h = \alpha n^{-1/(2\beta+1)}$ gives the rate of convergence

$$\operatorname{MSE}\big(\widehat{f}_n^{(r)}(x)\big) \lesssim n^{-\frac{2\beta-2r}{2\beta+1}}. \tag{3.7.1}$$

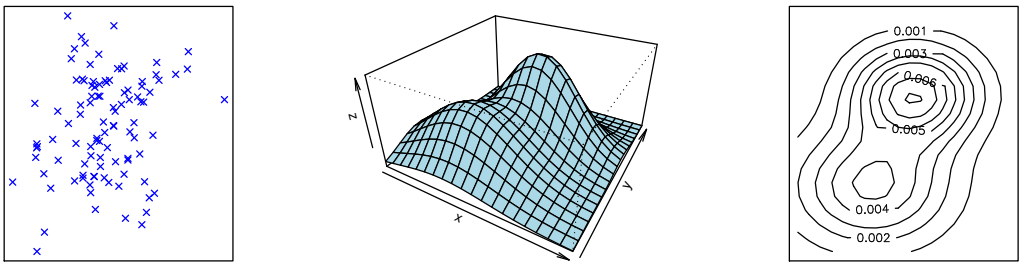


Figure 10: Bivariate density estimation. Raw data (left), perspective plot (middle) and contour plot (right) of a bivariate density

Here $\lesssim$ means that the r.h.s. is an upper bound up to a constant factor. Observe that the convergence rate of the MSE deteriorates if we estimate higher derivatives. We will see later, that $n^{-\frac{2\beta-2r}{2\beta+1}}$ is the fastest rate that can be obtained. It is also interesting to notice that the optimal order $n^{-1/(2\beta+1)}$ of the bandwidth does not depend on r. If we pick the bandwidth of this order, the kernel density estimator converges therefore with the optimal rate of convergence to $f(x)$ but at the same time also the derivatives converge with the fastest possible convergence rate.

An application: Exponential deconvolution

Consider the deconvolution model (1.3.6) with exponential errors, that is, $f_\xi(x) = e^{-x}\mathbf{1}(x \geq 0)$. Assume that $f' \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$. By Lemma 14 and Lemma 13,

$$g(t) = \int_0^\infty f(t-x)e^{-x}dx = f(t) + \int_0^\infty f'(t-x)e^{-x}dx = f(t) + g'(t).$$

Therefore, $f(t) = g(t) - g'(t)$. This motivates the estimator that replaces g and g' by the corresponding kernel density estimators

$$\widehat{f}_n(t) = \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{Y_i - t}{h}\Big) + \frac{1}{nh^2}\sum_{i=1}^n K'\Big(\frac{Y_i - t}{h}\Big)$$

for bandwidth $h > 0$. The estimator for the density f can therefore be seen as an estimator for a combination of g and g'.

To derive the rate of convergence, we assume that $f \in \mathcal{C}^\beta(L)$. The observations $(Y_i)_i$ come, however, from the density g. We need therefore the Hölder index of g. Integrating the inversion formula above gives $g(t) = G(t) - F(t)$. The right hand side can also be written as $\int_0^\infty F(t-x)e^{-x}dx - F(t)$. With $f \in \mathcal{C}^\beta(L)$ it follows that $g \in \mathcal{C}^{\beta+1}(2L)$.

The rate of convergence for estimation of the derivative of a $\beta+1$-smooth density is therefore driving the rate of convergence for f. By (3.7.1) the rate of convergence for the MSE is hence $n^{-2\beta/(2\beta+3)}$ provided that the bandwidth has been chosen of the order $h \asymp n^{1/(2\beta+3)}$.

3.8 Estimation of a multivariate density

Another interesting generalization is to estimate multivariate densities. For simplicity, we consider the bivariate case where $f : \mathbb{R}^2 \to [0, \infty)$. A random element drawn from this density is a random vector (X, Y) in $\mathbb{R}^2$. The statistical task is to estimate f from observing n i.i.d. copies of (X, Y). An example is provided in Figure 10.

In this case we have to impose smoothness constraints on the bivariate density $(x, y) \mapsto f(x, y)$. One way to do that is to assume that for any fixed $x, y \in \mathbb{R}$, the univariate functions $f(x, \cdot)$ and $f(\cdot, y)$ are in the Hölder ball $\mathcal{F}^\beta(L)$.

As estimator, we can take a product kernel and estimate $f(x, y)$ by

$$\widehat{f}_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) K\Big(\frac{Y_i - y}{h}\Big). \tag{3.8.1}$$

Instead of a formal proof, we give a heuristic argument for the rate of convergence. Recall that the sample size is n. The number of observations that fall into a neighborhood of radius h around (x, y) is therefore of the order nh^2. By the i.i.d. assumption, the variance is thus of order $1/(nh^2)$. The bias has order h^β, which is the same as in the univariate case. Therefore, $\mathrm{MSE}(x, y) \lesssim h^{2\beta} + 1/(nh^2)$. The rate in n is optimized for the bandwidth $h = \alpha n^{-1/(2\beta+2)}$ and

$$\mathrm{MSE}(\widehat{f}_n(x, y)) \lesssim n^{-\frac{2\beta}{2\beta+2}}.$$

A similar argument can be employed for d-dimensional density estimation. In this case the rate of convergence of the kernel density estimator with bandwidth $h = \alpha n^{-1/(2\beta+d)}$ is

$$\mathrm{MSE}\left(\widehat{f}_n(x, y)\right) \lesssim n^{-\frac{2\beta}{2\beta+d}}.$$

The rate of convergence becomes worse as the dimension d increases. This effect is often referred to as *curse of dimensionality.*

3.9 Bandwidth selection by cross-validation

Bandwidth selection is a delicate problem. If we chose the bandwidth too small the reconstruction becomes highly oscillating and unstable. In this case, we say that the kernel density estimator *undersmoothes*. The other extreme is *oversmoothing* where we pick a bandwidth that is too large. In this case, we get a very slowly varying reconstruction smoothing out relevant features (see also Figure 11).

In Section **??**, we considered optimization of the bandwidth using the theoretical value of the first order approximation of the MSE. In this section we study a similar procedure. Instead of estimating the unknown quantities in the first order approximation, we discuss a general idea to estimate the risk. The advantage is that this formula does not require the restrictive conditions which are necessary to establish Theorem **??**.

Instead of the MSE, we consider in this section the mean integrated squared error (MISE) corresponding to $L^2(\mathbb{R})$-loss. Let $\widehat{f}_{nh}$ denote the kernel density estimator (3.3.2). We add the

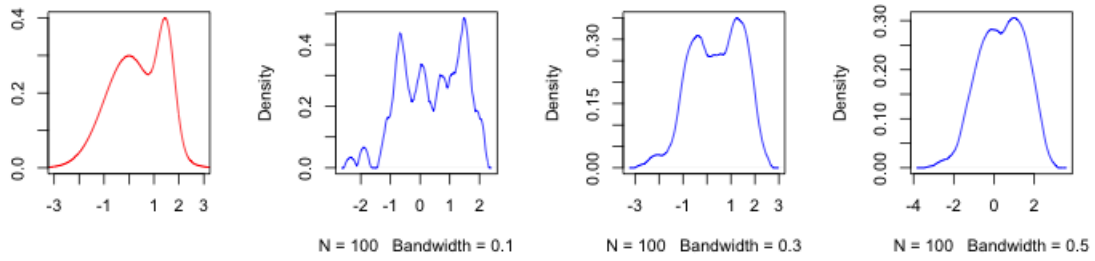


Figure 11: True density (left) and kernel density estimates for sample size 100, Epanechnikov kernel, and bandwidths $h \in \{0.1, 0.3, 0.5\}$ (left to right). The reconstruction for $h = 0.1$ shows typical undersmoothing behavior. The reconstruction with bandwidth $h = 0.5$ oversmoothes.

index h to make the dependence on the bandwidth explicit. The MISE is then

$$\begin{aligned}\text{MISE} &= E_f\Big[\int \big(\widehat{f}_{nh}(x) - f(x)\big)^2\,dx\Big] \\ &= E_f\Big[\int \big(\widehat{f}_{nh}(x)\big)^2\,dx - 2\int \widehat{f}_{nh}(x) f(x)\,dx + \int \big(f(x)\big)^2\,dx\Big]\end{aligned}$$

The last term does not depend on the bandwidth. It is therefore enough to estimate the functional

$$J(h) = E_f\Big[\int \big(\widehat{f}_{nh}(x)\big)^2\,dx - 2\int \widehat{f}_{nh}(x) f(x)\,dx\Big].$$

The first term in the expectation does not depend on f and can be directly computed from the data. In order to estimate the second term, define the kernel density estimator

$$\widehat{f}_{nh,-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\Big(\frac{X_j - x}{h}\Big),$$

where we leave the i-th observation out. The sample size is then $n-1$ and this explains the scaling. Consider now

$$\widehat{G} = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{nh,-i}(X_i).$$

Lemma 6.

$$E_f[\widehat{G}] = E_f\Big[\int \widehat{f}_{nh}(x) f(x)\,dx\Big].$$

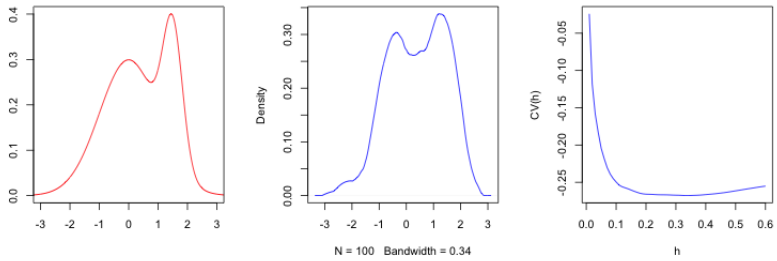


Figure 12: True density (left), kernel density estimate based on bandwidth selection using cross-validation (middle) and $CV(h)$ (right).

Proof.

$$\begin{aligned} E_f[\widehat{G}] &= E_f\big[\widehat{f}_{nh,-1}(X_1)\big] \\ &= E_f\Big[E_f\Big[\frac{1}{(n-1)h}\sum_{j\neq i} K\Big(\frac{X_j - X_1}{h}\Big)\Big|X_1\Big]\Big] \\ &= E_f\Big[\frac{1}{h}\int K\Big(\frac{u-X_1}{h}\Big) f(u)\,du\Big] \\ &= \int \frac{1}{h}\int K\Big(\frac{u-x}{h}\Big) f(u)\,du f(x)\,dx. \end{aligned}$$

On the other hand, by Fubini theorem

$$\begin{aligned} E_f\big[\int \widehat{f}_{nh}(x) f(x)\,dx\big] &= \int E_f\big[\widehat{f}_{nh}(x)\big] f(x)\,dx \\ &= \int \frac{1}{h}\int K\Big(\frac{u-x}{h}\Big) f(u)\,du f(x)\,dx \\ &= E_f[\widehat{G}]. \end{aligned}$$

□

This suggests to estimate $J(h)$ by

$$\mathrm{CV}(h) = \int \big(\widehat{f}_{nh}(x)\big)^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{nh,-i}(X_i).$$

Here, CV stands for *cross-validation*. Cross-validation is a general principle in statistics where we split the data and use one part for estimation and the other part for prediction. In our case, we have applied a version of *leave-one-out cross-validation*. The bandwidth choice is now a minimizer of $\mathrm{CV}(h)$, that is,

$$\widehat{h} \in \mathrm{argmin}_{h>0}\,\mathrm{CV}(h).$$

Kernel density estimation with this data-driven bandwidth choice works reasonably well. For a simulation example see Figure 12. The only drawback is that it tends to undersmooth the density. Many modifications and refinements of this method exist.

3.10 Exercises

Ex. 3.1 — Consider nonparametric density estimation with random sample size. The model is $N \sim \text{Poisson}(n)$ and given N, we observe $X_1, \dots, X_N \sim f$, i.i.d. Show that $Y_t = \sum_{i=1}^{N} \mathbf{1}(X_i \leq t)$, $t \in \mathbb{R}$, is a Poisson process on $\mathbb{R}$ with time-varying intensity $t \mapsto nf(t)$.

Ex. 3.2 — Check the following properties of the kernel density estimator $\widehat{f}_n$.
(i) $\int_{-\infty}^{\infty} \widehat{f}_n(x)\,dx = 1$
(ii) if K is a kernel of order ℓ, then $\int_{-\infty}^{\infty} x^\ell \widehat{f}_n(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} X_i^\ell$.

Ex. 3.3 — Consider the kernel density estimator $\widehat{f}_{nh}(x)$. For a sequence of probability measures $(\Pi_n)_n$ on $(0, \infty)$, we now define a new density estimator

$$\widehat{f}_n(x) = \int_0^\infty \widehat{f}_{nh}(x) d\Pi_n(h).$$

This estimator takes a weighted average over all bandwidths. Show that if $d\Pi_n(h) = a_n^2 h e^{-a_n h} dh$ with $a_n = n^{1/(2\beta+1)}$, then, under the conditions of Theorem 1, $\widehat{f}_n(x)$ has convergence rate $n^{-2\beta/(2\beta+1)}$ for MSE loss. By taking weighted averages, it seems that one can circumvent the bandwidth but a new hyperparameter occurs that depends - if chosen optimally - again on β.

Ex. 3.4 — Let $h^* > 0$. Consider the histogram estimator with $h = h^*$ and the kernel density estimator $\widehat{f}_n$ with rectangular kernel and bandwidth $h = h^*/2$. Prove that for any $k \in \mathbb{Z}$,

$$\widehat{f}^{\text{hist}}_{nh^*a}\big(a + \big(k + \tfrac{1}{2}\big)h^*\big) = \widehat{f}_n\big(a + \big(k + \tfrac{1}{2}\big)h^*\big),$$

with a the constant in the construction of the histogram (cf. also Figure 8).

Ex. 3.5 — Use the previous exercise and Theorem 1 to show that for the histogram estimator with $h = \alpha n^{-\frac{1}{2\beta+1}}$ and $\beta \leq 1$,

$$\sup_{f \in \mathcal{F}^\beta(L)} \sup_{x \in \mathbb{R}} \text{MSE}\left(\widehat{f}^{\text{hist}}_{nha}(x)\right) \leq C n^{-\frac{2\beta}{2\beta+1}},$$

for some constant C which does not depend on n.

Ex. 3.6 — Show that the histogram estimator (3.2.1) is the MLE if the parameter space consists of all piecewise constant densities on the intervals $(a + kh, a + k(h+1)]$.

Ex. 3.7 — Prove (**??**).

Ex. 3.8 — Determine the order of the kernels $K_0 = \frac{1}{2}\mathbf{1}(\cdot \in (-1, 1])$, $K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}(\cdot \in (-1, 1])$ and $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$.

Ex. 3.9 — Show that the kernel density estimator with rectangular kernel jumps almost surely $2n$ times. The total variation distance is therefore almost surely $1/h$ and this tends to infinity if $h \to 0$. The kernel density reconstruction with rectangular kernel is therefore much rougher than the true density. This might be misleading in application, since the reconstruction will show many local features such as local extrema which are pure artifacts.

Ex. 3.10 — Prove Lemma 2.

Ex. 3.11 — Denote by $P(u)$ the $L^2[0,1]$-projection of the function $\mathbf{1}(u \in [0,1])$ on the linear subspace spanned by $\{u, \dots, u^\ell\}$. Show that

$$K(u) = \frac{1 - P(u)}{\int (1 - P(v))\, dv}$$

is an ℓ-th order kernel.

Ex. 3.12 — Define the (weighted) Legendre polynomials as $\phi_0(x) = 2^{-1/2}$ and

$$\phi_m(x) = \sqrt{m + \tfrac{1}{2}} \frac{1}{2^m m!} \frac{d^m}{(dx)^m}(x^2 - 1)^m, \quad m = 1, 2, \dots,$$

for $|x| \le 1$ and $\phi_m(x) = 0$ for $|x| > 1$. These functions are orthonormal in $L^2(\mathbb{R})$. This means that $\int \phi_m(x)\phi_n(x)dx$ equals zero if $m \neq n$ and equals one if $m = n$. Show that $K(u) = \sum_{m=0}^{\ell} \phi_m(0)\phi_m(u)$ is a kernel of order ℓ.

* **Ex. 3.13 —** Consider the functions $f_\alpha : (0,1] \to \mathbb{R}$, $f_\alpha(x) = x^\alpha$ with $\alpha \in \mathbb{R}$. For any α, determine the largest Hölder index β, such that f_α lies in the Hölder space $\mathcal{C}^\beta$.

Ex. 3.14 — [Connection of kernel density estimation and MLE]

(i) Let $a_1, \dots, a_n$ be non-negative numbers. Show that the function

$$f(x) = \max_i \left(a_i - \frac{|X_i - x|}{h} \right)_+$$

is the unique pointwise minimizer over all non-negative functions g with $|g|_{\mathcal{C}^1} \le 1/h$ and $g(X_i) = a_i$.

(ii) Consider nonparametric density estimation for a sample $X_1, \dots, X_n \in \{r_1, \dots, r_k\}$ for some real numbers $r_1, \dots, r_k$ and some positive integer k. Show that if $|r_i - r_j| \ge 2h$ for all $i \neq j$, then, the MLE over all densities f with $|f|_{\mathcal{C}^1} \le 1/h$ is given by the kernel density estimator with triangular kernel. In particular, the result holds if $|X_i - X_j| \ge 2h$ for all $i \neq j$.

Hint: Apply the arithmetic mean - geometric mean inequality to

$$\prod_{i=1}^{n} \frac{f(X_i)}{\#\{j : X_j = X_i\}}$$

(iii) Show that $|r_i - r_j| \ge 2h$ for all $i \neq j$ is also a necessary condition.

Chapter 4

Nonparametric regression

In the previous chapter, we discussed kernel density estimators. In a first part of this section, we transfer the idea of kernel smoothing to the nonparametric regression model introduced in Section 1.3. Many of the ideas for kernel density estimation carry over immediately, but the nonparametric regression model leads to more technical proofs because of the discrete observation scheme.

4.1 Nonparametric regression with uniform random design

Consider the nonparametric regression model with uniform random design. Our sample consists of n i.i.d. random vectors (U_i, Y_i) $i = 1, \dots, n$ with $U_i \sim \mathrm{Unif}[0,1]$ and

$$Y_i = f(U_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,1), \text{i.i.d.}, \quad i = 1, \dots, n.$$

Given a kernel function $K : \mathbb{R} \to \mathbb{R}$, we can now define the estimator,

$$\widetilde{f}_{nh}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\Big(\frac{U_i - x}{h}\Big). \tag{4.1.1}$$

Notice the slightly different form, if compared with the kernel density estimator (3.3.2). Recall that the MSE can be decomposed in bias and variance. For the bias, conditioning on U_i and using that $E[Y_i|U_i] = f(U_i)$ yields

$$\begin{aligned}
\text{Bias}\left[\widetilde{f}_{nh}(x)\right] &= \frac{1}{nh} \sum_{i=1}^{n} E\Big[E\Big[Y_i K\Big(\frac{U_i - x}{h}\Big)\Big|U_i\Big]\Big] - f(x) \\
&= \frac{1}{nh} \sum_{i=1}^{n} \int_0^1 K\Big(\frac{u - x}{h}\Big) f(u) du - f(x) \\
&= \frac{1}{h} \int_0^1 K\Big(\frac{u - x}{h}\Big) f(u) du - f(x).
\end{aligned}$$

Except for the different integration domain, this is the same as the bias for kernel density estimators derived in (3.4.1). For the variance, we use that (U_i, Y_i) are independent,

$$\begin{aligned}\operatorname{Var}\left(\widetilde{f}_{nh}(x)\right) &= \frac{1}{nh^2}\operatorname{Var}\Big(Y_1 K\Big(\frac{U_1 - x}{h}\Big)\Big) \\ &\le \frac{1}{nh^2} E\Big[E[Y_1^2|U_1]K^2\Big(\frac{U_1 - x}{h}\Big)\Big] \\ &= \frac{1}{nh^2} E\Big[(1 + f^2(U_1))K^2\Big(\frac{U_1 - x}{h}\Big)\Big] \\ &\le \frac{(1 + \|f\|_\infty^2)\|K\|_2^2}{nh}.\end{aligned}$$

Theorem 2. *Work in the nonparametric regression model with uniform design. Let $\beta > 0$ and L be a positive constant. Consider the kernel estimator $\widetilde{f}_n(x)$ for a kernel K of order $\lfloor\beta\rfloor$ that satisfies $\|K\|_2^2 < \infty$. If $h = \alpha n^{-\frac{1}{2\beta+1}}$, then, for any fixed $x \in (0,1)$,*

$$\sup_{f \in \mathcal{C}^\beta(L)} \operatorname{MSE}\left(\widetilde{f}_n(x)\right) \le C n^{-\frac{2\beta}{2\beta+1}},$$

for some constant C which does not depend on n.

The theorem can be proved along the lines of Theorem 1. Details are left as an exercise.

4.2 Nonparametric regression with arbitrary random design

We might ask now, whether the estimator $\widetilde{f}_n$ defined in (4.1.1) is still a good estimator if instead of the uniform design, we consider the random design regression model (1.3.2), where we observe i.i.d. (X_i, Y_i), $i = 1, \dots, n$, with

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

The answer is no. To see that the estimator is heavily biased, we denote by p the marginal density of the X_i and follow essentially the same heuristic as for the bias above. This yields,

$$\begin{aligned}E\big[\widetilde{f}_{nh}(x)\big] &= E\Big[E\big[\widetilde{f}_{nh}(x)\big|X_1, \dots, X_n\big]\Big] = E\Big[\frac{1}{nh}\sum_{i=1}^n f(X_i)K\Big(\frac{X_i - x}{h}\Big)\Big] \\ &= \frac{1}{h}\int f(u)K\Big(\frac{u - x}{h}\Big)p(u)\,du \approx f(x)p(x).\end{aligned}$$

The estimator $\widetilde{f}_n$ estimates therefore $f(x)p(x)$ in the nonparametric regression model with random design. This only gives the right answer in the case that $p(x) = 1$ which would be assured for instance if the design points X_i are drawn from a uniform distribution on $[0, 1]$. If the density p is unknown, we could simply divide by it and obtain an estimator for the regression function at x. From an applied point of view, this is, however, unrealistic. A better method is to use the kernel density estimator (3.3.2) to estimate $p(x)$ from $X_1, \dots, X_n$.

Definition 6. *Denote by* $\widetilde{f}_{nh}(x)$ *and* $\widehat{p}_n(x)$ *the estimator* (4.1.1) *and the kernel density estimator of* $p(x)$*, respectively. The Nadaraya-Watson estimator with kernel* $K : \mathbb{R} \to \mathbb{R}$ *and bandwidth* $h > 0$ *is defined as*

$$\widehat{f}_n(x) = \frac{\widetilde{f}_{nh}(x)}{\widehat{p}_n(x)} = \frac{\sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} \tag{4.2.1}$$

if $\widehat{p}_n(x) \neq 0$ *and* $\widehat{f}_n(x) = 0$ *otherwise.*

This estimator can be generalized further to so called locally polynomial estimators. To define this class, rewrite the ℓ-th order Taylor approximation

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \ldots + \frac{f^{(\ell)}(x_0)}{\ell!}(x - x_0)^\ell$$

Definition 7. *Let* $K : \mathbb{R} \to \mathbb{R}$ *a kernel,* $h > 0$ *a bandwidth parameter and* $\ell = 0, 1, 2, \ldots$ *Consider a vector* $\widehat{\theta}_n = \widehat{\theta}_n(x) \in \mathbb{R}^{\ell+1}$ *satisfying*

$$\widehat{\theta}_n \in \operatorname{argmin}_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n \left[Y_i - \sum_{j=0}^{\ell} \theta_j (X_i - x)^j\right]^2 K\left(\frac{X_i - x}{h}\right)$$

Taking the first component of the vector $\widehat{\theta}_n$ *gives an estimator for* $f(x)$ *and this estimator is called locally polynomial estimator of order* ℓ *of* $f(x)$*.*

The idea is that we fit in a weighted least squares sense the best polynomial of order ℓ to the data.

4.3 Exercises

Ex. 4.1 — Prove Theorem 2.

Ex. 4.2 — Verify that the Nadaraya-Watson estimator with non-negative kernel coincides with the locally polynomial estimator of order zero.

Chapter 5

Function estimation in the Gaussian white noise model

5.1 Equivalence of the Gaussian white noise model and the sequence space model

Let $(W_t)_{t\in[0,1]}$ be a Brownian motion and recall the definition and properties of the integral $\int \phi(t)\,dW_t$ in Section 13.2. The Gaussian white noise model was introduced in Section 1.3. Recall that in this model, we observe the path of the process $(Z_t)_{t\in[0,1]}$ with

$$Z_t = \int_0^t f(u)\,du + \frac{1}{\sqrt{n}} W_t, \quad t \in [0,1], \tag{5.1.1}$$

where f is the unknown regression function and W is a Brownian motion. Throughout the following we always assume that the parameter space is a subspace of $L^2[0,1]$, implying that $f \in L^2[0,1]$. The motivation of this model is its interpretation as a continuous version of the nonparametric regression model with uniform fixed design. Indeed we will see that the Gaussian white noise model leads to an elegant theory avoiding discretization effects that occur in the regression model. The practical use is rather limited but includes the important problem of image denoising, see Section 5.6 below.

This model is equivalent to observing

$$\int_0^1 \phi(t)\,dZ_t = \int_0^1 \phi(t) f(t)\,dt + \frac{1}{\sqrt{n}} \int_0^1 \phi(t)\,dW_t, \tag{5.1.2}$$

for all simple functions ϕ, that is, ϕ is a linear combination of indicator functions. Since $f \in L^2[0,1]$, we can take limits and find that given (5.1.1), we also can observe (5.1.2) for all $\phi \in L^2[0,1]$. Here $\int_0^1 \phi(t)\,dW_t$ has to be understood as Wiener integral.

The Gaussian white noise model is therefore often written in the differential form

$$dZ_t = f(t)dt + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0,1], \tag{5.1.3}$$

which just means that we have access to all scalar products $\int \phi(t)\, dZ_t$ with $\phi \in L^2[0,1]$. Since the indicator functions $\phi_s(t) = \mathbf{1}(t \le s)$ are in $L^2[0,1]$ we can recover from the differential form the integral form (5.1.1) of the Gaussian white noise model, proving that the differential and integral form of the Gaussian white noise model are equivalent.

With these properties of the Wiener integral, another equivalent representation of the Gaussian white noise model can be derived. Let $(\phi_k)_{k=1,2,\dots}$ be an orthonormal basis of $L^2[0,1]$, then,

$$\varepsilon_k := \int \phi_k(t)\, dW_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad k = 1, 2, \dots$$

and therefore, the Gaussian white noise model implies that we can observe the random variables $(Z_k)_k$ with

$$Z_k = f_k + \frac{1}{\sqrt{n}}\varepsilon_k, \quad \varepsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad k = 1, 2, \dots \tag{5.1.4}$$

and $f_k := \int_0^1 \phi_k(t) f(t)\, dt$. The scalar product f_k is the k-th basis coefficient of f with respect to the basis $(\phi_k)_k$. In the statistics literature, the f_k's are sometimes referred to as Fourier coefficient even if $(\phi_k)_k$ is not the Fourier basis. Model (5.1.4) is the so called *sequence model*. It says that we can observe the Fourier coefficients of the regression functions subject to a Gaussian error with variance $1/n$. Since

$$f = \sum_{k=1}^{\infty} f_k \phi_k, \quad \text{with convergence in } L^2[0,1], \tag{5.1.5}$$

this gives us a method to reconstruct f from $(Z_k)_k$.

From the sequence model, we can also recover the Gaussian white noise model. Since for any $\phi \in L^2[0,1]$, there exist c_k with $\phi = \sum_{k=1}^{\infty} c_k \phi_k$, we must have $\int \phi(t)\, dZ_t = \sum_{k=1}^{\infty} c_k Z_k$. This shows that sequence model and Gaussian white noise model are equivalent. Nevertheless, the sequence model depends on the choice of an orthonormal basis of $L^2[0,1]$.

5.2 Estimation in the sequence model

What is a good estimator of the regression function f based on the sequence model? A naive estimation strategy would be to use the basis expansion (5.1.5) and to estimate each f_k by Z_k. This would then give the estimator

$$\widehat{f} = \sum_{k=1}^{\infty} Z_k \phi_k \tag{5.2.1}$$

which turns out not to be well-defined since $(Z_k)_k \notin \ell^2$, that is, $\sum_k Z_k^2 = \infty$, a.s. (this should be checked as an exercise). In order to better understand the problem, we could ask which basis coefficients carry relevant information about the signal f? The answer is, that smoothness of the signal can be typically expressed as decay of the Fourier coefficients in the sense that a very smooth functions corresponds to a rapidly decreasing sequence $(f_k)_k$ of basis coefficients. Consequently,

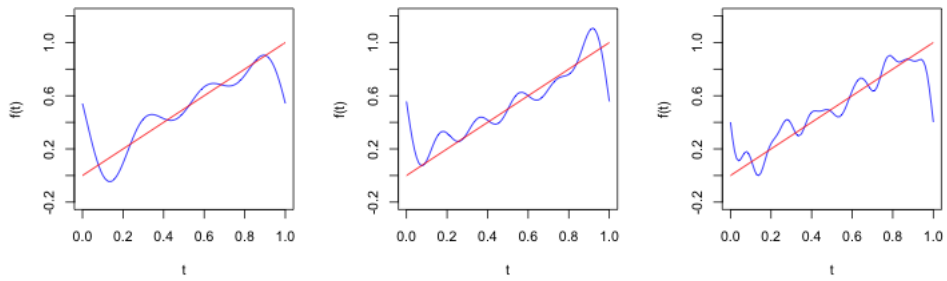


Figure 13: True regression function (red) and reconstructions (blue) based on the Fourier series estimator and cut-off $M_n = 3, 5, 10$ (from left to right). Sample size $n = 2000$.

only the first few coefficients are relevant and the rate at which the difference $f - \sum_{k=1}^{M_n} f_k \phi_k$ converges to zero, as $M_n \to \infty$, depends on the regularity of f. The idea is therefore, to modify the naive estimator (5.2.1) by introducing a cut-off.

Definition 8. *Let $(M_n)_n$ be a sequence of positive integers. The (Fourier) series estimator with cut-off level M_n is*

$$\widehat{f}_n = \sum_{k=1}^{M_n} Z_k \phi_k. \tag{5.2.2}$$

Notice that $E\widehat{f}_n(t) = \sum_{k=1}^{M_n} f_k \phi_k(t)$, $\mathrm{Bias}(\widehat{f}_n(t)) = -\sum_{k=M_n+1}^{\infty} f_k \phi_k(t)$, and $\mathrm{Var}(\widehat{f}_n(t)) = \sum_{k=1}^{M_n} \mathrm{Var}(Z_k)\phi_k^2(t) = \sum_{k=1}^{M_n} \phi_k^2(t)/n$. Using (2.3.1) and that $(\phi_k)_k$ is an orthonormal basis, the mean integrated squared error (MISE) for this estimator can therefore be computed explicitly

$$\begin{aligned} \mathrm{MISE}\left(\widehat{f}_n\right) &= \int_0^1 \mathrm{MSE}\left(\widehat{f}_n(t)\right) dt \\ &= \int_0^1 \left(\mathrm{Bias}^2\left(\widehat{f}_n(t)\right) + \mathrm{Var}\left(\widehat{f}_n(t)\right)\right) dt \\ &= \sum_{k=M_n+1}^{\infty} f_k^2 + \frac{M_n}{n}. \end{aligned} \tag{5.2.3}$$

What we see from this formula is that there is again a bias-variance tradeoff. Making M_n large, the variance becomes big and for small M_n we have a large bias. To find a good balance between these two types of errors requires knowledge of the decay of the basis coefficients $(f_k)_k$.

Although the theoretical analysis of the series estimator is quite straightforward, these estimators have many drawbacks in practice. Some of them are displayed in the reconstructions in Figure 13. For this plot, we have used the sequence model with true regression function $f(x) = x$

and $(\phi_k)_k$ the trigonometric basis

$$\phi_0 := 1, \quad \phi_{2k}(t) = \sqrt{2}\cos(2k\pi t), \quad \phi_{2k+1}(t) = \sqrt{2}\sin(2k\pi t), \quad k = 1, 2, \dots$$

It is well-known that $(\phi_k)_k$ is an orthonormal basis of $L^2[0,1]$, cf. Lemma 16. The reconstructions show that the Fourier series estimator leads to *boundary problems*. Indeed because of the periodicity $\phi_k(0) = \phi_k(1)$ of the basis functions, the Fourier series estimator will also have this property irrespective of what the true regression function f is. The Fourier series estimator does therefore not converge pointwise to the true function at the boundary points. Another drawback are the oscillations that are due to the cut-off in the Fourier domain and clearly visible in the reconstruction. Unless one has some periodic signal, these oscillations might be misleading and are an artifact of the method.

The problem with the Fourier series is that the signal is localized in the frequency domain but not in the coordinate space. This also becomes apparent if one studies the Fourier series of a step functions which leads to highly oscillating behavior at the jump point also known as Gibbs phenomenon.

There are other orthonormal bases of $L^2[0,1]$ which localize signals in the frequency domain and the coordinate space simultaneously. Using these methods, some of the problems of Fourier series estimators can be removed.

5.3 * Boundary correction of series estimators

One possibility to overcome boundary effects is by adding functions to the function system. In the case of the trigonometric basis, we can for instance add the function $g(x) = x$. Since $g(0) \neq g(1)$, this breaks the periodicity.

By adding functions to an orthonormal basis we obtain non-orthogonal function systems and the theory from above does not apply anymore. There are different options. First, we can use Gram-Schmidt orthonormalization to make the functions orthogonal. Gram-Schmidt orthonormalization works as follows. Given a function system $\{\phi_j : j \geq 0\}$, define the functions $\widetilde{\phi}_j$ successively via

$$\widetilde{\phi}_j = \frac{\phi_j - \sum_{\ell<j}\langle \phi_j, \widetilde{\phi}_\ell\rangle \widetilde{\phi}_\ell}{1 - \sum_{\ell<j}\langle \phi_j, \widetilde{\phi}_\ell\rangle^2}.$$

Then, $\{\widetilde{\phi}_j : j \geq 0\}$ forms an orthonormal system and we can apply the theory above.

5.4 Haar wavelet

Let $\phi = \mathbf{1}(\cdot \in [0,1])$ be the indicator function on $[0,1]$ and let $\psi = \mathbf{1}(\cdot \in [0,1/2]) - \mathbf{1}(\cdot \in [1/2,1])$ be the difference between the indicator function on $[0,1/2]$ and $[1/2,1]$. Further define

$$\psi_{j,k} = 2^{j/2}\psi(2^j \cdot - k), \quad j = 0, 1, \dots, \quad k = 0, \dots, 2^j - 1$$

and observe that the support of $\psi_{j,k}$ is $[k/2^j, (k+1)/2^j]$. The functions $\psi_{j,k}$ thus become more and more concentrated for large j, see Figure 14.

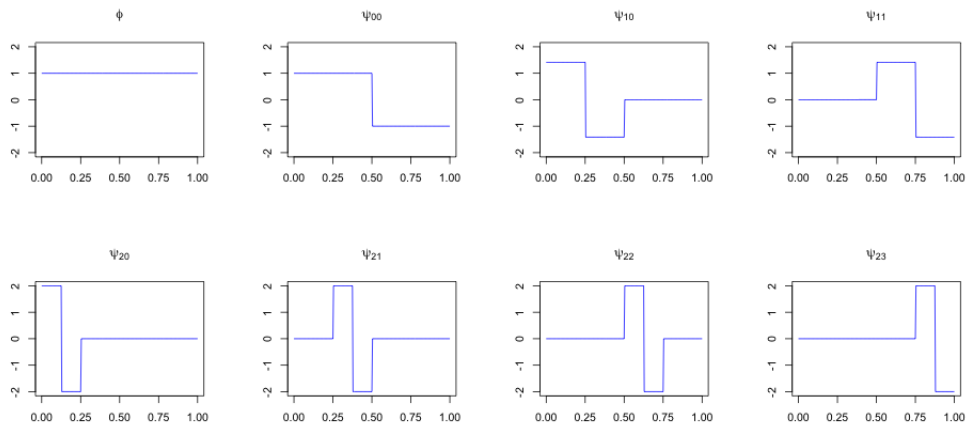


Figure 14: First eight Haar wavelet functions

Definition 9. *The set of functions* $\{\phi\} \cup \{\psi_{j,k} : j = 0, 1, \dots, \ k = 0, \dots, 2^j - 1\}$ *is called the Haar basis or Haar wavelet.*

Theorem 3. *The Haar basis forms an orthonormal basis of* $L^2[0, 1]$.

As a consequence of the theorem we have that

$$f = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}, \quad \text{with } c_0 := \int_0^1 f(u)\,du, \ \ d_{j,k} := \int_0^1 f(u)\psi_{j,k}(u)\,du, \quad (5.4.1)$$

and convergence is in $L^2[0, 1]$. The scalar product c_0 is called *(Haar) scaling coefficient* and we refer to the $(d_{j,k})_{j,k}$ as *(Haar) wavelet coefficients.* In contrast to Fourier series, the Haar wavelet expansion has two indices k and j that induce translation and dilation of the wavelet function. The index j is called *resolution level*.

Proof of Theorem 3. We verify the two conditions in Exercise 11.1. It is easy to see that the basis functions are L^2-normalized. Moreover, $\int \psi_{j,k}(u)\,du = 0$ and therefore ϕ is orthogonal to $\psi_{j,k}$ for all j, k. The support of $\psi_{j,k}$ and $\psi_{j,k'}$ is disjoint if $k \neq k'$ and therefore the basis functions with the same j and different k are all pairwise orthogonal. The scalar product of $\psi_{j,k}$ and $\psi_{j',k'}$ with $j < j'$ must also vanish since the function $\psi_{j,k}$ is constant on the support of the function $\psi_{j',k'}$ and $\int \psi_{j',k'}(u)\,du = 0$. We have therefore proved that the Haar wavelet forms an orthonormal system of functions.

It remains to check the second condition in Exercise 11.1. This means that if $\int_0^1 f(u)\,du = 0$ and $\int_0^1 f(u)\psi_{j,k}(u)\,du = 0$ for all j, k, then $f = 0$.

Let $F(s) = \int_0^s f(u)\,du$. Observe that $\int_0^1 f(u)\,du = 0$ implies $F(1) = 0$. Moreover, $\int_0^1 f(u)\psi_{0,1}(u)\,du = 0$ and $F(1) = 0$ show that $F(1/2) = -F(1/2)$ and thus $F(1/2) = 0$.

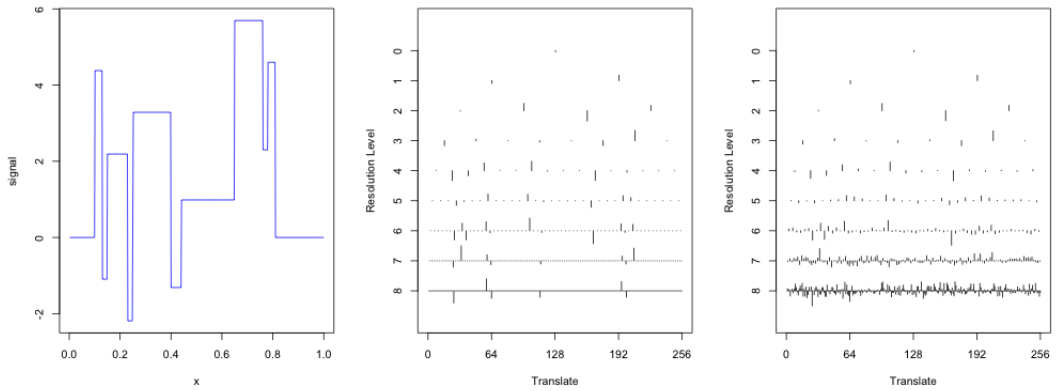


Figure 15: Signal (left), scaled Haar wavelet coefficients (middle) and noisy scaled Haar wavelet coefficients (right). In the two plots on the right, the j-th horizontal line displays the 2^j wavelet coefficients on the j-th resolution level. The wavelet coefficients on the j-th resolution level are multiplied by a factor $2^{j/2}$ to better display the wavelet coefficients on the larger resolution levels. This also explains why the noise becomes more visible for large j.

From $\int_0^1 f(u)\psi_{1,1}(u)\,du = 0$ we then obtain that $F(1/4) = 0$ and from $\int_0^1 f(u)\psi_{1,2}(u)\,du = 0$ that $F(3/4) = 0$. Continuing, we derive $F(k/2^j)$ for all $j = 0, 1, \ldots$ and $k = 0, \ldots, 2^j$.

We apply the Lebesgue differentiation theorem (Theorem 15) for $d = 1$. Thus for a Lebesgue integrable function f, $\lim_{\varepsilon\to 0}(2\varepsilon)^{-1}\int_{x-\varepsilon}^{x+\varepsilon} f(u)\,du$ exists and is equal to $f(x)$, almost everywhere. This shows $f(x) = 0$ almost everywhere which completes the proof. □

As mentioned above, smoothness can be translated into decay conditions of the basis coefficients. The next lemma provides a bound on the decay of Haar wavelet coefficients for Hölder functions.

Lemma 7. *If $f \in \mathcal{C}^\beta(L)$ with $\beta \le 1$, then $|c_0| \le L$ and $|d_{j,k}| \le L2^{-\frac{j}{2}(2\beta+1)}$.*

Proof. $|c_0| \le \|f\|_\infty \le L$ and

$$|d_{j,k}| = \left|2^{\frac{j}{2}}\int_{k/2^j}^{(k+\frac{1}{2})/2^j}\left(f(u) - f(u+2^{-j-1})\right)du\right| \le 2^{j/2}L2^{-\beta(j+1)}2^{-j-1} \le L2^{-\frac{j}{2}(2\beta+1)}.$$

□

The lemma only covers smoothness index up to $\beta = 1$. For $\beta > 1$ one obtains because of the embedding $\mathcal{C}^\beta(L) \subset \mathcal{C}^1(L)$, that $|d_{j,k}| \le L2^{-\frac{3j}{2}}$. Indeed, this bound cannot be improved much. Consider for instance the function $f(x) = x$ that lies in any Hölder space. The wavelet coefficients can be computed explicitly as $d_{j,k} = 2^{-\frac{3j}{2}-2}$. Haar wavelets are therefore not able to capture higher order smoothness in the signal. The situation is comparable with higher order kernels which also capture the right bias only up to a certain upper bound.

Due to the two indices in the definition of the wavelet functions, we need to rewrite the sequence model accordingly. If the Gaussian white noise model (5.1.3) is applied to the Haar basis, we observe Z_0 and $(Z_{j,k})_{j,k}$ with

$$\begin{aligned} Z_0 &= c_0 + \frac{1}{\sqrt{n}}\varepsilon_0 \\ Z_{j,k} &= d_{j,k} + \frac{1}{\sqrt{n}}\varepsilon_{j,k}, \quad j = 0, 1, \ldots;\ k = 0, 1, \ldots, 2^j - 1. \end{aligned}$$

where ε_0 and $(\varepsilon_{j,k})_{j,k}$ are i.i.d. standard normal. As an estimator for f, we can consider the wavelet series truncated at resolution level M_n,

$$\widehat{f}_n = Z_0 + \sum_{j=0}^{M_n} \sum_{k=0}^{2^j-1} Z_{j,k}\psi_{j,k}. \tag{5.4.2}$$

Theorem 4. *Suppose that $\beta \leq 1$ and M_n is chosen such that $2^{M_n} \asymp n^{1/(2\beta+1)}$, then,*

$$\sup_{f \in \mathcal{C}^\beta(L)} \mathrm{MISE}\left(\widehat{f}_n\right) \lesssim n^{-\frac{2\beta}{2\beta+1}}.$$

Proof. Following the same argument as in (5.2.3) and using Lemma 7 (now we have 2^{M_n+1} coefficients that we keep)

$$\begin{aligned} \mathrm{MISE}\left(\widehat{f}_n\right) &= \sum_{j=M_n+1}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k}^2 + \frac{2^{M_n+1}}{n} \\ &\leq L^2 \sum_{j=M_n+1}^{\infty} 2^{-2j\beta} + \frac{2^{M_n+1}}{n} \\ &= L^2 \frac{2^{-2(M_n+1)\beta}}{1 - 2^{-2\beta}} + \frac{2^{M_n+1}}{n} \\ &\lesssim n^{-\frac{2\beta}{2\beta+1}}. \end{aligned}$$

□

5.5 Adaptive wavelet thresholding for Haar wavelet

So far, we have discussed various nonparametric estimators which all require prior knowledge of the smoothness of the true signal in order to choose a suitable bandwidth or truncation level. From a statistical point of view, this is a serious restriction, since the smoothness is typically unknown. In this section, we provide a wavelet estimator *which is independent of* β and gives the good rate $(n/\log n)^{-\beta/(2\beta+1)}$ under supremum norm loss as long as $0 < \beta \leq 1$. Such procedures are called *adaptive* because they adapt to the true smoothness.

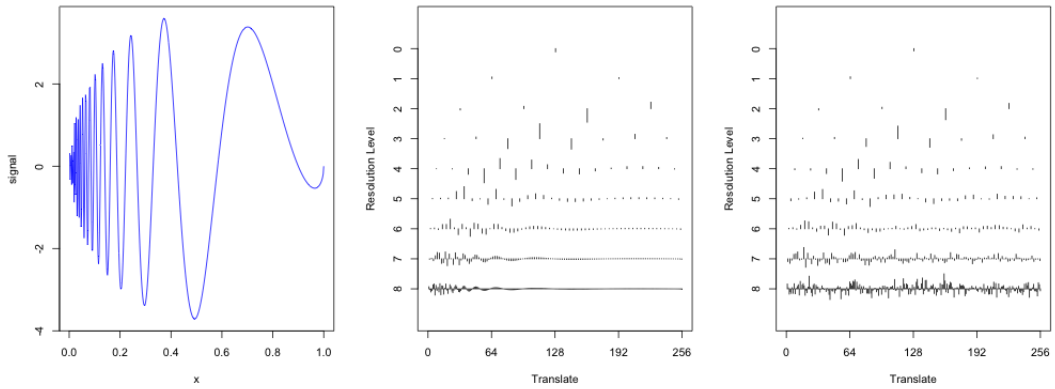


Figure 16: Same as Figure 15 with Doppler signal.

The motivation for the wavelet truncation estimator (5.4.2) was that the wavelet coefficients are small on large resolution levels. Only the first resolution levels contain therefore relevant signal and on large resolution levels the noise dominates. This can be seen for instance in Figure 15 and Figure 16 (to make the wavelet coefficients on high resolution levels better visible, the scaled coefficients $2^{j/2}d_{j,k}$ are displayed). If the smoothness of the signal varies also the wavelet coefficients on one resolution level can be of different magnitude. This can be seen for instance for the Doppler function in Figure 16. The Doppler function oscillates more and more as $x \downarrow 0$. This is reflected in the inhomogeneous size of the wavelet coefficients on the same resolution level. Instead of the estimator (5.4.2) a better idea is therefore to truncate wavelet coefficients individually. We keep the large coefficients since these are the ones that cannot be only due to the noise and which therefore must contain some signal. The small coefficients might just be caused by the measurement noise and are set to zero to reduce the variance of the estimator. The *wavelet thresholding estimator with threshold* $\tau > 0$ is then defined as

$$\widehat{f}_{n,\tau} = Z_0 + \sum_{j=0}^{J} \sum_{k=0}^{2^j-1} Z_{j,k} \mathbf{1}\big(|Z_{j,k}| > \tau\big) \psi_{j,k}. \tag{5.5.1}$$

To compute the estimator it is more convenient to have finitely many terms instead of an infinite sum. Therefore, a maximal resolution level J is introduced. In most applications J is chosen such that $2^J \asymp n$. Wavelet coefficients on resolution levels $j > J$ can typically be shown to be negligible compared with the convergence rate.

For the performance of the estimator, it is important to pick a good threshold value τ. One of the most common choices is the *universal threshold*

$$\tau_U = \sqrt{\frac{2 \log n}{n}}. \tag{5.5.2}$$

Let us briefly motivate this value. Recall that the idea of the threshold is to select empirical wavelet coefficients $Z_{j,k} = d_{j,k} + n^{-1/2}\varepsilon_{j,k}$ for which we can be certain that they contain information

about the signal, that is, $d_{j,k} \neq 0$. This can be phrased as hypotheses test, where a coefficient is selected if the null hypothesis $d_{j,k} \neq 0$ can be rejected with probability $1-\alpha$. Since there are in total $O(n)$ simple tests, this would still lead to selecting an α-fraction of the noisy coefficients which, if α is fixed, does not help to improve the variance of the estimator. A better way is to think of threshold value selection as multiple testing problem. Then, we should choose τ in such a way that the probability of making one false rejection becomes small and we therefore can be certain that all selected coefficients contain signal. We should thus pick τ such that

$$P\big(\max_{j\leq J}\max_{k}|\varepsilon_{j,k}| > \sqrt{n}\tau\big) = \text{"small"}$$

Since the maximum over m i.i.d. standard normal r.v.'s behaves like $\sqrt{2\log m}$ and $\max_{j\leq J}\max_k$ is a maximum over $O(n)$ indices, this motivates the universal threshold (5.5.2).

Now, we are ready to state the main result of this section. To simplify the proof, we work with a multiple of the universal threshold.

Theorem 5. *Consider the wavelet thresholding estimator $\widehat{f}_{n,\tau}$ in (5.5.1) with $\tau = 4\sqrt{\log n/n}$ and maximal resolution level 2^J such that $n/2 < 2^J \leq n$. Then, for any $0 < \beta \leq 1$,*

$$\sup_{f\in\mathcal{C}^\beta(L)} E_f\big[\|\widehat{f}_{n,\tau} - f\|_\infty\big] \lesssim (n/\log n)^{-\beta/(2\beta+1)}.$$

Proof. Let J^* be the smallest integer such that $L2^{-\frac{j}{2}(2\beta+1)} < \tau/2$ for all $j > J^*$. Observe that $L2^{-\frac{J^*+1}{2}(2\beta+1)} < \tau/2 \leq L2^{-\frac{J^*}{2}(2\beta+1)}$ and therefore,

$$2^{J^*} \asymp \tau^{-2/(2\beta+1)} \asymp (n/\log n)^{1/(2\beta+1)}. \tag{5.5.3}$$

By Lemma 7, $|d_{j,k}| \leq L2^{-\frac{j}{2}(2\beta+1)}$. Together with the bound in Exercise 5.1,

$$\begin{aligned}
&\big\|\widehat{f}_{n,\tau} - f\big\|_\infty \qquad (5.5.4)\\
&\leq |Z_0 - c_0| + \sum_{j=0}^{J} 2^{j/2}\max_{k=0,\ldots,2^j-1}\big|Z_{j,k}\mathbf{1}\big(|Z_{j,k}|>\tau\big) - d_{j,k}\big| + \sum_{j=J+1}^{\infty} 2^{j/2}\max_{k=0,\ldots,2^j-1}|d_{j,k}|\\
&= \frac{|\varepsilon_0|}{\sqrt{n}} + \sum_{j=0}^{J} 2^{j/2}\max_k\Big(\frac{|\varepsilon_{j,k}|}{\sqrt{n}}\mathbf{1}\big(|Z_{j,k}|>\tau\big) + |d_{j,k}|\mathbf{1}\big(|Z_{j,k}|\leq\tau\big)\Big) + L\sum_{j=J+1}^{\infty}2^{-j\beta}. \qquad (5.5.5)
\end{aligned}$$

Because of

$$\{|Z_{j,k}|>\tau\} \subset \Big\{|d_{j,k}| \geq \frac{\tau}{2} \cup |\varepsilon_{j,k}| \geq \sqrt{n}\frac{\tau}{2}\Big\} \subset \Big\{j \leq J^* \cup |\varepsilon_{j,k}| \geq \sqrt{n}\frac{\tau}{2}\Big\}$$

and

$$\{|Z_{j,k}|\leq\tau\} \subset \Big\{|d_{j,k}| \leq \frac{3\tau}{2} \cup |\varepsilon_{j,k}| \geq \sqrt{n}\frac{\tau}{2}\Big\}.$$

we can further bound (5.5.5) from above by

$$\|\widehat{f}_{n,\tau} - f\|_\infty \le \frac{|\varepsilon_0|}{\sqrt{n}} + \sum_{j=0}^{J} 2^{j/2} \max_k \Big(\frac{|\varepsilon_{j,k}|}{\sqrt{n}} \mathbf{1}(j \le J^*) + \Big(\frac{|\varepsilon_{j,k}|}{\sqrt{n}} + L \Big) \mathbf{1}\big(|\varepsilon_{j,k}| \ge \sqrt{n}\tau/2\big) + |d_{j,k}| \mathbf{1}\big(|d_{j,k}| \le 3\tau/2\big) \Big) + L \sum_{j=J+1}^{\infty} 2^{-j\beta}. \tag{5.5.6}$$

using also $|d_{j,k}| \le L$. Now, we control the expectation of the terms in the sum on the r.h.s. With Lemma 20 and Jensen's inequality,

$$E\big[\max_{k=0,\dots,2^j-1} |\varepsilon_{j,k}| \big] \le \sqrt{2 \log 2^j} + 1 \le \sqrt{2 \log n} + 1 \le 2\sqrt{\log n},$$

for all $j \le J$ and all $n \ge 2$. With (5.5.3),

$$E\Big[\sum_{j=0}^{J} 2^{j/2} \max_k \frac{|\varepsilon_{j,k}|}{\sqrt{n}} \mathbf{1}(j \le J^*) \Big] \le 2\sqrt{\frac{\log n}{n}} \sum_{j=0}^{J^*} 2^{j/2} = 2\sqrt{\frac{\log n}{n}} \frac{2^{(J^*+1)/2} - 1}{\sqrt{2} - 1} \lesssim \Big(\frac{n}{\log n} \Big)^{-\frac{\beta}{2\beta+1}}.$$

For $\xi \sim \mathcal{N}(0,1)$ it follows that $E[|\xi| \mathbf{1}(|\xi| \ge u)] = \sqrt{2/\pi} \exp(-u^2/2)$, cf. Exercise 14.1. Recall also the Gaussian tail bound in Lemma 19. For the second term in the sum in (5.5.6), we bound the $\max_k \le \sum_k$ and

$$\begin{aligned} &E\Big[\sum_{j=0}^{J} 2^{j/2} \max_k \Big(\frac{|\varepsilon_{j,k}|}{\sqrt{n}} + L \Big) \mathbf{1}\big(|\varepsilon_{j,k}| \ge \sqrt{n}\tau/2\big) \Big] \\ &\le \sum_{j=0}^{J} 2^{3j/2} \Big(\frac{1}{\sqrt{n}} E\big[|\xi| \mathbf{1}(|\xi| \ge 2\sqrt{\log n}) \big] + L P\big(|\xi| \ge 2\sqrt{\log n} \big) \Big) \\ &\le \frac{4}{n^2} \sum_{j=0}^{J} 2^{3j/2} = \frac{4}{n^2} \cdot \frac{2^{3(J+1)/2} - 1}{2^{3/2} - 1} \le \frac{4 \cdot 2^{3/2}}{2^{3/2} - 1} n^{-1/2}. \end{aligned}$$

Finally, for the last two terms in (5.5.6), using (5.5.3) again,

$$\begin{aligned} &\sum_{j=0}^{J} 2^{j/2} \max_k |d_{j,k}| \mathbf{1}\big(|d_{j,k}| \le 3\tau/2\big) + L \sum_{j=J+1}^{\infty} 2^{-j\beta} \\ &\le \frac{3\tau}{2} \sum_{j=0}^{J^*} 2^{j/2} + L \sum_{j=J^*+1}^{\infty} 2^{-j\beta} = \frac{3\tau}{2} \cdot \frac{2^{(J^*+1)/2} - 1}{\sqrt{2} - 1} + L \frac{2^{-(J^*+1)\beta}}{1 - 2^{-\beta}} \lesssim \Big(\frac{n}{\log n} \Big)^{-\frac{\beta}{2\beta+1}}. \end{aligned}$$

Combining the bounds for the individual terms in (5.5.6) gives $E_f[\|\widehat{f}_{n,\tau} - f\|_\infty] \lesssim (n/\log n)^{-\beta/(2\beta+1)} + n^{-1/2} \lesssim (n/\log n)^{-\beta/(2\beta+1)}$ uniformly over $f \in \mathcal{C}^\beta(L)$, which proves the assertion. □

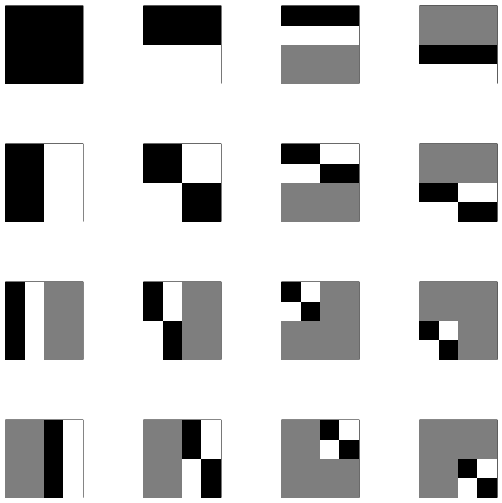


Figure 17: First 2d Haar wavelet basis functions. Representation is color coded (black $\to 1$, grey $\to 0$, white $\to -1$) .

5.6 * Image denoising using the 2d Haar wavelet

In this section, the Haar wavelet is extended to two dimensions. As application, we present an algorithm for image denoising.

In the two-dimensional nonparametric regression model with fixed uniform design, we observe a vector $\mathbf{Y}_n = (Y_{i,j,n})_{i,j=1,\dots,n}$ with

$$Y_{i,j,n} = f\left(\tfrac{i}{n}, \tfrac{j}{n}\right) + \varepsilon_{i,j,n}, \quad \varepsilon_{i,j,n} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

We might interpret the regression function f as the true image with colors converted into numerical values. The observations $Y_{i,j,n}$ are then the noisy pixel values.

The 2d Haar wavelet is the tensor product of the one-dimensional Haar wavelet functions. For simplicity define $\psi_{-1,0} := \phi$ and let $I_1 := \{(-1,0), (j,k), j = 0,1,\dots; k = 0,\dots,2^j - 1\}$. With this notation, the 1d Haar wavelet functions can be written as $\{\psi_{j,k} : (j,k) \in I_1\}$. For the 2d extension, define $I_2 := \{(j,k,j',k') : (j,k), (j',k') \in I_1\}$ and let

$$\psi_{j,k,j',k'}(x,y) := \psi_{j,k}(x)\psi_{j',k'}(y), \quad \text{for all } x,y \in [0,1], \ \text{ for all } (j,k,j',k') \in I_2.$$

Definition 10 (2d Haar wavelet)**.** *The set of functions $\{\psi_\lambda : \lambda \in I_2\}$ is called the 2d Haar basis or the 2d Haar wavelet.*

In the definition, λ stands for an arbitrary index $(j,k,j',k') \in I_2$. The following theorem can be proved similarly as Theorem 3.

Theorem 6. *The 2d Haar basis $\{\psi_\lambda : \lambda \in I_2\}$ is an orthonormal $L^2([0,1]^2)$-basis.*

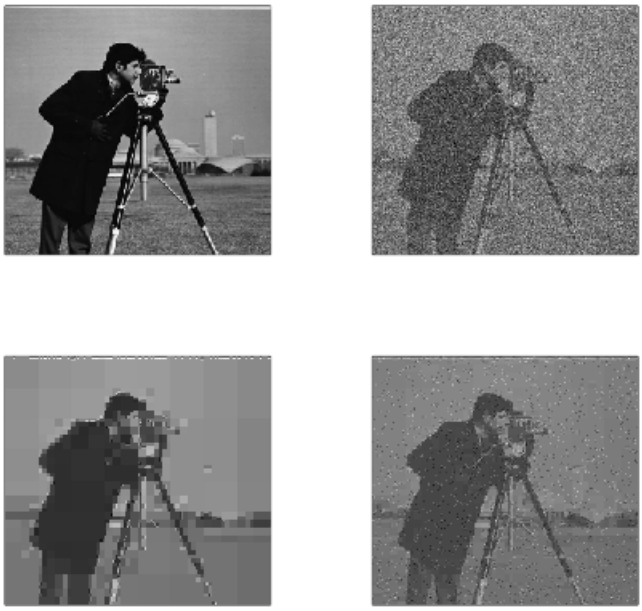


Figure 18: *Top:* True image (left) and observed noisy image (right). *Bottom:* Reconstructed image using hard thresholding with $\tau = \tau_U$ (left) and reconstructed image using hard thresholding with $\tau = \tau_U/2$ (right).

Any function $f \in L^2([0,1]^2)$ can consequently be written as

$$f = \sum_{\lambda \in I_2} d_\lambda \psi_\lambda, \quad \text{with} \quad d_\lambda = \int_0^1 \int_0^1 f(x,y)\psi_\lambda(x,y)\,dx\,dy,$$

where the sum converges in $L^2([0,1]^2)$. The next step is to estimate the wavelet coefficients from the data. The route that we took in the 1d case is to start with the Gaussian white noise model and to project the observed path onto a basis, which gives then the empirical basis coefficients. A similar approach could also be worked out in two dimensions. Alternatively, we can extend the discrete observations to a stochastic process $Y^n = (Y^n_{s,t})_{s,t\in[0,1]}$ via

$$Y^n_{s,t} = \sum_{i,j=1}^{n} Y_{i,j,n} \mathbf{1}\big(s \in \big[\tfrac{i-1}{n}, \tfrac{i}{n}\big), t \in \big[\tfrac{j-1}{n}, \tfrac{j}{n}\big)\big).$$

The basis coefficient estimates are then given by replacing the function f by the process Y^n,

$$\widehat{d}_\lambda = \int_0^1 \int_0^1 Y^n_{x,y}\psi_\lambda(x,y)\,dx\,dy.$$

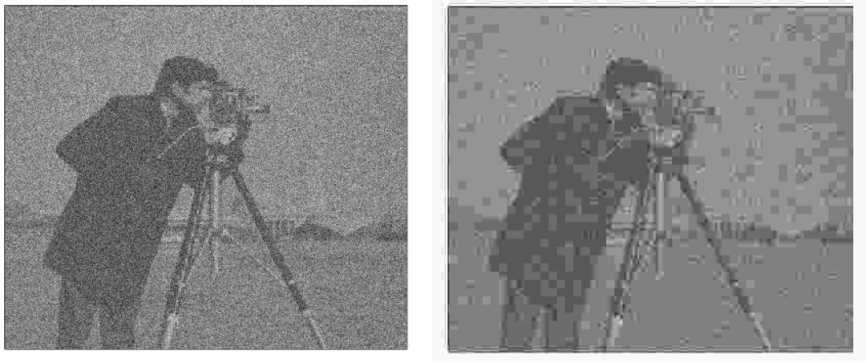


Figure 19: Image compression has a denoising effect. Noisy image (left) and noisy image after JPEG compression with quality parameter 5% (right).

The wavelet thresholding estimator is thus defined as

$$\widehat{f_\tau} = \sum_\lambda \widehat{d}_\lambda \mathbf{1}(|\widehat{d}_\lambda| > \tau)\psi_\lambda.$$

The effective sample size is here the number of pixels or n^2. The universal threshold is therefore

$$\tau_U = \sqrt{2\log(n^2)} = 2\sqrt{\log(n)}.$$

In Figure 18, the hard wavelet thresholding estimator is shown for grayscale images with threshold levels τ_U and $\tau_U/2$. For the universal threshold the method removes the noise in the reconstruction but destroys also several finer details in the image. This is consistent with the choice of the universal threshold which by construction only takes wavelet coefficients into account if we know that they contain relevant signal with high probability.

As a second method, we consider wavelet thresholding with threshold $\tau_U/2$. Due to the smaller threshold value, additional wavelet coefficients are incorporated into the reconstruction compared with the universal threshold. The reconstruction shows that there is still a bit of noise in the image. On the contrary, details such as the face of the cameraman are better preserved.

* Connection to digital image compression

Digital images on computers can be stored quite efficiently using compression algorithms. The different file formats such as .png and .jpg stand for different compression methods. Denoising and compression are closely related if not the same and this will be discussed in more detail in this section.

The JPEG/JPG compression for digital images works similar than the thresholding method. For JPEG compression the image is divided into smaller blocks. On each of the blocks a discrete cosine transform is applied. This is the discrete version of the expansion with respect to the cosine basis. The computed coefficients are then quantized, which has a similar effect as thresholding

and puts many small coefficients to zero. These coefficients are then stored in an efficient way. If a JPEG image is displayed, the transformation is inverted. Because of the small coefficients that are assigned to zero, higher frequencies are cut-off. The reconstructed JPEG image is therefore similar to the series estimator.

In JPEG, there is a quality parameter that determines the amount of compression. For high quality parameters, almost nothing is thresholded and the reconstruction can hardly be distinguished from the original image. If the quality parameter is taken to be low, more of the empirical coefficients are set to zero. The denoising therefore becomes more pronounced. In Figure 19, the JPEG reconstruction of a nosy image is shown with low quality parameter. The JPEG reconstruction is completely denoised. One should also observe that more coefficients are retained in areas with smaller details.

The compression behind the JPEG 2000 file format works similarly as JPEG but uses wavelets instead of the cosine expansion.

5.7 Exercises

Ex. 5.1 — Let $f = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}$ be the Haar wavelet decomposition of f. Show that the uniform norm on $[0,1]$ can be bounded as follows

$$\|f\|_\infty \le |c_0| + \sum_{j=0}^{\infty} 2^{j/2} \max_{k=0,\dots,2^j-1} \big|d_{j,k}\big|.$$

Ex. 5.2 — Consider pointwise estimation in the Gaussian white noise model (5.1.1). For $h > 0$ and a kernel $K \in L^2(\mathbb{R})$ define the kernel smoothing estimator

$$\widehat{f}(x) = \frac{1}{h} \int_0^1 K\Big(\frac{t-x}{h}\Big)\, dZ_t.$$

(i) Show that

$$\operatorname{Bias}\big(\widehat{f}(x)\big) = \frac{1}{h} \int_0^1 K\Big(\frac{t-x}{h}\Big) f(t)\, dt - f(x)$$

and

$$\operatorname{Var}\big(\widehat{f}(x)\big) \le \frac{\int K^2(u)\, du}{nh}.$$

(ii) Let $\beta > 0$. If K is a kernel of order $\ell = \lfloor \beta \rfloor$, conclude that for $x \in (0,1)$,

$$\sup_{f \in C^\beta(L)} E_f\big[\big(\widehat{f}(x) - f(x)\big)^2\big] \lesssim n^{-\beta/(2\beta+1)},$$

provided that the bandwidth $h = h_n$ is chosen such that $h_n \asymp n^{-1/(2\beta+1)}$.

Ex. 5.3 — Work in the Gaussian white noise model (5.1.1). For $0 < \Delta \leq 1$ with $1/\Delta$ a positive integer, define the piecewise constant estimator

$$\widehat{f} = \sum_{j=1}^{1/\Delta} \frac{Z_{\Delta j} - Z_{\Delta(j-1)}}{\Delta} \mathbf{1}\big(\cdot \in [(j-1)\Delta, j\Delta)\big).$$

Denote the space of monotone functions with values between zero and one by $\mathcal{M}(R) := \{f : [0,1] \to [0,1] \ : f \text{ monotone increasing}\}$ and let $\|g\|_1 := \int_0^1 |g(t)|\, dt$. Show that

$$\sup_{f \in \mathcal{M}(R)} E_f[\|\widehat{f} - f\|_1] \leq \Delta + (n\Delta)^{-1/2}$$

and find the value of $\Delta = \Delta_n$ that minimizes the expression.

Ex. 5.4 — Prove Theorem 6.

Chapter 6

Shrinkage in the sequence model

6.1 Introduction

So far we were mainly concerned with upper bounds on the estimation risk. In this chapter, we look at a criterion called *admissibility*. It turns out that already for parametric settings the MLE is not admissible.

Throughout this chapter, we study the following version of the sequence model: Suppose we observe the vector $\mathbf{Y} = (Y_1, \dots, Y_d)$ with

$$Y_i = \theta_i + \sigma\varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, \dots, d \tag{6.1.1}$$

for some fixed $d \in \{1, 2, \dots\}$. Here, σ is known and $\theta = (\theta_1, \dots, \theta_d)$ is the unknown parameter vector. The loss is again the squared ℓ^2-loss, $\ell(\theta, \theta') = \sum_{i=1}^d (\theta_i - \theta_i')^2 = \|\theta - \theta'\|_2^2$. As parameter space, we take $\Theta = \mathbb{R}^d$. Equivalently, we can also state the model as observing $\mathbf{Y} = (Y_1, \dots, Y_d)$ with independent components $Y_i \sim \mathcal{N}(\theta_i, \sigma^2)$.

This model is slightly different compared to the sequence model (5.1.4) in the sense that it is only d-dimensional. It also allows for a more general variance σ^2. In principle, we can always normalize the model dividing the observations Y_i by σ or $\sigma\sqrt{n}$ in order to obtain noise level $\sigma = 1$ or $\sigma = 1/\sqrt{n}$, respectively. The assumption that σ is known is very important. For unknown σ^2 there is no way to distinguish anymore the contribution of the mean vector from the noise. In particular, no consistent estimator for σ^2 exists. In Section 6.5 we provide an examples where the noise variance is known.

The sequence model (5.1.4) comes from nonparametric estimation in the Gaussian white noise model with θ_i being the series coefficients of the unknown regression function. Model (6.1.1) can be interpreted as observing a number of scalars $\theta_1, \dots, \theta_d$ subject to Gaussian noise. Our approach in this chapter is completely non-asymptotic in the sense that none of the quantities will be taken to infinity. This also explains why there is no n in the model. In particular, the dimension is called d and not n.

Since d is finite, estimation of the vector $\theta = (\theta_1, \dots, \theta_d)$ in model (6.1.1) is a parametric problem. The result are however mainly useful for high-dimensional problems that we will study in subsequent chapters.

The simplest version of model (6.1.1) is the case if $d = 1$. In this model, we observe one Gaussian random variable $Y \sim \mathcal{N}(\theta, \sigma^2)$ with unknown mean $\theta \in \Theta = \mathbb{R}$. The "best" estimator for θ is given by the MLE $\widehat{\theta} = Y$. "Best" can here mean many things in particular being minimax. For $d = 2$, the estimator $\widehat{\theta} = \mathbf{Y}$ is also best . For $d > 2$, the situation changes and $\widehat{\theta} = \mathbf{Y}$ is for certain criteria not the optimal estimator anymore.

The chapter is structured as follows. In Section 6.2, we introduce admissibility and discuss its relation to minimax risk.

6.2 Admissibility

Definition 11. *An estimator $\widehat{\theta}$ is called* inadmissible *for estimation of $\theta \in \Theta$ and with respect to the loss ℓ, if there exists another estimator $\widetilde{\theta}$, such that*

$$E_\theta\big[\ell(\widetilde{\theta}, \theta)\big] \leq E_\theta\big[\ell(\widehat{\theta}, \theta)\big], \quad \textit{for all } \theta \in \Theta$$

and

$$E_{\theta^*}\big[\ell(\widetilde{\theta}, \theta^*)\big] < E_{\theta^*}\big[\ell(\widehat{\theta}, \theta^*)\big], \quad \textit{for some } \theta^* \in \Theta.$$

If an estimator is not inadmissible, the estimator is called admissible.

Inadmissibility therefore means that there exists another estimator $\widetilde{\theta}$ with smaller risk. The second condition in the definition requires that there is at least one parameter θ for which the risk is strictly smaller. Inadmissible estimators should therefore be avoided and we should better work with $\widetilde{\theta}$.

Often one can prove that an estimator is inadmissible, but proving admissibility is an extremely difficult problem and only few results are known.

Admissible estimators are not always better. In model (6.1.1) with $d = 1$, we could consider the estimator $\widehat{\theta} = 0$ that irrespectively of the observation always assigns the value zero. This estimator has risk zero for $\theta = 0$. If this estimator would be inadmissible, there would exist an estimator $\widetilde{\theta}$ with smaller risk. In particular, the risk of $\widetilde{\theta}$ at $\theta = 0$ must also be zero which is only possible if $\widetilde{\theta} = 0$. This shows that $\widehat{\theta} = 0$ is an admissible estimator, but it is in general not a good estimator. In particular, admissibility does not imply that an estimator is minimax.

Also the opposite is not true and there are inadmissible minimax estimators. This is plausible since minimax estimators only need to be good in a worst case sense and admissibility requires that an estimator cannot be improved anywhere. The following surprising result shows that even the MLE in parametric models can be inadmissible.

Theorem 7 (Stein's phenomenon)**.** *Work in model (6.1.1) with $d > 2$. Then the estimator $\widehat{\theta} = \mathbf{Y}$ is inadmissible.*

A proof of Theorem 7 is given in Section 6.4. It is based on an explicit construction of an estimator $\widetilde{\theta}$ with smaller risk and application of Definition 11. The estimator with smaller risk is the *James-Stein estimator*

$$\widehat{\theta}_{\mathrm{JS}} = \Big(1 - \frac{\sigma^2(d-2)}{\|\mathbf{Y}\|_2^2}\Big)\mathbf{Y}, \quad \text{for } d > 2$$

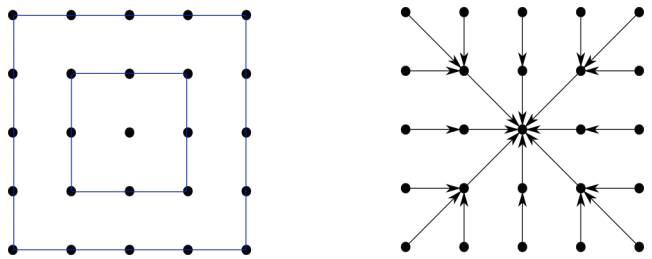


Figure 20: Shrinkage in 2d. The number of grid points at $\|\cdot\|_\infty$-distance $= r$ from the origin grows with r (left). Shrinkage to the closest grid point with smaller $\|\cdot\|_\infty$-norm leads to the effect that several less likely events are added up (right).

with $\|\mathbf{Y}\|_2^2 = \sum_{i=1}^d Y_i^2$.

One should observe that the James-Stein estimator is not a linear estimator because of the scaling factor which depends on $\|\mathbf{Y}\|_2^2$. Since $1 - \sigma^2(d-2)/\|\mathbf{Y}\|_2^2$ is less than one, the estimator shrinks the values $Y_1, \ldots, Y_d$ towards zero. The shrinkage becomes stronger if $\|\mathbf{Y}\|_2^2$ is small. It can also happen that the scaling factor becomes negative. In this case one can also modify the estimator and put the factor to zero. The so modified estimator is also known as *positive part James-Stein estimator.*

Before we give a proof of Theorem 7, we try to provide some intuition why shrinkage appears in higher dimensions.

6.3 Intuition for shrinkage

Intuitively, it is not clear why the James-Stein estimator could be better than the MLE. In particular, one might wonder why this effect kicks-in for $d > 2$ and whether Stein's phenomenon is just a weird side-effect of working with the squared ℓ^2-loss. In this section, we provide some intuition and try to show that inadmissibility of the MLE is a general phenomenon in high-dimensional spaces.

The reason for the inadmissibility of the maximum likelihood estimator in high dimensions has something to do with the "speed" at which balls grow. To illustrate this, we might work with a simplified model, where the parameter set Θ consists of all grid points $(m_1, \ldots, m_d)$ with $m_1, m_2, \ldots, m_d$ integers. To simplify the exposition, we might also consider the componentwise maximum loss $\ell_\infty(\theta, \theta') = \max_i |\theta_i - \theta_i'| = \|\theta - \theta'\|_\infty$. There are $(2r+1)^d - (2r-1)^d = 2d(2r)^{d-1} + O(r^{d-2})$ grid points which are at $\|\cdot\|_\infty$-distance $= r$ from the origin. This means that the number of grid points grows polynomially with the distance r. In Figure 20, this is displayed for dimension $d = 2$.

In this simplified model, consider the maximum likelihood estimator $\widehat{\theta}_{\mathrm{MLE}}$ that estimates the parameter by the most likely grid point. Consider also a shrinkage estimator $\widetilde{\theta}$ that picks (one of) the grid point(s) θ that is at minimal maximum-norm distance to $\widehat{\theta}_{\mathrm{MLE}}$ and satisfies $\|\theta\|_\infty = (\|\widehat{\theta}_{\mathrm{MLE}}\|_\infty - 1)_+$. Obviously, this shrinks the estimate towards zero. Now, we might

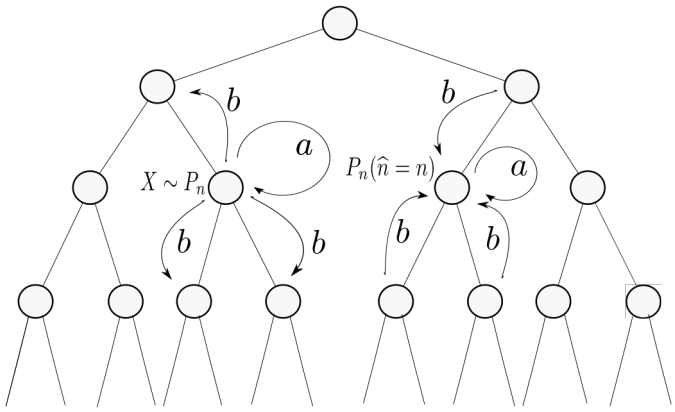


Figure 21: Shrinkage on a tree

wonder which estimator has smaller risk. The shrinkage estimator takes a point that is less likely. On the contrary due to the shrinkage, the grid points at $\|\cdot\|_\infty$-distance $= r$ from the origin are mapped to the grid points at $\|\cdot\|_\infty$-distance $= r - 1$ from the origin. Therefore, several less likely events have to be added up, cf. Figure 20. The number of the events depends on the growth of the number of grid points which are at distance $= r$ from the origin as r grows.

The last argument is a bit 'handwaving'. To make the argument precise we will instead work on a graph where the nodes take the role of the grid points. We consider a graph were except for the so called root node, all nodes have three edges. One edge connects a node with a parent node which is closer to the root node. The remaining two edges connect a node to children nodes which are further away from the root of the tree. In graph theory terminology, we consider a rooted, infinite complete binary tree, cf. Figure 21 for a visualization. On this binary tree we define a random variable that takes values on the nodes of the tree. Let m denote a node, $p(m)$ the parent node and $c_1(m)$ and $c_2(m)$ the two children nodes. Write $X \sim P_m$, if $P_m(X = m) = a$ and $P_m(X = p(m)) = P_m(X = c_1(m)) = P_m(X = c_2(m)) = b$ where a and b are positive numbers such that $a + 3b = 1$ ensuring that P_m is indeed a probability distribution. If m is the root of the tree, we set $P_m(X = m) = a + b$ and $P_m(X = c_1(m)) = P_m(X = c_2(m)) = b$. Assume further that $0 < b < a < 2b$. Suppose, we observe $X \sim P_n$ and want to estimate n with respect to the loss function $\mathbf{1}(m \neq m')$.

This model is very similar to the grid point model before if we link the origin to the root of the tree and the grid points at distance $= r$ from the origin with the nodes at level r (distance from the root) in the tree, cf. Figure 20. In the tree model, the number of nodes on each level doubles and therefore shrinkage should help.

Indeed, since $b < a$ the maximum likelihood estimator in the tree model is $\widehat{m}_{\mathrm{MLE}} = X$ and this estimator has risk $P_m(\widehat{m} \neq m) = 1 - a$. Consider now the estimator $\widetilde{m} = p(X)$ which estimates m by the parent node of X and thus shrinks everything one step towards the root of the tree. If m is the root, we have $P_m(\widetilde{m} \neq m) = 0$ and otherwise the risk is $P_m(\widetilde{m} \neq m) =$

$1 - P_m(p(X) = m) = 1 - P_m(X \in \{c_1(m), c_2(m)\}) = 1 - 2b < 1 - a$ using the condition $a < 2b$. The shrinkage estimator has thus smaller risk.

To understand the model, one can think of each node as a scientist. The root of the tree represents the founding father of a research field. Except for the founder, a scientist has one adviser (parent) and two students (children). Suppose we have an article written by one scientist in the tree and we know the following rule: If a scientist has an publishable idea then with probability a it will be published by the same scientist. The probabilities that the idea is published by the adviser of the scientist or any of the two students are b. Given an article, we want to find out who had the idea originally. Since $a > b$ the intuitive estimator is to guess that the author itself came up with the idea. The analysis tells us, however, that it is more likely that the idea stems from the adviser of the author.

Finally let us comment on the connection between transient random walks and inadmissibility of the MLE. It is well-known that the d-dimensional random walk becomes transient for $d > 2$ (this means a random walk started in the origin returns to a neighborhood of the origin only finitely often) which is exactly the same condition as we have for the inadmissibility of the MLE in Stein's phenomenon in Theorem 7. There is indeed a deeper relation which was worked out in the seminal article [2]. A similar phenomenon can be observed on the tree model. A random walk $(X_i)_{i\geq 0}$ on the nodes with transition probability given by $X_{i+1}|X_i \sim P_{X_i}$ will be transient for all $b > 0$. For the shrinkage estimators we need, however, $2b > a$ and thus there is no perfect one-to-one correspondence. One should notice that transience/recurrence is a general concept, while shrinkage depends on the choice of the loss function.

Inadmissibility and shrinkage depends on the choice of the loss function. Loss functions which are large if the estimator is far away from the truth might require less or no shrinkage. Shrinkage as discussed above improves the original estimator for many realizations of the data but might also lead occasionally to estimates that are quite bad and worse than the MLE. This is not a contradiction to the notion of admissibility since admissibility is stated in terms of the averaged loss or risk. A loss that is large if the estimator is far away from the true parameter gives more weight to those events which makes shrinkage less favorable. As illustration consider the binary tree model above with loss function the length of the (shortest) path between two nodes. Compared to the loss function $\mathbf{1}(m \neq m')$ the graph distance results in higher loss if an estimator is far off. Assume $b < a$. The MLE returns the right node with probability a and otherwise results in loss one. The shrinkage estimator returns the right node with probability $1 - 2b$ and has loss one with probability a and loss two with probability b. The risk for the estimator $\widehat{m}_{\mathrm{MLE}} = X$ therefore is $1 - a$ whereas the risk for the estimator $\widetilde{m} = p(X)$ is $1 - b$. The MLE has then smaller risk and shrinkage does not help.

* Motivation for the correction term in the James-Stein estimator

The previous section shows that shrinkage helps in higher dimensions. In this section we give some motivation for the specific form of the correction term $\sigma^2(d-2)\|\mathbf{Y}\|_2^{-1}\mathbf{Y}$ in the James-Stein estimator.

The loss function is $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$. This means that if θ and θ' are close, then, $\|\theta\|_2^2 \approx \|\theta'\|_2^2$. The MLE $\widehat{\theta}^{\mathrm{MLE}} = \mathbf{Y}$ is unbiased for estimation of θ. It is, however, not unbiased for

estimation of $\|\theta\|_2^2$. Indeed,

$$E_\theta\big[\|\widehat{\theta}^{\mathrm{MLE}}\|_2^2\big] = E_\theta\big[\|\mathbf{Y}\|_2^2\big] = \|\theta\|_2^2 + \sigma^2 d$$

implying that the estimated θ has in expectation a larger $\|\cdot\|_2$-norm than the mean vector θ. The $\|\cdot\|_2$-projection of the MLE on the sphere $\{\theta' : \|\theta'\|_2 = \|\theta\|_2\}$ is $\widetilde{\theta} = (1 - \sigma^2 d/(\|\theta\|_2^2 + \sigma^2 d))\mathbf{Y}$. Not surprisingly, we find the same expression if we search for the estimator in the class $\{t\mathbf{Y} : t \in \mathbb{R}\}$ with smallest $\|\cdot\|_2^2$-risk, see Exercise 6.1. Since θ is unknown, $\widetilde{\theta}$ is not an estimator. With $\|\mathbf{Y}\|_2^2 \approx \|\theta\|_2^2 + \sigma^2 d$, it is therefore natural to consider

$$\Big(1 - \frac{\sigma^2 d}{\|\mathbf{Y}\|_2}\Big)\mathbf{Y},$$

which agrees with the James-Stein estimator up to a factor $1 - 2/d$ in the MLE correction term. For any parameter, this estimator has a strictly larger risk than the James-Stein estimator, see Exercise 6.8.

Again one should notice the important role of the loss played in the argument. A similar reasoning already appeared in Stein's orginial article [25], see also [4].

6.4 Stein's lemma

Suppose that $X \sim \mathcal{N}(0,1)$ and $f'(X)$ is absolutely continuous. If $E[|f'(X)|] < \infty$ then also $E[(1+|X|)|f(X)|] < \infty$, see Exercise 6.2

Lemma 8 (Stein's lemma in one dimension)**.** *If $X \sim \mathcal{N}(0,1)$, f is absolutely continuous and $E[|f'(X)|] < \infty$, then*

$$E\big[Xf(X)\big] = E\big[f'(X)\big].$$

Proof. Let K be a smooth function with support on $[-2,2]$, $K(y) = 1$ for all $y \in [-1,1]$ and $\|K\|_\infty \le 1$, $\|K'\|_\infty < \infty$. Define the function $f_m(x) = f(x)K(x/m)$. For $m \ge 1$, $|f_m(x)| \le |f(x)|$ and $|f_m'(x)| \le |f'(x)| + |f(x)| \cdot \|K'\|_\infty$. Moreover, $f_m'(x) = 0$ for $|x| > m$.

By assumption $\int |f'(x)|e^{-x^2/2}dx < \infty$ and therefore also $\int |xf(x)|e^{-x^2/2}dx$. By Lemma 14,

$$\int x f_m(x) e^{-x^2/2}dx = \int f_m'(x)e^{-x^2/2}dx$$

for any m. Sinc for $m \to \infty$, $f_m(x) \to f(x)$ and $f_m'(x) \to f'(x)$ almost everywhere with respect to the Lebesgue measure, we find using dominated convergence,

$$\int x f(x) e^{-x^2/2}dx = \int f'(x)e^{-x^2/2}dx.$$

□

In order to derive the risk of the James-Stein estimator, we need to consider the multivariate case where $g : \mathbb{R}^d \to \mathbb{R}^d$. Recall that ∂_i denotes the partial derivative with respect to the i-th component. Write $\nabla \cdot g = \sum_{j=1}^d \partial_j g_j$ for the divergence of g and define $\mathbf{x}_{-j} := (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$.

Lemma 9 (Stein's lemma)**.** *Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be as in (6.1.1) and suppose that $g = (g_1, \ldots, g_d) : \mathbb{R}^d \to \mathbb{R}^d$. If for any j, $x_j \mapsto g_j(\mathbf{x})$ is absolutely continuous for all $x_j \in \mathbb{R}$ and almost all $\mathbf{x}_{-j} \in \mathbb{R}^{d-1}$ with respect to the Lebesgue measure and*

$$E\big[|\partial_i g_i(\mathbf{Y})|\big] < \infty, \quad \textit{for all } i = 1, \ldots, d, \tag{6.4.1}$$

then, Stein's identity

$$E\big[(\mathbf{Y} - \theta)^\top g(\mathbf{Y})\big] = \sigma^2 E\big[\nabla \cdot g(\mathbf{Y})\big]. \tag{6.4.2}$$

holds and

$$E\big[\big\|\mathbf{Y} + g(\mathbf{Y}) - \theta\big\|_2^2\big] = \sigma^2 d + E\big[2\sigma^2 \nabla \cdot g(\mathbf{Y}) + \|g(\mathbf{Y})\|_2^2\big]. \tag{6.4.3}$$

Proof. Identity (6.4.3) is a direct consequence of (6.4.2). Define $\mathbf{X} = \sigma^{-1}(\mathbf{Y} - \theta)$ and $\widetilde{g} = g(\sigma \cdot + \theta)$. Observe that $\widetilde{g}(\mathbf{X}) = g(\mathbf{Y})$ and $\partial_i \widetilde{g}_i = \sigma \partial_i g_i(\sigma \cdot + \theta)$. The latter implies $\nabla \cdot \widetilde{g}(\mathbf{X}) = \sigma \nabla \cdot g(\mathbf{Y})$, where on the left hand side of the equation, the derivatives are with respect to $\mathbf{X}$ and on the left hand side with respect to $\mathbf{Y}$. Therefore it is sufficient to show

$$E\big[\mathbf{X}^\top \widetilde{g}(\mathbf{X})\big] = E\big[\nabla \cdot \widetilde{g}(\mathbf{X})\big], \quad \mathbf{X} \sim \mathcal{N}(0, I_d), \tag{6.4.4}$$

where $\mathcal{N}(0, I_d)$ denotes a multivariate normal distribution with the $d \times d$ identity matrix I_d as covariance.

We also have that $E[|\widetilde{\partial}_i g_i(\mathbf{X})|] < \infty$ for all $i = 1, \ldots, d$. With $\mathbf{X}_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$, we also have that

$$E\Big[|\widetilde{\partial}_i g_i(\mathbf{X})| \,\Big|\, \mathbf{X}_{-i}\Big] < \infty$$

for almost all $\mathbf{X}_{-i}$ with respect to the Lebesgue measure on $\mathbb{R}^{d-1}$ and for all $i = 1, \ldots, d$. The bound $E[|\widetilde{\partial}_i g_i(\mathbf{X})| \,|\, \mathbf{X}_{-i}] < \infty$ implies by Lemma 8, that

$$\int_{-\infty}^{\infty} \partial_i g_i(\mathbf{x}) e^{-x_i^2/2} dx_i = \int_{-\infty}^{\infty} x_i g_i(\mathbf{x}) e^{-x_i^2/2}\, dx_i$$

for almost all $\mathbf{x}_{-i}$, which is equivalent to $E[X_i \widetilde{g}(\mathbf{X}) \,|\, \mathbf{X}_{-i}] = E[\partial_i g_i(\mathbf{X}) \,|\, \mathbf{X}_{-i}]$. By taking expectation on both sides and summing over the index i, we finally obtain (6.4.4). This completes the proof. □

To check the conditions of Stein's lemma for the James-Stein estimator we need the following result.

Lemma 10. *If* $\mathbf{Y} = (Y_1, \dots, Y_d)$ *is as in* (6.1.1) *and* $d > 2$, *then,*

$$E\big[\|\mathbf{Y}\|_2^{-2}\big] = E\Big[\frac{1}{d+2N-2}\Big] \geq \frac{1}{d+\|\theta\|_2^2/\sigma^2-2}.$$

For $d > 2$, *we have moreover that* $E[\|\mathbf{Y}\|_2^{-2}] \leq 1/(d-2)$.

Proof. We start with proving the first equality. Observe that

$$E_\theta\Big[\frac{1}{\|\mathbf{Y}\|_2^2}\Big] = \frac{1}{\sigma^2} E_\theta\Big[\frac{1}{\|\sigma^{-1}\theta + \varepsilon\|_2^2}\Big] \tag{6.4.5}$$

and $\|\sigma^{-1}\theta + \varepsilon\|_2^2 = \sum_{j=1}^d (\sigma^{-1}\theta_j + \varepsilon_j)^2$. If $X_i \sim \mathcal{N}(\mu_i, 1)$ are independent, then $\sum_{i=1}^d X_i^2$ follows a non-central χ_d^2 distribution with non-centrality parameter $\sum_{i=1}^d \mu_i^2$. This distribution can also be written as χ^2_{d+2N} distribution where N denotes a Poisson distribution with intensity $\frac{1}{2}\sum_{i=1}^d \mu_i^2$. For the one-dimensional case ($d = 1$) this can be checked using characteristic functions. The multivariate case can be reduced to $d = 1$ with the transformation formula, cf. Exercise 6.4. Thus, $\|\sigma^{-1}\theta + \varepsilon\|_2^2 \sim \chi^2_{d+2N}$ with $N \sim \text{Poisson}(\|\theta\|_2^2/(2\sigma^2))$ and we can further rewrite (6.4.5) with $U \sim \chi^2_{d+2N}$ as

$$E_\theta\Big[\frac{1}{\|\mathbf{Y}\|_2^2}\Big] = E\Big[\frac{1}{U}\Big] = E\Big[E\Big[\frac{1}{U}|N\Big]\Big] = E\Big[\frac{1}{d+2N-2}\Big],$$

where we used for the last step that for the inverse moment of a χ_k^2-distribution, $E[1/\chi_k^2] = 1/(k-2)$ whenever $k > 2$, cf. Exercise 6.6.

By Cauchy-Schwarz inequality, $1 \leq E[X]E[1/X]$ provided the right hand side is well-defined and this yields

$$E_\theta\Big[\frac{1}{\|\mathbf{Y}\|_2^2}\Big] \geq \frac{1}{d+2E[N]-2} = \frac{1}{d+\|\theta\|_2^2/\sigma^2-2}.$$

□

From the previous result, we can now analyse the James-Stein estimator $\widehat{\theta}_{\text{JS}} = (1 - \sigma^2(d-2)/\|\mathbf{Y}\|_2^2)\mathbf{Y}$. Let

$$g(\mathbf{y}) = -\sigma^2(d-2)/\|\mathbf{y}\|_2^2\mathbf{y} \tag{6.4.6}$$

The function g is everywhere differentiable except for the point $\mathbf{y} = (0, 0, \dots, 0)$ and the partial derivatives are given by

$$\partial_i g_i(\mathbf{y}) = -\sigma^2(d-2)\Big(\frac{1}{\|\mathbf{y}\|_2^2} - \frac{2y_i^2}{\|\mathbf{y}\|_2^4}\Big)$$

for $i = 1, \dots, d$. The divergence is thus $\nabla \cdot g(\mathbf{y}) = \sum_{i=1}^d \partial_i g_i(\mathbf{y}) = -\sigma^2(d-2)^2/\|\mathbf{y}\|_2^2$. Observe that $\widehat{\theta}_{\text{JS}} = \mathbf{Y} + g(\mathbf{Y})$. To apply Stein's lemma, we need to check the moment conditions (6.4.1).

risk	general	$\|\theta\|_2$ small	$\|\theta\|_2 \to \infty$
$\widehat{\theta} = Y$	d	d	d
$\widehat{\theta}_{\mathrm{JS}}$	$d - \frac{(d-2)^2}{d-2+\|\theta\|_2^2}$	$2 + \|\theta\|_2^2$	d

Table 61: Risk gain by James-Stein shrinkage for $\sigma = 1$ and $d > 2$.

This is left as an exercise. Using Stein's lemma,

$$\begin{aligned} E_\theta\big[\|\widehat{\theta}_{\mathrm{JS}} - \theta\|_2^2\big] &= \sigma^2 d + E_\theta\Big[-2\sigma^4 \frac{(d-2)^2}{\|\mathbf{Y}\|_2^2} + \frac{\sigma^4(d-2)^2}{\|\mathbf{Y}\|_2^2}\Big] \\ &= \sigma^2 d - \sigma^4(d-2)^2 E_\theta\Big[\frac{1}{\|\mathbf{Y}\|_2^2}\Big] \\ &\le \sigma^2 d - \frac{\sigma^4(d-2)^2}{\sigma^2(d-2) + \|\theta\|_2^2} \\ &\le \sigma^2 d \wedge \big(2\sigma^2 + \|\theta\|_2^2\big), \end{aligned} \tag{6.4.7}$$

where for the last inequality we used that $-(1+x)^{-1} \le (x-1) \wedge 0$, for $x \ge 0$, and

$$\sigma^2 d - \frac{\sigma^4(d-2)^2}{\sigma^2(d-2) + \|\theta\|_2^2} = \sigma^2 d - \sigma^2(d-2)\frac{1}{1 + \|\theta\|_2^2/(\sigma^2(d-2))} \le 2\sigma^2 + \|\theta\|_2^2.$$

Since the MLE $\widehat{\theta} = \mathbf{Y}$ has risk $E_\theta[\|\widehat{\theta} - \theta\|_2^2] = \mathbb{E}[\sum_{i=1}^d \sigma^2 \varepsilon_i^2] = \sigma^2 d$, we conclude that the James-Stein estimator has smaller risk. This proves Theorem 7.

As a first remark, let us mention that the James-Stein estimator is not admissible either (cf. for instance [23]).

By (6.4.7), the risk of the James-Stein estimator is bounded from above by $\sigma^2 d - \sigma^4(d-2)^2/(\sigma^2(d-2) + \|\theta\|_2^2)$. In Table 61, we analyze the risk bound for different scenarios assuming that the standard deviation of the noise σ equals one. Obviously the risk of the maximum likelihood estimator is always d and the James-Stein procedure has much smaller risk if the signal is small. In the specific case that $\theta = 0$, the risk of the James-Stein estimator is 2 and does not depend on the dimension d. For large signal, Stein shrinkage does improve the risk only by a small amount.

6.5 * An example for sports data

Suppose we want to predict the number of goals of all soccer teams in a league given the number of goals from the previous season. A good model is that the number of goals N_i of team i in one season follows a Poisson distribution with intensity λ_i, that is, $N_i \sim \mathrm{Poisson}(\lambda_i)$. Using the variance stabilizing transform (cf. ???), we find that $\sqrt{N_i} \approx \sqrt{\lambda_i} + \frac{1}{2}\varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0,1)$, i.i.d. If M_i is the number of goals in the next season, $\sqrt{M_i} \approx \sqrt{\lambda_i} + \frac{1}{2}\varepsilon_i'$ with independent $\varepsilon_i' \sim \mathcal{N}(0,1)$. Consequently, $\sqrt{N_i} - \sqrt{M_i} \sim \mathcal{N}(0, 1/2)$. Treating M_i now as fixed, we arrive at the model $\sqrt{N_i} = \sqrt{M_i} + 2^{-1/2}\eta_i$, with $\eta_i \sim \mathcal{N}(0,1)$, i.i.d.

To estimate $\sqrt{M_i}$ we compare the MLE with the James-Stein estimator. The data are the total number of goals of all 17 teams playing premier league in the seasons 2016/17 and 2017/18. With respect to the loss $\ell(\theta, \theta') = \sum_i (\sqrt{\theta_i} - \sqrt{\theta'_i})^2$ the risk of the MLE is 10.02 and the risk of the James-Stein estimator is 9.6783. For the loss $\ell(\theta, \theta') = \sum_i |\theta_i - \theta'_i|$ the MLE has risk 160 and the risk of the James-Stein estimator is 157.06. In both cases the James-Stein estimator is slightly better.

The theoretical improvement of the James-Stein estimator over the MLE is $\sigma^4(d-2)^2 E[\|\mathbf{Y}\|_2^{-2}]$, see (6.4.7). We can estimate the size of the term by $\sigma^4(d-2)^2\|\mathbf{Y}\|_2^{-2}$. In the case above this is 0.058 suggesting a minor improvement of the James-Stein estimator. The improvement is small since the signal is quite large compared to the variance in the model. This is also referred to as low signal-to-noise ratio (SNR). The James-Stein estimator outperforms the MLE more significantly if only the total number of goals for the first few games in each season are used for the prediction. If the Poisson counts are too small, the normal approximation will not work well anymore and this might affect the performance of the estimator.

The James-Stein estimator has been applied to baseball data in [7, 21].

6.6 Exercises

Ex. 6.1 — For any real number t, consider the rescaled MLE $\widehat{\theta}_t^{\mathrm{MLE}} := t\mathbf{Y}$ in the sequence model (6.1.1). Show that

$$1 - \frac{\sigma^2 d}{\|\theta\|_2^2 + \sigma^2 d} = \mathrm{argmin}_t \, E_\theta\big[\|\widehat{\theta}_t^{\mathrm{MLE}} - \theta\|_2^2\big].$$

Ex. 6.2 — Suppose that $X \sim \mathcal{N}(0,1)$ and $f'(X)$ is absolutely continuous. If $E[|f'(X)|] < \infty$ then also $E[|Xf(X)|] < \infty$ and $E[|f(X)|] < \infty$.
Hint: Show first that $E[|Xf(X)|\mathbf{1}(X \geq 0)] < \infty$ using $|f(x)| \leq \int_0^x |f'(y)|dy$.

Ex. 6.3 — Let $X \sim \mathcal{N}(0,1)$ and k be a positive integer. Using Stein's lemma, prove that

$$E[X^{2k}] = \prod_{j=1}^{k}(2j-1).$$

Ex. 6.4 — Assume that $X_i \sim \mathcal{N}(\mu_i, 1)$ $i = 1, \ldots, d$ are independent and define $S_d = \sum_{i=1}^d X_i^2$. Consider the following two step procedure: First generate a random variable N from a Poisson distribution with intensity $\frac{1}{2}\sum_{i=1}^d \mu_i^2$. Given N draw $V_d \sim \chi^2_{d+2N}$. Show that S_d and V_d have the same distribution.

Ex. 6.5 — If $\mathbf{Y} = (Y_1, \ldots, Y_d)$ is as in (6.1.1) and $d > 2$, then,

$$E\Big[\|\mathbf{Y}\|_2^{-2}(\mathbf{Y} - \theta)^\top \mathbf{Y}\Big] = (d-2)E\big[\|\mathbf{Y}\|_2^{-2}\big].$$

Ex. 6.6 — Let ξ_i, $i = 1, \dots, d$ be i.i.d. $\mathcal{N}(0,1)$ random variables. Show that for $d > 2$,

$$E\Big[\Big(\sum_{i=1}^{d} \xi_i^2\Big)^{-1}\Big] = \frac{1}{d-2}.$$

Hint: Use the equality in the previous exercise.

Ex. 6.7 — For the function g as defined in (6.4.6), check that the conditions of Lemma 9 are satisfied.

Ex. 6.8 — For $d > 2$ and fixed $a \geq 0$ consider the estimator

$$\widehat{\theta}_a = \Big(1 - \frac{a\sigma^2(d-2)}{\|\mathbf{Y}\|_2^2}\Big)\mathbf{Y}.$$

In particular, $\widehat{\theta}_0$ is the MLE and $\widehat{\theta}_1$ is the James-Stein estimator. Show that $E[\|\widehat{\theta}_1 - \theta\|_2^2] < E[\|\widehat{\theta}_a - \theta\|_2^2]$ for all $a \neq 1$ and all $\theta \in \mathbb{R}^d$.

Ex. 6.9 — Use integration by parts to prove Mill's ratio

$$P\big(X \geq \tau\big) \leq E\Big[\Big(1 + \frac{1}{X^2}\Big)\mathbf{1}(X \geq \tau)\Big] = \frac{\phi(\tau)}{\tau},$$

with $X \sim \mathcal{N}(0,1)$ and ϕ the p.d.f. of X.

Ex. 6.10 — Show that if $g : \mathbb{R} \to \mathbb{R}$ satisfies the assumptions of Lemma 9 and (X, Y) follows a bivariate normal distribution, then

$$\mathrm{Cov}(g(X), Y) = \mathrm{Cov}(X, Y)E[g'(X)].$$

Chapter 7

High-dimensional models and structural constraints

7.1 Introduction

Recall the definition of a high-dimensional statistical model in (1.2). One of the differences between nonparametric and high-dimensional statistics is that for high-dimensional problems we also take computational complexity into account. Before, we were interested in estimators, which are optimal in some statistical sense, achieving for instance the minimax estimation rate. In high-dimensional statistics, we ask for estimators which are both, statistically optimal and computationally feasible. The algorithmic constraint is very important in order to account for the typically large datasets. For high-dimensional estimation problems, we often define estimators as minimizers of some functional. Whether an estimator is computational feasible depends then on the algorithmic complexity to compute a minimizer. For this reason, we prefer estimator which are minimizers of convex functionals.

Below, we introduce high-dimensional linear regression, which is the most important high-dimensional model.

7.2 High-dimensional regression

The high-dimensional regression model has been introduced in Section 1.3. In this model, we observe a vector $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and a $n \times p$ design matrix X with

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{7.2.1}$$

Here, $\boldsymbol{\beta}$ is an unobserved p-dimensional coefficient vector. Moreover, I_n denotes the $n \times n$ identity matrix and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n)$ is a centered random vector from a multivariate normal distribution with covariance matrix I_n. The design matrix X is known and the goal is to estimate the unknown coefficient vector $\boldsymbol{\beta}$. This model is high-dimensional if $p > n$, since in this case we have more parameters than observations.

Given the parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$, we can define *the support of* $\boldsymbol{\beta}$ or *active set* as the non-zero components $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$ and the *sparsity (index)* of $\boldsymbol{\beta}$ as $s_{\boldsymbol{\beta}} = |S_{\boldsymbol{\beta}}| =$

the cardinality of $S_{\boldsymbol{\beta}}$. Because the parameter changes with n, the active set and the sparsity are also n dependent but again this dependence is omitted in the notation.

We say that *a model is sparse if the number of non-zero parameters is of smaller order than the number of observations* which is equivalent to $s_{\boldsymbol{\beta}} \ll n$.

Sparsity constraints are not so different from smoothness assumptions on the regression function. Recall that smoothness corresponds to decay conditions on the (Fourier) series coefficients. In the sequence model, the structural constraint is therefore the decay of the coefficients. This means that most series coefficients are very small which is closely related to imposing sparsity. A difference is that in function estimation, we expect that the first series coefficients are the largest ones and for sparse vectors there is no ordering of the non-zero coefficients.

7.3 Estimation in the sequence model under sparsity

As a toy model, we consider in a first step the sequence model with n observations, that is, we observe $\mathbf{Y}^n = (Y_1, \ldots, Y_n)^\top$ with

$$Y_i = \beta_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n,$$

and β_i the unknown coefficients. This version of the sequence model is a special case of model (7.2.1) with $p = n$ and $X = I_n$. Although this model is not truly high-dimenional, it is very useful as a toy model in high-dimensional statistics.

In order to estimate the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^\top$, we apply the same strategy as for wavelet thresholding estimator defined in (5.5.1): We keep large coefficients for which we are sure that they contain signal and the remaining coefficients are estimated by zero. Consider the thresholding estimator

$$\widehat{\beta}_k = Y_k \mathbf{1}(|Y_k| \geq \sqrt{2 \log n}).$$

For squared ℓ^2-loss $\ell(\boldsymbol{\beta}, \boldsymbol{\beta}') = \sum_{k=1}^n (\beta_k - \beta_k')^2 = \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2$,

$$E_{\boldsymbol{\beta}}\big[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\big] \leq 8(s_{\boldsymbol{\beta}} + 1) \log n, \tag{7.3.1}$$

cf. Exercise 7.2. We observe that the rate of convergence also depends on $s_{\boldsymbol{\beta}}$ and becomes better if the true signal is sparse. Thus sparsity helps. This is not surprising since we already know that in function estimation smoother functions are easier to reconstruct and are more sparse because of the faster decaying series coefficients. As we will see later, the rate $(s_{\boldsymbol{\beta}} + 1) \log n$ is optimal over the parameter space of $s_{\boldsymbol{\beta}}$-sparse vectors.

The factor $\log n$ is the amount we have to pay in the estimation rate for not knowing the support $S_{\boldsymbol{\beta}}$. Indeed, if $S_{\boldsymbol{\beta}}$ would be known, we could simply estimate $\boldsymbol{\beta}$ by $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_n)$ with $\widetilde{\beta}_k = Y_k \mathbf{1}(k \in S_{\boldsymbol{\beta}})$ and this estimator would have risk $s_{\boldsymbol{\beta}}$.

It is also interesting to compare the result with the risk bound obtained for the James-Stein estimator in (6.4.7). Since $d = n$ and $\sigma = 1$, the James-Stein estimator has risk smaller than

$$n - \frac{(n-2)^2}{(n-2) + \|\boldsymbol{\beta}\|_2^2} \leq n \wedge \big(2 + \|\boldsymbol{\beta}\|_2^2\big).$$

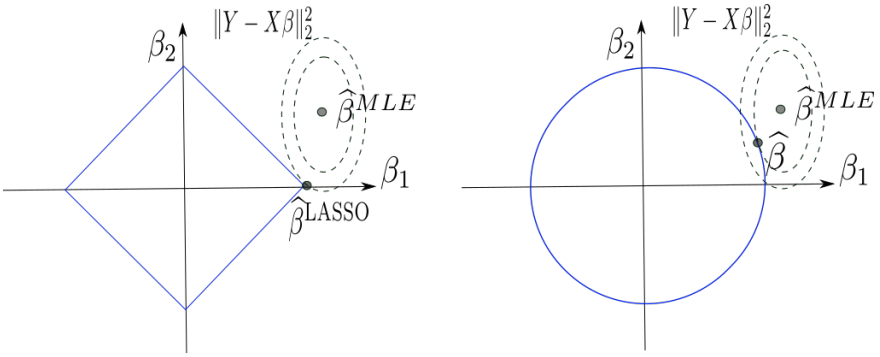


Figure 22: Heuristic for the LASSO. The dashed lines are the contour lines of the least squares functional $\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2$. Because of the corners of the ℓ^1-ball, the LASSO solution sets many coefficients to zero (left). The ℓ^2-ball does not induce sparsity (right).

This shows that the James-Stein estimator has comparable risk if $\|\boldsymbol{\beta}\|_2$ is not too big. In the minimax sense the James-Stein estimator does however not perform well since we could have a sparse signal with few extremely large non-zero coefficients making the risk bound as large as n.

As conclusion of this section, we should remember that sparsity constraints improve estimators. In the next section we study estimation in the high-dimensional regression model.

7.4 The LASSO

Recall the high-dimensional regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon.$$

In this model, the estimator for $\boldsymbol{\beta}$ cannot be expressed as explicit formula. Instead we will consider an estimator that is the minimizer of a functional. To motivate this, consider the likelihood $p(\mathbf{y}|\boldsymbol{\beta}) = (2\pi)^{-n/2}\exp(-\frac{n}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2)$ and observe that the maximum likelihood estimator for $\boldsymbol{\beta}$ is the least squares estimator that also can be written as

$$\widehat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2.$$

This expression does not incorporate the sparsity assumption yet. Because of $p > n$, there is an affine subspace of dimension $p - n$ which perfectly fits the data in the sense that $\|\mathbf{Y} - X\boldsymbol{\beta}\|_2 = 0$. There is a close connection between shrinkage and penalization. This suggests to replace the least squares solution by a penalized version of the form

$$\widehat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda\,\mathrm{pen}(\boldsymbol{\beta}),$$

where $\mathrm{pen}(\boldsymbol{\beta})$ is a penalty or regularisation term that is large if $\boldsymbol{\beta}$ is not sparse. The positive weight $\lambda > 0$ measures how much influence the penalization term has. It is also known as *regularization parameter* and plays a similar role as the bandwidth for kernel estimators. If λ is

large the penalization term gets more weight which leads to sparse reconstructions. The other extreme is to set $\lambda = 0$ in which case we recover the least squares problem. In practice, the regularization parameter needs to be chosen very carefully as we will see later.

What could be a good penalty in our problem? One possibility would be to penalize the number of non-zero entries and to consider the estimator

$$\widehat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda s_{\boldsymbol{\beta}}.$$

This would clearly push the solution to be sparse. Unfortunately, this estimator is computationally infeasible since minimization of the functional results in a non-convex optimization problem. It turns out that the ℓ^1-penalty $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ is sparsity inducing and leads to a convex optimization problem if combined with the least-squares data misfit term.

The *LASSO* (Least Absolute Shrinkage and Selection Operator) is any estimator satisfying

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{7.4.1}$$

The LASSO was proposed by Tibshirani in his 1996 article [27]. It can be computed very efficiently even for large n and p. There is a simple heuristic why the LASSO induces sparsity. We can rewrite (7.4.1) as

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p\,:\,\|\boldsymbol{\beta}\|_1\le R} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 \tag{7.4.2}$$

(cf. [1], Chapter 5.3). Here R is a constant depending on the regularization parameter λ, the design matrix X and the data. The minimization problem is displayed in Figure 22. The MLE $\widehat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ minimizes the unconstrained least squares problem. The additional constraint $\widehat{\boldsymbol{\beta}} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \le R\}$ forces the LASSO estimator to lie on the first intersection of a contour line with the ℓ^1-ball. For many configurations, the minimizer will be in one of the corners of the ℓ^1-ball. This means that at least one of the components of $\boldsymbol{\beta}$ is zero and shows why the LASSO induces sparsity. An ℓ^2-penalty on the contrary would lead to projection on a ℓ^2-ball which has no corners and therefore leads to non-sparse reconstructions.

Figure 22 is commonly used to illustrate that ℓ^1-penalization induces sparse solutions. But it is a bit misleading. The argument suggests that if $(Y_1, Y_2) \neq (0, 0)$, then also $\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} \neq 0$. It seems thus that the LASSO prefers the corners by setting exactly one of the two components to zero but that it cannot shrink to the sparse vector $(0, 0)$. This is, however, wrong. Exercise 7.3 shows for instance that for orthogonal design the LASSO does soft shrinkage. It sets all small component to zero and thus shrinks many vectors to $(0, 0)$. The explanation is that in (7.4.2), the value of R is zero in these cases. The heuristic is also not strong enough to generalize to other problems. One could for instance wonder what would happen if an ℓ^∞-penalty is used. The corresponding unit ball has 2^d corners.

In a next step, we analyze the LASSO estimator. For recovery of $\boldsymbol{\beta}$, we need a local invertibility assumption on the design matrix X. It could indeed happen that $X\boldsymbol{\beta} = 0$. In this case, we cannot distinguish whether the data were generated by the parameter $\boldsymbol{\beta}$ or the zero vector. The condition below ensures that this does not happen. For an arbitrary index set S, write $\boldsymbol{\beta}_S = (\beta_i)_{i\in S}$ and define $\|X\| := (\max_{i=1,\dots,p}(X^\top X)_{i,i})^{1/2}$.

Theorem 8. *Let*

$$\phi(S,L) := \inf\Big\{\frac{\|X\boldsymbol{\beta}\|_2\sqrt{|S|}}{\|\boldsymbol{\beta}\|_1\|X\|} \,:\, \|\boldsymbol{\beta}_{S^c}\|_1 \le L\|\boldsymbol{\beta}_S\|_1\Big\}.$$

If $\lambda \ge 4\|X\|\sqrt{2\log p}$, *then as* $p\to\infty$,

$$P_{\boldsymbol{\beta}}\Big(\|X(\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}-\boldsymbol{\beta})\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}-\boldsymbol{\beta}\|_1 > \frac{4\lambda^2 s_{\boldsymbol{\beta}}}{\|X\|^2\phi(S_{\boldsymbol{\beta}},3)^2}\Big)\to 0.$$

The condition $\phi(S_{\boldsymbol{\beta}},L)>0$ is also known as $(S_{\boldsymbol{\beta}},L)$-compatibility condition. Unfortunately, given observation from the model, it cannot be checked whether the compatibility condition holds. The condition should rather be viewed as another constraint on the support of $\boldsymbol{\beta}$, cf. Exercise 7.6 for an example.

If the $(S_{\boldsymbol{\beta}},L)$-compatibility condition holds and $\lambda \asymp \|X\|\sqrt{\log p}$, the convergence rate for the squared *prediction error* $\|X(\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}-\boldsymbol{\beta})\|_2^2$ is $s_{\boldsymbol{\beta}}\log p$. This rate should be compared with (7.3.1). Notice that if p is large, for instance if $p=e^n$, the $\log p$-factor can become extremely big. For the ℓ^1-error we have under compatibility and $\lambda \asymp \|X\|\sqrt{\log p}$,

$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}-\boldsymbol{\beta}\|_1 \lesssim \frac{s_{\boldsymbol{\beta}}\sqrt{\log p}}{\|X\|}$$

with high probability.

The result suggests that a good choice for the regularization parameter is $\lambda = 4\|X\|\sqrt{2\log p}$. In practice, this choice turns out to be too conservative making the reconstruction extremely sparse. Instead the regularization parameter is selected by some data-driven method. A popular procedure which is implemented in standard software is ten-fold cross-validation, cf. [10]. Another concept is to study the LASSO estimator as a function in dependence of λ.

Proof of Theorem 8. Write $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$. By the definition of the LASSO, we have $\|\mathbf{Y}-X\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \|\mathbf{Y}-X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ and so

$$\|X(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le 2\epsilon^\top X(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1.$$

Define the event $\mathcal{A} = \{\max_{i=1,\dots,p}|(X^\top\epsilon)_i| \le \|X\|\sqrt{2\log p}\}$. Observe that $(X^\top\epsilon)_i \sim \mathcal{N}(0,(X^\top X)_i)$ and therefore with the union bound (see Section 14.1) and Mill's ratio (cf. Exercise 6.9),

$$P_{\boldsymbol{\beta}}(\mathcal{A}^c) \le \sum_{i=1}^p P_{\boldsymbol{\beta}}\big(|(X^\top\epsilon)_i| \ge \|X\|\sqrt{2\log p}\big)\to 0, \quad \text{as } p\to\infty.$$

On $\mathcal{A}$, we have thanks to $\lambda \ge 4\|X\|\sqrt{2\log p}$,

$$\|X(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{\lambda}{2}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 + \lambda\|\boldsymbol{\beta}\|_1.$$

Denote by $S = S_{\boldsymbol{\beta}}$ the true support. Triangle inequality gives, $\|\widehat{\boldsymbol{\beta}}\|_1 = \|\widehat{\boldsymbol{\beta}}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \geq \|\boldsymbol{\beta}_S\|_1 - \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c}\|_1$. Together with $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|_1 + \|\widehat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}_{S^c}\|_1$ and $\boldsymbol{\beta}_{S^c} = 0$,

$$\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 + \frac{\lambda}{2}\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \leq \frac{3\lambda}{2}\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|_1,$$

on the event $\mathcal{A}$. In particular, we can deduce from the previous inequality that $\|\widehat{\boldsymbol{\beta}}_{S^c}\|_1 \leq 3\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|_1$ and by definition of the compatibility condition, $\phi(S,3)\|X\|\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq \|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2\sqrt{s_{\boldsymbol{\beta}}}$. Since $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ holds for all real numbers a, b, this shows

$$\begin{aligned}\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 + \frac{\lambda}{2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq 2\lambda\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 &\leq 2\lambda\frac{\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2\sqrt{s_{\boldsymbol{\beta}}}}{\|X\|\phi(S_{\boldsymbol{\beta}}, 3)} \\ &\leq \frac{1}{2}\|X(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 + 2\lambda^2\frac{s_{\boldsymbol{\beta}}}{\|X\|^2\phi(S_\beta, 3)^2}\end{aligned}$$

on the event $\mathcal{A}$. Since $P_{\boldsymbol{\beta}}(\mathcal{A}^c) \to 0$, the result follows. □

Even if convex relaxation still leads to the same minimax rates one typically has to pay a price. To illustrate this consider again the special case of the sequence model ($n = p$ and X the $n \times n$ identity matrix). The minimizer $\widehat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2 + \mu s_{\boldsymbol{\beta}}$ can then be minimized for each component $\widehat{\beta}_i \in \operatorname{argmin}_{\beta_i\in\mathbb{R}}(Y_i - \beta_i)^2 + \mu\mathbf{1}(\beta_i \neq 0)$ and this gives the hard-thresholding estimator $\widehat{\beta}_i = Y_i\mathbf{1}(|Y_i| > \sqrt{\mu})$. On the contrary, by Exercise (7.3), the LASSO with regularization parameter λ is given by $\widehat{\beta}_i^{\mathrm{LASSO}} = \big(Y_i - \lambda\operatorname{sign}(Y_i)/2\big)\mathbf{1}\big(|Y_i| \geq \lambda/2\big)$. To account for sparsity Y_i has to be put to zero whenever $|Y_i|$ is below $\sqrt{2\log n}$. This implies that $\sqrt{\mu}$ and $\lambda/2$ have to be chosen at least as $\geq \sqrt{2\log n}$ and $\sqrt{\mu} = \lambda/2 = \sqrt{2\log n}$ is the optimal choice to avoid thresholding of large coefficients. This shows that the size of the regularization parameter needs to be adapted under convex relaxation. Suppose that for given s_0 the true regression vector is $\beta_i = \sqrt{8\log n}\mathbf{1}(i \leq s_0)$ for $i = 1, \ldots, n$. This means that $\boldsymbol{\beta} = (\beta_i)_i$ is s_0-sparse. One can show (cf. Exercise ??? [to do ???]) that for the hard-thresholding estimator with regularization parameter $\sqrt{\mu} = \sqrt{2\log n}$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim s_0$. For the LASSO with $\lambda/2 = \sqrt{2\log n}$, we have the lower bound $\|\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} - \boldsymbol{\beta}\|_2^2 \gtrsim s_0\log n$. This shows that for this parameter the LASSO performs worse by (at least) a $\log n$ factor. Although this does not affect the minimax estimation rate, it shows that on large parts of the parameter space different rates can be obtained.

The LASSO for nonparametric regression

Recall the nonparametric regression model with uniform fixed design, that is, we observe $\mathbf{Y} = (Y_{1n}, \ldots, Y_{nn})^\top$ with

$$Y_{in} = f\Big(\frac{i}{n}\Big) + \varepsilon_{i,n}, \quad i = 1, \ldots, n, \;\; \varepsilon_{i,n} \sim \mathcal{N}(0,1), \text{ independent.}$$

If the regression function is piecewise constant $f(t) = \sum_{k=1}^n \beta_k\mathbf{1}(t \leq k/n)$, the problem can be rewritten as a linear model with $p = n$ and sparse regression vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. For

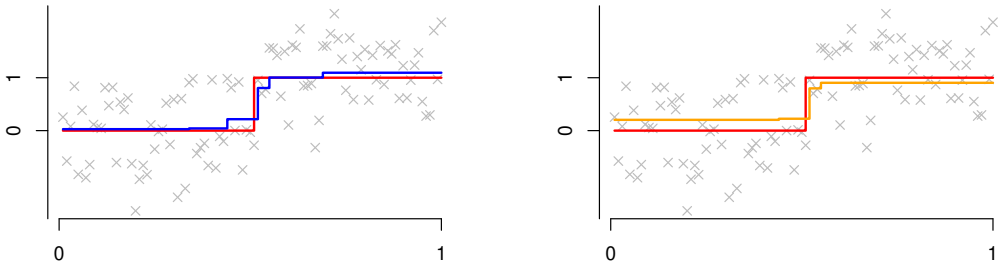


Figure 23: Nonparametric regression with piecewise constant regression function interpreted as high-dimensional linear model. Data (gray), true regression function (red), LASSO reconstructions with regularization parameter chosen by 10-fold cross validation (blue) and regularization parameter $\lambda = 0.25$ (orange). The LASSO adds spurious jumps to the reconstruction around the location of the true jump. Large regularization parameters reduce this effect but add bias.

details see Exercise 7.6. The estimated sparsity corresponds then to the number of linear pieces of the reconstruction and the ℓ_1-penalty is the total variation. The LASSO does therefore total variation penalization in this case. Figure 23 shows that the LASSO reconstruction builds in additional jumps around the true jump and therefore overshoots the true model dimension. To decrease the model dimension, we can choose a larger regularization parameter. In that case, the shrinkage of the LASSO will correspond to a bias in the reconstruction of the heights of the linear pieces; see the plot on the right in Figure 23.

We can now wonder why the spurious jumps of the LASSO reconstruction mainly occur around the true jump locations. The reason for this phenomenon comes from the columns in the design matrix. Denote the p columns of the design matrix X by $\mathbf{X}_1, \dots, \mathbf{X}_p$. Suppose that for some K and some coefficients γ_i, $i = 1, \dots, K$, $\mathbf{X}_p = \sum_{i=1}^K \gamma_i \mathbf{X}_i$ and $\sum_{i=1}^K |\gamma_i| < 1$. If $\beta_p \neq 0$, the LASSO functional decreases if instead of the p-th variable with regression coefficient β_p, the variables $1, \dots, K$ are selected with regression coefficients $\beta_p \gamma_i$. This can be checked by looking at the data misfit term $\|\mathbf{Y} - X\boldsymbol{\beta}\|_2^2$ and the LASSO penalty $\lambda\|\boldsymbol{\beta}\|_1$ separately. This should be contrasted with the behavior of the complexity penalty which always increases if a larger model larger is selected.

To overcome this issue, one possibility is to reparametrize the model. For any $p \times p$ diagonal matrix D, $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X'\boldsymbol{\beta}' + \boldsymbol{\varepsilon}$ with $X' = XD$ and $\boldsymbol{\beta}' = D^{-1}\boldsymbol{\beta}$. By choosing a suitable D, we can therefore assume that all column vectors of the design matrix have the same Euclidean norm. The argument above does then not work anymore. This reparametrization does not really help to remove the artificially created jumps by the LASSO. In this case, additional variables around the true jump will increase the ℓ^1-penalty a little and on the same time decrease the data misfit term by a much larger amount.

To conclude, straightforward application of the LASSO does not perform well for this problem. As the LASSO can be quickly computed and produces models that are too large but contain the

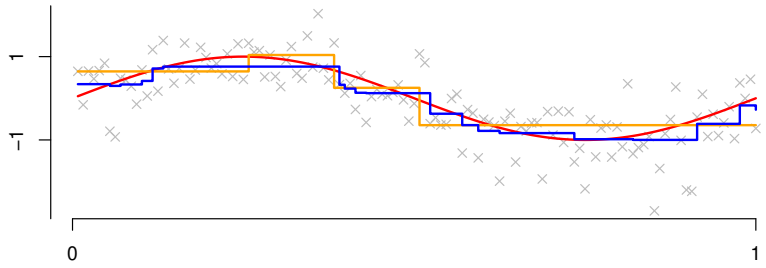


Figure 24: Nonparametric regression using the LASSO. Data (gray), true regression function (red), LASSO reconstructions with regularization parameter chosen by 10-fold cross validation (blue), hard thresholding estimator using Haar wavelets and universal threshold (orange).

true model, it still can be used as an initial estimator, that is then further appended by a more careful selection of a submodel and a de-biasing step for the estimated jump heights.

Whereas ℓ_0-penalization is in general computationally infeasible, for this specific model, there is an algorithm computing the ℓ_0-solution in polynomial time (more here ???).

It is also possible to use the LASSO for nonparametric regression if the regression function is not piecewise constant; cf. Figure 24 for an example. Since the LASSO leads to a piecewise constant reconstruction it is natural to contrast it with the reconstruction using Haar wavelets and hard thresholding with universal threshold. In this case the LASSO reconstruction seems to be better.

7.5 Sparse Gaussian graphical models

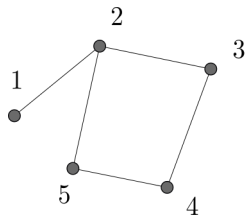
A natural question is whether ℓ^1-penalization works in general if the parameter exhibits some form of sparsity. In this section, we study graphical models under sparsity constraints and show how ℓ^1-penalization can be used.

Suppose we observe n i.i.d. copies $\mathbf{X}_1, \dots, \mathbf{X}_n$ of a multivariate normal random vector $\mathbf{X} = (X_1, \dots, X_p)^\top \sim \mathcal{N}(0, \Sigma)$ with Σ an invertible $p \times p$ covariance matrix. The inverse $\Omega = \Sigma^{-1}$ is called the *precision matrix* or *concentration matrix*. A classical property of the multivariate normal distribution is that $\Sigma_{i,j} = 0$ if and only if X_i and X_j are independent. The precision matrix decodes the conditional dependence structure. For convenience, set $[p] := \{1, \dots, p\}$.

Lemma 11. *Let $i, j \in [p]$, $i \neq j$. Given $(X_k)_{k \in [p] \setminus \{i,j\}}$, the variables X_i, X_j are independent if and only if $\Omega_{i,j} = 0$.*

Proof. We prove the following more general statement. If $\mathbf{X} = (\mathbf{X}_A, \mathbf{X}_{A^c})^\top$ for some subset $A \subset [p]$, and $\Omega_{AB} := (\Omega_{i,j})_{i \in A, j \in B}$ then,

$$\mathbf{X}_A | \mathbf{X}_{A^c} \sim \mathcal{N}\big(-\Omega_{AA}^{-1} \Omega_{AA^c} \mathbf{X}_{A^c}, \Omega_{AA}^{-1} \big). \tag{7.5.1}$$



$$\Omega = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 1 & 0 & 1 & 3 \end{pmatrix}$$

Figure 25: Precision matrix Ω and corresponding graphical model.

With the formula for the multivariate Gaussian likelihood

$$\begin{aligned} \mathbf{x}_A \mapsto f_{\mathbf{X}_A|\mathbf{X}_{A^c}}(\mathbf{x}_A|\mathbf{x}_{A^c}) &= \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}_{A^c}}(\mathbf{x}_{A^c})} \\ &\propto \exp\Big(-\frac{1}{2}\mathbf{x}^\top\Omega\mathbf{x}\Big) \\ &\propto \exp\Big(-\frac{1}{2}\mathbf{x}_A^\top\Omega_{AA}\mathbf{x}_A - \mathbf{x}_A^\top\Omega_{AA^c}\mathbf{x}_{A^c}\Big) \\ &\propto \exp\Big(-\frac{1}{2}(\mathbf{x}_A-\mu)^\top\Omega_{AA}(\mathbf{x}_A-\mu)\Big) \end{aligned}$$

with $\mu = -\Omega_{AA}^{-1}\Omega_{AA^c}\mathbf{x}_{A^c}$. This shows (7.5.1). The assertion of the lemma follows since for $A = \{i, j\}$,

$$\Omega_{AA}^{-1} = \frac{1}{\Omega_{i,i}\Omega_{j,j} - \Omega_{i,j}\Omega_{j,i}} \begin{pmatrix} \Omega_{j,j} & -\Omega_{i,j} \\ -\Omega_{j,i} & \Omega_{i,i} \end{pmatrix}.$$

□

The *graphical model* or *conditional independence graph* associated to the $\mathcal{N}(0, \Sigma)$ distribution is the graph with p nodes and an edge between node i and node j if X_i and X_j are dependent given $(X_k)_{k\in[p]\setminus\{i,j\}}$. Besides the pairwise conditional (in)dependence relations, the graph decodes many other statements. An example of a precision matrix and the corresponding conditional independence graph is displayed in Figure 25.

The statistical challenge is to estimate the precision matrix and to reconstruct the graphical model. We are interested in the situation where p is potentially large and the conditional independence graph is sparse. Denote by $|\Omega|$ the determinant of Ω. The likelihood for the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is

$$p_\Omega(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \prod_{i=1}^n p_\Omega(\mathbf{X}_i) = \prod_{i=1}^n \frac{|\Omega|^{1/2}}{(2\pi)^{p/2}} \exp\Big(-\frac{1}{2}\mathbf{X}_i^\top\Omega\mathbf{X}_i\Big).$$

Observe that $\mathbf{X}_i^\top\Omega\mathbf{X}_i = \operatorname{tr}(\Omega\mathbf{X}_i\mathbf{X}_i^\top)$. With $\widehat{\Sigma} := n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top$, we have for the log-likelihood

$$\frac{\log p_\Omega(\mathbf{X}_1, \ldots, \mathbf{X}_n)}{n} = \text{const.} + \frac{1}{2}\log|\Omega| - \frac{1}{2}\operatorname{tr}\big(\Omega\widehat{\Sigma}\big).$$

The MLE for Ω is thus the minimizer

$$\widehat{\Omega}^{\mathrm{MLE}} \in \operatorname{argmin}_{\Omega \succ 0} - \log |\Omega| + \operatorname{tr}\left(\Omega \widehat{\Sigma}\right),$$

where $\Omega \succ 0$ means that we minimize over the cone of positive semi-definite matrices. To incorporate sparsity, the natural penalty is the elementwise ℓ^1-norm of the off-diagonal entries $\|\Omega\|_1 := \sum_{i<j} |\Omega_{i,j}|$. The constrained estimator is therefore the minimizer of the following convex optimization problem

$$\widehat{\Omega} \in \operatorname{argmin}_{\Omega \succ 0} - \log |\Omega| + \operatorname{tr}\left(\Omega \widehat{\Sigma}\right) + \lambda \|\Omega\|_1.$$

This estimator is also called the GLasso or graphical LASSO estimator. The analysis of this estimator is somehow similar to the LASSO. In particular, it requires an analog of the compatibility condition. For precise results, we refer to [18]. There are also other estimators which are easier to analyze theoretically, cf. [6, 19].

7.6 Matrix completion

In the matrix completion model with measurement errors, we observe few, noisy entries of a matrix and want to recover the full matrix under a low-rank constraint.

As before, define $[p] := \{1, \ldots, p\}$ and $[q] := \{1, \ldots, q\}$. Let $A = (a_{\ell,k})_{\ell \in [p], k \in [q]}$ be a $p \times q$ matrix with real valued entries. Suppose, we observe n independent variables (Y_i, L_i, K_i), where the indices L_i and K_i are independently drawn uniformly at random from $[p]$ and $[q]$, respectively and

$$Y_i = a_{L_i, K_i} + \varepsilon_i,$$

with $\varepsilon_i \sim \mathcal{N}(0, 1)$ independent of (L_i, K_i). In the matrix completion model, we thus see n entries which are perturbed by measurement noise. The MLE in this model is the minimizer of

$$\widehat{A} \in \operatorname{argmin}_{A=(a_{\ell,k})_{\ell \in [p], k \in [q]}} \sum_{i=1}^{n} (Y_i - a_{L_i, K_i})^2.$$

As long as not all entries are observed, the MLE is not uniquely defined.

From an applied point, it is natural to assume that the true matrix has low rank. This assumption is slightly different to the entrywise sparsity considered before. Nevertheless, a similar approach can be followed. A first attempt would be to penalize the rank of the matrix. If σ_j denotes the j-th singular value of A, the rank can also be written as $\#\{j : \sigma_j(A) \neq 0\}$. Penalization of the rank is therefore the same as ℓ^0-penalization of the singular values. Earlier, we argued that ℓ^0-penalization of the coefficient vector in high-dimensional regression does not lead to a convex optimization problem. The same argument applies here for ℓ^0-penalization of the singular value. A convex relaxation that is still sparsity inducing is to take the ℓ^1-norm of the singular values and to consider

$$\widehat{A} \in \operatorname{argmin}_{A=(a_{\ell,k})_{\ell \in [p], k \in [q]}} \sum_{i=1}^{n} (Y_i - a_{L_i, K_i})^2 + \lambda \sum_{j=1}^{p \wedge q} \sigma_j(A).$$

Theory for this estimator has been developed in [20].

7.7 Exercises

Ex. 7.1 — Suppose we observe $Y \sim \mathcal{N}(\theta, 1)$ with parameter space $\Theta = \mathbb{R}$. Let $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ be the density of a standard normal random variable. Prove that the risk of the hard thresholding estimator $\widehat{\theta}_\tau = Y\mathbf{1}(|Y| > \tau)$ with positive threshold value τ satisfies

(a) $E_\theta[(\widehat{\theta}_\tau - \theta)^2] \le 2(\tau^2 + 1)$,

(b) and for $\theta = 0$,

$$E_0\big[\widehat{\theta}_\tau^2\big] = 2\tau\phi(\tau) + 2\int_\tau^\infty \phi(x)dx \le 2\big(\tau + \tfrac{1}{\tau}\big)\phi(\tau).$$

Hint: For (b) use Stein's lemma and Exercise 6.9.

Ex. 7.2 — Using the previous exercise, prove (7.3.1) for $n > 1$. Compare the risk bound to the risk of the estimators $\widehat{\boldsymbol{\beta}}^{(1)} = 0$ and $\widehat{\boldsymbol{\beta}}^{(2)} = \mathbf{Y}^n$.

Ex. 7.3 — If $n = p$ and X is a $n \times n$ orthogonal matrix, the LASSO does soft-thresholding of $X^\top Y$. Show that for $i = 1, \dots, n$,

$$\widehat{\beta}_i^{\text{LASSO}} = \big((X^\top Y)_i - \lambda \operatorname{sign}((X^\top Y)_i)/2\big)\mathbf{1}\big(|(X^\top Y)_i| \ge \lambda/2\big).$$

Ex. 7.4 — Consider the LASSO estimator with penalty $\lambda > 2\max_{i=1,\dots,p} |X_i^\top \mathbf{Y}|$, where X_i denotes the i-th column of the design matrix X. Show that the LASSO functional has the unique solution

$$\widehat{\beta}^{\text{LASSO}} = 0.$$

Ex. 7.5 — Show that if $\lambda \ge 4\|X\|\sqrt{2\log p}$, then

$$P_{\boldsymbol{\beta}}\Big(\big\|X(\widehat{\boldsymbol{\beta}}^{\text{LASSO}} - \boldsymbol{\beta})\big\|_2^2 > \frac{3}{2}\lambda\|\boldsymbol{\beta}\|_1\Big) \to 0.$$

Hint: Argue as in the proof of Theorem 8.

Ex. 7.6 — Consider the space of piecewise constant functions $f_{\boldsymbol{\beta}}(t) = \sum_{k=1}^n \beta_k \mathbf{1}(t \le k/n)$. Recall the nonparametric regression model, where we observe $\mathbf{Y} = (Y_{1n}, \dots, Y_{nn})$ with

$$Y_{in} = f_{\boldsymbol{\beta}}\Big(\frac{i}{n}\Big) + \varepsilon_{i,n}, \quad i = 1, \dots, n, \;\; \varepsilon_{i,n} \sim \mathcal{N}(0,1), \text{ independent.}$$

(a) Show that this model can be rewritten as high-dimensional regression model $\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$ with $X = (\mathbf{1}(i \le j))_{i,j=1,\dots,n}$.

(b) Show that $X^\top X = (i \wedge j)_{i,j=1,\dots,n}$.

(c) Fix a sparse vector $\boldsymbol{\beta}$ and denote by $s = s_{\boldsymbol{\beta}}$ the sparsity and by $i_1 < i_2 < \dots < i_s$ the ordered indices in $S_{\boldsymbol{\beta}}$. Show that

$$\begin{aligned}\boldsymbol{\beta}^\top X^\top X \boldsymbol{\beta} &= i_1\big(\beta_{i_1} + \dots + \beta_{i_s}\big)^2 + (i_2 - i_1)\big(\beta_{i_2} + \dots + \beta_{i_s}\big)^2 + \dots + (i_s - i_{s-1})\beta_{i_s}^2 \\ &= n\|f_{\boldsymbol{\beta}}\|_{L^2[0,1]}^2.\end{aligned}$$

(d) Now, we study the compatibility number $\phi(S_{\boldsymbol{\beta}}, L)$. Suppose that there is a positive constant C, such that $|i_\ell - i_{\ell-1}| \leq Cn/s_{\boldsymbol{\beta}}$ for all $\ell = 1, \ldots, s$ with $i_0 := 0$. Show that for any $L \geq 0$, $\phi(S_{\boldsymbol{\beta}}, L)^2 \leq C/s_{\boldsymbol{\beta}}$.

Ex. 7.7 — Let $(W_t)_{t\geq 0}$ be a Brownian motion and consider the Gaussian vector $X = (W_1, \ldots, W_n)$. Show that the precision matrix is tridiagonal, that is, all non-zero entries are on the main diagonal and first off-diagonal. *Hint:* Write $W_j = (W_{j+1} + W_{j-1})/2 + 2^{-1/2}\eta_j$ with η_j being independent of $(W_1, \ldots, W_{j-1}, W_{j+1}, \ldots, W_n)$.

Chapter 8

Neural networks and deep learning

8.1 Machine learning and statistics

The main focus in machine learning is on methods for prediction and classification given a large number of covariates. Beyond that it is difficult to give a precise characterization of the difference between machine learning and statistics. The difference is often explained via examples. A standard machine learning example is the classification of an email as "spam" or "not spam". The dataset contains say n emails and for each email the corresponding label "spam" or "not spam". The statistical challenge is to estimate (in machine learning wording: to learn) a spam filter. The optimal spam filter is a function that takes as input an email and outputs the correct label.

To convert the problem into a statistical model, a first approach is to represent the i-th email in the dataset as a list of words $\mathbf{X}_i$. If Y_i denotes the corresponding label, the dataset consists then of pairs $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$. Written in this form, the covariates are the words. Words such as "we, today, ..." have none or very little influence on a message being spam or not. On the other hand there are words such as "casino, winner, ..." that often occur in spam messages. This means that we have a high-dimensional design vector of covariates but only a small subset is relevant.

Another distinctive property of a machine learning task is prediction. The spam filter will see a new email and has to predict whether this is spam. Viewed as a statistical problem, this means that the loss function should be the prediction loss. We will discuss this in more detail below. As for the spam example, machine learning applications have often difficult data structures such as text data.

To analyze machine learning methods, the non-parametric regression model is often taken as a benchmark. Recall that in this model, we observe n i.i.d. pairs $(\mathbf{X}_i, Y_i)$ with d-dimensional design vectors $\mathbf{X}_i$, real valued response variables Y_i and

$$Y_i = f(\mathbf{X}_i) + \tau\varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0,1), \quad i = 1, \ldots, n. \tag{8.1.1}$$

The regression function $f : \mathbb{R}^d \to \mathbb{R}$ is unknown. To account for small noise level in machine learning applications we introduced the noise level $\tau > 0$ and moreover allow for τ to decrease to zero with the sample size n.

In Theorem 2, we derived the rate of convergence of the kernel smoothing estimator for $d = 1$. Now, we are interested in the multivariate case with d potentially large.

The prediction problem is that we observe a new $\mathbf{X}$ with the same distribution as $\mathbf{X}_1, \dots, \mathbf{X}_n$ and want to predict the corresponding output Y. If $\widehat{f}$ is the estimator of f based on the data $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, the predicted value is $\widehat{Y} := \widehat{f}(\mathbf{X})$ and

$$E\big[(Y - \widehat{Y})^2\big] = \sigma^2 + E\big[(f(\mathbf{X}) - \widehat{f}(\mathbf{X}))^2\big],$$

where the expectation should be taken with respect to $(\mathbf{X}, Y)$ and $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$. This means that an estimator $\widehat{f}$ should be evaluated under the prediction risk

$$R(\widehat{f}, f) := E\big[(f(\mathbf{X}) - \widehat{f}(\mathbf{X}))^2\big].$$

8.2 Shallow neural networks

Neural networks are functions of a specific type. We first describe the function class and derive the approximation theory. Fitting neural networks to data is discussed later.

In this section we start with a specific form of neural networks, so called shallow neural networks. In a first step we need to choose a function $\sigma : \mathbb{R} \to \mathbb{R}$. This function is called the *activation function.* There are a few activation functions that are widely used in practice. In the early literature, the sign function $\sigma(x) = \operatorname{sign}(x)$ has been studied. This function has good approximation theoretic properties. Similar as ℓ^0-penalization it leads, however, to computationally infeasible problems.

In the eighties/nineties, smoother activation functions have been used that imitate the sign function. The class of sigmoidal activation functions consists of all functions σ that are continuous, strictly increasing and satisfy $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to+\infty} \sigma(x) = 1$. Within this class, the standard example is the logistic activation function $\sigma(x) = 1/(1 + e^{-x})$.

Recently, the so called ReLU activation function $\sigma(x) = \max(x, 0)$ became popular.

A shallow neural network with one output is a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(\mathbf{x}) = \sum_{j=1}^{m} c_j \sigma\big(\mathbf{a}_j^\top \mathbf{x} + b_j\big), \quad \mathbf{a}_j \in \mathbb{R}^d,\ b_j, c_j \in \mathbb{R}. \tag{8.2.1}$$

A shallow neural network depends on the number of terms in the sum m which is also called the number of hidden units. It moreover depends on the so called weight vectors $\mathbf{a}_j \in \mathbb{R}^d$, the shifts/biases b_j, and the coefficients c_j.

The activation function and the number of hidden units m will be fixed and $\mathbf{a}_j, b_j, c_j$ are considered as free parameters that will be estimated from the data. It is therefore natural to study the neural network function class of all functions that can be generated by varying the free parameters,

$$\mathcal{F}_{m,\sigma} := \Big\{ f = \sum_{i=1}^{m} c_j \sigma\big(\mathbf{a}_j^\top \cdot + b_j\big) \ :\ \mathbf{a}_j \in \mathbb{R}^d, b_j, c_j \in \mathbb{R} \Big\}. \tag{8.2.2}$$

For non-linear activation functions the function class $\mathcal{F}_{m,\sigma}$ is non-convex, see Exercise 8.1.

The MLE is the best least-squares fit to the data, that is,

$$\widehat{f} \in \operatorname{argmin}_{f \in \mathcal{F}_{m,\sigma}} \sum_{i=1}^{n} \big(Y_i - f(\mathbf{X}_i)\big)^2. \tag{8.2.3}$$

The minimization problem does not have an explicit solution. Since $\mathcal{F}_{m,\sigma}$ is non-convex the argmin is computationally intractable. Gradient descent algorithms are popular in practice that aim to decrease the objective functional $\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2$. The expression $\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f(\mathbf{X}_i)\big)^2$ is called the *empirical risk* since it is an estimator for the risk $R(\widehat{f}, f) := E[(f(\mathbf{X}) - \widehat{f}(\mathbf{X}))^2]$. The estimator $\widehat{f}$ is therefore also called the *empirical risk minimizer*.

If also the parameters $\mathbf{a}_j, b_j$, are fixed, the only parameters that need to be estimated from the data are the weights c_j. The least-squares estimator has then an explicit form. The problem is indeed very similar to series estimation as discussed in Section 5.2. [make Exercise ???] The main difference is that the basis functions for shallow networks are not necessarily orthogonal.

By varying $\mathbf{a}_j, b_j$ as well, the least-squares fit contains also a data driven search for good basis functions $\mathbf{x} \mapsto \sigma\big(\mathbf{a}_j^\top \mathbf{x} + b_j\big)$. Here, good means that a linear combination of these functions resembles f. Series estimation with data-driven (subset) selection of the basis function is often referred to as *dictionary learning*.

8.3 The universal approximation theorem

How large is the function space $\mathcal{F}_{m,\sigma}$? Differently speaking, how well can we approximate functions of a specific smoothness or the function $f(x_1, x_2) = x_1 x_2$ with functions in $\mathcal{F}_{m,\sigma}$?

Functions in the class $\mathcal{F}_{m,\sigma}$ have $m(d+2)$ real parameters since $\mathbf{a}_j$ are d-dimensional vectors. Moreover, the function classes are obviously nested in the sense that $\mathcal{F}_{m,\sigma} \subseteq \mathcal{F}_{m',\sigma}$ whenever $m' \geq m$.

In this section, we study whether for large m, any continuous function on $[0,1]^d$ can be eventually approximated up to a small sup-norm error.

Definition 12 (Universal approximation property)**.** *Shallow networks with activation function σ have the universal approximation property if for any $\varepsilon > 0$ and any continuous function f on $[0,1]^d$, there exists an integer $m = m(f, \varepsilon)$, such that*

$$\inf_{g \in \mathcal{F}_{m,\sigma}} \|f - g\|_{L^\infty([0,1]^d)} \leq \varepsilon.$$

For $d = 1$, the universal approximation property can be established for the ReLU and sigmoidal activation functions. For $d > 1$, this is much more difficult. It is for instance not clear whether a shallow network can approximate the multiplication of two numbers $f(x_1, x_2) = x_1 x_2$.

It is not very difficult to show that any $f \in L^2([0,1]^d)$ can be approximated up to L^2-error ε by a shallow network with activation function $\sigma = \cos(\cdot)$.

Lemma 12. *Let $f \in L^2([0,1]^d)$. Then, there exist $\widetilde{\mathbf{w}}_j \in \mathbb{R}^d$ and $\widetilde{c}_j \in \mathbb{R}$, $j \geq 1$, such that*

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} \widetilde{c}_j \cos(\widetilde{\mathbf{w}}_j^\top \mathbf{x})$$

and convergence of the sum is in L^2.

Proof. For univariate functions $h_1, \ldots, h_d$, we write $\bigotimes_{k=1}^d h_k$ for the function $(x_1, \ldots, x_d) \mapsto \prod_{k=1}^d h_k(x_k)$. If $\{\phi_j : j \geq 0\}$ is an ONB for $L^2[0,1]$, then, $\{\bigotimes_{k=1}^d h_{j_k} : j_1, \ldots, j_d \geq 0\}$ is an ONB for $L^2([0,1]^d)$.

By Lemma 16 (ii), any function $f \in L^2[0,1]^d$ can therefore be expanded in the tensorized cosine basis and $f(x_1, \ldots, x_d) = \sum_{(i_1,\ldots,i_d)\in\mathbb{N}^d} a_{i_1\ldots i_d} \prod_{j=1}^d \cos(i_j \pi x_j)$. Recall the addition theorem $\cos(u)\cos(v) = \frac{1}{2}(\cos(u+v) + \cos(u-v))$. Using this together with the fact that any function $f \in L^2[0,1]^d$ can be expanded in the tensorized cosine basis, we obtain $f(x_1, \ldots, x_d) = \sum_{(i_1,\ldots,i_d)\in\mathbb{N}^d} a_{i_1\ldots i_d} \prod_{j=1}^d \cos(i_j \pi x_j) = \sum_j \widetilde{c}_j \cos(\widetilde{\mathbf{w}}_j^\top \mathbf{x})$ for suitable $\widetilde{\mathbf{w}}_j \in \mathbb{R}^d$ and $\widetilde{c}_j \in \mathbb{R}$. □

There are many different proofs for the universal approximation theorem for shallow networks. One can argue via the Hahn-Banach theorem, the Radon transform, radial basis functions and the Fourier transform. The latter can be explained as follows.

Denote by $\mathcal{F}$ the Fourier transform, that is, $\mathcal{F}f(\boldsymbol{\xi}) = \int e^{-i\boldsymbol{\xi}^\top \mathbf{x}} f(\mathbf{x})\, d\mathbf{x}$. The inverse Fourier transform $\mathcal{F}^{-1}$ for a d-variate function is then $\mathcal{F}^{-1}f(\boldsymbol{\xi}) = (2\pi)^{-d} \int e^{i\mathbf{x}^\top \boldsymbol{\xi}} f(\boldsymbol{\xi})\, d\boldsymbol{\xi}$. In particular, we have for $f \in L^2$, that $f = \mathcal{F}^{-1}\mathcal{F}f$.

For any function $f \in L^2(\mathbb{R}^d)$, and any function $\phi \in L^1(\mathbb{R})$, with $\mathcal{F}\phi(1) \neq 1$,

$$f(x_1, \ldots, x_d) = \frac{1}{(2\pi)^d \mathcal{F}\phi(1)} \int_{\mathbb{R}^{d+1}} \phi\big(\boldsymbol{\xi}^\top \mathbf{x} + v\big) \mathcal{F}f(\boldsymbol{\xi}) e^{-iv}\, dv d\boldsymbol{\xi}.$$

The result follows from

$$\int_{\mathbb{R}} \phi\big(\boldsymbol{\xi}^\top \mathbf{x} + v\big) e^{-iv}\, dv = \mathcal{F}\phi(1) e^{i\boldsymbol{\xi}^\top \mathbf{x}}$$

and Fourier inversion.

By discretization of the integral we obtain an approximate representation of f as shallow network with activation function ϕ. Most activation functions are, however, not in $L^1(\mathbb{R})$. Instead we use that for sigmoidal activation functions $\sigma - \sigma(\cdot + \Delta)$ is typically in L^1 for any $\Delta > 0$. For the ReLU, $\sigma(x) - 2\sigma(x-1) + \sigma(x-2)$ is in L^1.

We state the next result without proof.

Theorem 9. *If the activation function σ is infinitely differentiable on an open neighborhood of a point, then the universal approximation property for shallow networks with activation function σ holds.*

8.4 Statistical analysis

We now provide some intuition for the empirical risk minimizer (8.2.3). Assume that the data are generated from the nonparametric regression model (8.1.1). For any $f^* \in \mathcal{F}_{m,\sigma}$, we have that

$$\sum_{i=1}^n (Y_i - \widehat{f}(\mathbf{X}_i))^2 \leq \sum_{i=1}^n (Y_i - f^*(\mathbf{X}_i))^2.$$

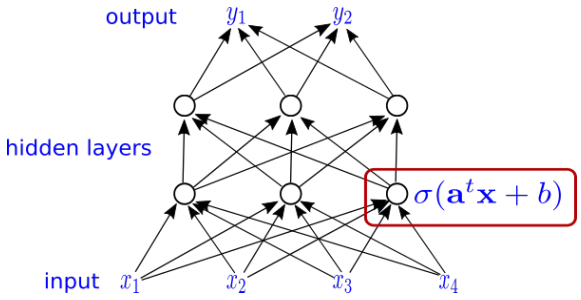


Figure 26: Representation as a direct graph of a network with two hidden layers $L = 2$ and width vector $\mathbf{p} = (4, 3, 3, 2)$.

Define the norm $\|g\|_n := (\frac{1}{n}\sum_{i=1}^n g(X_i)^2)^{1/2}$. Rewriting $Y_i - \widehat{f}(\mathbf{X}_i) = \tau\varepsilon_i + f(\mathbf{X}_i) - \widehat{f}(\mathbf{X}_i)$ and $Y_i - f^*(\mathbf{X}_i) = \tau\varepsilon_i + f(\mathbf{X}_i) - f^*(\mathbf{X}_i)$, the previous inequality leads to

$$\|f - \widehat{f}\|_n^2 \leq \|f^* - f\|_n^2 + \frac{2\tau}{n}\sum_{i=1}^n \varepsilon_i\big(\widehat{f}(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big).$$

[Explained in the lecture. Not yet in lecture notes] If the activation function is Lipschitz, we find that for some constant C

$$\|f - \widehat{f}\|_n^2 \leq C\|f^* - f\|_n^2 + C\frac{md\log n}{n},$$

with high probability. The first part can be viewed as the bias and the second as the stochastic error. It requires some extra effort to show that there exists a constant C', such that

$$R(\widehat{f}, f) \leq C'\|f^* - f\|_n^2 + C'\frac{md\log n}{n}$$

with high probability. Results of this type are also called *oracle inequalities.*

To bound the first term, the universal approximation property is not enough. Instead we need bounds on the approximation error in dependence on properties of the regression function f, such as smoothness.

8.5 Deep neural networks

After having discussed shallow networks, we will now continue with deep networks or multilayer neural networks.

Fitting a multilayer neural network requires again the choice of an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and a network architecture. For $\mathbf{v} = (v_1, \ldots, v_r) \in \mathbb{R}^r$, define the shifted activation

function $\sigma_{\mathbf{v}} : \mathbb{R}^r \to \mathbb{R}^r$ as

$$\sigma_{\mathbf{v}} \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_r - v_r) \end{pmatrix}.$$

The network architecture $(L, \mathbf{p})$ consists of a positive integer L called the *number of hidden layers* or *depth* and a *width vector* $\mathbf{p} = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$. A neural network with network architecture $(L, \mathbf{p})$ is then any function of the form

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad \mathbf{x} \mapsto f(\mathbf{x}) = W_L \sigma_{\mathbf{v}_L} W_L \sigma_{\mathbf{v}_{L-1}} \cdots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x}, \tag{8.5.1}$$

where W_i is a $p_{i+1} \times p_i$ weight matrix and $\mathbf{v}_i \in \mathbb{R}^{p_i}$ is a shift vector. Network functions are therefore build by alternating matrix-vector multiplications with the action of the non-linear activation function σ. In (8.5), it is also possible to omit the shift vectors by considering the input $(\mathbf{x}, 1)$ and enlarging the weight matrices by one row and one column with appropriate entries. To fit networks to data generated from the d-variate nonparametric regression model we must have $p_0 = d$ and $p_{L+1} = 1$.

In computer science, neural networks are more commonly introduced via their representation as directed acyclic graph, cf. Figure 26. Using this equivalent definition, the nodes in the graph (also called *units*) are arranged in layers. The input layer is the first layer and the output layer the last layer. The layers that lie in between are called hidden layers. The number of hidden layers corresponds to L and the number of units in each layer generates the width vector $\mathbf{p}$. Each node/unit in the graph representation stands for a scalar product of the incoming signal with a weight vector which is then shifted and applied to the activation function.

A shallow network with one output as defined in (8.2.1) is a special case of with $L = 1$, $p = (1, m, 1)$, $\mathbf{a}_i^\top$ the i-th column of W_1, b_i the i-th entry of v_1 and c_i the i-th entry of $W_2^\top$.

8.6 Exercises

Ex. 8.1 — A function class $\mathcal{F}$ is called convex if for any $f, g \in \mathcal{F}$ and any $\lambda \in [0, 1]$, $\lambda f + (1 - \lambda) g \in \mathcal{F}$. Show that for the ReLU activation function, (8.2.2) is not convex.

Ex. 8.2 — For $\sigma(x) = \max(x, 0)$ and $d = 1$, prove the universal approximation property.

Chapter 9

Lower bounds

It is natural to wonder about the optimality of the derived methods. For nonparametric methods, we have seen that different methods achieve the convergence rate $n^{-2\beta/(2\beta+1)}$ for squared loss. Is there a statistical procedure that does converge with a faster rate? We discuss a general strategy in this chapter to derive lower bounds for the estimation rate.

9.1 Superefficiency and minimax risk

One has to be a bit careful, what exactly one means with a lower bound on the rate of convergence since there are always estimators that are 'superefficient' for some values in the parameter space. A classical example due to Hodge illustrates this. Consider the parametric model, where we observe $Y_i \sim \mathcal{N}(\mu, 1)$, i.i.d. $i = 1, \ldots, n$. As an estimator for μ, we take $\widehat{\mu} = \overline{Y} = n^{-1} \sum Y_i \sim \mathcal{N}(\mu, 1/n)$ which has mean squared error $E_\mu[(\widehat{\mu} - \mu)^2] = 1/n$. It is somehow obvious that $1/n$ should be the best rate for any estimator. For each μ^* we can, however, "improve" the rate for $\mu = \mu^*$ using the estimator

$$\widetilde{\mu} = \begin{cases} \mu^*, & \text{if } |\overline{Y} - \mu^*| \le n^{-1/4}, \\ \overline{Y}, & \text{if } |\overline{Y} - \mu^*| > n^{-1/4}. \end{cases} \tag{9.1.1}$$

This estimator is a version of hard thresholding, returning μ^* whenever $\overline{Y}$ is in a $n^{-1/4}$-neighborhood of μ^*. From the construction, it is not surprising that $\widetilde{\mu}$ has a faster rate of convergence in μ^*. Using Exercise 9.1 (a), we obtain for the risk at $\mu = \mu^*$, $E_{\mu^*}[(\widetilde{\mu} - \mu^*)^2] \le n^{-3/4} e^{-\sqrt{n}/2}$. The risk at μ^* decays thus exponentially fast in n. On the contrary, for fixed $\mu \neq \mu^*$, $E_\mu[(\widetilde{\mu} - \mu)^2] = \frac{1}{n} + o(n^{-1})$ as $n \to \infty$, cf. Exercise 9.1 (b). Therefore, the rate of convergence for all $\mu \neq \mu^*$ is still n^{-1}. Hodges' estimator is only superefficient for $\mu = \mu^*$. But with a more refined construction, estimators can be obtained which are superefficient for many parameter values. In parametric problems, the set of superefficient parameters has Lebesgue measure zero. In nonparametric models, the situation is worse since estimators exist that are superefficient for every fixed parameter, cf. [3].

This suggests that the rate of convergence can be improved for nonparametric estimators, but a close inspection shows that superefficient estimators are not better. The gain of the convergence

rate in one parameter value comes at the expense of a deterioration of the rate in a shrinking neighborhood around this parameter. Studying the rate of convergence for fixed $\theta \in \Theta$ only, makes us blind to this effect since for large sample size each fixed parameter will eventually be outside of the shrinking neighborhood of any superefficient point and then the rate is unaffected by the superefficiency. If a parameter is very close to a superefficient parameter value, it takes however larger sample sizes until this happens. For such parameters the corresponding risk is large, until at some point the rate of convergence kicks in. Another way to formulate the effect of superefficiency is therefore that it improves the rate of convergence but it also causes existence of parameter values for which the risk decays with the correct rate only for arbitrary large sample size.

One way around superefficiency is to study the worst case risk, that is, the risk uniformly over the parameter space

$$\sup_{\theta\in\Theta} R(\widehat{\theta},\theta) = \sup_{\theta\in\Theta} E_\theta[\ell(\widehat{\theta},\theta)].$$

By introducing the supremum over the parameter space, we rule out that asymptotics might kick in for some parameter values arbitrarily late. Obviously, for the estimator $\widehat{\mu} = \overline{Y}$, the worst case risk is $\sup_\mu E_\mu[(\widehat{\mu}-\mu)^2] = 1/n$. But Hodges' estimator has a slower rate of convergence. Indeed,

$$\begin{aligned}
\sup_\mu E[(\widetilde{\mu}-\mu)^2] &\geq E_{\mu^*+n^{-1/4}}[(\widetilde{\mu}-\mu^*-n^{-1/4})^2] \\
&\geq \frac{1}{\sqrt{n}} P_{\mu^*+n^{-1/4}}\big(|\overline{Y}-\mu^*| \leq n^{-1/4}\big) \\
&= \frac{1}{\sqrt{n}}\big[\Phi(0)-\Phi(-2n^{1/4})\big] \\
&= \frac{1}{2\sqrt{n}} + o(1), \quad \text{as } n\to\infty.
\end{aligned}$$

The worst case risk of Hodges' estimator is thus at least of the order $n^{-1/2}$ and therefore converges much slower to zero than the rate n^{-1} obtained for the estimator $\widehat{\mu} = \overline{Y}$.

To avoid superefficiency phenomena, it is standard in nonparametric statistics to derive risk bounds uniformly over the parameter space. In particular, the upper bounds derived so far, are all worst case risk bounds, the only exception being the exact MSE in Section **??**.

The worst case risk does not only have advantages and might be misleading in some situations. It only says that somewhere on the parameter space a certain rate is attained. But it might indeed also be true that large subsets of parameter values are easier to estimate. This information is lost if the worst case risk is considered.

Definition 13. *Given a loss function ℓ, the best worst case risk or minimax risk is defined as*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\widehat{\theta},\theta) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} E_\theta[\ell(\widehat{\theta},\theta)],$$

where the infimum is taken over all estimators of θ.

(put the definition earlier???) The name minimax comes from the $\inf \sup$ in the definition. It also emphasizes the close connection to game theory. We might interpret the minimax risk as a two persons game, with payment $E_\theta[\ell(\widehat{\theta}, \theta)]$. The opponent picks a θ in the parameter space that maximizes the payment. The best strategy for the statistician is then to find the estimator minimizing $\sup_\theta E_\theta[\ell(\widehat{\theta}, \theta)]$.

The exact minimax risk is extremely difficult to compute. Explicit expressions have been found only for very few estimation problems. Instead we study the rate for models with sample size n growing to infinity. Then, for each n, we define the minimax risk as $R_n^* = \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} R(\widehat{\theta}_n, \theta)$ and the *minimax rate of convergence* as any sequence $(\psi_n)_n$ such that

$$R_n^* \asymp \psi_n.$$

To establish the minimax rate, we first construct an estimator $\widetilde{\theta}_n$ and derive its worst case risk $\sup_{\theta \in \Theta} R(\widetilde{\theta}_n, \theta)$. If this risk can be bounded from above by $C\overline{\psi}_n$ for a constant C that does not depend on n, then also

$$R_n^* \lesssim \overline{\psi}_n.$$

It remains to show that there is also a lower bound

$$R_n^* \gtrsim \psi_n.$$

In particular, if $\psi_n \asymp \overline{\psi}_n$ then, $\psi_n \asymp R_n^*$ is the minimax rate. Since Markov's inequality $P(X \geq \varepsilon) \leq E[X]/\varepsilon$ gives

$$P_\theta\big(\ell(\widehat{\theta}_n, \theta) \geq \psi_n\big) \leq \frac{E_\theta[\ell(\widehat{\theta}_n, \theta)]}{\psi_n}.$$

it is therefore enough to prove for the lower bound that there is a constant $c > 0$, such that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}_n, \theta) \geq \psi_n\big) \geq c > 0. \tag{9.1.2}$$

Here and in the following, we assume that the loss is such that $\{\ell(\widehat{\theta}_n, \theta) \geq \psi_n\}$ are measurable events.

In this chapter, we discuss general strategies how to establish an inequality of type (9.1.2).

9.2 The minimax error function

For probability measures $P_1, \ldots, P_M$ all defined on the same probability space $(\Omega, \mathcal{A})$, define the *minimax error function* as

$$T(P_1, \ldots, P_M) := \inf_{A_j \in \mathcal{A} \text{ disjoint}} \max_{j=1,\ldots,M} P_j(A_j^c), \tag{9.2.1}$$

where the infimum is taken over all disjoint sets $A_1, \dots, A_M$, that is, $A_j \cap A_k = \varnothing$ whenever $j \neq k$. Obviously, the infimum will be attained on an $\mathcal{A}$-measurable partition of Ω. Recall the definition of the total variation distance (Definition 1). For $M = 2$,

$$T(P,Q) = \inf_A \max(P(A^c), Q(A)) \geq \inf_A \max\left(\frac{P(A^c)+Q(A^c)}{2}, \frac{P(A)+Q(A)}{2}\right) - \frac{V(P,Q)}{2}$$
$$\geq \frac{1-V(P,Q)}{2}. \tag{9.2.2}$$

This inequality is often quite sharp. In Exercise 9.2 a condition on P and Q is given that ensures equality.

9.3 * Connection to cake division problems

Suppose we want to divide a divisible good among M persons. To be more concrete we think of this good as a cake. To be consistent with the previous notation we denote the cake by $\mathcal{X}$. All possible pieces that we can cut out from the cake are supposed to form a σ-algebra $\mathcal{A}$ on $\mathcal{X}$. As a utility function, we associate to the i-th person a probability measure P_i that quantifies how much person i likes any specific piece $A \in \mathcal{A}$. To get the full cake will therefore have utility one and to obtain nothing from the cake has utility zero. A larger piece of the cake will moreover result in higher utility. The fact that every person has its own utility function allows to model individual preferences. One person might not like the cherries on the cake and would therefore give more utility to a piece with more chocolate.

Define the partitions of $\mathcal{X}$ in M disjoint sets by $\mathcal{P} = \{(A_1, \dots, A_M) : \cup_j A_j = \mathcal{X}, A_i \cap A_j = \varnothing \text{ for all } i \neq j\}$. In cake division problems we ask how we need to split the cake such that it is in some sense fair. The answer depends heavily on the meaning of the word fair. The utilitarian approach is to maximize the total utility. This means that we search for the partition $(A_1, \dots, A_M)$ that maximizes $\sum_j P_j(A_j)$. This can be, however, very unfair in the sense that some persons might get very little or, if $M > 2$, even nothing. An alternative is to maximize the smallest utility, that is,

$$\sup_{(A_1,\dots,A_M)\in\mathcal{P}} \min_j P_j(A_j) = 1 - T(P_1, \dots, P_M),$$

where the equality follows directly from the definition of $T(P_1, \dots, P_M)$ in (9.2.1). This provides us with another interpretation of the quantity $T(P_1, \dots, P_M)$. We will derive several bounds for $T(P_1, \dots, P_M)$ which together with the identity can be used to derive bounds on the smallest utility.

The standard procedure for two agents is that person one splits the cake in two halves A and A^c with $P_1(A) = P_1(A^c) = 1/2$. The second person then picks piece A if $P_2(A) \geq P_2(A^c)$ and otherwise A^c. In the smallest utility sense, this method is suboptimal since person one always gets $1/2$ of the cake. Under the assumption of Example 9.2, both agents could obtain utility $(1 + V(P,Q))/2$.

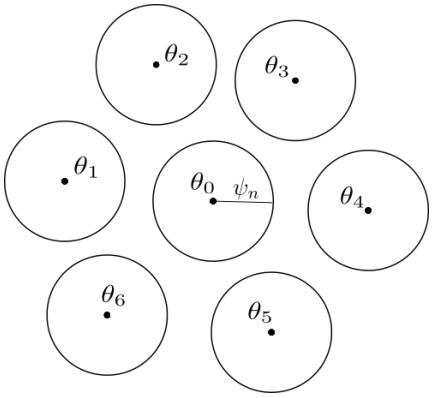


Figure 27: Lower bound argument

9.4 Reduction of lower bounds to information theoretic properties

A pseudo-distance on Θ is a map $d : \Theta \times \Theta \to [0, \infty)$ such that $d(\theta, \theta') \geq 0$, $d(\theta, \theta') = d(\theta', \theta)$ and $d(\theta, \theta') \leq d(\theta, \theta'') + d(\theta'', \theta')$ for all $\theta, \theta', \theta'' \in \Theta$. This notion is slightly weaker than a distance since it does not require that $d(\theta, \theta') = 0$ implies $\theta = \theta'$.

Theorem 10. *Suppose there are parameters $\theta_0, \theta_1, \ldots, \theta_M \in \Theta$, such that $\ell(\theta_j, \theta_k) \geq 2\psi_n$ for any non-equal pair $j, k \in \{0, 1, \ldots, M\}$, $j \neq k$. Further, assume that the loss function ℓ is a pseudo-distance. Then,*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \geq \psi_n\big) \geq T(P_{\theta_0}, \ldots, P_{\theta_M}).$$

Proof. By definition of a statistical experiment, all probability measures are defined on the same probability space $(\mathcal{X}, \mathcal{A})$. For any estimator $\widehat{\theta}$, we can study the measurable function $\mathcal{X} \to \Theta$, $\omega \mapsto \widehat{\theta}$. The sets $A_j = \{\omega : \ell(\widehat{\theta}, \theta_j) < \psi_n\}$ are measurable and disjoint since $\ell(\theta_i, \theta_j) \geq 2\psi_n$ and ℓ is a semi-metric, cf. Figure 27. The result thus follows from

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \geq \psi_n\big) \geq \inf_{\widehat{\theta}} \max_{j=0,\ldots,M} P_{\theta_j}\big(\ell(\widehat{\theta}, \theta_j) \geq \psi_n\big)$$

and the definition of $T(P_{\theta_0}, \ldots, P_{\theta_M})$ in (9.2.1). □

One should think of θ_j as all possible local perturbations of θ_0 in directions that affect the loss function. If we estimate a functional such as a function f at a fixed point x, we can only perturb it in finitely many directions such that the loss changes. In the case of pointwise estimation, we can put $\theta_0 = f_0(x)$ and $\theta_1 = f_0(x) + 2\psi_n$. In this case the following simplified lower bound is often sufficient.

Corollary 1. *Suppose that the loss function ℓ is a pseudo-distance. If there are parameters $\theta_0, \theta_1 \in \Theta$ with $\ell(\theta_0, \theta_1) \geq 2\psi_n$, then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \geq \psi_n\big) \geq \frac{1 - V\big(P_{\theta_0}, P_{\theta_1}\big)}{2} \geq \frac{1}{2} - \sqrt{\frac{K(P_{\theta_0}, P_{\theta_1})}{8}}.$$

Proof. The first inequality follows immediately from Theorem 10 and (9.2.2). The second inequality in the assertion is a consequence of Pinsker's inequality (2.6.1). □

In most cases, the second lower bound $\geq \frac{1}{2} - \sqrt{K(P_{\theta_0}, P_{\theta_1})/8}$ is typically easier to compute and from a practical point of view therefore more useful. Relating the statement of the corollary to constraint optimization, gives the following simple strategy to derive a lower bound. Take parameters $\theta_0, \theta_1 \in \Theta$ which are far apart with respect to the loss function, but such that $K(P_{\theta_0}, P_{\theta_1}) \leq 1$. The best choice is then given by the modulus of continuity

$$\omega(\Theta, \ell) = \sup_{\theta_0, \theta_1 \in \Theta} \big\{ \ell(\theta_0, \theta_1) : K(P_{\theta_0}, P_{\theta_1}) \leq 1 \big\}. \tag{9.4.1}$$

Under the assumptions of Corollary 1, $\omega(\Theta, \ell)/2$ is a lower bound and

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta \Big(\ell(\widehat{\theta}, \theta) \geq \frac{\omega(\Theta, \ell)}{2} \Big) \geq \frac{1}{2} - \frac{1}{2\sqrt{2}} > 0.$$

9.5 Lower bound for pointwise estimation

In this section, we derive a lower bound based on Corollary 1. For that we consider the Gaussian white noise model

$$dY_t = f(t)dt + n^{-1/2} dW_t, \quad t \in [0, 1] \tag{9.5.1}$$

(cf. (5.1.3)) and squared pointwise loss function $\ell^2(f, g) = (f(x) - g(x))^2$ with $x \in (0, 1)$ fixed. As parameter space, we consider the Höder ball $\mathcal{C}^\beta(L)$. In Exercise 5.2, it is shown that under standard assumptions, the mean squared error is $n^{-2\beta/(2\beta+1)}$.

Now, we prove that this is also the minimax rate of convergence. The first problem that we encounter is that the loss function $\ell^2(f, g) = (f(x) - g(x))^2$ does not satisfy the triangle inequality and is thus not a pseudo-distance. Due to $E^2 X \leq EX^2$,

$$\inf_{\widehat{f}} \sup_{f \in \Theta} E_\theta \big[\ell^2(\widehat{f}, f) \big] \geq \inf_{\widehat{f}} \sup_{f \in \Theta} E_\theta^2 \big[|\widehat{f}(x) - f(x)| \big]. \tag{9.5.2}$$

and it is therefore enough to prove that $n^{-\beta/(2\beta+1)}$ is a lower bound for the minimax rate with respect to the loss $\ell(f, g) = |f(x) - g(x)|$. Observe that ℓ is a pseudo-distance but not a distance since $\ell(f, g) = 0$ does not necessarily imply $f = g$.

Coming back to our lower bound, we know that the loss function is the pointwise loss $\ell(f, g) = |f(x) - g(x)|$ and the Kullback-Leibler is the squared $L^2[0, 1]$ norm scaled by a factor n. We noticed earlier that a good pair of parameters for the lower bound is far apart with respect to the loss and has Kullback-Leibler divergence at most one. A good ansatz is therefore $f_0 = 0$ and $f_1 = r_n(\cdot - x)$ for some function r_n. Then, $\ell(f_0, f_1) = |r_n(0)|$ and with Lemma 18, $K(P_{f_0}, P_{f_1}) = n\|r_n\|^2_{L^2[0,1]}$. We study functions r_n of the form, $r_n = L h_n^\beta K(x/h_n)$ with $h_n = cn^{-1/(2\beta+1)}$ for some constant $c > 0$.

Theorem 11. *Consider the Gaussian white noise model (9.5.1). Then, there exists a positive constant $A = A(\beta, L)$, such that*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{C}^\beta(L)} P_f\big(|\widehat{f}(x) - f(x)| \geq An^{-\frac{\beta}{2\beta+1}}\big) \geq \frac{1}{8}. \tag{9.5.3}$$

In particular, $n^{-2\beta/(2\beta+1)}$ is the minimax rate of convergence for squared pointwise loss function $\ell^2(f,g) = (f(x) - g(x))^2$.

Proof. We apply Corollary 1. Let $K \in L^2(\mathbb{R}) \cap \mathcal{C}^\beta(L)$ and $K(0) > 0$. Here, we require that K viewed as a function $\mathbb{R} \to \mathbb{R}$ should be in the Hölder ball $\mathcal{C}^\beta(L)$. An example of a function K with these properties should be found as exercise. Consider the two functions

$$f_0 = 0, \quad \text{and} \quad f_1 = h_n^\beta K\Big(\frac{\cdot - x}{h_n}\Big),$$

where $h_n = cn^{-1/(2\beta+1)}$ with $c := 1 \wedge \|K\|_{L^2(\mathbb{R})}^{-2/(2\beta+1)}$.

Notice that $f_0, f_1 \in \mathcal{C}^\beta(L)$. For f_0 this is trivial and for f_1 this follows from Lemma 2 (ii). With Lemma 18,

$$K(P_{f_0}, P_{f_1}) = \frac{n}{2}\|f_0 - f_1\|_{L^2[0,1]}^2 \leq \frac{n}{2}h_n^{2\beta+1}\|K\|_{L^2(\mathbb{R})}^2 = \frac{1}{2}c^{2\beta+1}\|K\|_{L^2(\mathbb{R})}^2 = \frac{1}{2}$$

and $\ell(f_0, f_1) = h_n^\beta K(0) = K(0)c^\beta n^{-\beta/(2\beta+1)}$. Thus, we can apply Corollary 1 with $A = K(0)(1 \wedge \|K\|_{L^2(\mathbb{R})}^{-2\beta/(2\beta+1)})/2$. This yields (9.5.3).

Using (9.1.2) and (9.5.2), we deduce that $n^{-\beta/(2\beta+1)}$ is a lower bound of the minimax rate for squared pointwise loss ℓ_2. Since this rate is also an upper bound, it must be the minimax rate. □

Similar constructions can be used to show that $n^{-2\beta/(2\beta+1)}$ is a lower bound of the minimax rate in density estimation for squared pointwise loss. Together with the upper bound in Theorem 1, this shows then that $n^{-2\beta/(2\beta+1)}$ is the minimax rate in this model and that the kernel density estimator with correctly chosen bandwidth is rate optimal.

9.6 Lower bounds with $M > 1$

The two parameter argument employed in the previous section only works for linear functionals. On the contrary, if we estimate for instance a function with respect to L^p-loss, we can construct a growing number of local perturbations around θ_0. To achieve the sharpest rate in the lower bound, the growth of the number of alternatives has to be taken into account and Theorem 10 needs to be applied with $M = M_n \to \infty$.

Theorem 12. *Let P_j, $j = 0, 1, \ldots, M$ be probability measures on the same probability space satisfying $P_j \ll P_0$. Then,*

$$T(P_0, \ldots, P_M) \geq \frac{M}{\kappa + M}\Big[1 - \frac{1}{M}\sum_{j=1}^{M} P_j\Big(\frac{dP_j}{dP_0} > \kappa\Big)\Big].$$

Proof. We prove the assertion by contradiction. If the statement is not true, then there exist disjoint sets $A_0, \dots, A_m$ such that $P_i(A_i^c) < B$ for all i, where $B = \frac{M}{\kappa+M}(1 - \frac{1}{M}\sum_{j=1}^M P_j(\frac{dP_j}{dP_0} > \kappa))$. We might assume that $B > 0$ since otherwise this is already a contradiction. Considering the cases $dP_j/dP_0 > \kappa$ and $dP_j/dP_0 \le \kappa$, separately and using $\cup_{j=1}^M A_j \subset A_0^c$,

$$\begin{aligned} M(1-B) < \sum_{j=1}^M P_j(A_j) &= \sum_{j=1}^M E_0\Big[\frac{dP_j}{dP_0}\mathbf{1}(A_j)\Big] \le \kappa E_0\big[\mathbf{1}(A_0^c)\big] + \sum_{j=1}^M E_0\Big[\frac{dP_j}{dP_0}\mathbf{1}\Big(\frac{dP_j}{dP_0} > \kappa\Big)\Big] \\ &< \kappa B + \sum_{j=1}^M P_j\Big(\frac{dP_j}{dP_0} > \kappa\Big) = M(1-B). \end{aligned}$$

Since $M(1-B) < M(1-B)$ is a contradiction, the assertion follows. □

Corollary 2. *Suppose there are parameters $\theta_0, \theta_1, \dots, \theta_M \in \Theta$, such that $\ell(\theta_j, \theta_k) \ge 2\psi_n$ for any non-equal pair $j, k \in \{0, 1, \dots, M\}$, $j \ne k$. Further, assume that the loss function ℓ is a pseudo-distance and that $P_{\theta_j} \ll P_{\theta_0}$, for $j \in \{1, \dots, M\}$. Then for any $\kappa > 0$,*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \ge \psi_n\big) \ge \frac{M}{\kappa + M}\Big[1 - \frac{1}{M}\sum_{j=1}^M P_{\theta_j}\Big(\frac{dP_{\theta_j}}{dP_{\theta_0}} > \kappa\Big)\Big].$$

Proof. The result follows from Theorem 10 and Theorem 12. □

Next, we relate lower bounds for $M > 1$ to the Kullback-Leibler divergence. As a preliminary result, we need the following version of Pinsker's second inequality (cf. p. 88 in [28])

$$\int \Big(\log\frac{dP}{dQ}\Big)_+ dP \le K(P,Q) + \sqrt{\frac{K(P,Q)}{2}} \le \frac{5}{4}K(P,Q) + \frac{1}{2} \tag{9.6.1}$$

where for the last inequality, we used that $\sqrt{a} \le (a+1)/2$ for all $a > 0$.

Theorem 13. *Suppose there are parameters $\theta_0, \theta_1, \dots, \theta_M \in \Theta$, such that $\ell(\theta_j, \theta_k) \ge 2\psi_n$ for any non-equal pair $j, k \in \{1, \dots, M\}$, $j \ne k$. Further, assume that the loss function ℓ is a pseudo-distance and that $P_{\theta_j} \ll P_{\theta_0}$, for $j \in \{1, \dots, M\}$. Then, for $M \ge 9$,*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \ge \psi_n\big) \ge \frac{3}{8}\Big(1 - 5\frac{\sum_{j=1}^M K(P_{\theta_j}, P_{\theta_0})}{M\log M}\Big).$$

Proof. This is another consequence of Theorem 10. Write $P_j = P_{\theta_j}$ and use Markov's inequality as well as (9.6.1),

$$P_j\Big(\frac{dP_j}{dP_0} > \sqrt{M}\Big) = P_j\Big(\log\frac{dP_j}{dP_0} > \frac{1}{2}\log M\Big) \le \frac{2}{\log M}\int\Big(\log\frac{dP_j}{dP_0}\Big)_+ dP_j \le \frac{2.5K(P_j, P_0) + 1}{\log M}.$$

Applying Theorem 10 with $\kappa = \sqrt{M}$ gives

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \ge \psi_n\big) \ge \frac{\sqrt{M}}{1+\sqrt{M}}\Big[1 - \frac{1}{\log M} - 2.5\frac{\sum_{j=1}^M K(P_j, P_0)}{M\log M}\Big].$$

Recall that $M \geq 9$. Thus, $\log(M) \geq 2$. Since $x \mapsto x/(1+x)$ is increasing for $x > 0$, we have also $\sqrt{M}/(1+\sqrt{M}) \geq 3/4$ and this shows that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \geq \psi_n\big) \geq \frac{3}{4}\Big(\frac{1}{2} - 2.5\frac{\sum_{j=1}^M K(P_j, P_0)}{M \log M}\Big) = \frac{3}{8}\Big(1 - 5\frac{\sum_{j=1}^M K(P_j, P_0)}{M \log M}\Big).$$

□

The restriction that $M \geq 9$ is not severe, since M is typically chosen as a sequence that tends to $+\infty$ as $n \to \infty$. The theorem provides us with the lower bound ψ_n on the minimax rate, whenever

$$\frac{1}{M}\sum_{j=1}^M K(P_{\theta_j}, P_{\theta_0}) \leq \gamma \log M, \quad \text{for some } \gamma < \frac{1}{5}.$$

It is also enough to ask that $K(P_{\theta_j}, P_{\theta_0}) \leq \gamma \log M$ for all $1 \leq j \leq M$. This should be compared to Corollary 1, which states that ψ_n is a lower bound if $K(P_{\theta_1}, P_{\theta_0}) < \sqrt{2}$. Obviously, if M tends to infinity, $\sqrt{2} \ll \gamma \log M$ and therefore the lower bound based on Theorem 13 is in some cases weaker.

9.7 Lower bounds in supremum norm

In the Gaussian white noise model, the wavelet thresholding estimator achieves by Theorem 5 the rate of convergence $(n/\log n)^{-\beta/(2\beta+1)}$, uniformly over the Hölder ball $\mathcal{C}^\beta(L)$ and with respect to the sup-norm loss. In this section we establish a lower bound based on Theorem 13 which shows that $(n/\log n)^{-\beta/(2\beta+1)}$ is also the minimax estimation rate. For this setting, wavelet thresholding is therefore a rate optimal estimator in the minimax sense and the additional $\log n$ factor in the rate of convergence for supremum norm loss is unavoidable.

Theorem 14. *Consider the Gaussian white noise model (9.5.1). Then, there exists a positive constant $A = A(\beta, L)$, such that for all sufficiently large n,*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{C}^\beta(L)} P_f\Big(\|\widehat{f} - f\|_{L^\infty[0,1]} \geq A\Big(\frac{n}{\log n}\Big)^{-\frac{\beta}{2\beta+1}}\Big) \geq \frac{1}{16}. \tag{9.7.1}$$

In particular, $(n/\log n)^{-\beta/(2\beta+1)}$ is the minimax rate of convergence for supremum norm loss.

Proof. We apply Theorem 13 and write $P_j = P_{\theta_j}$. Take a kernel $K \in L^2(\mathbb{R}) \cap \mathcal{C}^\beta(L)$ with $\|K\|_{L^\infty} > 0$ and support of K in $[-1/2, 1/2]$. Denote by $\lceil y \rceil$ the smallest integer larger than x. Let

$$M_n = \Big\lceil c\Big(\frac{n}{\log n}\Big)^{\frac{1}{2\beta+1}} \Big\rceil$$

with $c^{-1} = 1 \wedge (6(2\beta+2)\|K\|_2^2)^{-1/(2\beta+1)}$ and $h_n = 1/M_n$. Set $x_j = (j - \frac{1}{2})h_n$ and consider the functions

$$f_0 = 0 \quad \text{and} \quad f_j = h_n^\beta K\Big(\frac{\cdot - x_j}{h_n}\Big), \quad j = 1, \dots, M_n.$$

By Lemma 2 (ii), $f_j \in \mathcal{C}^\beta(L)$ for $j = 0, 1, \dots, M$. For the Kullback-Leibler divergence, observe that for any sufficiently large n,

$$\log(M_n) \geq \log\Big(c\Big(\frac{n}{\log n}\Big)\Big) = \log c + \frac{\log n - \log\log n}{2\beta+1} \geq \frac{\log n}{2\beta+2}.$$

Toegether with Lemma 18 and substitution

$$K(P_j, P_0) = n\|f_j\|_2^2 = nh_n^{2\beta+1}\int K^2(u)\,du \leq \Big(\frac{1}{c}\Big)^{2\beta+1}\int K^2(u)\,du \log n \leq \frac{\log n}{6(2\beta+2)} \leq \frac{1}{6}\log M_n$$

for all sufficiently large n.

By construction, the functions f_j $j = 1, \dots, M_n$ have all disjoint support. For $j \neq k$, and n large enough,

$$\ell(f_j, f_k) = \sup_x |f_j(x) - f_k(x)| \geq \sup_x |f_j(x)| \geq \Big(\frac{1}{2c}\Big)^\beta \|K\|_\infty \Big(\frac{n}{\log n}\Big)^{-\frac{\beta}{2\beta+1}}.$$

Thus, we can apply Theorem 13 and obtain (9.7.1) with $A = (2c)^{-\beta}\|K\|_\infty/2$. Together with the upper bound in Theorem 5, this yields the minimax estimation rate $(n/\log n)^{-\beta/(2\beta+1)}$. □

9.8 Exercises

Ex. 9.1 — Consider the Hodge estimator (9.1.1).

(a) Using Exercise 14.1, show that for the risk at $\mu = \mu^*$,

$$E_{\mu^*}[(\widetilde{\mu} - \mu^*)^2] = 2n^{-3/4}\phi(n^{1/4}) - 2n^{-1}\Phi(-n^{1/4}) \leq n^{-3/4}e^{-\sqrt{n}/2},$$

with Φ the c.d.f. of the standard normal distribution and $\phi = \Phi'$.

(b) For $\mu \neq \mu^*$, prove

$$E_\mu[(\widetilde{\mu} - \mu)^2] \leq \frac{1}{n} + (\mu^* - \mu)^2 P_\mu(|\overline{Y} - \mu^*| \leq n^{-1/4}) = \frac{1}{n} + o(n^{-1}), \quad \text{as } n \to \infty,$$

using that $P(|\xi + a| \leq b) = \Phi(b - a) - \Phi(-b - a)$ for $\xi \sim \mathcal{N}(0, 1)$ and real numbers a, b.

Ex. 9.2 — Let P, Q be two probability measures on the same measurable space with $\mu = (P + Q)/2$-densities p and q, respectively. Prove that if $P(q(X) \geq p(X)) = Q(p(X) > q(X))$ then there is equality in (9.2.2), that is,

$$T(P, Q) = \frac{1 - V(P, Q)}{2}.$$

Show that the condition holds for $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(\mu', 1)$ with $\mu \neq \mu'$.

Ex. 9.3 — Let $M \geq 1$ and $\Theta = \{\theta_0, \ldots, \theta_M\}$. If $\ell(\theta, \theta') = \mathbf{1}(\theta \neq \theta')$, then

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(\ell(\widehat{\theta}, \theta) \geq 1\big) = T(P_{\theta_0}, \ldots, P_{\theta_M}).$$

Ex. 9.4 — Let $P_1, \ldots, P_M$ be probability measures on the same probability space. Show that

$$T(P_1, \ldots, P_M) \geq T(P_1, \ldots, P_{M-1}).$$

Part II

Appendix

Chapter 10

A quick introduction to measure theory

10.1 Measures and measurable functions

Measure theory underlies probability theory which is the mathematical basis of statistics. This section is intended to give a very brief introduction of the relevant concepts from measure theory. This section has been added to provide a short and concise description of measure theory for statisticians. There are many excellent references including Chapter 1 in [22] and the book [26].

In the beginning of the 20th century it has been observed that it is impossible to find a function that assigns to each subset of the Euclidean space $\mathbb{R}^d$ a number that measures the size/volume of the set. The Banach-Tarski paradox says that one can decompose a ball B in three dimensional space in finitely many pieces such that the pieces can be reassembled to yield two identical copies of B. This means that we can double the volume of B by re-arranging the pieces. This highly non-intuitive result is based on a decomposition into pathological sets. To circumvent this problem, measures can only be defined on sets which are not too irregular. On the contrary, the system of measurable sets should be rich enough and should obey some structural assumptions. It is now common to define measures on σ-algebras (also called σ-fields). In the following Ω will always denote the underlying space and $\mathcal{A}$ is the collection of measurable sets. The empty set is denoted by $\varnothing$.

Definition 14 (σ-algebra)**.** *Let $\mathcal{A}$ be a collection of subsets of Ω. We call $\mathcal{A}$ a σ-algebra (on Ω) if the following three conditions hold:*

(i) $\varnothing \in \mathcal{A}$

(ii) $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$

(iii) $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \cup_i A_i \in \mathcal{A}$.

This concept is quite similar to the definition of a topology on a space. Indeed, a collection of subsets τ (called open sets) defines a topology of Ω if $\varnothing, \Omega \in \tau$ and τ is stable under countable unions and finite intersections. Given a topological space (Ω, τ), we define the Borel σ-algebra $\mathcal{B}(\Omega)$ as the smallest σ-algebra which contains the open sets. A running example is the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ generated from the open sets on $\mathbb{R}$.

If $\mathcal{A}$ is a σ-algebra on the space Ω, the pair $(\Omega, \mathcal{A})$ is called a *measurable space* and the sets $A \in \mathcal{A}$ are called *measurable sets.* A measurable space can be equipped with a measure.

Definition 15 (measure)**.** *A function $m : \mathcal{A} \to [0, \infty]$ is called a measure if*

(i) $m(\varnothing) = 0$

(ii) For any sequence $A_1, A_2, \ldots \in \mathcal{A}$ of disjoint sets, we have $m(\cup_i A_i) = \sum_i m(A_i)$.

An example is the Lebesgue measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. It generalizes the notion of length, area and volume in $\mathbb{R}$, $\mathbb{R}^2$ and $\mathbb{R}^3$. The Lebesgue measure is defined for any hyperrectangle $A := (a_1, b_1) \times \ldots \times (a_d, b_d)$ with $a_i \leq b_i$,

$$\lambda(A) = \prod_{i=1}^{d} (b_i - a_i)$$

and is then extended to all Borel sets $\mathcal{B}(\mathbb{R}^d)$. As a second example of a measure consider the indicator function defined as $\delta_x(A) = \mathbf{1}(x \in A)$ for all $A \in \mathcal{A}$.

A measurable set A is called a *m-null set* (or just a null set if it is clear to which measure m is referred to) if $m(A) = 0$. Any discrete set of finitely many points is for instance a null set with respect to the Lebesgue measure. Null sets can be large, for the measure δ_x above, $\delta_x(A) = 0$ whenever $x \notin A$.

A statement that holds up to a m-null set is said to hold *almost everywhere*. Under the Lebesgue measure λ on $\mathbb{R}$, it holds that x is irrational for almost every x since the set of rational points is a λ-null set.

The triple $(\Omega, \mathcal{A}, m)$ is called a *measure space.* Notice the difference between measure space and measurable space. In statistics, Ω will be the sample space. A probability measure is a measure m satisfying $m(\Omega) = 1$. A measure space with m a probability measure is called a probability space. If a measure is a probability measure, we often write $P, Q, \ldots$. For a probability measure P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ we define the cumulative distribution function (c.d.f.) as

$$x \mapsto F(x) = P((-\infty, x]). \tag{10.1.1}$$

Besides σ-algebras and measures, the third pillar of measure theory are measurable functions. For probability measures, these functions correspond to random elements.

For two measurable spaces $(\Omega, \mathcal{A})$ and $(E, \mathcal{A}')$ consider the function $f : \Omega \to E$. Even if f is not invertible in the classical sense, we can define the inverse image as the set of all elements in Ω that results in a specific value of f, that is, $f^{-1}(A') := \{\omega \in \Omega : f(\omega) \in A'\}$, for all $A' \in \mathcal{A}'$.

Definition 16 (measurable function)**.** *Given two measurable spaces $(\Omega, \mathcal{A})$ and $(E, \mathcal{A}')$, a function $f : \Omega \to E$ is called measurable if $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{A}'$.*

In statistics, we call such a measurable function an *E-valued random variable.* A measurable function is called a *random vector* if $(E, \mathcal{A}') = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-algebra on $\mathbb{R}^d$. In the specific case $d = 1$, measurable functions are called *random variables* or Borel measurable functions. Random elements are typically denoted by capital letters $X, Y, \ldots$.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and X a random element. Then, we can define a probability measure P_X on the measurable space $(E, \mathcal{A}')$ via $P_X(A) = P(X^{-1}(A))$ for all $A \in \mathcal{A}'$. We then call P_X the distribution of X and write $X \sim P_X$.

10.2 Lebesgue integration

Let $(\Omega, \mathcal{A}, m)$ be a measure space. If $\phi = \sum_{j=1}^{k} a_j \mathbf{1}_{A_j}$ with $A_j \in \mathcal{A}$ and $a_j \in \mathbb{R}$, then, we can define

$$\int \phi \, dm := \sum_{j=1}^{k} a_j m(A_j).$$

For a non-negative function f the *Lebesgue integral* is defined as

$$\int f dm := \sup \Big\{ \int \phi \, dm \, : \, \phi = \sum_{j=1}^{k} a_j \mathbf{1}_{A_j} \text{ with } a_j \geq 0, \phi \leq f, k = 1, 2, \dots \Big\}.$$

For an arbitrary real-valued function f, we can define the Lebesgue integral via the decomposition in a negative and a non-negative part $f = f_+ - f_-$ with $f_+, f_- \geq 0$. If m is the Lebesgue measure, we also write $\int \phi(x) dx$. It is also common to make the dependence on Ω explicit by writing $\int_\Omega f dm$ for $\int f dm$. We also define $\int_A f dm := \int_\Omega \mathbf{1}_A f dm$ for $A \in \mathcal{A}$.

There are many functions f such that $\int f dm = 0$. It can be shown that all these functions can, however, only differ from each other on a set of measure zero with respect to m. For given measure m we can define norms if we look at equivalence classes, where two functions are in the same class if they only differ on a m-null set. The following should be understood in terms of these equivalence classes. Let

$$\|f\|_{L^p(\Omega,m)} := \Big(\int |f|^p dm \Big)^{1/p}.$$

It can be shown that for $1 \leq p < \infty$ this defines a norm on the so called *Lebesgue spaces*

$$L^p(\Omega, m) = \big\{ f \, : \, \|f\|_{L^p(\Omega,m)} < \infty \big\}.$$

In many cases m is the Lebesgue measure and Ω is an interval or $\mathbb{R}^d$. It is common to shorten then the notation. More generally, whenever it is obvious which Ω and m is used we omit it from the notation and write L^p or $L^p(\Omega)$ instead of $L^p(\Omega, m)$ and $\|f\|_{L^p}$ or $\|f\|_{L^p(\Omega)}$ instead of $\|f\|_{L^p(\Omega,m)}$. It is moreover common to write $L^P\Omega$ for $L^p(\Omega)$ if Ω is an interval or a hypercube.

Consider a probability space $(\Omega, \mathcal{A}, P)$ and let X be a random variable. If $X \in L^1(\Omega, P)$, the expectation of X is $E[X] := \int X(\omega) dP(\omega)$ and if $X \in L^2(\Omega, P)$, the variance is $\mathrm{Var}(X) := E[(X - E[X])^2] = \int (X(\omega) - E[X])^2 dP(\omega)$.

Below, we mention several results that we need to develop the statistical theory. The first result extends Leibniz' rule to possibly improper integrals.

Lemma 13 (Interchanging differentiation and integration in parameter integrals)**.** *Let $A \in \mathcal{B}(\mathbb{R})$ and $f : A \times [a, b] \to \mathbb{R}$. If*

(i) for any $\theta \in [a,b]$ the function $E \ni x \mapsto f(\theta, x)$ is Borel measurable,

(ii) $x \mapsto f(\theta, x)$ is integrable over A for some $\theta \in [a,b]$,

(iii) the partial derivatives $\partial_\theta f(x,\theta)$ exist almost everywhere on $[a,b] \times A$

(iv) there exists a function h such that $|\partial_\theta f(\theta, x)| \leq h(x)$ for all $x \in A$ and all $\theta \in [a,b]$ with $\int_A h(x)dx < \infty$.

Then, we can interchange integration and differentiation in the following sense

$$\frac{d}{d\theta} \int_A f(\theta, x)dx = \int_A \partial_\theta f(\theta, x)dx.$$

Proof. This follows from the dominated convergence theorem. For a full proof, cf. [24], Proposition 11.2.4 and Notation 10.1.12. □

The integration by parts formula is typically stated for bounded intervals. The next result provides also an useful extension to the real line.

Lemma 14 (Integration by parts)**.** *Suppose that $a < b$ and $f, g \in L^1([a,b])$. Then for any F, G with $F' = f$ and $G' = g$,*

$$\int_a^b F(x)g(x)\,dx = F(x)G(x)\Big|_a^b - \int_a^b f(x)\,G(x).$$

If $f, g \in L^1(\mathbb{R})$, then also

$$\int_{-\infty}^{\infty} F(x)g(x)\,dx = FG(\infty) - FG(-\infty) - \int_{-\infty}^{\infty} f(x)G(x)dx.$$

Proof. The first statement follows from [24], Theorem 7.5.15. For the second statement, observe that $f, g \in L^1(\mathbb{R})$ implies that $F, G \in L^\infty(\mathbb{R})$ and thus $Fg, fG \in L^1(\mathbb{R})$. Moreover, since for $x < y$, $|F(x) - F(y)| \leq \int_x^y |f(u)|du$, the limits $F(-\infty) := \lim_{x\to-\infty} F(x)$ and $F(\infty) := \lim_{x\to\infty} F(x)$ exist and are finite. The same holds of course also for G. The second statement can then be obtained from the first by letting $a \to -\infty$ and $b \to \infty$. □

Theorem 15 (Lebesgue differentiation theorem)**.** *Let $f \in L^1(\mathbb{R}^d, \lambda)$ with λ the Lebesgue measure on $\mathbb{R}^d$. Denote by $B_r(\mathbf{x}) \subset \mathbb{R}^d$ the Euclidean ball centered at $\mathbf{x} \in \mathbb{R}^d$ with radius $r > 0$ and let $\mathrm{Vol}(B_\varepsilon(\mathbf{x}))$ be the corresponding volume. Then,*

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\mathrm{Vol}(B_\varepsilon(\mathbf{x}))} \int_{B_\varepsilon(\mathbf{x})} f(\mathbf{y})\,d\mathbf{y}$$

exists and is equal to $f(\mathbf{y})$ almost everywhere.

For more details see [26], Section 1.6.2.

10.3 The Radon-Nikodym derivative

Let $(\Omega, \mathcal{A})$ be a measure space and consider two measures μ, ν on this space. Suppose that for any $A \in \mathcal{A}$, $\nu(A) = 0$ if $\mu(A) = 0$. Then we say that μ is a *dominating measure* (of ν) and write $\nu \ll \mu$.

A measure m is called a *σ-finite measure* if Ω can be represented as countable union of measurable sets with finite measure. For the Lebesgue measure $\lambda(\mathbb{R}) = \infty$ but $\mathbb{R} = \cup_{k\in\mathbb{Z}}[k, k+1]$ and the Lebesgue measure is consequently σ-finite.

Theorem 16 (Radon-Nikodym derivative)**.** *If ν, μ are σ-finite measures and $\nu \ll \mu$, then there exists a non-negative Borel function f such that*

$$\nu(A) = \int_A f \, d\mu, \quad \textit{for all } A \in \mathcal{A}.$$

The function f is referred to as *Radon-Nikodym derivative* and we write for f also $d\nu/d\mu$. From the definition of the Lebesgue integral it is obvious that $\nu \ll \mu$ is also a necessary condition.

Chapter 11

Hilbert spaces

11.1 Definition

A metric space $\mathcal{F}$ is called complete if every Cauchy sequence has a limit in $\mathcal{F}$. If $\mathcal{F}$ is a complete space and $\|\cdot\|$ a norm on $\mathcal{F}$, then $(\mathcal{F}, \|\cdot\|)$ is called Banach space.

Definition 17 (Hilbert space)**.** *If $\mathbb{H}$ is a complete vector space and $\langle\cdot,\cdot\rangle$ is a scalar product on $\mathbb{H}$, then $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ is called a Hilbert space.*

The scalar product induces the norm $\|f\|_{\mathbb{H}} := \sqrt{\langle f, f\rangle}$. Using Cauchy-Schwarz for the triangle inequality, it can be checked that this is indeed a norm. In particular, if $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ is a Hilbert space, then $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is a Banach space.

Definition 18. *Let $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ be a Hilbert space and I be an index set. If $\{\phi_k\}_{k\in I} \subset \mathbb{H}$ satisfies*
(i) $\langle\phi_j, \phi_k\rangle = \mathbf{1}(j \neq k)$ *for all* $j, k \in I$,
(ii) $\overline{\text{span}}\{\phi_k\}_{k\in I} = \mathbb{H}$,
then, $\{\phi_k\}_{k\in I}$ is called an orthonormal basis (ONB) of $\mathbb{H}$.

A topological space is called separable, if it contains a countable and dense subset. It is well-known that a Hilbert space is separable if and only if it has an orthonormal basis with countable index set. If $\{\phi_k\}_{k\in I}$ is an orthonormal basis of a separable $\mathbb{H}$ then, for any $f \in \mathbb{H}$, there exists a unique $(c_k)_{k\in I} \in \mathbb{R}^{|I|}$ with

$$f = \sum_{k\in I} c_k\phi_k \quad \text{and} \quad c_k = \langle f, \phi_k\rangle. \tag{11.1.1}$$

If I is an infinite set, the right hand side converges with respect to the norm $\|\cdot\|_{\mathbb{H}}$.

11.2 L^2-spaces

Lemma 15. *For any $-\infty \leq a < b \leq \infty$, define the inner product $\langle f, g\rangle = \int_a^b f(x)g(x)dx$. Then, $(L^2[a,b], \langle\cdot,\cdot\rangle)$ is a separable Hilbert space.*

Lemma 16. *Each of the following collection of functions is an orthonormal basis of the Hilbert space* $(L^2[0,1], \langle\cdot,\cdot\rangle)$.

(i) $\{1, \sqrt{2}\cos(2k\pi\cdot), \sqrt{2}\sin(2k\pi\cdot), k = 1, 2, \ldots\}$,

(ii) $\{1, \sqrt{2}\cos(k\pi\cdot), k = 1, 2, \ldots\}$,

(iii) $\{1, \sqrt{2}\sin(k\pi\cdot), k = 1, 2, \ldots\}$.

The bases are referred as trigonometric basis (i), cosine basis (ii) and sine basis (iii).

11.3 Exercises

Ex. 11.1 — Assume that $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ is a separated Hilbert space. Prove that $\{\phi_k\}_{k\in I} \subset \mathbb{H}$ is an orthonormal basis of $\mathbb{H}$ if

(i)$\langle\phi_j, \phi_k\rangle = \mathbf{1}(j \neq k)$ for all $j, k \in I$,

(ii)$\langle f, \phi_k\rangle = 0$ for all $k \in I$ implies $f = 0$.

Chapter 12

Distributions, densities and the maximum likelihood principle

There are various specific classes of distributions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that appear frequently in statistics. One option is to introduce them via the c.d.f. in (10.1.1). Because of nicer expressions, it is however more convenient to take the Radon-Nikodym derivative of F with respect to an appropriate measure.

For probability measures P supported on integers we can take the Radon-Nikodym derivative with respect to the measure $m(A) = \sum_{i=-\infty}^{\infty} \mathbf{1}(i \in A)$ which yields $dP/dm(i) = P(\{i\}) =: p_i$ for all $i \in \mathbb{Z}$. The sequence $(p_i)_i$ is called *probability mass function* (p.m.f.). Observe that the p_i are probabilities and $\sum_i p_i = 1$.

An example is the *Bernoulli distribution* modelling a possibly unfair coin flip. A random variable from the Bernoulli distribution with parameter p attains the value 0 with probability $1-p$ and the value 1 with probability p. The p.m.f. is therefore $p_0 = 1-p, p_1 = p$ and $p_i = 0$ for $i \in \mathbb{Z} \setminus \{0, 1\}$.

A second example is the *Poisson distribution* $\mathrm{Poisson}(\lambda)$ with parameter $\lambda > 0$. It can be defined via its p.m.f. $p_i = e^{-\lambda}\lambda^i/i!$ for $i = 0, 1, \ldots$ and $p_i = 0$ for $i < 0$. This distribution occurs in many applications where a number of waiting times is observed. For instance counting the number of photons in some detector is typically Poisson distributed.

A random variable X is said to have a continuous distribution if P_X is dominated by the Lebesgue measure on $\mathbb{R}$. By the Radon-Nikodym theorem, this means that there is a function f such that $P_X(A) = \int_A f(x)dx$. The function f is called the *probability density function* (p.d.f.) of X. The p.d.f. is non-negative and integrates to one.

An example of a continuous random variable is the *normal distribution* (or Gaussian distribution) $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The p.d.f. of the normal distribution is $f(x) = (2\pi\sigma^2)^{-1/2} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Another example of a continuous random variable is the *uniform distribution* on an interval $[a, b]$. We write $\mathcal{U}[a, b]$ and the corresponding p.d.f. is $f(x) = (b-a)^{-1}\mathbf{1}(x \in [a, b])$.

A random vector $(X_1, \ldots, X_n)$ is said to have a continuous distribution if $P_{(X_1,\ldots,X_n)}$ is dominated by the Lebesgue measure on $\mathbb{R}^n$. An example is the multivariate normal distribution

with mean vector $\mu \in \mathbb{R}^n$ and positive definite $n \times n$-covariance matrix Σ. We write $\mathcal{N}(\mu, \Sigma)$ and the p.d.f. is $f(\mathbf{x}) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu))$.

Given a random vector $(X_1, \ldots, X_n)$, we say that $X_1, \ldots, X_n$ are jointly (or mutually) independent if $P(X_1 \leq t_1, \ldots, X_n \leq t_n) = P(X_1 \leq t_1) \cdot \ldots \cdot P(X_n \leq t_n)$. If $X_1, \ldots, X_n$ also have the same distribution than we say that $X_1, \ldots, X_n$ are *independent and identically distributed* (i.i.d.). The p.d.f. of the random vector $(X_1, \ldots, X_n)$ is in this case $\prod_{i=1}^n f(x_i)$ with f the p.d.f. of a single component.

12.1 The maximum likelihood principle

Suppose that we observe n coin flips of a coin that returns 1 (heads) with probability p and 0 (tails) with probability $1-p$. Given the data $X_1, \ldots, X_n$ we are interested in recovering the unknown parameter p. To build an estimator (reconstruction method based on the data), we can use the maximum likelihood principle. It suggests to take the most likely parameter value given the data provided this exists. Suppose we observe a random element X with distribution P_θ where $\theta \in \Theta$ is the unknown parameter and Θ is the parameter space, for instance the real numbers. Let p_θ be the Radon-Nikodym derivative with respect to a dominating measure μ, for example the p.m.f. or the p.d.f. The maximum likelihood estimator (MLE) $\widehat{\theta}$ is then the maximizer

$$\widehat{\theta} = \operatorname{argmax}_{\theta \in \Theta} p_\theta(X).$$

This requires that the MLE exists and is unique.

As an example consider the setting above with i.i.d. X_i following a Bernoulli distribution with unknown parameter p. The p.m.f. can be written as $\prod_{i=1}^m p^{x_i}(1-p)^{1-x_i}$ since $p^{x_i}(1-p)^{1-x_i} = p$ for $x_i = 1$ and $p^{x_i}(1-p)^{1-x_i} = 1-p$ for $x_i = 0$. The MLE is then the minimizer of $p^{\sum_i X_i}(1-p)^{n-\sum_i X_i}$ which is $\widehat{p} = n^{-1}\sum_i X_i$.

Chapter 13

Brownian motion

13.1 Definition and basic properties of Brownian motion

Brownian motion appeared at the beginning of the 20th century in the work by Bachelier on finance. Around the same time Einstein used it as a model for particle movement. A mathematical treatment including a proof of existence of Brownian motion is due to Norbert Wiener. Brownian motion is therefore also sometimes called Wiener process.

Definition 19 (Brownian motion)**.** *A collection of random variables $(W_t)_{t\in[0,T]}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ is called a Brownian motion (on $[0,T]$) if*

(i) $W_0 = 0$ *P-almost surely*

(ii) *W has independent increments. This means that for increasing indices $0 \leq t_0 < t_1 < \ldots < t_m \leq T$, the random variables $W_{t_i} - W_{t_{i-1}}$ $i = 1, \ldots, m$ are mutually independent.*

(iii) *For any $0 \leq s \leq t < T$, $W_t - W_s \sim \mathcal{N}(0, t-s)$*

(iv) *sample paths $t \mapsto W_t$ are continuous functions P-almost surely.*

Existence of the Brownian motion is non-trivial. It turns out that condition (iv) can be derived from $(i) - (iii)$. Brownian motion has many interesting properties. A consequence of the definition is that Brownian motion is a Gaussian process, that is, the random vectors $(W_t)_{t\in I}$ follows a multivariate normal distribution for all finite subsets $I \subset [0,T]$. Moreover,

$$\mathrm{Cov}(W_s, W_t) = s \wedge t.$$

There is also an inversion of this. If a Gaussian process has this covariance for all $0 \leq s, t \leq T$ then it is a Brownian motion. Moreover, it can be shown that the realizations or paths $t \mapsto W_t$ have Hölder index α for any $\alpha < 1/2$, P-almost surely.

13.2 Integration with respect to Brownian motion

In this section we aim to define an integral with respect to Brownian motion. This means that for a Brownian motion $(W_t)_{t\in[0,T]}$ on the probability space $(\Omega, \mathcal{F}, P)$ we want to give an interpretation to the expression $\int_0^T f(s)\,dW_s$. One option is the Itô integral that defines $\int X_s\,dW_s$ for a class of stochastic processes $(X_s)_{s\in[0,T]}$. For our purpose it is enough to define $\int f(s)\,dW_s$ for deterministic integrands in $L^2[0,T]$. This is sometimes called the Wiener integral (but there is another integral that is also called Wiener or Paley-Wiener integral).

Construction of the Wiener integral: For simple functions $f = \sum_{j=1}^k a_j \mathbf{1}(\cdot \in [t_{j-1}, t_j])$ with real coefficients a_j and $0 \le t_0 < t_1 < \ldots < t_k \le T$, the Wiener integral is defined as $\int_0^T f(s)\,dW_s = \sum_{j=1}^k a_j(W_{t_j} - W_{t_{j-1}})$. For two simple functions f, g, one can check that by the properties of Brownian motion,

$$E\Big[\int_0^T f(s)\,dW_s \int_0^T g(s)\,dW_s\Big] = \int_0^T f(s)g(s)\,ds. \tag{13.2.1}$$

Denote by $L^2(\Omega, \mathcal{F}, P)$ the space of square integrable random variables on the probability space $(\Omega, \mathcal{F}, P)$ with norm $\|X\|_{L^2(\Omega,\mathcal{F},P)} = E^{1/2}[X^2]$. The map $f \mapsto \int_0^T f(s)\,dW_s$ that sends simple functions $f \in L^2[0,T]$ to $\int_0^T f(s)\,dW_s \in L^2(\Omega, \mathcal{F}, P)$ is therefore an isometry, that is, $\|f\|_{L^2[0,T]} = \|\int_0^T f(s)\,dW_s\|_{L^2(\Omega,\mathcal{F},P)}$. Simple functions are dense in $L^2[0,T]$ (reference ???). For any function $f \in L^2[0,T]$ there exists thus a sequence of simple functions $(f_k)_k$ with $\|f - f_k\|_{L^2[0,T]} \to 0$. This is then a Cauchy sequence and therefore $\int_0^T f_k(s)\,dW_s$ is a Cauchy sequence in $L^2(\Omega, \mathcal{F}, P)$. Therefore, we can define $\int_0^T f(s)\,dW_s$ as the $L^2(\Omega, \mathcal{F}, P)$-limit of $\int_0^T f_k(s)\,dW_s$.

Basic properties: Let us mention some basic properties of the Wiener integral $\int_0^T f(s)\,dW_s$ for a deterministic integrand $f \in L^2[0,T]$.

Lemma 17. *If $f, g \in L^2[0,T]$, then,*

(i) $\int_0^T f(s)\,dW_s \sim \mathcal{N}(0, \|f\|^2_{L^2[0,T]})$

(ii) $\mathrm{Cov}\big(\int_0^T f(s)\,dW_s, \int_0^T g(s)\,dW_s\big) = \int_0^T f(s)g(s)\,ds$

(ii) implies that for two L^2-orthogonal function f, g the corresponding integrals $\int f(s)\,dW_s$ and $\int g(s)\,dW_s$ are uncorrelated and, because they are Gaussian, also independent.

13.3 Girsanov's formula

Denote by $\mathcal{C}[0,T]$ the space of continuous functions on the interval $[0,T]$. By property (iv) in the definition of a Brownian motion, we know that a realization or sample path of $(W_t)_{t\in[0,T]}$ lies in $\mathcal{C}[0,T]$ almost surely. Now, we can equip $\mathcal{C}[0,T]$ with the maximum norm $\|f\|_\infty = \max_{t\in[0,T]} |f(t)|$. This norm generates a topology and we denote by $\mathcal{B}(\mathcal{C}[0,T])$ the corresponding Borel σ-algebra on the space of continuous functions $\mathcal{C}[0,T]$. For any $f \in L^2[0,T]$,

consider the process $Y = (Y_t)_{t\in[0,T]}$ with $Y_t = \int_0^t f(s)\,ds + W_t$. On the measure space $(\mathcal{C}[0,T], \mathcal{B}(\mathcal{C}[0,T]), P)$ define the probability measure $P_f(A) = P(Y_t \in A)$ for all $A \in \mathcal{B}(\mathcal{C}[0,T])$. Girsanov's formula gives an explicit expression for the Radon-Nikodym derivative dP_f/dP_0, where P_0 denotes the distribution with f being identical to zero.

Theorem 17 (Girsanov's theorem)**.** *For any $f, g \in L^2[0,T]$, $P_f \ll P_g$ and*

$$\frac{dP_f}{dP_0}(Y) = \exp\Big(\int_0^T f(s)dY_s - \frac{1}{2}\|f\|^2_{L^2[0,T]}\Big).$$

13.4 Kullback-Leibler divergence for Gaussian white noise model

Based on Girsanov's formula, we can derive an explicit formula for the Kullback-Leibler divergence in the Gaussian white noise model.

Lemma 18. *Denote by P_f the distribution of $Y = (Y_t)_{t\in[0,1]}$ with $dY_t = f(t)dt + n^{-1/2}dW_t$, $t \in [0,1]$ on the measurable space $(\mathcal{C}[0,1], \mathcal{B}(\mathcal{C}[0,1]))$. If $f, g \in L^2[0,1]$, then*

$$K(P_f, P_g) = \frac{n}{2}\|f-g\|^2_{L^2[0,1]}.$$

Proof. By Theorem 17,

$$\frac{dP_f}{dP_0}(Y) = \exp\Big(n\int_0^1 f(t)\,dY_t - \frac{n}{2}\int_0^1 f^2(t)\,dt\Big).$$

We can now rewrite dP_f/dP_g using standard properties of the Radon-Nikodym derivative. Using $P_f \ll P_g \ll P_0$ for the first identity and $P_g \ll P_0 \ll P_g$ for the second, we obtain

$$\frac{dP_f}{dP_g} = \frac{dP_f}{dP_0}\frac{dP_0}{dP_g} = \frac{dP_f}{dP_0}\Big(\frac{dP_g}{dP_0}\Big)^{-1},$$

which yields

$$\begin{aligned}
K(P_f, P_g) &= \int \log\Big(\frac{dP_f}{dP_g}\Big)\,dP_f \\
&= \int \log\Big(\frac{dP_f}{dP_0}\Big)\,dP_f - \int \log\Big(\frac{dP_g}{dP_0}\Big)\,dP_f \\
&= E_f\Big[\log\Big(\frac{dP_f}{dP_0}\Big)\Big] - E_f\Big[\log\Big(\frac{dP_g}{dP_0}\Big)\Big] \\
&= E_f\Big[n\int_0^1 f(t)\,dY_t - \frac{n}{2}\int_0^1 f(t)^2\,dt - n\int_0^1 g(t)\,dY_t + \frac{n}{2}\int_0^1 g(t)^2\,dt\Big] \\
&= n\int_0^1 (f(t)-g(t))f(t)\,dt - \frac{n}{2}\int_0^1 f(t)^2\,dt + \frac{n}{2}\int_0^1 g(t)^2\,dt \\
&= n\int_0^1 \big(f(t)-g(t)\big)^2\,dt \\
&= n\big\|f-g\big\|^2_{L^2[0,1]}.
\end{aligned}$$

□

13.5 Exercises

Ex. 13.1 — Prove (13.2.1) for simple functions f, g.

Chapter 14

Concentration inequalities and tail bounds

14.1 The union bound

The union bound says that for measurable events A_j,

$$P\Big(\bigcup_j A_j\Big) \leq \sum_j P(A_j).$$

The inequality follows directly from the axioms of a measure. In statistics, the union bound is often applied to bound the probability of a maximum over random variables. More precisely,

$$P\big(\max_j V_j \geq t\big) = P\Big(\bigcup_j \{V_j \geq t\}\Big) \leq \sum_j P(V_j \geq t).$$

14.2 Tail bounds for the normal distribution

Lemma 19. *For $\xi \sim \mathcal{N}(0,1)$, we have $P(|\xi| \geq u) \leq 2e^{-u^2/2}$.*

Proof. By Markov's inequality, $P(\xi \geq u) = P(e^{u\xi} \geq e^{u^2}) \leq E[e^{u\xi}]/e^{u^2} = e^{-u^2/2}$. Consequently, $P(|\xi| \geq u) = P(\{\xi \geq u\} \cup \{-\xi \geq u\}) \leq 2P(\xi \geq u) \leq 2e^{-u^2/2}$. □

Together with Mill's ratio (cf. Exercise (6.9)), we obtain the tail bound

$$P(|\xi| \geq u) \leq \min\Big(2, \frac{\sqrt{2}}{\sqrt{\pi}u}\Big)e^{-u^2/2}.$$

Lemma 20. *If η_j, $j = 1, \ldots, M$ are centered normal random variables, then,*

$$E\big[\max_{1\leq j\leq M} |\eta_j|\big] \leq \big(\sqrt{2\log M} + 1\big) \max_{1\leq j\leq M} \sqrt{\mathrm{Var}(\eta_j)}.$$

Proof. By rescaling, it is enough to consider the case that $\max_{1\le j\le M} \mathrm{Var}(\eta_j) = 1$. With Exercise 14.1 and $|\eta_j| \le \sqrt{2\log M} + |\eta_j|\mathbf{1}(|\eta_j| > \sqrt{2\log M})$,

$$E\big[\max_{1\le j\le M} |\eta_j|\big] = \sqrt{2\log M} + \sum_{j=1}^{M} E\big[|\eta_j|\mathbf{1}(|\eta_j| \ge \sqrt{2\log M})\big] \le \sqrt{2\log M} + 1.$$

□

14.3 Exercises

Ex. 14.1 — Prove that for $\xi \sim \mathcal{N}(0,1)$, Φ the c.d.f. of ξ and $\phi = \Phi'$,

$$E[|\xi|\mathbf{1}(|\xi| \ge u)] = 2\phi(u)$$

and

$$E[\xi^2\mathbf{1}(|\xi| \ge u)] = 2u\phi(u) + 2\Phi(-u).$$

Chapter 15

Miscellaneous

15.1 Order symbols

To derive convergence rates, for instance, we often need to bound the growth of a sequence of non-negative real numbers by another sequence. Since we are interested in the decay of the sequence, the bounds need to hold up to constants only. Below, we summarize mathematical notation that is relevant for these type of arguments.

Consider two non-negative sequences $(a_n)_n$ and $(b_n)_n$ with n running through the positive integers. We write $a_n \lesssim b_n$ if there is a constant $C > 0$ such that $a_n \leq C b_n$ for all n. The symbol $\lesssim$ means therefore smaller up to constants. The expression $b_n \gtrsim a_n$ is equivalent to $a_n \lesssim b_n$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$ we write $a_n \asymp b_n$. The symbol $\asymp$ thus means that the two sequences are of the same order.

If it is clear from the context, we do not need to define the sequence. For instance, we can write $2n^{-1} \lesssim n$ since it is obvious that we mean the sequences $(2n^{-1})_n$ and $(n)_n$.

Bibliography

[1] Bertsekas, D. *Nonlinear Programming*. Athena Scientific, 1999.

[2] Brown, L. D. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist. 42*, 3 (1971), 855–903.

[3] Brown, L. D., Low, M. G., and Zhao, L. H. Superefficiency in nonparametric function estimation. *Ann. Statist. 25*, 6 (12 1997), 2607–2625.

[4] Brown, L. D., and Zhao, L. H. A geometrical explanation of Stein shrinkage. *Statist. Sci. 27*, 1 (02 2012), 24–30.

[5] Bühlmann, P., and van de Geer, S. *Statistics for high-dimensional data*. Springer, 2011.

[6] Cai, T., Liu, W., and Luo, X. A constrained ℓ_1-minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 494 (2011), 594–607.

[7] Efron, B., and Morris, C. Stein's paradox in statistics. *Scientific American 236*, 5 (1977), 119–127.

[8] Emery, M., Nemirovski, A., and Voiculescu, D. *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVIII*. Springer, 2000.

[9] Evans, L. C. *Partial differential equations*, second ed., vol. 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010.

[10] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1 (2010), 1–22.

[11] Giné, E., and Nickl, R. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[12] Györfi, L., Kohler, M., Krzyzak, and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[13] Haerdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. *Wavelets, Approximation, and Statistical Applications*. Springer, 1998.

[14] Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity*. Chapman & Hall, 2015.

[15] Johnstone, I. Gaussian estimation: Sequence and wavelet models, September 2015.

[16] Korostelev, A., and Tsybakov, A. *Minimax Theory of Image Reconstruction*. Springer, 1993.

[17] Pagan, A., and Ullah, A. *Nonparametric econometrics*. Cambridge university press, 1999.

[18] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist. 5* (2011), 935–980.

[19] Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist. 43*, 3 (06 2015), 991–1026.

[20] Rohde, A., and Tsybakov, A. B. Estimation of high-dimensional low-rank matrices. *Ann. Statist. 39*, 2 (04 2011), 887–930.

[21] Samworth, R. Small confidence sets for the mean of a spherically symmetric distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol. 67*, 3 (2005), 343–361.

[22] Shao, J. *Mathematical statistics*, second ed. Springer Texts in Statistics. Springer-Verlag, New York, 2003.

[23] Shao, P. Y.-S., and Strawderman, W. E. Improving on the James-Stein positive-part estimator. *Ann. Statist. 22*, 3 (1994), 1517–1538.

[24] Sohrab, H. *Basic real analysis (Second ed.)*. Birkhäuser, 2014.

[25] Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Berkeley, Calif., 1956), University of California Press, pp. 197–206.

[26] Tao, T. *An introduction to measure theory*, vol. 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

[27] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*, 1 (1996), 267–288.

[28] Tsybakov, A. *Introduction to nonparametric estimation*. Springer, 2009.

[29] Wand, M., and Jones, M. *Kernel Smoothing*. Chapman & Hall, 1995.

[30] Wasserman, L. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006.