Jump estimation for noisy blurred step functions

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Leif Boysen
aus Bremen

Göttingen, 2006

D7
Referent: Prof. Dr. Axel Munk
Koreferent: Prof. Dr. Lutz Dümbgen
Tag der mündlichen Prüfung: 09.05.2006

Contents

Chapter 1

Introduction

A central topic in statistics is regression analysis. It deals with estimation of the impact of some independent quantity X on a dependent quantity Y. For example, it may be of interest to examine the dependence of the air temperature on the date at several locations, the dependence of plant growth on the amount of fertilizer used or the dependence of the size of infants on their age.

These examples show that in general it does not make sense to assume an exact dependency of the quantities of interest. Clearly, infants of the same age can differ in size. Hence, one tries to estimate the mean of the variable of interest Y given X. The deviation of the observations from this mean is modeled by an error term ε. The general approach is to assume that the mean of the variable of interest depends on X through some function f. The corresponding model can be written as

$$Y = f(X) + \varepsilon\,,$$

where the error term ε is assumed to have mean zero.

When trying to fit a function f to some given data points $(X_1, Y_1), \ldots, (X_n, Y_n)$, one usually chooses f from a special function class. Examples are the class of all linear functions, the class of spline functions, or the class of functions with bounded second derivative. Many commonly used function classes consist of continuous functions. However, one frequently faces situations where the most striking features of the regression function are sharp transitions or structural changes. A location where such a break or jump occurs is called a change-point.

There are several applications for which it is interesting to estimate the location of such a change-point. This problem occurs in dose finding studies, where there may be some minimal amount to be taken before any effect occurs. Another example is provided by quality control. If a change of quality in some continuous production process is suspected, it is important to find out whether and when the quality started to deteriorate. A further example comes from geology. If there is a significant change in measurements of core samples obtained from different geological sites, the geologist may want to know where this change took place. (For example to determine the size of some deposit of a natural resource like oil or gold.) In general the analysis of change-points has two main aspects. The first is to decide whether a change has occurred. The second is estimation of the location, the magnitude of change and the number of change-points.

In regression analysis, a jump of a one-dimensional function f can be estimated at a rate of n^{-1}, if n is the number of observations and the regression function f is Lipschitz continuous elsewhere (cf. Korostelëv, 1987). The n^{-1} rate is optimal in a minimax sense. Roughly speaking, this means that there exists no estimator which uniformly, in the function class of interest, converges at a faster rate (a more rigid definition can be found in Section 4.2).

The distribution of the estimate of the jump location depends on the estimator used. If f is piecewise constant with one jump, Hinkley (1970) showed that after multiplication with n the least squares estimate of the jump location converges to the minimum of a certain two-sided asymmetric random walk. This was generalized by Yao and Au (1989) to the case where f is piecewise constant with finitely many jumps. In this thesis we focus on function spaces of this type. Figure 1.1 shows an example of noisy observations of a piecewise constant function.

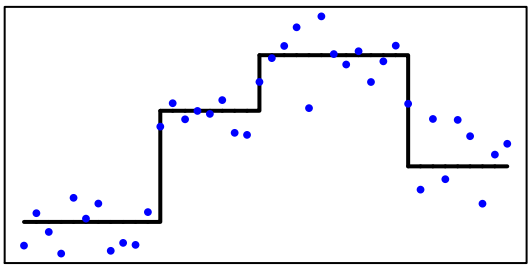

Figure 1.1.: Noisy observations of a step function. The blue dots represent the observations and the black line the step function f.

Indeed, there is a number of examples, for which it is reasonable to use a step function to model the observations in a regression setting. An example concerning the segmentation of gene data is given by Braun et al. (2000). Fredkin and Rice (1992) use step functions to model the neuronal flow in calcium channels (these are ion-pores in the plasma membrane of electrically excitable cells). Christensen and Rudemo (1996) use this model for disease incidence data. Noisy observations of blocked data also occur in mass-spectroscopy. Typically the location of the jumps, their height and, in certain cases, the number of jumps are unknown.

Let us now turn to a different aspect of regression analysis. There are several settings, where one does not observe the function of interest directly but only a blurred or diffused version of it. To reconstruct the original function, one generally assumes that the way the function is blurred is known. This is the case if the data collection mechanism is known to transform the signal in a certain way. For example, if the observation of an image is taken through a lens with known curvature or if a sensor does not collect the information at a certain point, but only some weighted mean of the neighborhood of this point. The function of interest then can be reconstructed by inverting the transformation mechanism. However, in many situations small changes in the observations can cause major changes in the reconstruction of the original function. This can lead to instabilities if noise is present. Consequently, it may be much harder to reconstruct the original function than the blurred

one. In this case the corresponding problem is called ill-posed. In the presence of random noise, one generally speaks of a statistical inverse problem. In regression analysis the model is written as

$$Y = \mathrm{K}f(X) + \varepsilon\,.$$

Here the error term ε is assumed to have mean zero and K is some known operator. A standard example of a statistical inverse problem is estimation of the derivative of a regression function. The rate of convergence of estimating f in an inverse problem is in general slower than in the direct problem. Moreover, this rate usually depends on the eigenvalues in the spectral decomposition of the operator, determining the ill-posedness of the problem (cf. Fan, 1991; Abramovich and Silverman, 1998; Hall et al., 2003).

The topic of this thesis is the reconstruction of step functions from noisy blurred observations (see Figure 1.2). Special attention is given to the reconstruction of the jumps of the step function. The investigation of the jumps (or change-points) focuses on estimation of the location of the jumps. It turns out that the estimation problem is harder in the inverse setting than in the direct problem. This can be stated in the following way. Suppose K is some integral operator with bounded integral kernel K, i.e.

$$\mathrm{K}f(x) = \int K(x,y)f(y)dy\,.$$

Then the estimate of the jump location converges at a rate of $n^{-1/2}$ to the true location, compared to a rate of n^{-1} in the direct problem. Surprisingly, the rate does not depend on the ill-posedness of the problem. Furthermore, this rate is minimax.

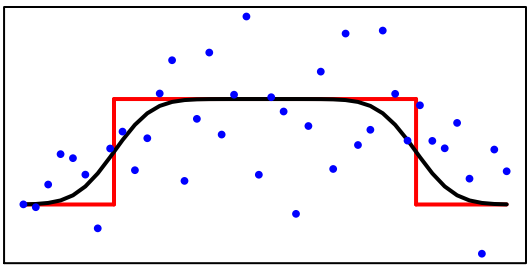

Figure 1.2.: Noisy observations of a blurred step function. The blue dots represent the observations and the black line the blurred function. The red line shows the original step function f, which is to be estimated.

Now we discuss the results in more detail. For estimation we use the least squares estimator if the number of jumps is known. If this number is unknown, we use a penalized version of this estimator with penalty term proportional to the number of jumps of the reconstruction. If K is an integral operator with a bounded integral kernel, we show $n^{-1/2}$ consistency of the estimates of the jump locations. Furthermore, we derive asymptotic normality of the parameter estimates, given that the number of jumps is known and the distribution of the error has a finite second moment. This can be done under rather general assumptions on the design, which cover both the random and the fixed design. The

asymptotic result can be used to derive confidence sets for the jump locations and heights. A confidence set covers the true function with a given probability of, say, $1-\alpha$, provided the number of jumps has been correctly specified. Figure 1.3 shows an example of such a confidence sets for a step function with one jump.

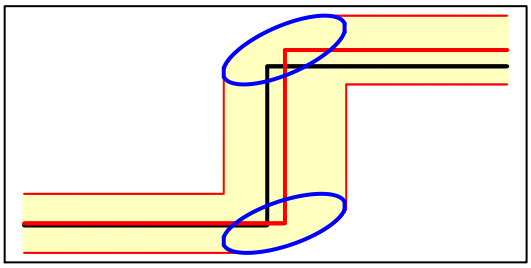

Figure 1.3.: Confidence band for a step function with one jump. The black line represents the true function and the thick red line the estimate. The blue ellipses show the confidence sets for the jump location and the first and second level of the step function, respectively. The thin red lines show the resulting confidence band for the step function. A description of the construction of these confidence sets can be found in Chapter 6.

If the number of jumps is unknown, we show that – under the additional assumption of subgaussian tails of the error distribution – the number of jumps can be asymptotically estimated correctly with probability one.

We find that the $n^{-1/2}$ rate does not depend on the spectral information of the operator. This may be surprising at a first glance because one might suspect that the degree of ill-posedness determines the rate convergence, as it does for piecewise smooth functions (see Goldenshluger et al., 2006). The reason for this is that the space of step functions allows for a finite-dimensional parameterization.

We give general conditions for integral kernels, which are sufficient to deduce the $n^{-1/2}$ rate. These conditions covers supersmooth functions such as the Gauss-kernel, polynomial kernels $K(x,y) = (x-y)^p\, 1_{[0,\infty)}(x-y)$ with $p = 0, 1, \ldots$ and convolution kernels $K(x,y) = \Phi(x-y)$ with Φ some continuous symmetric function, which has a Fourier transform satisfying $|\widehat{\Phi}(x)| \geq c(1+|x|)^n$ for some $n \in \mathbb{N}$ and $c > 0$.

A comparison of the $n^{-1/2}$ rate to the n^{-1} rate of the direct case clearly shows that the direct and indirect problem are substantially different. Also, the asymptotic distribution differs fundamentally from the direct case. In the inverse setting, the estimates are normal and not distributed according to a minimum of some two sided random walk. It is particularly interesting that the parameter estimates are in general no longer asymptotically independent.

Furthermore, we show that faster rates can be obtained if the assumption of the kernel's boundedness is dropped. If K is an integral operator with an Abel-type integral kernel $K(x,y) = (x-y)^{-\alpha}\, 1_{(0,\infty)}(x-y), 0 < \alpha < 1$, we prove that a jump location can be estimated at a rate of $\min(n^{-1/2}, n^{-1/(3-2\alpha)})$ and that this rate is optimal in a minimax sense. Thus it becomes apparent that a more spiky kernel improves the rate because it

allows for better localization of the jump.

We close the description of the results by some technical remarks and an outline of the main proof.

1. Assume the number of jumps is known. An entropy argument yields consistency of the least squares estimator in this setting. It is possible to represent the estimator as the minimizer of a stochastic process, which allows for a local stochastic expansion. This can be used to derive asymptotic normality.

2. Introduce a penalty for the number of jumps. An imitation of techniques from empirical process theory shows that for a suitable choice of the smoothing parameter the case of an unknown number of jumps can asymptotically be reduced to the case where this number is known.

Difficulties arise in Step 1 because some terms of the stochastic process are not differentiable. Step 2 is rather technical. We cannot apply standard techniques from empirical process theory because the used penalty is not a pseudo-norm on the space of step functions.

Aside from the proof of the main result, we use techniques from approximation theory to verify the assumptions on the operator for several different kinds of integral kernels. Some theory of reproducing kernel Hilbert spaces, positive definite functions and extended sign regular functions is used to achieve this.

Next, we compare our results to the existing literature. The problem of change-point estimation in inverse problems was first studied by Neumann (1997). He investigated the estimation of a change-point in a density deconvolution model. Here one tries to estimate the density of a random variable X which cannot be observed directly, but only $X + \xi$, where ξ is an error term with known density f_ξ. In deconvolution problems the quality of the reconstruction generally depends on the tails of the Fourier transform of f_ξ. Neumann (1997) treated the case that the density of X is bounded, has one jump at τ and is Lipschitz continuous elsewhere. He showed that τ can be estimated at a rate of $\min(n^{-1/(2\beta+1)}, n^{-1/(\beta+3/2)})$, provided the tails of the Fourier transform $\widehat{f_\xi}(x)$ decrease at a rate of $|x|^{-\beta}$. Moreover, he proved that these rates are optimal in a minimax sense. This result was extended by Goldenshluger et al. (2006) (in a white noise model) to classes of functions f which can be written as a sum of a step function and a function with smooth m-th derivative. They showed that in this case the minimax rates are of order $\min(n^{-1/(2\beta+1)}, n^{-(m+1)/(2\beta+2m+1)})$. We remark that the setting used in this thesis can be seen as limit case of the model given by Goldenshluger et al. (2006) for $m \to \infty$.

A classical model which fits into our framework was given by Quandt (1958). He introduced a linear regression model which obeys two separate regimes and where the point at which the switch from one regime to the other occurs is not known. This model is also called two-phase regression and inference in this setting was studied by Quandt (1960), Sprent (1961), Hinkley (1969) and more recently by Yakir et al. (1999) and Koul et al. (2003), among others. In two-phase regression the objective function f is assumed to be piecewise linear with two different slopes. There exist two different versions of two-phase regression, that differ in whether or not f is allowed to have a jump at the point where the

slope changes. If f is assumed to be continuous, two-phase regression can be modeled by an inverse regression model with a polynomial kernel with $p = 0$, i.e. $K(x, y) = 1_{[0,\infty)}(x - y)$. In this setting the $n^{-1/2}$ rate and the asymptotic distribution were derived by Hinkley (1969) and – for more general segmented regression models – by Feder (1975b).

We generalize the known results on the estimation of the intersection in two phase regression to the case where the objective function is piecewise polynomial of order $p + 1$, with p continuous derivatives and a $(p + 1)$-th derivative, which is a step function. The somewhat surprising result is that the rate of estimating the intersection does not depend on p, whereas in general nonparametric regression settings, the convergence rates for estimating a jump in the p-th derivative become slower as p grows (cf. Raimondo, 1998).

This thesis is organized as follows: We start by introducing the model and some notation in Chapter 2. Then, in Chapter 3, we present the known results for the direct case to permit comparison with our results. We indicate some of the differences to the inverse problem and give the proofs of some nonstandard results. Chapter 4 introduces the model assumptions and presents the asymptotic normality of the parameter estimates, the results on the penalized least squares estimator for an unknown number of jump locations and a lower bound for estimating the jump locations. Afterwards, Chapter 5 verifies the assumptions on the operator for several well known classes of integral kernels. In Chapter 6, we evaluate the speed of convergence and quality of the approximation by the asymptotic distribution for two examples of integral kernels. In addition, we describe a method to construct confidence bands for f and $\mathrm{K}f$. Chapter 7 contains the proof of the main result and Chapter 8 provides some auxiliary results, particularly on Abel-type kernels. Finally, in Chapter 9 shortcomings, feasibility and possible extensions of the given results are discussed. The appendix summarizes results from empirical process theory, the theory of minimax estimation and approximation theory, which are used in the proofs.

Acknowledgments

I am grateful to my advisor Axel Munk for proposing the problem, constant encouragement and being open for discussion and questions. I wish to thank Lutz Duembgen for taking the Koreferat, interesting discussions and a warm welcome during my time in Bern. During my time as a Ph.D. student I was a member of the Ph.D. Program “Applied Statistics and Empirical Methods” and the Graduiertenkolleg “Identifikation in mathematischen Modellen: Synergie stochastischer und numerischer Methoden”, and I would like to thank for financial support and the possibility of interesting scientific discourse. Also I would like to thank the DAAD for financial support of my stay in Bern and Philadelphia and a summer course in Laredo.

Moreover, special thanks go to Robert Schaback for helpful discussions on reproducing kernel Hilbert spaces, Larry Brown for inviting me for a stay at the statistics department of the Wharton School in Philadelphia and being open for questions; Manfred Denker for interesting remarks and regular demonstrations that no one knows everything; Thorsten Hohage for helpful discussions; my family for support and the regular attempts at understanding what I am working on; the Mensa crew Carola, Karin, Sven and Moritz for keeping

the good spirit; the people at the Institute for Mathematical Stochastics in Göttingen for good working atmosphere, especially Hajo and Nico for regular "Kaffee und Brötchen" breaks and of course Achim, Bernd, Denise, Janis, Gudrun, Marina, Michael, Susanne K., Susanne P. and all the others; Nick for lots of coffee and British humor during my stay in Phili; Angelika, Kaspar and Thomas for the nice atmosphere during my time in Bern; my volleyball team for distraction, good team spirit and losing nearly all matches during my stays in Philadelphia and Bern and thus making me feel important. The last and most important thanks go to Stephie. Words are not enough.

Chapter 2

The inverse regression model

In the following we introduce the general model applicable in the direct and indirect case. More specific assumptions will be made in Chapter 3 for the direct case and in Chapter 4, Section 4.1 for the indirect case, which is the main topic of this thesis.

We assume that we are in a regression model where the objective function f is a step function with finitely many jumps, i.e. we can write f as

$$f(x) = \sum_{i=1}^{k+1} b_i 1_{[\tau_{i-1},\tau_i)}(x)\,, \tag{2.1}$$

with $-\infty \leq \tau_{low} = \tau_0 \leq 0 < \tau_1 < \ldots < \tau_k < 1 \leq \tau_{k+1} = \tau_{up} \leq \infty$ and τ_{low}, τ_{up} known. Note that instead of choosing a right-continuous function f, we could also use a left-continuous function. However, it is common to use a right-continuous function.

Assume that we cannot observe the function directly, but only the image $\mathrm{K}f$ sampled at points $x_1, \ldots, x_n$ contained in some compact interval $I \subset \mathbb{R}$ plus some additional noise $\varepsilon_1, \ldots, \varepsilon_n$, where K is a known operator. Without loss of generality we set $I = [0, 1]$. Given some observations Y, we write our model as

$$Y = \Big((\mathrm{K}f)(x_i) + \varepsilon_i\Big)_{i=1}^{n}\,. \tag{2.2}$$

In the following (2.2) will be called inverse regression model. To be precise, we should use $x_{i,n}$ and $\varepsilon_{i,n}$ instead of x_i and ε_i, as both the design points and the errors are in fact given by triangular arrays. However, for ease of notation, we will suppress the dependency on n throughout this thesis and write $\varepsilon_1, \ldots, \varepsilon_n$ and $x_1, \ldots, x_n$ instead of $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ and $x_{1,n}, \ldots, x_{n,n}$.

Note that the choice of the lower and the upper bound τ_{low} and τ_{up} depends on what we wish to model. Typically we will have $\tau_{low} \in \{-\infty, 0\}$ and $\tau_{up} \in \{1, \infty\}$ depending on the operator K. Here the values $0, 1$ are the bounds of the interval I, which contains the design points $x_1, \ldots, x_n$. To model convolution with some probability density function, set $\tau_{low} = -\infty$ and $\tau_{up} = \infty$. Otherwise f is zero outside of $[\tau_{low}, \tau_{up}]$, which leads to the image $\mathrm{K}f$ being drawn towards zero at the boundaries of the observed interval (see Figure 2.1).

Moreover, if we want to model piecewise linear functions, $\mathrm{K}f$ is the convolution of f

with $1_{[0,\infty)}$. This gives

$$\mathrm{K}\, 1_{[\tau_{low},\tau_1)}(x) = \int_{\tau_{low}}^{\tau_1} 1_{[0,\infty)}(x-y)dy = \begin{cases} x-\tau_{low} & \tau_{low} < x < \tau_1\,, \\ \tau_1 - \tau_{low} & \tau_1 > x\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly, we have to take $\tau_{low} > -\infty$ for this to make sense.

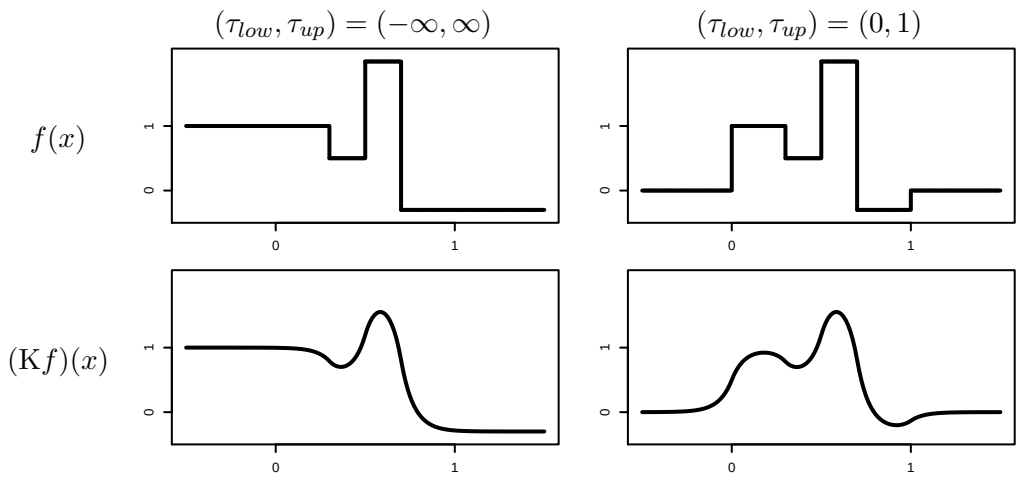

Figure 2.1.: Impact of the location of the boundary points. The first row displays the signal f and the second row the convolution of f with the Laplace kernel.

2.1. Notation

A list of symbols is given on page 103. This section introduces the main notation used in this thesis.

Spaces of step functions

Define

$$\Gamma_k(\tau_{low}, \tau_{up}) := \{(\gamma_0, \gamma_1, \ldots, \gamma_{k+1}) \,:\, \tau_{low} = \gamma_0 \le 0 < \gamma_1 < \ldots < \gamma_k < 1 \le \gamma_{k+1} = \tau_{up}\}$$

as the set of possible jumps of our function f, and denote the corresponding function space by

$$T_k(\tau_{low}, \tau_{up}) := \Big\{ \sum_{i=1}^{k+1} b_i 1_{[\tau_{i-1},\tau_i)}(x) \,:\, \tau \in \Gamma_k(\tau_{low}, \tau_{up})\,,\, b_i \in \mathbb{R} \Big\}\,.$$

Write $T_\infty(\tau_{low}, \tau_{up}) := \bigcup_{k=1}^{\infty} T_k(\tau_{low}, \tau_{up})$ for the set of all step functions on $[\tau_{low}, \tau_{up}]$ and

$$T_{k,R}(\tau_{low}, \tau_{up}) = \{g \in T_k(\tau_{low}, \tau_{up}) : \|g\|_\infty < R\}$$

as well as

$$T_{\infty,R}(\tau_{low},\tau_{up}) := \bigcup_{k=1}^{\infty} T_{k,R}(\tau_{low},\tau_{up})$$

for the corresponding spaces of uniformly bounded functions.

Empirical and L_2 norms

As usual, $\|\cdot\|_2$ will denote the $L_2(\mathbb{R})$ norm, $\langle\cdot,\cdot\rangle_2$ the corresponding inner product and $\|\cdot\|_\infty$ the supremum norm. Additionally define the empirical norm $\|\cdot\|_n$ and the empirical inner product $\langle\cdot,\cdot\rangle_n$ by

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2 \qquad \text{as well as} \qquad \langle f,g\rangle_n := \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(x_i)$$

where $x_1,\dots,x_n$ are the design points and

$$\|y\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} y_i^2 \qquad \text{as well as} \qquad \langle y,z\rangle_n := \frac{1}{n}\sum_{i=1}^{n} y_i z_i$$

for $y,z\in\mathbb{R}^n$.

Sets of jumps and Hausdorff distance

We will also need some notions concerning the jumps of functions and sets of piecewise constant functions.

Definition 2.1. *Write $g(t_+) := \lim_{x\searrow t} g(x)$ for the right limit of g in t and $g(t_-) := \lim_{x\nearrow t} g(x)$ for the corresponding left limit. For some function $g:\mathbb{R}\to\mathbb{R}$ define the set of jump points of g as*

$$\mathcal{J}(g) := \{t\in[0,1] : g(t_-)\neq g(t_+)\}. \tag{2.3}$$

Finally, write

$$h(g) := \min\{|g(t_+)-g(t_-)| : t\in\mathcal{J}(g)\}$$

for the minimal jump height of g.

Furthermore, introduce the following notation for the distances of real sets.

Definition 2.2. *Define the distance of some point $a\in\mathbb{R}$ to the set $B\subset\mathbb{R}$ as*

$$\mathrm{d}(a,B) = \inf_{b\in B}|a-b|$$

and, slightly abusing notation, the Hausdorff distance of two sets A,B as

$$\mathrm{d}(A,B) = \max\{\sup_{a\in A}\mathrm{d}(a,B)\,,\sup_{b\in B}\mathrm{d}(b,A)\}\,.$$

Convergence in probability and distribution

For a sequence of random vectors $X_n \in \mathbb{R}^d$, some constant $c \in \mathbb{R}^d$, random vectors $X \in \mathbb{R}^d$ and a distribution function F on $\mathbb{R}^d$, write

$$X_n \xrightarrow{p} c \qquad \text{and} \qquad X_n \xrightarrow{p} X\,,$$

if X_n converges in probability to c and X, respectively. Moreover, if X has a distribution function given by F, write

$$X_n \xrightarrow{\mathcal{L}} F\,,$$

if X_n converges in distribution to X. We use the stochastic order notation and write $X_n = O_P(a_n)$ if

$$\lim_{C\to\infty} \limsup_{n\to\infty} P(|X_n| > Ca_n) = 0\,.$$

Similarly write $X_n = o_P(a_n)$ if

$$a_n^{-1}|X_n| \xrightarrow{p} 0.$$

Measures and Kullback-Leibler distance

For two measures P and Q, write $P \ll Q$ if P is absolutely continuous with respect to Q. Moreover define the Kullback-Leibler distance of P and Q as

$$d_K(P,Q) = \begin{cases} \displaystyle\int \log\Big(\frac{dP}{dQ}\Big)dP & P \ll Q\,, \\ \infty & \text{otherwise}\,. \end{cases}$$

Here log means the natural logarithm.

Notation for empirical processes

Given a measure Q, a set of Q-measurable functions $\mathcal{G}$ and a real number $\delta > 0$, define the δ-covering number $N(\delta, \mathcal{G}, Q)$ as the smallest value of N for which there exist functions $g_1, \ldots, g_N$ such that for every $g \in \mathcal{G}$ there is a $j \in 1, \ldots N$ with

$$\int (g - g_j)^2 dQ \le \delta\,.$$

Moreover, define the δ-entropy H of $\mathcal{G}$ as

$$H(\delta, \mathcal{G}, Q) = \log N(\delta, \mathcal{G}, Q)\,.$$

If Q is the Lebesgue measure we will write $H(\delta, \mathcal{G})$ and $N(\delta, \mathcal{G})$ instead of $H(\delta, \mathcal{G}, Q)$ and $N(\delta, \mathcal{G}, Q)$. Given design points $x_1, \ldots, x_n \in \mathbb{R}$, the empirical measure will be denoted by $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$. Note that $\|\cdot\|_n$ is the norm corresponding to the space $L_2(\mathbb{R}, Q_n)$.

Finally, define the entropy integral

$$J(\delta, \mathcal{G}, Q) := \max\Big(\delta\,, \int_0^\delta H^{1/2}(u, \mathcal{G}, Q)du\Big)\,.$$

Additional notation

For $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, denote by $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Define

$$\mathcal{I}(a, b) = [a \wedge b, a \vee b]$$

as the interval with endpoints a and b. If $(cond)$ is a boolean expression set $1_{(cond)}$ to one if $(cond)$ is true and to zero if $(cond)$ is false. For any real valued function f we will write $f(x)_+ = f(x)1_{(f(x)>0)}$ for the positive part of f and $\text{supp}(f) = \{x : f(x) \neq 0\}$ for the support of f. For any real number $a \in \mathbb{R}$ set $\lfloor a \rfloor := \max\{z \in \mathbb{Z} : z \leq a\}$ as the largest integer smaller than or equal to a and $\lceil a \rceil := \min\{z \in \mathbb{Z} : z \geq a\}$ as the smallest integer larger than or equal to a. Finally, for any vector β denote by β^t the transpose of β.

Chapter 3

Rates and asymptotic distribution for the direct problem

If K is the identity, the rates and the asymptotic distribution of the jump estimates differ fundamentally from the case where $\mathrm{K}f = \int K(x,y)f(y)dy$ for some bounded integral kernel $K(x,y)$. This chapter gives the classical results for $\mathrm{K} = I$. The estimation of the jump of a step function has been first studied by Hinkley (1970) for the case $k = 1$. Yao (1988) and Yao and Au (1989) define a penalized estimate for a step function with an unknown number of jumps. Their estimate asymptotically finds the right number of jumps with probability one, given some known upper bound for this number. They derive the rates and asymptotic distribution of the parameter estimates. Their results have been generalized to over-dispersion models by Braun et al. (2000).

The model was extended by Boysen et al. (2005), who derive the same rate without the constraint of a known upper bound for the number of jumps. Further they give consistency and rates of convergence in Skorokhod topology and in the case where the true regression is not a step function. As it turns out their estimate is adaptive and nearly attains optimal rates of convergence if the objective function is in C^1 or less smooth.

Model and estimator

Throughout this section we will assume that the observations Y are given by equation (2.2), where the operator K is the identity and f is a step function with k jumps between zero and one. This means that f is given by equation (2.1) with $b_i \neq b_{i+1}$ for $i = 1, \ldots, k$. For simplicity assume that the design points are fixed and equidistant with $x_i = (i-1)/n$. In addition, assume that the ε_i are independent identically distributed with mean zero and $\mathrm{E}(\varepsilon_1^6) < \infty$. Moreover, set $\tau_{low} = 0$ and $\tau_{up} = 1$, since the values of f outside this interval do not play a role.

A characteristic of the least squares estimator to be defined later in the direct case is that the residual sum of squares does not change if an estimated jump location is varied in-between the design points next to it. (This is in general not true for the indirect case.) Since we work with right-continuous functions, we define the estimated jump location to be the next largest design point.

Define

$$\Gamma_{k,n} := \{\tau \in \Gamma_k(0,1) \,:\, n\tau_i \in \mathbb{N} \text{ for } i = 1, \ldots, k\},$$

where $\Gamma_k(a,b)$ is defined in Section 2.1. For $\tilde{\tau} \in \Gamma_{k,n}$ set

$$b_j(\tilde{\tau}) = \frac{1}{n\tilde{\tau}_j - n\tilde{\tau}_{j-1}} \sum_{i=n\tilde{\tau}_{j-1}}^{n\tilde{\tau}_j - 1} Y_i \qquad \text{and} \qquad f_{\tilde{\tau}}(x) = \sum_{j=1}^{k+1} b_j(\tilde{\tau}) 1_{[n\tilde{\tau}_{j-1}, n\tilde{\tau}_j)}(x)\,.$$

Then the estimator of Yao and Au (1989) is given by

$$\hat{f}_{\text{Yao}} = \operatorname*{argmin}_{f_{\tilde{\tau}}\,:\,\tilde{\tau}\in\Gamma_{k,n}\,;\,k\le K_u} \log\left(\|Y - f_{\tilde{\tau}}\|_n^2\right) + \rho_n \#\mathcal{J}(f_{\tilde{\tau}})\,, \tag{3.1}$$

where ρ_n is some smoothing parameter. Note that at most K_u jumps are allowed. The functional in (3.1) is motivated by the Schwarz model selection criterion (cf. Yao, 1988). By definition of the estimator it is clear that the upper bound for the number of jumps is needed. If g with $\#\mathcal{J}(g) = n$ would be allowed in minimization of the right hand side of equation (3.1), the minimizer would always interpolate the data and the log term would be $-\infty$. Thus, the choice of K_u is important. Simulations show, that a choice of $\sqrt{n}/2$ works reasonably well in most cases.

An estimator which overcomes the obstacle of choosing an upper bound for the number of jumps is introduced by Boysen et al. (2005) as minimizer of the so called Potts functional. The estimate is given by

$$\hat{f}_{\text{Potts}} = \operatorname*{argmin}_{f_{\tilde{\tau}}\,:\,\tilde{\tau}\in\Gamma_{k,n}} \|Y - f_{\tilde{\tau}}\|_n^2 + \lambda_n \#\mathcal{J}(f_{\tilde{\tau}})\,, \tag{3.2}$$

where again λ_n is some smoothing parameter. The estimators $\hat{f}_{\text{Yao}}$ and $\hat{f}_{\text{Potts}}$ coincide if $\#\mathcal{J}(\hat{f}_{\text{Yao}}) = \#\mathcal{J}(\hat{f}_{\text{Potts}})$. The name Potts functional refers to a model which is well-known in statistical mechanics and was introduced by Potts (1952) as a generalization of the Ising model from Ising (1925) for a binary spin system to more than two states. The original model is a Gibbs field with energy equal to the above penalty.

Note that both estimators defined above exist but are not necessarily unique. However there can only be finitely many solutions of (3.1) and (3.2) (see Kempe, 2004). If multiple minimizers exist, define the estimate as an arbitrary member in the set of minimizers.

If $\hat{f}_{\text{Yao}}$ has jumps in i/n and j/n for $0 \le i < j < n$ and no jump in $(i/n, j/n)$, then $\hat{f}_{\text{Yao}}|_{[i/n,j/n)}$ is constant and equal to the mean $m_{i,j} = (j-i)^{-1}\sum_{r=i+1}^{j} Y_r$. The $m_{i,j}$ and the corresponding empirical losses $n^{-1}\sum_{r=i+1}^{j}(Y_r - m_{i,j})^2$ can be calculated in time $O(n^2)$ for all $i < j$. Winkler and Liebscher (2002) give an algorithm to use these values to compute the minimizer $\hat{f}_{\text{Potts}}$ (and thus $\hat{f}_{\text{Yao}}$) in $O(n^2)$ time for any given k such that $\#\mathcal{J}(\hat{f}_{\text{Potts}}) = k$. As k takes only $n+1$ possible different values the estimates for all values of k can be calculated in $O(n^3)$ steps. Winkler and Liebscher (2002) additionally show that these estimates can be used to determine a partition $0 = \gamma_{n+1} < \gamma_n < \ldots < \gamma_1 < \gamma_0 = \infty$ such that for $\lambda_n \in (\gamma_{k+1}, \gamma_k)$ the corresponding estimate $\hat{f}_{\text{Potts}}$ satisfies $\#\mathcal{J}(\hat{f}_{\text{Potts}}) = k$. (For $\lambda_n = \gamma_k$ the estimate is not unique.) A similar result can be shown for the estimate $\hat{f}_{\text{Yao}}$. Thus both estimators can be calculated in time $O(n^3)$ for all possible λ_n.

Rates and asymptotic distribution

The following theorem gives the convergence rates for the estimates of the jump locations in Hausdorff distance and for the function estimate in L_2 norm.

Theorem 3.1. *Suppose that* $Y_i = f((i-1)/n) + \varepsilon_i$, $f \in T_{K_u}(0,1)$ *and that* $\varepsilon_1, \dots, \varepsilon_n$ *are i.i.d. with mean zero and* $\mathrm{E}(\varepsilon_1^6) < \infty$. *Then*

(i) If $\rho_n \to 0$ *and* $\rho_n n^{1/3} \to \infty$ *then* $P(\#\mathcal{J}(\hat{f}_{Yao}) = \#\mathcal{J}(f)) \to 1$ *as* $n \to \infty$,

(ii) if $\lambda_n \to 0$ *and* $\lambda_n n^{1/3} \to \infty$ *then* $P(\#\mathcal{J}(\hat{f}_{Potts}) = \#\mathcal{J}(f)) \to 1$ *as* $n \to \infty$,

(iii) $\|\hat{f}_{Yao} - f\|_{L_2} = \mathrm{O}_P(n^{-1/2})$,

(iv) $d(\mathcal{J}(\hat{f}_{Yao}), \mathcal{J}(f)) = O_P(1/n)$.

Note that $P(\#\mathcal{J}(\hat{f}_{\mathrm{Yao}}) = \#\mathcal{J}(f)) \to 1$ and $P(\#\mathcal{J}(\hat{f}_{\mathrm{Potts}}) = \#\mathcal{J}(f)) \to 1$ means that the estimates defined by (3.1) and (3.2) asymptotically coincide with probability one. Consequently, (iii) and (iv) also hold if $\hat{f}_{\mathrm{Yao}}$ is replaced by $\hat{f}_{\mathrm{Potts}}$.

The fact that discontinuities can be estimated at a rate of $O_P(n^{-1})$ is well known and has been shown in various settings for the regression function f (cf. Korostelëv, 1987, for f in a Lipschitz class). A good overview on change points estimation is given by Carlstein and Müller (1994).

The proof of Theorem 3.1 is rather technical and the same rates of convergence can be observed in more general settings, therefore we omit most of the proof and refer the interested reader to the corresponding papers. However, Boysen et al. (2005) show part (ii) only under a subgaussian assumption on $\varepsilon_1, \dots, \varepsilon_n$, thus we give the proof of this part in the following.

Proof of Theorem 3.1, (ii). For any $\gamma \in \Gamma_{k,n}$ define the vector $\varepsilon|_\gamma$ by its entries

$$(\varepsilon|_\gamma)_i = \frac{1}{n\gamma_j - n\gamma_{j-1}} \sum_{r=n\gamma_{j-1}+1}^{n\gamma_j} \varepsilon_r \,,$$

where j is chosen such that $(i-1)/n \in [\gamma_{j-1}, \gamma_j)$.

By definition of $\hat{f}_{\mathrm{Potts}}$ we have that

$$\|\hat{f}_{\mathrm{Potts}} - Y\|_n^2 + \lambda_n \#\mathcal{J}(\hat{f}_{\mathrm{Potts}}) \le \|f - Y\|_n^2 + \lambda_n \#\mathcal{J}(f) \,,$$

which implies

$$\|\hat{f}_{\mathrm{Potts}} - f\|_n^2 + \lambda_n(\#\mathcal{J}(\hat{f}_{\mathrm{Potts}}) - \#\mathcal{J}(f)) \le 2\langle \hat{f}_{\mathrm{Potts}} - f, \varepsilon\rangle_n \,. \tag{3.3}$$

Now assume that $\gamma \in \Gamma_{s,n}$ such that $(\gamma_1, \dots, \gamma_s)$ are the ordered jumps of $\hat{f}_{\mathrm{Potts}} - f$. Note that $s \le \#\mathcal{J}(\hat{f}_{\mathrm{Potts}}) + \#\mathcal{J}(f)$. Since $\hat{f}_{\mathrm{Potts}} - f$ is constant on $[\gamma_{i-1}, \gamma_i)$ we have

$$\langle \hat{f}_{\mathrm{Potts}} - f, \varepsilon\rangle_n = \langle \hat{f}_{\mathrm{Potts}} - f, \varepsilon|_\gamma\rangle_n \,.$$

Moreover,

$$\langle \hat{f}_{\mathrm{Potts}} - f, \varepsilon|_\gamma\rangle_n \le \|\hat{f}_{\mathrm{Potts}} - f\|_n \|\varepsilon|_\gamma\|_n \le \frac{1}{8}\|\hat{f}_{\mathrm{Potts}} - f\|_n^2 + 2\|\varepsilon|_\gamma\|_n^2 \,.$$

With the help of (3.3), we obtain

$$\lambda_n(\#\mathcal{J}(\hat{f}_{\text{Potts}}) - \#\mathcal{J}(f)) \leq -\frac{3}{4}\|\hat{f}_{\text{Potts}} - f\|_n^2 + 4\|\varepsilon|_\gamma\|_n^2 \leq 4\|\varepsilon|_\gamma\|_n^2$$

and

$$\frac{3}{4}\|\hat{f}_{\text{Potts}} - f\|_n^2 \leq 2\|\varepsilon|_\gamma\|_n^2 + \lambda_n(\#\mathcal{J}(f) - \#\mathcal{J}(\hat{f}_{\text{Potts}}))\,. \tag{3.4}$$

If $\mathrm{E}(\varepsilon_1^6) < \infty$, Lemma 1 of Yao and Au (1989) gives that

$$\max_{0\leq i<j\leq n} \frac{1}{j-i}\Big(\sum_{r=i+1}^{j} \varepsilon_r\Big)^2 = O_P(n^{2/3})\,,$$

which implies

$$\begin{aligned} \|\varepsilon|_\gamma\|_n^2 &\leq n^{-1}\big(\#\mathcal{J}(\hat{f}_{\text{Potts}}) + \#\mathcal{J}(f)\big) \max_{0\leq i<j\leq n} \frac{1}{j-i}\Big(\sum_{r=i+1}^{j} \varepsilon_r\Big)^2 \\ &= O_P(n^{-1/3})(\#\mathcal{J}(\hat{f}_{\text{Potts}}) + \#\mathcal{J}(f))\,. \end{aligned} \tag{3.5}$$

We arrive at

$$\lambda_n(\#\mathcal{J}(\hat{f}_{\text{Potts}}) - \#\mathcal{J}(f))/(\#\mathcal{J}(\hat{f}_{\text{Potts}}) + \#\mathcal{J}(f)) \leq O_P(n^{-1/3})\,,$$

which gives $(\#\mathcal{J}(\hat{f}_{\text{Potts}}) - \#\mathcal{J}(f)) = o_p(1)$ by $\lambda_n n^{1/3} \to \infty$.

This gives $\mathrm{P}(\#\mathcal{J}(\hat{f}_{\text{Potts}}) \leq \#\mathcal{J}(f)) \to 1$. Equation (3.5) then yields $\|\varepsilon|_\gamma\|_n^2 = O_P(n^{-1/3})$ and by (3.4) we obtain $\|\hat{f}_{\text{Potts}} - f\|_n^2 = o_P(1)$. As as step function cannot be consistently estimated by a sequence of step functions with fewer jumps this implies that $\mathrm{P}(\#\mathcal{J}(\hat{f}_{\text{Potts}}) \geq \#\mathcal{J}(f)) \to 1$. This proves the claim. □

The next theorem gives the asymptotic distribution of the jump estimates, as derived in Yao and Au (1989). Since the result is nonstandard, the proof is given for convenience of the reader.

Theorem 3.2. *Suppose the assumptions of Theorem 3.1 are met and $\hat{\tau}$ is the estimator of the jump points of f as given by (3.1). Moreover, suppose the number of jumps is correctly estimated as k. The estimates $n\hat{\tau}_1, \ldots, n\hat{\tau}_k$ are asymptotically independent as $n \to \infty$. For each $j = 1, \ldots, k$ the difference $n\hat{\tau}_j - \lceil n\tau_j \rceil$ converges in distribution to L_j, the location of the minimum of the random walk $\{\ldots, Z_{-1}^{(j)}, Z_0^{(j)}, Z_1^{(j)}, \ldots\}$, where $Z_0^{(j)} = 0$ and*

$$Z_r^{(j)} = \begin{cases} (b_{j+1} - b_j)\sum_{i=1}^{r}\Big(2U_i^{(j)} + (b_{j+1} - b_j)\Big) & r = 1, 2, \ldots\,, \\ (b_j - b_{j+1})\sum_{i=r+1}^{0}\Big(2U_i^{(j)} + (b_j - b_{j+1})\Big) & r = -1, -2, \ldots\,, \end{cases}$$

and the $U_i^{(j)}$ are i.i.d. according to the same distribution as ε_1 for all $j = 1, \ldots, k$.

Proof. Note that $Z_r^{(j)} = r(b_j - b_{j+1})^2 + O_P(\sqrt{r})$. This means the random walk converges to infinity for $|r| \to \infty$. Consequently, $|L_j| = O_P(1)$. Theorem 3.1 gives $|n\hat{\tau}_j - \lceil n\tau_j \rceil| = O_P(1)$. Thus for each $\epsilon > 0$ we can find an $M_\epsilon \in \mathbb{N}$ such that $\mathrm{P}(|L_j| \leq M_\epsilon) > 1 - \epsilon$ as well as $\mathrm{P}(|n\hat{\tau}_j - \lceil n\tau_j \rceil|) > 1 - \epsilon$ for all $j = 1, \ldots, k$. Hence it is enough to show convergence in distribution of the estimates

$$\hat{\tau}_M = \underset{(n\tilde{\tau}) \in \mathbb{N}^k : \|n\tilde{\tau} - n\tau\|_\infty < M}{\operatorname{argmin}} \|Y - f_{\tilde{\tau}}\|_n \tag{3.6}$$

to the location of the minima of the random walks $\{Z_{-M}^{(j)}, \ldots, Z_M^{(j)}\}$ for all fixed $M \in \mathbb{N}$.

Now, suppose $n\tilde{\tau}_j \in \mathbb{N}$ and $|n\tilde{\tau}_j - \lceil n\tau_j \rceil| \leq M$ for all $j = 1, \ldots, k$. Observe that

$$\sup_{\tilde{\tau} : \|\tilde{\tau} - \tau\|_\infty \leq M} \max_{j=1,\ldots,k+1} |b_j - b_j(\tilde{\tau})| = O_P(n^{-1/2}),$$

and for $b(\tau) := b(\lceil n\tau \rceil / n)$ that

$$\sup_{\tilde{\tau} : \|\tilde{\tau} - \tau\|_\infty \leq M} \max_{j=1,\ldots,k+1} |b_j(\tau) - b_j(\tilde{\tau})| = O_P(n^{-1}).$$

Denote by

$$\mathcal{I}_j = \big[(\lceil n\tau_{j-1} \rceil \vee n\tilde{\tau}_{j-1}), (\lceil n\tau_j \rceil \wedge n\tilde{\tau}_j) \big)$$

for $j = 1, \ldots, k+1$ the intervals with $\|(f - f(\tilde{\tau}))|_{\mathcal{I}_j}\|_\infty = O_P(n^{-1/2})$ and by

$$\mathcal{M}_j = \big[(\lceil n\tau_j \rceil \wedge n\tilde{\tau}_j), (\lceil n\tau_j \rceil \vee n\tilde{\tau}_j) \big)$$

for $j = 1, \ldots, k$ the intervals with $\|(f - f(\tilde{\tau}))|_{\mathcal{M}_j}\|_\infty = |b_{j+1} - b_j| + O_P(n^{-1/2})$. Set $\mathcal{M}_{k+1} = \emptyset$. Note that $\#\{i : i \in \mathcal{M}_j\} = O(1)$ for $j = 1, \ldots, k+1$. Consequently,

$$\begin{aligned}
\|Y - f(\tilde{\tau})\|_n^2 &= \frac{1}{n} \sum_{j=1}^{k+1} \Bigg(\sum_{i \in \mathcal{I}_j} (b_j + \varepsilon_i - b_j(\tilde{\tau}))^2 + \sum_{i \in \mathcal{M}_j} (b_{j+1} + \varepsilon_i - b_j(\tilde{\tau}))^2 1_{(\tilde{\tau}_j \geq \tau_j)} + \\
&\qquad \sum_{i \in \mathcal{M}_j} (b_j + \varepsilon_i - b_{j+1}(\tilde{\tau}))^2 1_{(\tilde{\tau}_j < \tau_j)} \Bigg) \\
&= O_P(n^{-3/2}) + \frac{1}{n} \sum_{j=1}^{k+1} \Bigg(\sum_{i \in \mathcal{I}_j} \Big((b_j - b_j(\tilde{\tau}))^2 + 2\varepsilon_i (b_j - b_j(\tilde{\tau})) \Big) + \\
&\qquad \|\varepsilon\|_n^2 + \sum_{i \in \mathcal{M}_j} \Big((b_{j+1} - b_j)^2 + 2(b_{j+1} - b_j)\varepsilon_i \Big) 1_{(\tilde{\tau}_j \geq \tau_j)} + \\
&\qquad \sum_{i \in \mathcal{M}_j} \Big((b_{j+1} - b_j)^2 + 2(b_j - b_{j+1})\varepsilon_i \Big) 1_{(\tilde{\tau}_j < \tau_j)} \Bigg),
\end{aligned}$$

uniformly in $\{\tilde{\tau} : \|\tilde{\tau} - \tau\|_\infty \leq M\}$. The term $\|\varepsilon\|_n^2$ does not play a role when the expression above is minimized. The minimizer of the last two terms converges to the desired limit

distribution. Moreover,

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{I}_j}\Big((b_j-b_j(\tilde\tau))^2+2\varepsilon_i(b_j-b_j(\tilde\tau))^2\Big)\\
&\quad= O_P(n^{-3/2})+\frac{1}{n}\sum_{i\in\mathcal{I}_j}\Big((b_j-b_j(\tau))^2+\varepsilon_i(b_j-b_j(\tau))\Big)+\frac{(b_j(\tau)-b_j(\tilde\tau))}{n}\Big(\sum_{i\in\mathcal{I}_j}\varepsilon_i\Big)\\
&\quad= O_P(n^{-3/2})+\frac{1}{n}\sum_{i\in\mathcal{I}_j}\Big((b_j-b_j(\tau))^2+\varepsilon_i(b_j-b_j(\tau))\Big),
\end{aligned}$$

uniformly in $\{\tilde\tau : \|\tilde\tau-\tau\|_\infty \le M\}$. Since these terms do not play a role for the minimization of $\|Y-f(\tilde\tau)\|_n^2$ in $\tilde\tau$, this proves the claim. □

For the next theorem assume that the estimator $\hat f_{\text{Yao}}$ has the form

$$\hat f_{\text{Yao}}(x)=\sum_{i=1}^{k+1}\hat b_i 1_{[\hat\tau_{i-1},\hat\tau_i)}(x)\,.$$

The asymptotic distribution of the level estimates $\hat b_i$ is normal, and the estimates are asymptotically independent.

Theorem 3.3. *Suppose the assumptions of Theorem 3.2 are met and that both the true and the estimated number of jump locations are equal to k. For $n\to\infty$ the normalized estimates $\sqrt{n}(b_j-\hat b_j)$ are asymptotically independent for $j=1,\dots,k+1$ and normally distributed with means 0 and variances $\sigma^2/(\tau_j-\tau_{j-1})$, where $\sigma^2<\infty$ is the variance of ε_1. In addition, $n\hat\tau_j-n\lceil\tau_j\rceil,(j=1,\dots,k),\sqrt{n}(b_j-\hat b_j),(j=1,\dots,k+1)$ are asymptotically independent.*

Proof. By $\|\hat\tau-\tau\|_\infty=O_P(n^{-1})$ we have

$$\hat b_j=\sum_{i=n\hat\tau_{j-1}}^{n\hat\tau_j}Y_i=\sum_{i=\lceil n\tau_{j-1}\rceil}^{\lfloor n\tau_j\rfloor}Y_i+O_P(n^{-1}).$$

An application of the central limit theorem directly gives the claimed limit distribution for $\sqrt{n}(b_j-\hat b_j)$.

To show the asymptotic independence, set for $M\in\mathbb{N}$

$$\hat b_{j,M}=\sum_{i=\lceil n\tau_{j-1}\rceil+M+1}^{\lfloor n\tau_j\rfloor-M-1}Y_i.$$

By the same arguments as above $\hat b_{j,M}=\hat b_j+O_P(n^{-1})$ and thus $\sqrt{n}(\hat b_j-b_j)=\sqrt{n}(\hat b_{j,M}-b_j)+O_P(n^{-1/2})$. For $\hat\tau_M$ defined by (3.6) we clearly have that $n\hat\tau_M$ is independent of $\sqrt{n}(\hat b_{j,M}-b_j)$ for all $j=1,\dots,k+1$. Thus it is asymptotically independent of $\sqrt{n}(\hat b_j-b_j)$ for all $j=1,\dots,k+1$. Note that for all $\epsilon>0$ there exists some $M_\epsilon\in\mathbb{N}$ such that $P(\hat\tau_M=\hat\tau)>1-\epsilon$. The claim follows by taking $\epsilon\to 0$. □

Chapter 4

Rates and asymptotic distribution for the inverse regression model

This chapter introduces and shortly discusses the model assumptions for the model (2.2). We define the least squares estimator and present the main result of this thesis. As the assumption on the operator is rather unusual, we try to give an interpretation in Section 4.3.

4.1. Model assumptions

Assumptions on the error

We will assume that the error has second moments and is independent identically distributed with mean zero.

Assumption A. *The array* $(\varepsilon_1, \dots, \varepsilon_n)$ *consists of independent identically distributed random variables with mean zero for every* n*. Additionally assume*

$$\mathrm{E}(\varepsilon_1^2) = \sigma^2 < \infty\,.$$

If the number of jumps of the objective function is unknown, we will additionally need that the error satisfies the following subgaussian condition.

(A1) There exists some $\alpha > 0$ such that $\mathrm{E}(\exp(\varepsilon_1^2/\alpha)) < \infty$.

Assumptions on the operator

We consider linear integral operators K of the type $(\mathrm{K}f)(x) = \int K(x,y)f(y)dy$.

To estimate f from observations of $\mathrm{K}f$, it is necessary to assume that the operator K is one-to-one as a mapping from the space of step functions $T_k(\tau_{low}, \tau_{up})$ to $L_2([0,1])$. This means for every function f with $f(x) = \sum_{i=1}^{k+1} b_i 1_{[\tau_{i-1},\tau_i)}(x)$, we have to assume that the relation

$$0 = \left\|(\mathrm{K}f)(\cdot)\right\|_{L_2([0,1])} = \left\|\sum_{i=1}^{k+1} b_i (\mathrm{K}\,1_{[\tau_{i-1},\tau_i)})(\cdot)\right\|_{L_2([0,1])}$$

implies $f \equiv 0$ and hence $b_i = 0$ for all $i = 1, \dots, k+1$. Consequently, this is equivalent to assuming that the functions

$$(\mathrm{K}\,1_{[\tau_0,\tau_1)})(\cdot), (\mathrm{K}\,1_{[\tau_1,\tau_2)})(\cdot), \dots, (\mathrm{K}\,1_{[\tau_k,\tau_{k+1})})(\cdot)$$

are linearly independent as functions in $L_2([0,1])$ for every choice of $(\tau_0, \dots, \tau_{k+1}) \in \Gamma_k(\tau_{low}, \tau_{up})$.

Define

$$\Delta_K(x,a,b) := \begin{cases} \int_a^b K(x,y)dy & b \neq a\,, \\ K(x,a) & b = a\,. \end{cases} \tag{4.1}$$

Instead of assuming independence of the functions $\mathrm{K}\,1_{[\tau_i,\tau_{i+1})}(\cdot)$, we will assume independence of the functions $\Delta_k(\cdot, \tau_i, \tau_{i+1})$, which is slightly stronger. The reason for this is discussed in detail in section 4.3.

Assumption B. *The operator* K *is given by* $\mathrm{K}f(x) = \int_{\tau_{low}}^{\tau_{up}} K(x,y)f(y)dy$*, there exists some* $C_K > 0$ *such that* $\int_{\tau_{low}}^{\tau_{up}} |K(x,y)|dy \leq C_K$ *for all* x *and*

$$\sup_{x\in[0,1]\,,\,y\in(\tau_{low},\tau_{up})} |K(x,y)| < \infty\,.$$

In addition the integral kernel $K(x,y)$ *satisfies one of the following assumptions:*

(B1) $K : [0,1] \times [\tau_{low}, \tau_{up}] \to \mathbb{R}$ *is continuous.*

(B2) $K(x,y) = \Phi(x-y)$ *and* Φ *is piecewise continuous with finitely many jumps.*

Additionally the functions

$$\Delta_K(x,\tau_0,\tau_1)\,,\ \Delta_K(x,\tau_1,\tau_2)\,,\ \dots\,,\ \Delta_K(x,\tau_k,\tau_{k+1})$$

are linearly independent for every choice of $k \in \mathbb{N}$ *and*

$$\tau_{low} = \tau_0 \leq 0 < \tau_1 \leq \tau_2 \leq \ldots \leq \tau_k < 1 \leq \tau_{k+1} = \tau_{up}\,,$$

where only two subsequent τ_i *are allowed to be equal. Here* Δ_K *is defined by (4.1).*

In chapter 5 we list several classes of kernels $K(x,y)$ satisfying Assumptions B. One special case is $K(x,y) = \Phi(x-y)$ for some continuous symmetric Φ, with Fourier transform satisfying $|\widehat{\Phi}(x)| \geq c(1+|x|)^{-n}$ for some $n \in \mathbb{N}$ and $c > 0$. Other examples are convolution with a Gauss kernel and $K(x,y) = 1_{(-\infty,x)}(y)$, which leads to a linear regression model with different slopes (also called multi-phase linear regression).

Assumptions on the design points

Empirical process theory allows us to make inference in the empirical norm only. However, this is restricted to $\|\mathrm{K}f - \mathrm{K}\hat{f}_n\|_n$. Once we wish to draw a conclusion on $f - \hat{f}_n$ in any norm, we need some regularity assumption on the design points.

Assumption C. *There exists a function* $h : [0,1] \to [c_l, c_u]$ *with* $0 < c_l < c_u < \infty$ *and* $\int_0^1 h(x)dx = 1$*, such that*

$$\frac{i}{n} = \int_0^{x_{(i)}} h(x)dx + \delta_i$$

for all $i = 1, \dots, n$, *with*

$$\max_{i=1,\dots,n} |\delta_i| = O_P(n^{-1/2}) .$$

Moreover, the design points $x_1, \dots, x_n$ *are independent of the error terms* $\varepsilon_1, \dots, \varepsilon_n$. *Here* $x_{(i)}$ *denotes the* i*-th order statistic of* $x_1, \dots, x_n$.

If the design points $x_1, \dots, x_n$ are nonrandom, the $O_P(n^{-1/2})$ term above is to be understood as $O(n^{-1/2})$. Note that the above assumption covers random designs as well as fixed designs generated by a regular density in the sense of Sacks and Ylvisaker (1970).

Lemma 4.1. *Assume* $x_1, \dots, x_n$ *are independent, identically distributed observations with probability density* h. *If* $\infty > c_u \geq h(x) \geq c_l > 0$ *for all* $x \in [0, 1]$ *and* $\operatorname{supp}(h) = [0, 1]$ *then* $x_1, \dots, x_n$ *satisfy Assumption C.*

Proof. Set $H(x) = \int_0^x h(x)dx$, where h is given by Assumption C, and

$$H_n(x) = n^{-1} \sum_{i=1}^{n} 1_{[x_{(i)},\infty)}(x) .$$

By the classical Dvoretzky-Kiefer-Wolfowitz theorem (cf. Massart, 1990) we have

$$\mathrm{P}(\sup_{x\in\mathbb{R}} |H(x) - H_n(x)| > t) \leq 2\exp(-2nt^2) .$$

Since $H_n(x_{(i)}) = i/n$ it follows that $\max_{i=1,\dots,n} \mathrm{P}(|H(x_{(i)}) - i/n| > t) \leq 2\exp(-2nt^2)$, which implies $\max_{i=1,\dots,n} |H(x_{(i)}) - i/n| = O_P(n^{-1/2})$. □

Munk (2002) assumes that

$$\max_{i=2,\dots,n} \Big| \int_{x_{(i-1)}}^{x_{(i)}} h(x)dx - \frac{1}{n} \Big| = O(n^{-(1+\gamma)}) ,$$

where h is assumed to be Hölder-continuous of order $\gamma > 0$. For $\gamma \geq 1/2$ and h bounded from below and above, this implies Assumption C. However, this does not allow for the random design case.

Dümbgen and Johns (2004) use an assumption on the design points, which implies that the number of design points contained in a sequence of intervals of length a_n is of order na_n provided $a_n \geq cn^{-1+\epsilon}$ for some $\epsilon > 0$ and $c > 0$. In comparison Assumption C is less restrictive in the sense that a similar statement holds only for $\epsilon > 1/2$.

Lemma 4.2. *If the design points* $x_1, \dots, x_n$ *satisfy Assumption C, then for any two sequences* $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ *with* $0 \leq a_n < b_n \leq 1$ *we have*

$$n^{-1}\big(\#\{i : x_i \in [a_n, b_n]\}\big) = O_P(|b_n - a_n| + n^{-1/2}).$$

Moreover, if $(b_n - a_n) > cn^{-1/2+\epsilon}$ *for some* $\epsilon > 0$, $c > 0$ *and all* n, *then*

$$\frac{\#\{i : x_i \in [a_n, b_n]\}}{n(b_n - a_n)} \geq c_1 + o_P(1) ,$$

for some $c_1 > 0$.

Proof. Set $H(x) = \int_0^x h(x)dx$, where h is given by Assumption C. Note that H is strictly monotone and the inverse H^{-1} is well defined on $[0,1]$.

By Assumption C we have $H^{-1}(i/n - \delta_i) = x_{(i)}$ with $\max_{i=1,\dots,n} |\ \delta_i| = O_P(n^{-1/2})$. Therefore

$$\begin{aligned}
\#\{i : a_n \le x_i \le b_n\} &= \#\{i : a_n \le x_{(i)} \le b_n\} \\
&= \#\{i : a_n \le H^{-1}(i/n - \delta_i) \le b_n\} \\
&= \#\{i : H(a_n) + \delta_i \le i/n \le H(b_n) + \delta_i\} \\
&\le \#\{i : H(a_n) - \max_{i=1,\dots,n} |\delta_i| \le i/n \le H(b_n) + \max_{i=1,\dots,n} |\ \delta_i|\} \\
&= n\big(H(b_n) - H(a_n) + O_P(n^{-1/2})\big) \\
&= n\ \big(O(b_n - a_n) + O_P(n^{-1/2})\big)\,,
\end{aligned}$$

which proves the first claim. Similarly

$$\#\{i : a_n \le x_i \le b_n\} \ge n\big(H(b_n) - H(a_n) - 2 \max_{i=1,\dots,n} |\delta_i|\big) \ge n(c_l(b_n - a_n) - 2 \max_{i=1,\dots,n} |\delta_i|)\,,$$

where c_l is the lower bound of the design density given by Assumption C.

For $b_n - a_n > cn^{-1/2+\epsilon}$ we get

$$n^{-1}(b_n - a_n)^{-1} \max_{i=1,\dots,n} |\delta_i| \le cn^{-1/2-\epsilon} \max_{i=1,\dots,n} |\delta_i| = o_P(1)\,.$$

This proves the second claim. □

4.2. Estimate and asymptotic results

Estimate

Define the restricted least squares estimate $\hat{f}_n$ as approximate minimizer of the empirical L_2 distance to the data in the space of step functions bounded in supremum norm by R and restricted to functions with at most k jumps. More precisely assume that $\hat{f}_n \in T_{k,R}(\tau_{low}, \tau_{up})$ and

$$\|\,\mathrm{K}\hat{f}_n - Y\|_n \le \min_{g \in T_{k,R}(\tau_{low},\tau_{up})} \|\,\mathrm{K}g - Y\|_n^2 + o(n^{-1})\,. \tag{4.2}$$

The minimizer of the functional on the right hand side always exists (compare Corollary 7.7). It does not need to be unique. Note that we use an approximate minimizer, since for some operators K it may not be possible to find an explicit form of the minimizer. In this case it is necessary to use a numeric optimization algorithm to compute the minimizer. The restriction to functions with $\|f\|_\infty < R$ is a technical assumption, which requires that some upper bound of the supremum norm of the objective function is known beforehand.

In the following we will assume that $\hat{f}_n$ has the form

$$\hat{f}_n(x) = \sum_{i=1}^{k+1} \hat{b}_i 1_{[\hat{\tau}_{i-1},\hat{\tau}_i)}(x)\,, \tag{4.3}$$

i.e. the vectors $\hat{b} = (\hat{b}_1, \dots, \hat{b}_{k+1})^t$ and $\hat{\tau} = (\hat{\tau}_0, \dots, \hat{\tau}_{k+1})^t$ are the approximate (in the sense of (4.2)) least squares estimates of the true parameter vectors b and τ given by equation (2.1).

If the number of jumps is unknown, a different estimate is needed. In this case, assume that the penalized least squares estimate $\hat{f}_{\lambda_n}$ satisfies $\hat{f}_{\lambda_n} \in T_{\infty,R}(\tau_{low}, \tau_{up})$ and

$$\| \mathrm{K}\hat{f}_{\lambda_n} - Y\|_n + \lambda_n \#\mathcal{J}(f) \leq \min_{g \in T_{\infty,R}(\tau_{low},\tau_{up})} \| \mathrm{K}g - Y\|_n^2 + \lambda_n \#\mathcal{J}(f) + o(n^{-1}) \,, \tag{4.4}$$

where λ_n is some smoothing parameter. Again, it is not assumed that the minimum is attained, but only that the functional above can be minimized up to some term of order $o(n^{-1})$.

Asymptotic results

Now we give the main results of this thesis, namely the asymptotic normality of the parameter estimates, given that the number of jumps is known, and the fact that the number of jumps can be correctly estimated with probability one.

To state this result, first define

$$\nu(x) = \begin{pmatrix} \Delta_K(x, \tau_0, \tau_1) \\ (b_1 - b_2)\Delta_K(x, \tau_1, \tau_1) \\ \Delta_K(x, \tau_1, \tau_2) \\ \vdots \\ (b_k - b_{k+1})\Delta_K(x, \tau_k, \tau_k) \\ \Delta_K(x, \tau_k, \tau_{k+1})) \end{pmatrix} , \tag{4.5}$$

and the $(2k+1) \times (2k+1)$ matrix V by its entries

$$(V)_{ij} = \int_0^1 (\nu(x)\nu(x)^t)_{ij}\, h(x)dx \,. \tag{4.6}$$

Here h is the design density given by Assumption C. The parameter estimates are asymptotically normally distributed, with covariance matrix $\sigma^2 V^{-1}$.

Theorem 4.3. *Suppose the Assumptions A, B and C are met. Let f, $\hat{f}_n$ and V be given by (2.1), (4.3) and (4.6), respectively. Set $\theta = (b_1, \tau_1, b_2, \tau_2, \dots, b_k, \tau_k, b_{k+1})$ as the parameter vector of f, and $\hat{\theta}_n$ as the corresponding vector of estimates given by $\hat{f}_n$. Given model (2.2) the following holds true:*

(i) *V is positive definite.*

(ii) $\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2 V^{-1})$.

(iii) $\mathrm{d}(\mathcal{J}(f), \mathcal{J}(\hat{f}_n)) = O_P(n^{-1/2})$.

(iv) $\|(\mathrm{K}f - \mathrm{K}\hat{f}_n)|_{[0,1]}\|_2 = O_P(n^{-1/2})$.

(v) $\|(f - \hat{f}_n)|_{[0,1]}\|_2 = O_P(n^{-1/4})$.

(vi) Suppose in addition to Assumption A, the condition (A1) is satisfied, i.e. the error is subgaussian. Let $\lambda_n \to 0$ and $\lambda_n n^{1/(1+\epsilon)} \to \infty$ for some $\epsilon > 0$ as $n \to \infty$, then

$$\lim_{n\to\infty} P\big(\#\mathcal{J}(\hat{f}_{\lambda_n}) = \#\mathcal{J}(f)\big) = 1\,.$$

The proof is given in several steps in Chapter 7.

To assess this result, let us introduce the notion of minimax rates. For any $n \in \mathbb{N}$ assume that $\{\mathrm{P}_{\theta,n}, \theta \in \Theta\}$ is a family of probability measures on the measurable spaces $(\Omega_n, \mathcal{B}_n)$ associated with the observations. Moreover, assume d defines some distance on Θ. We say that a sequence of estimators $\hat{\theta}_n$ attains the minimax rate for the class Θ and the distance d if there exists a sequence $(a_n)_{n\in\mathbb{N}}$ and some $c > 0$ with

$$\lim_{C\to\infty} \lim_{n\to\infty} \sup_{\theta\in\Theta} \mathrm{P}_{\theta,n}(d(\hat{\theta}_n, \theta) > Ca_n) = 0$$

and

$$\lim_{n\to\infty} \inf_{\tilde{\theta}_n} \sup_{\theta\in\Theta} \mathrm{P}_{\theta,n}(d(\tilde{\theta}_n, \theta) > ca_n) > 0\,.$$

Then, a_n is called minimax rate.

This allows us to make the following observation.

Theorem 4.4. *Suppose the assumptions of Theorem 4.3 are met. If $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed normal random variables with zero mean and positive variance, the rates given by Theorem 4.3 are minimax (in the space $T_{k,R}(\tau_{low}, \tau_{up})$).*

The proof is given in Section 7.5.

4.3. An interpretation of the assumption on the operator

The condition of the linear independence of the functions

$$\Delta_K(x, \tau_0, \tau_1)\,,\ \Delta_K(x, \tau_1, \tau_2)\,,\ \dots\,,\ \Delta_K(x, \tau_k, \tau_{k+1})$$

in Assumption B might seem a bit unusual at first. The aim of this section is to give a better understanding of this assumption.

Interpretation via the asymptotic variance

In view of the asymptotic distribution of the estimates in Theorem 4.3, it can be shown, that the assumption of linear independence in B is necessary for the asymptotic covariance matrix to be invertible.

Suppose that the assumption of linear independence in $L_2([0,1])$ does not hold and that $b_{i+1} \neq b_i$ for all $i = 1, \dots, k$ (i.e. that f has k jumps). In this case for $\nu(x)$ given by (4.5) there exist $\beta \in \mathbb{R}^{2k+1}$ with $\beta \neq 0$ such that

$$\int_0^1 (\beta^t \nu(x))^2 dx = 0\,,$$

which implies

$$0 = \int_0^1 (\beta^t \nu(x))^2 h(x)dx = \int_0^1 \beta^t (\nu(x)\nu(x)^t)\beta h(x)dx = \beta^t V \beta\,.$$

This means V does not have full rank and V is not invertible.

Interpretation via rates and approximation theory

There is another interpretation for the condition of linear independence inspired by approximation theory. Classical estimation theory provides rates of convergence for the estimate $\mathrm{K}\hat{f}_n$. The question is, how to obtain the rate of convergence of $\|\hat{f}_n - f\|_2$ from the rate of $\|\mathrm{K}f - \mathrm{K}\hat{f}_n\|_2$. To answer this, we first have a closer look at the function $\hat{f}_n - f$. Assume f is given by (2.1) and $\hat{f}_n$ by (4.3). Moreover, assume that $\tau_{i-1} < \hat{\tau}_i < \tau_{i+1}$ for all $i = 1, \dots, k$. In this case, $\mathrm{K}f - \mathrm{K}\hat{f}$ can be written as

$$(\mathrm{K}f - \mathrm{K}\hat{f})(x) = \sum_{i=1}^{k+1} (b_i - \hat{b}_i) \int_{\tau_{i-1}\vee\hat{\tau}_{i-1}}^{\tau_i\wedge\hat{\tau}_i} K(x,y)dy + \sum_{i=1}^{k} c_i \int_{\tau_i\wedge\hat{\tau}_i}^{\tau_i\vee\hat{\tau}_i} K(x,y)dy\,,$$

where $c_i = b_i - \hat{b}_{i+1}$ for $\tau_i > \hat{\tau}_i$ and $c_i = b_{i+1} - \hat{b}_i$ for $\tau_i \le \hat{\tau}_i$. Define $G_i(x,\tau,\hat{\tau}) := \int_{\tau_{i-1}\vee\hat{\tau}_{i-1}}^{\tau_i\wedge\hat{\tau}_i} K(x,y)dy, i = 1,\dots,k+1$ and $H_i(x,\tau,\hat{\tau}) := \int_{\tau_i\wedge\hat{\tau}_i}^{\tau_i\vee\hat{\tau}_i} K(x,y)dy, i = 1,\dots,k$. Fix j and assume $\hat{b}_j \neq b_j$. We obtain

$$\begin{aligned}
&\|\mathrm{K}f - \mathrm{K}\hat{f}\|_2^2 \\
&= (b_j - \hat{b}_j)^2 \Big\| G_j(x,\tau,\hat{\tau}) + \sum_{i=1, i\neq j}^{k+1} \frac{b_i - \hat{b}_i}{b_j - \hat{b}_j} G_i(x,\tau,\hat{\tau}) + \sum_{i=1}^{k} \frac{c_i}{b_j - \hat{b}_j} H_i(x,\tau,\hat{\tau}) \Big\|_2^2\,.
\end{aligned}$$

This means, if we were able to prove

$$\inf_{\tilde{\tau},\alpha,\beta} \Big\| G_j(x,\tau,\tilde{\tau}) + \sum_{i=1, i\neq j}^{k+1} \alpha_i G_i(x,\tau,\tilde{\tau}) + \sum_{i=1}^{k} \beta_i H_i(x,\tau,\tilde{\tau}) \Big\|_2^2 \ge C > 0\,,$$

we could conclude

$$(\hat{b}_j - b_j)^2 = O(\|\mathrm{K}f - \mathrm{K}\hat{f}\|_2^2)\,.$$

As argued in Section 4.1, the assumption that K is injective on the space $T_k(\tau_{low}, \tau_{up})$ implies that for any given $\tilde{\tau}$ with $\min_{i=1,\dots,k} |\tau_i - \tilde{\tau}_i| > 0$, the functions $G_i(x,\tau,\tilde{\tau})$ and $H_i(x,\tau,\tilde{\tau})$ are linearly independent. This means for any given $\tilde{\tau}$

$$\inf_{\alpha,\beta} \Big\| G_j(x,\tau,\tilde{\tau}) + \sum_{i=1, i\neq j}^{k+1} \alpha_i G_i(x,\tau,\tilde{\tau}) + \sum_{i=1}^{k} \beta_i H_i(x,\tau,\tilde{\tau}) \Big\|_2^2 \ge C_{\tilde{\tau}} > 0\,.$$

Hence, it would be sufficient to show that the infimum is attained. In terms of approximation theory, this means we have to show that the set

$$\Big\{ g \;:\; g(x) = \sum_{i=1, i\neq j}^{k+1} \alpha_i G_i(x,\tau,\hat{\tau}) + \sum_{i=1}^{k} \beta_i H_i(x,\tau,\hat{\tau})\,,\; \|\tau - \hat{\tau}\|_\infty < \epsilon\,,\; \alpha_i, \beta_i \in \mathbb{R}\ \forall i \Big\} \quad (4.7)$$

is an existence set.

A similar problem arises in the theory of approximation by exponential sums (see for example Braess, 1986, Chapter VI). From this theory it is known that sets of this type are not existence sets. The problem is the following. Assume $\hat{\tau}_i < \tau_i$. Choosing $\beta_i = (\tau_i - \hat{\tau}_i)^{-1}$ we get that

$$\lim_{\hat{\tau}_i \to \tau_i} (\tau_i - \hat{\tau}_i)^{-1} \int_{\hat{\tau}_i}^{\tau_i} K(x,y)dy = K(x,\tau_i)\,.$$

This means $K(x,\tau_i)$ can be arbitrarily well approximated, but is not in the span of our function system. The solution is to change the basis of the function system in (4.7) and to use $\tilde{H}_i(x,\tau,\hat{\tau}) = (\tau_i - \hat{\tau}_i)^{-1}\Delta_K(x,\tau_i,\hat{\tau}_i)$, instead of $H_i(x,\tau,\hat{\tau})$. In this context Assumption B assures that the functions $\tilde{H}_1,\ldots,\tilde{H}_k, G_1,\ldots,G_{k+1}$ are linearly independent. Under this assumption it is possible to show that, if we replace H_i by $\tilde{H}_i$ in (4.7), the corresponding set is an existence set. Then we can deduce $(b_i - \hat{b}_i)^2 \le C\|\,\mathrm{K}f - \mathrm{K}\hat{f}_n\|_2^2$. A similar argument gives $(\tau_i - \hat{\tau}_i)^2 \le C\|\,\mathrm{K}f - \mathrm{K}\hat{f}_n\|_2^2$.

Chapter 5

Classes of integral kernels with the desired properties

This chapter presents sufficient conditions for the operator K to satisfy Assumption B. The focus is on the case where the integral kernel $K(x, y)$ is of convolution type $K(x, y) = \Phi(x - y)$.

5.1. Positive definite kernels

In the following, assume that the operator K is of the form $\mathrm{K}f = \int \Phi(x-y)f(y)dy$ for some symmetric positive definite function Φ. Functions of this type are also called radial basis functions and have been thoroughly investigated in approximation theory. Most results used in the following originate from this area. An introduction is the recent book by Wendland (2005).

In order to verify Assumption B, we have to show that the functions

$$\Delta_K(x, \tau_0, \tau_1)\,,\ \Delta_K(x, \tau_1, \tau_2)\,,\ \ldots\,,\ \Delta_K(x, \tau_k, \tau_{k+1})$$

are linearly independent for every $k \in \mathbb{N}$ and every choice of $\tau_0 < \tau_1 \leq \tau_2 < \ldots \leq \tau_k < \tau_{k+1}$, where only two subsequent τ_i are allowed to be equal.

To this end, we will define the native Hilbert space $\mathcal{N}_\Phi$ of a positive definite function Φ and show that the elements of its dual space $\delta_x(f) = f(x)$ and $\rho_{x,y}(f) = \int_x^y f(t)dt$ are linearly independent, if Φ has certain properties. Then we will deduce that the functions $\Delta_K(\cdot, \tau_0, \tau_1), \ldots, \Delta_K(x, \tau_k, \tau_{k+1})$ are linearly independent.

We start by giving the required conditions on Φ.

Assumption D. *$\Phi \in C(\mathbb{R}) \cap L_1(\mathbb{R})$ is a symmetric real-valued function with $\widehat{\Phi}(x) \geq 0$. Moreover, there exists $n_0 \in \mathbb{N}$ and $C > 0$ such that*

$$C(1 + |x|^{n_0})^{-1} \leq |\widehat{\Phi}(x)| \qquad \textit{for all } x \in \mathbb{R}\,. \tag{5.1}$$

Here $\widehat{\Phi}(x) = \int \exp(itx)\Phi(t)dt$ denotes the Fourier transform of Φ. The requirement (5.1) basically means that Φ has finite smoothness or, in other words, has at most n_0 derivatives. The assumptions $\widehat{\Phi}(x) \geq 0$ and (5.1) imply that $\widehat{\Phi}$ is strictly positive. This means that Φ is positive definite. (For a definition and characterization of real-valued positive definite functions, compare Appendix B.1.)

For $\Omega \subset \mathbb{R}$ let $\mathcal{N}_\Phi(\Omega)$ denote the unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_\mathcal{H})$ of functions $f : \Omega \to \mathbb{R}$ satisfying $f(x) = \langle f, \Phi(x - \cdot) \rangle_\mathcal{H}$. $\mathcal{N}_\Phi(\Omega)$ is called native space for Φ and given by the closure of the span of the function set $\{\Phi(x - \cdot) : x \in \Omega\}$ under the inner product induced by $\langle \Phi(x - \cdot), \Phi(y - \cdot) \rangle = \Phi(x - y)$. A short introduction to native spaces along with some basic results of the theory can be found in Appendix B.2.

Denote by

$$\mathcal{S}(\mathbb{R}) = \left\{ f \in C^\infty(\mathbb{R}, \mathbb{C}) : \lim_{|x| \to \infty} |x^n f^{(m)}(x)| = 0 \text{ for all } n, m = 0, 1, 2, \dots \right\}$$

the Schwartz space, where $C^\infty(\mathbb{R}, \mathbb{C})$ is the set of smooth functions from $\mathbb{R}$ to $\mathbb{C}$. The first result is, that the native space $\mathcal{N}_\Phi(\Omega)$ contains all Schwartz functions which are compactly supported in Ω.

Lemma 5.1. *Assume $\Omega \subset \mathbb{R}$ and Φ satisfies Assumption D. Then all real Schwartz functions with support contained in Ω are elements of the native space $N_\Phi(\Omega)$, this means that*

$$\left\{ f \in \mathcal{S}(\mathbb{R}) : \operatorname{supp}(f) \subset \Omega \right\} \subset \mathcal{N}_\Phi(\Omega) \,.$$

Proof. We first proof the claim for $\Omega = \mathbb{R}$. Assume $f \in \mathcal{S}(\mathbb{R})$. Since Fourier transformation is a bijection from $\mathcal{S}(\mathbb{R})$ to $\mathcal{S}(\mathbb{R})$ (cf. Werner, 2000, Theorem V.2.8), $\widehat{f}$ and $\widehat{f}^2$ are also Schwartz functions. Hence for any $n_0 \in \mathbb{N}$, we can find a constant $c_1 > 0$ such that $|\widehat{f}(x)|^2 \leq c_1(1+|x|^{n_0+2})^{-1}$. By (5.1) there exist $c_2 > 0$ and $n_0 \in \mathbb{N}$ such that $(\widehat{\Phi}(x))^{-1} \leq c_2(1+|x|^{n_0})$. We arrive at

$$\int_\mathbb{R} \frac{|\widehat{f}(x)|^2}{\widehat{\Phi}(x)} dx \leq c_1 c_2 \int_\mathbb{R} \frac{1 + |x|^{n_0}}{1 + |x|^{n_0+2}} dx < \infty \,.$$

By Theorem B.7 the function f is in $\mathcal{N}_\Phi(\mathbb{R})$ if and only if $\int_\mathbb{R} |\widehat{f}(x)|^2 / \widehat{\Phi}(x) dx < \infty$. This proves the claim for $\Omega = \mathbb{R}$.

Now assume $\Omega \subset \mathbb{R}$ is arbitrary and $f \in \mathcal{S}(\mathbb{R})$ with $\operatorname{supp} f \subset \Omega$. We have shown $f \in \mathcal{N}_\Phi(\mathbb{R})$. By Theorem B.8 for $\Omega \subset \mathbb{R}$, $f \in \mathcal{N}_\Phi(\mathbb{R})$ implies $f|_\Omega \in \mathcal{N}_\Phi(\Omega)$. This proves the claim. □

Note that Lemma 5.1 implies that for any interval $(a, b) \subset \Omega$ there exists some test function $\psi \in \mathcal{N}_\Phi(\Omega)$ satisfying $\operatorname{supp}(\psi) = [a, b]$. One example is

$$\psi(x) = \begin{cases} 0 & x \leq a \,, \\ \exp((x-a)^{-1} + (b-x)^{-1}) & a < x < b \,, \\ 0 & x \geq b \,. \end{cases}$$

This observation can be used to show that point evaluation and integral mean are linearly independent as elements of the dual space of $\mathcal{N}_\Phi(\Omega)$.

Definition 5.2. *For $\gamma \in \mathbb{R}$ and $\gamma_1, \gamma_2 \in \mathbb{R} \cup \{-\infty, \infty\}$ with $\gamma_1 \leq \gamma_2$ define the point evaluation functional $\delta_\gamma : \mathcal{N}_\Phi(\Omega) \to \mathbb{R}$ by*

$$\delta_\gamma(f) := f(\gamma)$$

and the functional $\rho_{\gamma_1,\gamma_2} : \mathcal{N}_\Phi(\Omega) \to \mathbb{R}$ *by*

$$\rho_{\gamma_1,\gamma_2}(f) := \begin{cases} \int_{\gamma_1}^{\gamma_2} f(x)dx & \gamma_1 \neq \gamma_2\,, \\ \delta_{\gamma_1}(f) & \gamma_1 = \gamma_2\,. \end{cases}$$

Lemma 5.3. *Suppose* Φ *satisfies Assumption D. Assume* $\tau_0 < \ldots < \tau_{k+1}$, $\gamma_1 < \ldots < \gamma_r$ *and there exist an* $\epsilon > 0$ *such that* $(\tau_1 - \epsilon, \tau_k + \epsilon) \subset \Omega$ *as well as* $(\gamma_1 - \epsilon, \gamma_r + \epsilon) \subset \Omega$. *Then the functionals* $\rho_{\tau_0,\tau_1}, \rho_{\tau_1,\tau_2}, \ldots, \rho_{\tau_k,\tau_{k+1}}, \delta_{\gamma_1}, \ldots, \delta_{\gamma_r}$ *are linearly independent as elements of the dual space* $\mathcal{N}_\Phi(\Omega)'$.

Proof. Assume

$$\sum_{i=1}^{k+1} \alpha_i \rho_{\tau_{i-1},\tau_i}(f) + \sum_{j=1}^{r} \beta_j \delta_{\gamma_j}(f) = 0$$

for all $f \in \mathcal{N}_\Phi(\Omega)$. For each $i = 1, \ldots, k+1$ we can find an interval $J_i \subset [\tau_{i-1}, \tau_i] \cap \Omega$ such that $J_i \cap \gamma_j = \emptyset$ for all $j = 1, \ldots, r$. By Lemma 5.1 we can find a test function $f_i \in \mathcal{N}_\Phi(\Omega)$ with $\operatorname{supp}(f_i) \subset J_i$ and $\int_{\mathbb{R}} f_i(x)dx = 1$ for all $i = 1, \ldots, k+1$. We then have that $\rho_{\tau_{l-1},\tau_l}(f_i) = 1_{i=l}$ and $\delta_{\gamma_j}(f_i) = 0$ for all $i = 1, \ldots, k+1$ and $j = 1, \ldots, r$. This leads to

$$0 = \sum_{l=1}^{k+1} \alpha_l \rho_{\tau_{l-1},\tau_l}(f_i) + \sum_{j=1}^{r} \beta_j \delta_{\gamma_j}(f_i) = \alpha_i$$

for all $i = 1, \ldots, k+1$. Similarly we can find test functions $f_j \in \mathcal{N}_\Phi(\Omega)$ with $\delta_{\gamma_j}(f_i) = 1_{i=j}$ and deduce that $\beta_j = 0$ for all $j = 1, \ldots, r$. This proves the claim. □

Finally, we can prove that Assumption D implies Assumption B.

Theorem 5.4. *Assume* Φ *satisfies Assumption D and* $K(x,y) = \Phi(x-y)$. *Then the functions*

$$\Delta_K(x, \tau_0, \tau_1)\,,\ \Delta_K(x, \tau_1, \tau_2)\,,\ \ldots\,,\ \Delta_K(x, \tau_k, \tau_{k+1})$$

are linearly independent as functions in $L_2([0,1])$ *for every* $k \in \mathbb{N}$ *and every choice of*

$$-\infty \leq \tau_0 \leq 0 < \tau_1 \leq \tau_2 < \ldots \leq \tau_k < 1 \leq \tau_{k+1} \leq \infty\,,$$

where only two subsequent τ_i *are allowed to be equal.*

Proof. Assume

$$\Big\| \sum_{i=1}^{k+1} \alpha_i \Delta_K(\cdot, \tau_{i-1}, \tau_i) \Big\|_{L_2([0,1])} = 0\,. \tag{5.2}$$

By continuity of Φ, the function $\Delta_K(\cdot, \tau_{i-1}, \tau_i)$ and hence $\sum_{i=1}^{k+1} \alpha_i \Delta_K(\cdot, \tau_{i-1}, \tau_i)$ is continuous. Consequently, (5.2) implies

$$0 = \sum_{i=1}^{k+1} \alpha_i \Delta_K(x, \tau_{i-1}, \tau_i)\,,$$

for all $x \in [0,1]$. By definition of Δ_K (see (4.1))

$$0 = \sum_{i=1}^{k+1} \alpha_i \Delta_K(x, \tau_{i-1}, \tau_i) = \sum_{i=1}^{k+1} \alpha_i \rho_{\tau_{i-1},\tau_i}\big(\Phi(x - \cdot)\big)\,,$$

for all $x \in [0,1]$. Set $\Omega = [0,1]$. By Theorem B.6 the native space $\mathcal{N}_\Phi(\Omega)$ is the closure of the span of the set of functions $\{\Phi(x - \cdot) : x \in \Omega\}$. It follows that

$$0 = \sum_{i=1}^{k+1} \alpha_i \rho_{\tau_{i-1},\tau_i}(f)$$

for all $f \in \mathcal{N}_\Phi(\Omega)$. By Lemma 5.3 we know that $\rho_{\tau_0,\tau_1}, \ldots, \rho_{\tau_k,\tau_{k+1}}$ are linearly independent as elements of the dual space $\mathcal{N}_\Phi(\Omega)'$. Consequently, $\alpha_i = 0$ for all $i = 1, \ldots, k+1$, which proves the claim. □

The following corollary summarizes the results of this section.

Corollary 5.5. *Suppose $\Phi \in L_1(\mathbb{R})$ is symmetric, continuous and is of finite smoothness in the sense of (5.1). If $\widehat{\Phi}$ is strictly positive, the integral kernel $K(x,y) = \Phi(x-y)$ satisfies Assumption B.*

Examples of functions Φ, which satisfy these conditions, are

- the Laplace kernel $\Phi_L(x) = \exp(-|x|)/2$ with $\widehat{\Phi}_L(t) = (1+t^2)^{-1}$,
- the kernel $\Phi(x) = \exp(-|x|)\cos(x)$ with $\widehat{\Phi}(t) = (4+2t^2)(4+t^4)^{-1}$,

p-fold convolutions $\Phi_L * \ldots * \Phi_L$ of the Laplace kernel such as

- the kernel $\Phi(x) = \exp(-|x|)(|x|+1)/4$ with $\widehat{\Phi}(t) = (1+t^2)^{-2}$,
- the kernel $\Phi(x) = \exp(-|x|)(x^2+3|x|+3)/16$ with $\widehat{\Phi}(t) = (1+t^2)^{-3}$,
- the kernel $\Phi(x) = \exp(-|x|)(|x|^3+6x^2+15|x|+15)/96$ with $\widehat{\Phi}(t) = (1+t^2)^{-4}$,

kernels of the type $(1-|x|)_+^p$ for $p = 2, 3, \ldots$ such as

- the kernel $\Phi(x) = 3/2(1-|x|)_+^2$ with $\widehat{\Phi}(x) = \begin{cases} 6(x-\sin(x))/x^3 & x \neq 0\,, \\ 1 & x = 0\,, \end{cases}$
- the kernel $\Phi(x) = 2(1-|x|)_+^3$ with $\widehat{\Phi}(x) = \begin{cases} 12(x^2+2\cos(x)-2)/x^4 & x \neq 0\,, \\ 1 & x = 0\,, \end{cases}$
- the kernel $\Phi(x) = \big((p+1)/2\big)(1-|x|)_+^p$ for $p = 2, 3, \ldots$ with

$$\widehat{\Phi}(x) = \frac{(p+1)!}{x^{p+1}} \sum_{k=p+1}^{\infty} \frac{(-1)^{k-(p+1)} x^{2k-(p+1)}}{k!}\,,$$

to mention a few.

Given any two kernels Φ_1, Φ_2 which satisfy the assumptions of Corollary 5.5, one can construct a new kernel by $\Phi = \Phi_1 * \Phi_2$. Using $\widehat{\Phi}(x) = \widehat{\Phi}_1(x)\widehat{\Phi}_2(x)$ it is easy to check that this new kernel has the required properties. Another method of constructing new kernels is given by

$$\Phi(x) = (1-\lambda)\Phi_0(x) + \lambda/2\big(\Phi_0(x-\mu) + \Phi_0(x+\mu)\big)\,,$$

where $\mu \in \mathbb{R}$, $-\infty < \lambda < 1/2$ and Φ_0 is any of the kernels having the required properties. By

$$\widehat{\Phi}(x) = (1-\lambda(1-\cos(\mu x)))\widehat{\Phi}_0(x)\,,$$

it follows that $|\widehat{\Phi}(x)| \geq \min(1,(1-2\lambda))|\widehat{\Phi}_0(x)|$, which proves the lower bound (5.1) for Φ.

As the example $\exp(-|x|)\cos(x)$ shows, positive definite functions are not necessarily positive. Note that, if Φ satisfies the conditions of Corollary 5.5, then this is also true for $\Phi_h(x) := h^{-1}\Phi(h^{-1}x)$, since $\widehat{\Phi}_h(x) = \widehat{\Phi}(hx)$. Figure 5.1 shows some of the kernels given above and how the corresponding operator acts on step functions.

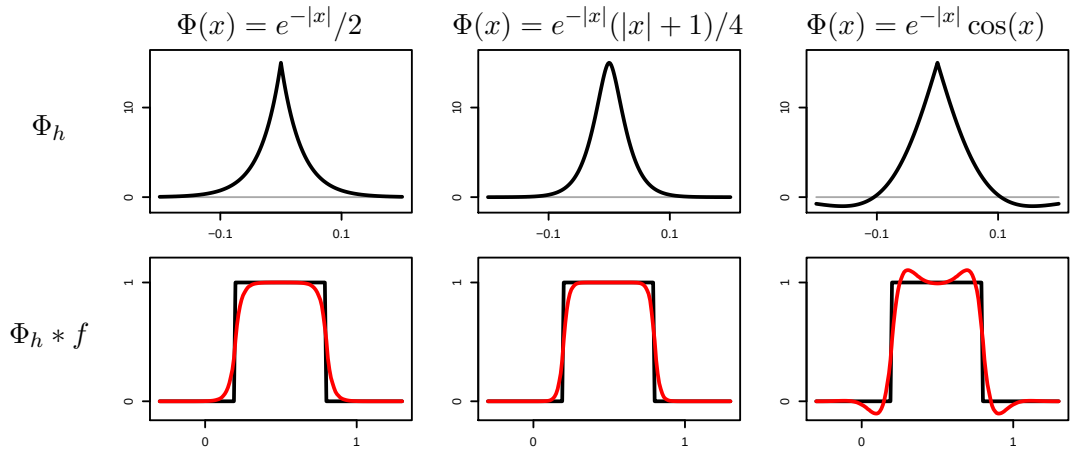

Figure 5.1.: Different kernels Φ and the image $\Phi_h * f$ for $f = 1_{[0.2,0.8)}$. The black line in the lower images shows the original function f and the red line $\Phi_h * f$. Here $\Phi_h(x) = h^{-1}\Phi(h^{-1}x)$ and h is chosen in such a way, that the values of Φ_h of the different kernels coincide at zero.

5.2. Extended sign regular kernels

The previous section established Assumption B for certain functions of finite smoothness. One aim of this section is to give an example of a supersmooth function, namely the Gauss kernel, which satisfies Assumption B.

To this end, we need a stronger concept than positive definiteness.

Definition 5.6. *For $\Phi \in C^{k-1}(\mathbb{R})$, $t_1,\ldots,t_k \in \mathbb{R}$ and $j = 1,\ldots,k$ define*

$$\Phi_{j,t_1,\ldots,t_k}(x) = \begin{cases} \Phi(x-t_j) & : \quad t_{j-1} < t_j \\ \Phi^{(r)}(x-t_j) & : \quad t_{j-r-1} < t_{j-r} = \ldots = t_j\,, \end{cases}$$

where t_0 is set to $-\infty$. Moreover, define

$$\Phi^* \begin{pmatrix} s_1, \ldots, s_k \\ t_1, \ldots, t_k \end{pmatrix} = \det \left(\Phi_{j,t_1,\ldots,t_k}(s_i) \right)_{i,j=1}^k .$$

The function Φ will be called extended sign regular of order k (ESR_k) on $\mathbb{R}$, provided that for each $r = 1, \ldots, k$ there exists $\varepsilon_r \in \{-1, 1\}$ such that

$$\varepsilon_r \Phi^* \begin{pmatrix} s_1, \ldots, s_r \\ t_1, \ldots, t_r \end{pmatrix} > 0 ,$$

for all choices of $s_1 < s_2 < \ldots < s_r$ and $t_1 \leq t_2 \leq \ldots \leq t_r$ with $s_i, t_i \in \mathbb{R}$.

As already indicated, one example for an extended sign regular function is the Gauss kernel.

Lemma 5.7. *The Gauss kernel $\Phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is extended sign regular of all orders on $\mathbb{R}$.*

Proof. See Karlin and Studden (1966), Section 3, Example 5. □

As shown below, extended sign regularity is sufficient for some function Φ to satisfy Assumption B.

Theorem 5.8. *Assume that Φ is extended sign regular of order $k+2$ on $\mathbb{R}$, with $0 < \int \Phi(x)dx < \infty$. Then the functions*

$$\Delta_K(x, \tau_0, \tau_1) , \Delta_K(x, \tau_1, \tau_2) , \ldots , \Delta_K(x, \tau_k, \tau_{k+1})$$

are linearly independent as functions in $L_2([0,1])$ for every choice of

$$-\infty \leq \tau_0 \leq 0 < \tau_1 \leq \tau_2 < \ldots \leq \tau_k < 1 \leq \tau_{k+1} \leq \infty ,$$

where only two subsequent τ_i are allowed to be equal.

Proof. It is equivalent to show that for any set $I_0 = \{0, \ldots, k'\}$ and $I_1 \subset I_0 \backslash \{0, k'\}$ satisfying $\#(I_0) + \#(I_1) = k+1$, and any choice of $0 < \tau_1 < \ldots < \tau_{k'} < 1$ (where $\tau_{k'+1} := \tau_{k+1}$) the functions

$$\Big\{ \Delta_K(x, \tau_i, \tau_{i+1}) : i \in I_0 \Big\} \cup \Big\{ \Delta_K(x, \tau_j, \tau_j) : j \in I_1 \Big\}$$

are linearly independent in $L_2([0,1])$. Assume

$$\Big\| \sum_{i \in I_0} \alpha_i \Delta_K(\cdot, \tau_i, \tau_{i+1}) + \sum_{j \in I_1} \beta_j \Delta_K(\cdot, \tau_j, \tau_j) \Big\|_{L_2([0,1])} = 0 .$$

Denote by $\Phi_0(x) = \int_{-\infty}^{x} \Phi(y)dy$ the primitive of Φ. Since the functions $\Delta_K(\cdot, \tau_i, \tau_{i+1})$ and $\Delta_K(\cdot, \tau_i, \tau_i)$ are continuous we have for all $x \in [0,1]$ that

$$\begin{aligned} 0 &= \sum_{i \in I_0} \alpha_i \Delta_K(x, \tau_i, \tau_{i+1}) + \sum_{j \in I_1} \beta_j \Delta_K(x, \tau_j, \tau_j) \\ &= \sum_{i \in I_0} \alpha_i \big(\Phi_0(x - \tau_i) - \Phi_0(x - \tau_{i+1})\big) + \sum_{j \in I_1} \beta_j \Phi(x - \tau_j) \\ &= \alpha_0 \Phi_0(x - \tau_0) + \sum_{i=1}^{k'} (\alpha_i - \alpha_{i-1}) \Phi_0(x - \tau_i) - \alpha_{k'} \Phi_0(x - \tau_{k'+1}) + \\ &\qquad \sum_{j \in I_1} \beta_j \Phi(x - \tau_j)\,. \end{aligned}$$

Consequently, this must also be true for the derivative, and the equation still holds if we replace Φ_0 by Φ and Φ by Φ'. Since the equality simultaneously holds for all $x \in [0,1]$, it holds for a choice of $k+2$ distinct points $x_0, \ldots, x_{k+1} \in [0,1]$. By the extended sign regularity of Φ we know that the vectors

$$\begin{pmatrix} \Phi(x_0 - \tau_i) \\ \vdots \\ \Phi(x_{k+1} - \tau_i) \end{pmatrix}_{i=0,\ldots,k'+1} \quad \text{and} \quad \begin{pmatrix} \Phi'(x_0 - \tau_j) \\ \vdots \\ \Phi'(x_{k+1} - \tau_j) \end{pmatrix}_{j \in I_1}$$

are linearly independent for $-\infty < \tau_0 \leq 0$ and $1 \leq \tau_{r+1} < \infty$. Hence, we immediately get that $\beta_j = 0$ for all $j \in I_1$. Moreover, $\alpha_0 = \alpha_{k'+1} = 0$ and $(\alpha_{i-1} - \alpha_i) = 0$ for all $i = 1, \ldots, k'$. This leads to $\alpha_i = 0$ for all $i \in I_0$.

Now assume $\tau_0 = -\infty$ and $\tau_{k+1} = \infty$. We then have

$$0 = \alpha_0 \int \Phi(x)dx + \sum_{i=1}^{k'} (\alpha_i - \alpha_{i-1}) \Phi_0(x - \tau_i) + \sum_{j \in I_1} \beta_j \Phi(x - \tau_j)\,.$$

By the same arguments as above we get $\beta_j = 0$ for all $j \in I_1$ and $(\alpha_{i-1} - \alpha_i) = 0$ for all $i = 1, \ldots, k'$. Since $\int \Phi(x)dx > 0$, it follows $\alpha_0 = 0$ and consequently $\alpha_i = 0$ for all $i \in I_0$.

If either τ_0 or τ_{r+1} is infinite and the other one finite, a similar argument applies. □

Remark. It would be sufficient to use a slightly weaker notion instead of extended sign regularity. In the proof we only require that

$$\Phi^* \begin{pmatrix} s_1, \ldots, s_r \\ t_1, \ldots, t_r \end{pmatrix}$$

differs from zero for choices of $s_1 < s_2 < \ldots < s_r$ and $t_1 \leq t_2 \leq \ldots \leq t_r$, where at most two subsequent t_i are allowed to be equal.

Remark. Karlin (1968) shows that for $K(x,y) = \Phi(x-y)$, the fact that Φ is is strictly sign regular of order $r+1$ already implies that Φ is extended sign regular of order r (cf. Karlin, 1968, Corollary 5.2, page 66). The definition for strictly sign regular kernels differs from Definition 5.6 in that $t_1 < t_2 < \ldots < t_r$ instead of $t_1 \leq t_2 \leq \ldots \leq t_r$ is required (consequently it is weaker).

Further examples for strictly sign regular kernels can be found in Karlin (1968).

The following corollary gives the main result of this section.

Corollary 5.9. *The integral Kernel* $K(x,y) = (2\pi)^{-1/2}\exp(-(x-y)^2/2)$ *satisfies Assumption B.*

5.3. Polynomial kernels

Until now, all investigated integral kernels were symmetric and continuous. In this section, we will examine a different class, namely kernels of the type $K(x,y) = \Phi(x-y)$, where Φ has the special form

$$\Phi(x) = (p+1)\, x_+^p\,, \tag{5.3}$$

with $p \in \{0,1,2,\ldots\}$. For positive p this is to be understood as $(x_+)^p$. Note that in the case $p = 0$ this leads to piecewise linear regression with unknown break points, also called multi-phase regression. For $p \geq 1$ the function $\mathrm{K}f$ is piecewise polynomial of order $p+1$, p-times continuously differentiable and has a $(p+1)$-th derivative, which is a step function. The image can be viewed as $(p+1)-th$ order spline without multiple knots. Figure 5.2 shows the kernel for $p = 1,2,3$ and how the corresponding operator act on a certain step function.

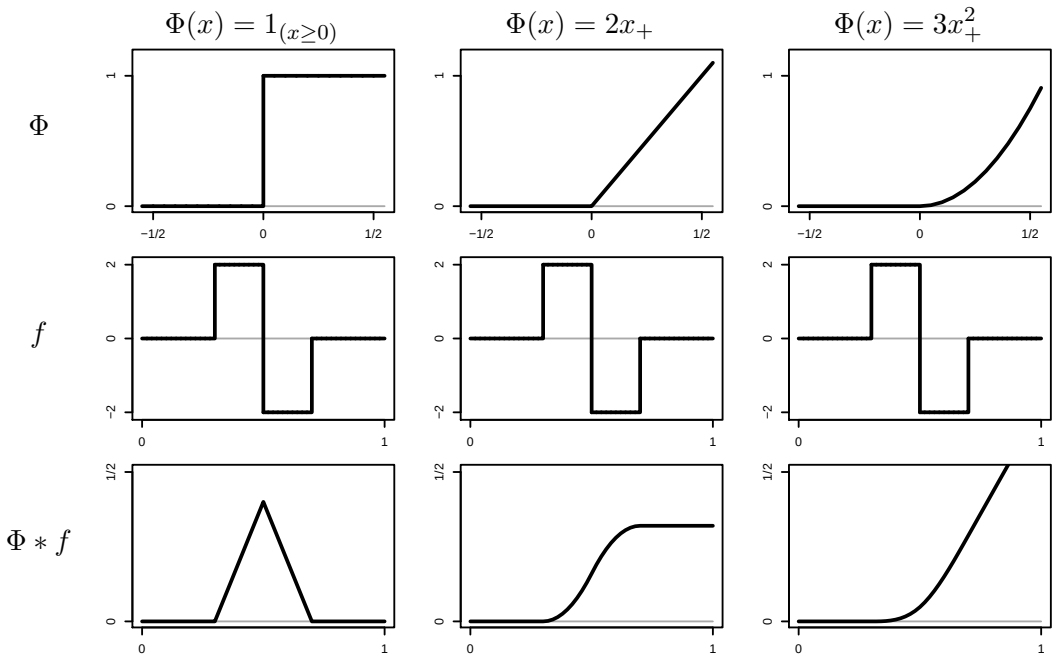

Figure 5.2.: Different kernels Φ of polynomial type and the image $\Phi * f$ for the step function $f(x) = 2(1_{[0.3,0.5)}(x) - 1_{[0.3,0.7)}(x))$.

The following theorem shows that integral operators of this type satisfy Assumption B.

Theorem 5.10. *Assume* $\mathrm{K}f = \int_0^1 \Phi(x-y)f(y)dy$ *and* Φ *is of type (5.3) with* $p \in \{0,1,2,\ldots\}$. *Then the functions*

$$\Delta_K(x,\tau_0,\tau_1)\,,\ \Delta_K(x,\tau_1,\tau_2)\,,\ \ldots\,,\ \Delta_K(x,\tau_k,\tau_{k+1})$$

are linearly independent as functions in $L_2([0,1])$ *for every choice of* $k \in \mathbb{N}$ *and*

$$\tau_0 = 0 < \tau_1 \le \tau_2 \le \ldots \le \tau_k < 1 = \tau_{k+1}\,,$$

where only two subsequent τ_i *are allowed to be equal.*

Proof. Assume

$$\Big\| \sum_{i=1}^{k+1} \alpha_i \Delta_K(\cdot,\tau_{i-1},\tau_i) \Big\|_{L_2([0,1])} = 0\,.$$

We will prove by induction that $\alpha_i = 0$ for all $i = 1,\ldots,k+1$. Compute that

$$\begin{aligned}(\mathrm{K}\,1_{[a,b)})(x) &= (p+1)\int_a^b (x-y)_+^p dy = (p+1)\int_{x-b}^{x-a} y_+^p dy \\ &= (x-a)_+^{p+1} - (x-b)_+^{p+1}\,.\end{aligned}$$

This gives

$$\sum_{i=1}^{k+1} \alpha_i \Delta_K(x,\tau_{i-1},\tau_i)\Big|_{[0,\tau_1]} = \alpha_1 (x-0)^{p+1}\,,$$

and hence $\alpha_1 = 0$.

Now assume $\alpha_j = 0$ for all $j < i$. For $\tau_{i-1} = \tau_i$, we have that $\tau_{i+1} > \tau_i$ and

$$\sum_{i=1}^{k+1} \alpha_i \Delta_K(x,\tau_{i-1},\tau_i)\Big|_{[\tau_{i-1},\tau_{i+1}]} = \alpha_i (x-\tau_i)^p + \alpha_{i+1}(x-\tau_i)^{p+1}\,.$$

As polynomials with different degrees are linearly independent (cf. Achieser, 1992) this gives that $\alpha_{i+1} = 0$ and $\alpha_i = 0$. For $\tau_{i-1} < \tau_i$, observe that

$$\sum_{i=1}^{k+1} \alpha_i \Delta_K(\cdot,\tau_{i-1},\tau_i)\Big|_{[\tau_{i-1},\tau_i]} = \alpha_i (x-\tau_i)^p\,,$$

which directly gives $\alpha_i = 0$. □

Chapter 6

Asymptotic and finite sample distribution for two examples

In this chapter we evaluate the speed of convergence and quality of the approximation by the asymptotic law given in Theorem 4.3 for two different kernels, namely the Laplace kernel (cf. Section 5.1) and the polynomial kernel, which defines the multi-phase regression model (cf. Section 5.3). Moreover, the empirical coverage probability of confidence bands for the estimate of the jump location is assessed. To construct confidence bands, it is necessary to estimate the variance of the parameter estimates. To this end we give an explicit expression of the asymptotic variance in these two examples. In contrast to the direct case, the estimates are not asymptotically independent.

In Section 6.2 we consider the case where K is the convolution with a Laplace kernel with bandwidth h. At the end of the section we investigate the dependence of the variance of the jump estimate on the bandwidth.

Note that, unlike in other chapters, many calculations in this chapter are shortened or even omitted. The reason for this is, that these calculations are mostly basic calculus, not very instructive and would cover several pages, when carried out in detail.

6.1. Multi-phase regression

In this section we consider the case $K(x,y) = 1_{[0,\infty)}(x-y)$ – the so called multi-phase regression. For simplicity, we treat an equidistant design density $h(x) = 1_{[0,1]}(x)$ only.

In a multi-phase regression model it is custom to define an additional intercept b_0 such that

$$Y = \mathrm{K}f + \varepsilon + b_0\,. \tag{6.1}$$

For many operators K the additional intercept leads to identifiability problems, however. If

$$\mathrm{K}\,1_{[\tau_{low},\tau_{up})}|_{[0,1]} \equiv 1\,,$$

then for all $x \in [0,1]$

$$\sum_{i=1}^{k+1} b_i\,\mathrm{K}\,1_{[\tau_{i-1},\tau_i)}(x) + b_0 = \sum_{i=1}^{k+1}(b_i + b_0)\,\mathrm{K}\,1_{[\tau_{i-1},\tau_i)}(x)\,.$$

An example for this is the case where K is the convolution with a symmetric probability density function (and $\tau_{low} = -\infty, \tau_{up} = \infty$). In that setting b_0 is not identifiable. For this

reason it was not included in the model (2.2). Note that $\mathrm{K}\,1_{[\tau_{low},\tau_{up})}|_{[0,1]} \not\equiv 1$ is sufficient for b_0 to be identifiable. This can be easily verified for the case of multi-phase regression.

It is straightforward to expand Theorem 4.3 to the case where the additional parameter b_0 has to be estimated.

Theorem 6.1. *Suppose the Assumptions A and C are met,* $\mathrm{K}f = \Phi * f$ *with* $\Phi = x_+^p$ *for* $p = 0, 1, 2, \ldots$ *and the observations* Y *are given by (6.1) with* $f \in T_{k,R}(0,1)$, $\#\mathcal{J}(f) = k$ *and*

$$f(x) = \sum_{i=1}^{k+1} b_i 1_{[\tau_{i-1},\tau_i)}(x)\,.$$

Suppose that $\hat{f}_n, \hat{b}_0$ *satisfy*

$$\|\,\mathrm{K}\hat{f}_n + \hat{b}_0 - Y\|_n \le \min_{g\in T_{k,R}(0,1), b\in\mathbb{R}} \|\,\mathrm{K}g + b - Y\|_n^2 + o(n^{-1})\,. \tag{6.2}$$

Let $\nu_0(x) = 1$, $\nu_j(x)$ *be defined by (4.5), and the* $(2k+2)\times(2k+2)$ *matrix* V *be defined by its entries (4.6) for* $i, j = 0, \ldots, 2k+2$. *Set* $\theta = (b_0, b_1, \tau_1, b_2, \tau_2, \ldots, b_k, \tau_k, b_{k+1})$ *as the parameter vector given by* f *and* b_0, *and* $\hat{\theta}_n$ *as the corresponding vector of estimates given by* $\hat{f}_n$ *and* $\hat{b}_0$. *Then*

$$\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2 V^{-1})\,.$$

The proof follows the same lines as the proof of Theorem 4.3. An outline is given in Section 7.3.1.

An explicit formula for the covariance matrix

To give an explicit formula for the covariance matrix V^{-1} in Theorem 6.1, the matrix V has to be computed.

Lemma 6.2. *Assume that the design density* h *given by Assumption C satisfies* $h(x) = 1_{[0,1]}(x)$. *Set* $h_i = b_{i+1} - b_i$ *for the jump heights and* $i = 1, \ldots, k$. *If* V *is the matrix given by Theorem 6.1 for the case* $\mathrm{K}f = \Phi * f$ *with* $\Phi(x) = 1_{[0,\infty)}(x)$, *then the entries of* V *are given by*

$$\begin{array}{lcll}
V_{0,0} & = & 1\,, & \\
V_{0,2j} & = & -h_j(1-\tau_j) & 1 \le j \le k\,, \\
V_{0,2j-1} & = & \frac{1}{2}(\tau_j - \tau_{j-1})^2 + (1-\tau_j)(\tau_j - \tau_{j-1}) & 1 \le j \le 2k+1\,, \\
V_{2i-1,2j-1} & = & (\tau_i - \tau_{i-1})\Big(\frac{1}{2}(\tau_j - \tau_{j-1})^2 + (1-\tau_j)(\tau_j - \tau_{j-1})\Big) & 1 \le i < j \le k\,, \\
V_{2i-1,2j-1} & = & V_{2j-1,2i-1} & 1 \le j < i \le k\,, \\
V_{2i-1,2i-1} & = & \Big(\frac{1}{3}(\tau_i - \tau_{i-1})^3 + (1-\tau_i)(\tau_i - \tau_{i-1})^2\Big) & 1 \le i \le k+1\,, \\
V_{2i,2j} & = & h_i h_j (1 - \max(\tau_i, \tau_j)) & 1 \le i, j \le k\,, \\
V_{2i,2j-1} & = & h_i\Big(\frac{1}{2}(\tau_j - \tau_{j-1})^2 + (1-\tau_j)(\tau_j - \tau_{j-1})\Big) & 1 \le i < j \le k+1, \\
V_{2i,2j-1} & = & h_j(1-\tau_j)(\tau_i - \tau_{i-1})\,. & 1 \le j \le i \le k\,.
\end{array}$$

Proof. We have to compute

$$\int_0^1 \nu_i(x)\nu_j(x)dx$$

for $i, j = 0, \ldots, 2k+1$.

First compute Kg for some indicator function g. For $\tau_{i-1} < \tau_i$

$$\begin{aligned}\Delta_K(x, \tau_{i-1}, \tau_i) &= \mathrm{K}\, 1_{[\tau_{j-1},\tau_j)}(x) = (x-\tau_{j-1})_+ - (x-\tau_j)_+ \\ &= (x-\tau_{j-1})1_{[\tau_j-1,\tau_j)}(x) + (\tau_j - \tau_{j-1})1_{[\tau_j,\infty)}(x)\,.\end{aligned}$$

For $j \le i$ we get

$$\begin{aligned}h_i^{-1}\int_0^1 \nu_{2i}(x)\nu_{2j-1}(x)dx &= \int_0^1 \Big(K(x,\tau_i)\,\mathrm{K}\, 1_{[\tau_{j-1},\tau_j)}(x)\Big)dx \\ &= \int_{\tau_i}^1 (\tau_j - \tau_{j-1})dx = (1-\tau_i)(\tau_j - \tau_{j-1})\,.\end{aligned}$$

Now assume $1 \le i < j \le k$. Then

$$\begin{aligned}(h_i)^{-1}\int_0^1 \nu_{2i}(x)\nu_{2j-1}(x)dx &= \int_0^1 \Big(K(x,\tau_i)\,\mathrm{K}\, 1_{[\tau_{j-1},\tau_j)}(x)\Big)dx \\ &= \int_{\tau_{j-1}}^{\tau_j} (x-\tau_{j-1})dx + \int_{\tau_j}^1 (\tau_j - \tau_{j-1})dx \\ &= \frac{1}{2}(\tau_j - \tau_{j-1})^2 + (1-\tau_j)(\tau_j - \tau_{j-1})\,.\end{aligned}$$

Compute

$$(h_i h_j)^{-1}\int_0^1 \nu_{2j}(x)\nu_{2i}(x)dx = \int_0^1 K(x,\tau_i)K(x,\tau_j)dx = 1 - \max(\tau_i,\tau_j)\,,$$

and for $1 \le i \le j-1 \le k$

$$\begin{aligned}\int_0^1 \nu_{2i-1}(x)\nu_{2j-1}(x)dx &= \int_0^1 \Big(\mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x)\,\mathrm{K}\, 1_{[\tau_{j-1},\tau_j)}(x)\Big)dx \\ &= (\tau_i - \tau_{i-1})\Big(\int_{\tau_{j-1}}^{\tau_j} (x-\tau_{j-1})dx + \int_{\tau_j}^1 (\tau_j - \tau_{j-1})dx\Big) \\ &= (\tau_i - \tau_{i-1})\Big(\frac{1}{2}(\tau_j - \tau_{j-1})^2 + (1-\tau_j)(\tau_j - \tau_{j-1})\Big)\,.\end{aligned}$$

Moreover, for $1 \le i \le k+1$

$$\begin{aligned}\int_0^1 \nu_{2i}(x)^2 dx &= \int_0^1 \big(\mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x)\big)^2 dx \\ &= \int_{\tau_{i-1}}^{\tau_i} (x-\tau_{i-1})^2 dx + \int_{\tau_i}^1 (\tau_i - \tau_{i-1})^2 dx \\ &= \frac{1}{3}(\tau_i - \tau_{i-1})^3 + (1-\tau_i)(\tau_i - \tau_{i-1})^2\,.\end{aligned}$$

The calculations for $\int_0^1 \nu_0(x)\nu_i(x)dx$ and $i = 0, \ldots, 2k+1$ are straightforward and thus omitted. □

The variance of $(\hat{b}_0, \hat{b}_1, \hat{\tau}_1, \hat{b}_2, \hat{\tau}_2, \ldots, \hat{\tau}_k, \hat{b}_{k+1})^t$ is given by the inverse of V times σ^2. By basic calculations it can be checked that the covariance matrix has entries defined by

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = \begin{cases} \sigma^2 \frac{4(\tau_{i+1}-\tau_{i-1})}{(b_{i+1}-b_i)^2(\tau_i-\tau_{i-1})(\tau_{i+1}-\tau_i)} & j = i\,, \\ \sigma^2 \frac{2}{(b_{i+1}-b_i)(b_{i+2}-b_{i+1})(\tau_{i+1}-\tau_i)} & j = i+1\,, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathrm{Cov}(\hat{\tau}_i, \hat{b}_j) = \begin{cases} \sigma^2 \frac{6}{(b_{i+1}-b_i)(\tau_{j+1}-\tau_j)^2} & j \in \{i, i+1\}\,, \\ \sigma^2 \frac{-2}{(b_2-b_1)\tau_1} & j = 0, i = 1\,, \\ 0 & \text{otherwise} \end{cases}$$

as well as

$$\mathrm{Cov}(\hat{b}_i, \hat{b}_j) = \begin{cases} \sigma^2 \frac{4}{\tau_1} & i = j = 0\,, \\ \sigma^2 \frac{12}{(\tau_i-\tau_{i-1})^3} & i = j > 0\,, \\ \sigma^2 \frac{-6}{\tau_1^2} & i = 0\,, j = 1\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

This means that in this special case, the estimates of the heights $\hat{b}_i$ depend only on the neighboring jump estimates, and the estimates of the jump points $\hat{\tau}_i$ depend only on the neighboring height estimates b_{i-1}, b_i and the neighboring jump estimates $\hat{\tau}_{i-1}, \hat{\tau}_{i+1}$. Note that the variance matrix is not symmetric with respect to $\hat{b}_0$. The estimated intercept depends only on $\hat{b}_1$ and $\hat{\tau}_1$ but not on $\hat{b}_{k+1}$ and $\hat{\tau}_k$. This is due to the fact that we defined the kernel by $K(x, y) = 1_{[0,\infty)}(x - y)$. If we had chosen $K(x, y) = 1_{(-\infty,0]}(x - y)$ instead the dependence would be different.

Application of the above results to the case $k = 1$ leads to

$$\sigma^{-2}\,\mathrm{Var}\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{\tau}_1 \\ \hat{b}_2 \end{pmatrix} = \begin{pmatrix} \frac{4}{\tau_1} & \frac{-6}{\tau_1^2} & \frac{2}{(b_1-b_2)\tau_1} & 0 \\ \frac{-6}{\tau_1^2} & \frac{12}{\tau_1^3} & \frac{-6}{(b_1-b_2)\tau_1^2} & 0 \\ \frac{2}{(b_1-b_2)\tau_1} & \frac{-6}{(b_1-b_2)\tau_1^2} & \frac{4}{(b_1-b_2)^2(1-\tau_1)\tau_1} & \frac{-6}{(b_1-b_2)(1-\tau_1)^2} \\ 0 & 0 & \frac{-6}{(b_1-b_2)(1-\tau_1)^2} & \frac{12}{(1-\tau_1)^3} \end{pmatrix}. \tag{6.3}$$

For $k = 2$ and $h_1 := (b_2 - b_1), h_2 := (b_3 - b_2)$ the covariance matrix has the following form

$$\sigma^{-2}\,\mathrm{Var}((\hat{b}_0, \hat{b}_1, \hat{\tau}_1, \hat{b}_2, \hat{\tau}_2, \hat{b}_3)^t) =$$

$$\begin{pmatrix} \frac{4}{\tau_1} & \frac{-6}{\tau_1^2} & \frac{-2}{h_1\tau_1} & 0 & 0 & 0 \\ \frac{-6}{\tau_1^2} & \frac{12}{\tau_1^3} & \frac{6}{h_1\tau_1^2} & 0 & 0 & 0 \\ \frac{-2}{h_1\tau_1} & \frac{6}{h_1\tau_1^2} & \frac{4\tau_2}{h_1^2(\tau_2-\tau_1)\tau_1} & \frac{6}{h_1(\tau_2-\tau_1)^2} & \frac{2}{h_1h_2(\tau_2-\tau_1)} & 0 \\ 0 & 0 & \frac{6}{h_1(\tau_2-\tau_1)^2} & \frac{12}{(\tau_2-\tau_1)^3} & \frac{6}{h_2(\tau_2-\tau_1)^2} & 0 \\ 0 & 0 & \frac{2}{h_1h_2(\tau_2-\tau_1)} & \frac{6}{h_2(\tau_2-\tau_1)^2} & \frac{4(1-\tau_1)}{h_2^2(\tau_2-\tau_1)(1-\tau_2)} & \frac{6}{h_2(1-\tau_2)^2} \\ 0 & 0 & 0 & 0 & \frac{6}{h_2(1-\tau_2)^2} & \frac{12}{(1-\tau_2)^3} \end{pmatrix}.$$

Some simulations

In order to evaluate the speed of convergence and quality of the approximation by the asymptotic law given in Theorem 6.1, we perform a small simulation study.

Test Bed 6.3. *Assume the observations Y are generated by*

$$Y_i = \mathrm{K}(-3 \cdot 1_{[0,0.5)} + 3 \cdot 1_{[0.5,1)})(i/n) + 0.5\,\varepsilon_i\,, \quad i = 1, \ldots, n\,,$$

where $\mathrm{K}f(x) = \int 1_{[0,\infty)}(x-y)f(y)dy$ *and* $\varepsilon_i \sim N(0,1)$ *for* $i = 1, \ldots, n$.

In this setting 10^5 simulation runs are performed. For each simulation run the least squares estimates $(\hat{b}_0, \hat{b}_1, \hat{\tau}, \hat{b}_2)$ are computed. Moreover, an estimate $\hat{\sigma}^2$ of σ^2 is computed by the mean of the squared residuals of the fitted model. By Theorem 6.1 and Equation (6.3) it holds

$$\sqrt{n}(\hat{\tau} - \tau) \overset{n\to\infty}{\longrightarrow} N\Big(0, \frac{4\sigma^2}{(b_1 - b_2)^2(1-\tau)\tau}\Big)\,.$$

As all estimates are consistent by Slutzky's theorem the asymptotic variance can be estimated by

$$\frac{4(\hat{\sigma})^2}{(\hat{b}_1 - \hat{b}_2)^2(1-\hat{\tau})\hat{\tau}}$$

without changing the limit law. This can be used to compute confidence intervals for $\hat{\tau}$.

Figure 6.1 shows the empirical and the asymptotic distribution of $\hat{\tau}$ for different sample sizes n. In the case $n = 100$ it is noticeable that the empirical distribution function is not smooth. This is due to the fact that the derivative of integral kernel is discontinuous at zero. As a consequence the speed of convergence of the asymptotic approximation is rather slow. For $n = 100$ the approximation is not very good, for $n = 1000$ it is quite reasonable and for $n = 10000$ the fit seems almost perfect.

The fact that the asymptotic law gives a poor approximation for small sample sizes was already noted by Hinkley (1969), who derived a finite sample size approximation for the case $k = 1$.

The quality of approximation by the asymptotic law is reflected in the empirical coverage of the confidence bands for $\hat{\tau}$ as displayed in Figure 6.2. For $n = 100$ the confidence bands are very anti-conservative. As n grows, the empirical coverage approaches the nominal coverage, and for $n = 1000$ the confidence bands are only slightly anti-conservative. For $n = 10000$ the nominal coverage is nearly obtained.

Finally, Figure 6.3 and 6.4 each show two exemplary simulated data sets for $n = 100$ and $n = 1000$, respectively. Note that Theorem 6.1 implies that for $\hat{\theta} = (\hat{b}_0, \hat{b}_1, \hat{\tau}, \hat{b}_2)^t$ and θ the corresponding vector of parameters of f, we have that the quadratic form

$$\sigma^{-2}(\hat{\theta} - \theta)^t V(\hat{\theta} - \theta)$$

is distributed according to a χ^2-distribution with four degrees of freedom. This still holds true if σ^{-2} and V are replaced by consistent estimates $\hat{\sigma}^2$ and $\hat{V}$, respectively. Hence, we obtain a $(1-\alpha)$-confidence ellipsoid for θ in $\mathbb{R}^4$ by

$$(\hat{\sigma})^{-2}(\hat{\theta} - \theta)^t(\hat{V})(\hat{\theta} - \theta) \le \chi^2_4(1-\alpha)\,, \tag{6.4}$$

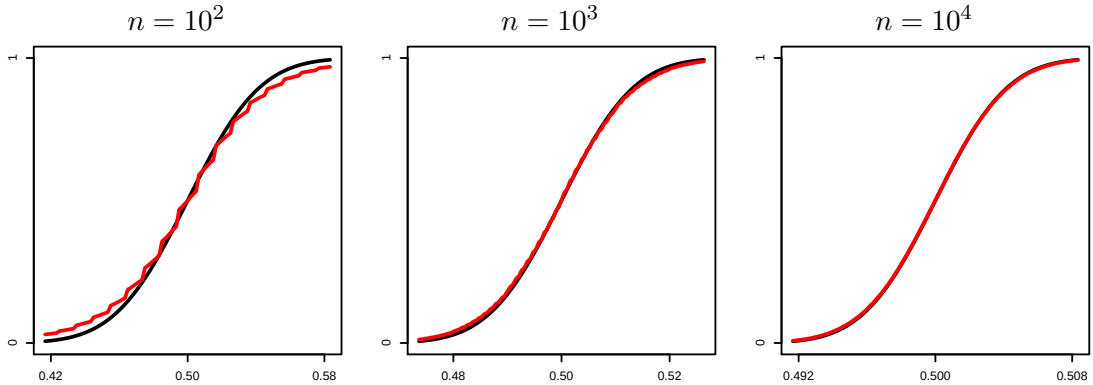

Figure 6.1.: Asymptotic and finite sample size distribution of the jump location in two-phase regression for different sample sizes n. 10^5 simulation runs with data generated according to Test Bed 6.3 were performed. The finite sample size distribution is given by the red line and the asymptotic distribution by the black line.

where $\hat{\sigma}^2$ is the estimate given above, $\chi^2_4(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with four degrees of freedom and $\hat{V}$ is obtained by inserting the least squares estimate $\hat{\theta}$ in the entries of V as given by Lemma 6.2. By projection of this confidence ellipsoid to the two dimensional subspaces containing (τ, b_1) and (τ, b_2) a uniform confidence band for the function f can be computed. Given τ the maxima and minima of b_1 and b_2 must be computed under the constraints imposed by (6.4). The corresponding confidence ellipses are shown in the lower rows of Figure 6.3 and 6.4.

The confidence set for θ clearly induces a confidence set for $\mathrm{K}f$. However, the boundary of the confidence ellipsoid for θ does not necessarily map to the boundary of the confidence band for $\mathrm{K}f$. For each x the maximum and minimum of

$$b_0 + b_1 \,\mathrm{K}\, 1_{[0,\tau)}(x) + b_2 \,\mathrm{K}\, 1_{[\tau,1)}(x)$$

has to be computed under the constraints imposed by (6.4). As the analytics are quite messy, this was done numerically. The results are given by the dashed lines in the first rows of Figure 6.3 and 6.4. Note that if the true parameter θ is not contained in the respective confidence band, this does not imply that $\mathrm{K}f$ is not contained in the respective confidence band. This is shown in the right column of Figure 6.3.

6.2. Convolution with Laplace

In this section we consider the quality of the normal approximation in Theorem 4.3. Let K be convolution with the Laplace kernel $\Phi_{L,h}(x) = 1/(2h)\exp(-|x|/h)$ and f be a step function with one jump. We first give an explicit formula for the matrix V defined by Theorem 4.3. For simplicity, we consider only the uniform design case.

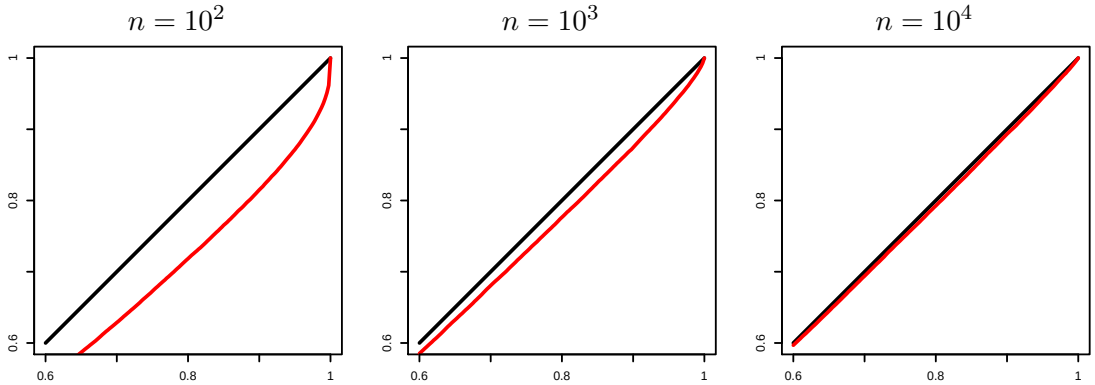

Figure 6.2.: Empirical coverage probability for different sample sizes n of confidence bands for the estimated jump location in two-phase regression (red lines). 10^5 simulation runs with data generated according to Test Bed 6.3 were performed. The x-axis shows the nominal and the y-axis the empirical coverage.

Lemma 6.4. *Assume that the design density h given by Assumption C satisfies $h(x) = 1_{[0,1]}(x)$. If V is the matrix given by Theorem 4.3 for the case where $\mathrm{K}f = \Phi_{L,h} * f$ and $k = 1$, then the entries of the symmetric matrix V are given by*

$$
\begin{aligned}
V_{1,1} &= h\Big(e^{\frac{-\tau}{h}} - \frac{1}{8}\big(e^{\frac{-2(1-\tau)}{h}} + e^{\frac{-2\tau}{h}}\big) - \frac{3}{4}\Big) + \tau\,,\\
V_{1,2} &= \frac{(b_1-b_2)}{8}\Big(e^{\frac{-2\tau}{h}} - e^{\frac{-2(1-\tau)}{h}} + 4\big(1 - e^{\frac{-\tau}{h}}\big)\Big)\,,\\
V_{1,3} &= \frac{h}{8}\Big(6 + e^{\frac{-2\tau}{h}} + e^{\frac{-2(1-\tau)}{h}} - 4\big(e^{\frac{-\tau}{h}} + e^{\frac{-(1-\tau)}{h}}\big)\Big)\,,\\
V_{2,2} &= \frac{(b_1-b_2)^2}{8h}\Big(2 - e^{\frac{-2\tau}{h}} - e^{\frac{-2(1-\tau)}{h}}\Big)\\
V_{2,3} &= \frac{(b_1-b_2)}{8}\Big(e^{\frac{-2(1-\tau)}{h}} - e^{\frac{-2\tau}{h}} + 4\big(1 - e^{\frac{-(1-\tau)}{h}}\big)\Big)\\
V_{3,3} &= h\Big(e^{\frac{-(1-\tau)}{h}} - \frac{1}{8}\big(e^{\frac{-2(1-\tau)}{h}} + e^{\frac{-2\tau}{h}}\big) - \frac{3}{4}\Big) + (1-\tau)\,.
\end{aligned}
$$

Proof. For $a < b$ compute that

$$
\begin{aligned}
&\Delta_K(x,a,b)\\
&= \frac{1}{2}\int_{(y-b)/h}^{(y-a)/h} e^{-|x|}dx = \frac{1}{2}\Big(\big(e^{(y-a)/h} - e^{(y-b)/h}\big)1_{(-\infty,a]}(y) +\\
&\qquad \big(2 - e^{-(y-a)/h} - e^{(y-b)/h}\big)1_{(a,b)}(y) + l\big(e^{(b-y)/h} - e^{(a-y)/h}\big)1_{[\tau_i,\infty)}(y)\Big)\,.
\end{aligned}
$$

This leads to

$$
V_{1,1} = \int_0^1 \nu_1^2(x)dx = \int_0^1 \Delta_K(x,-\infty,\tau)^2 dx
$$

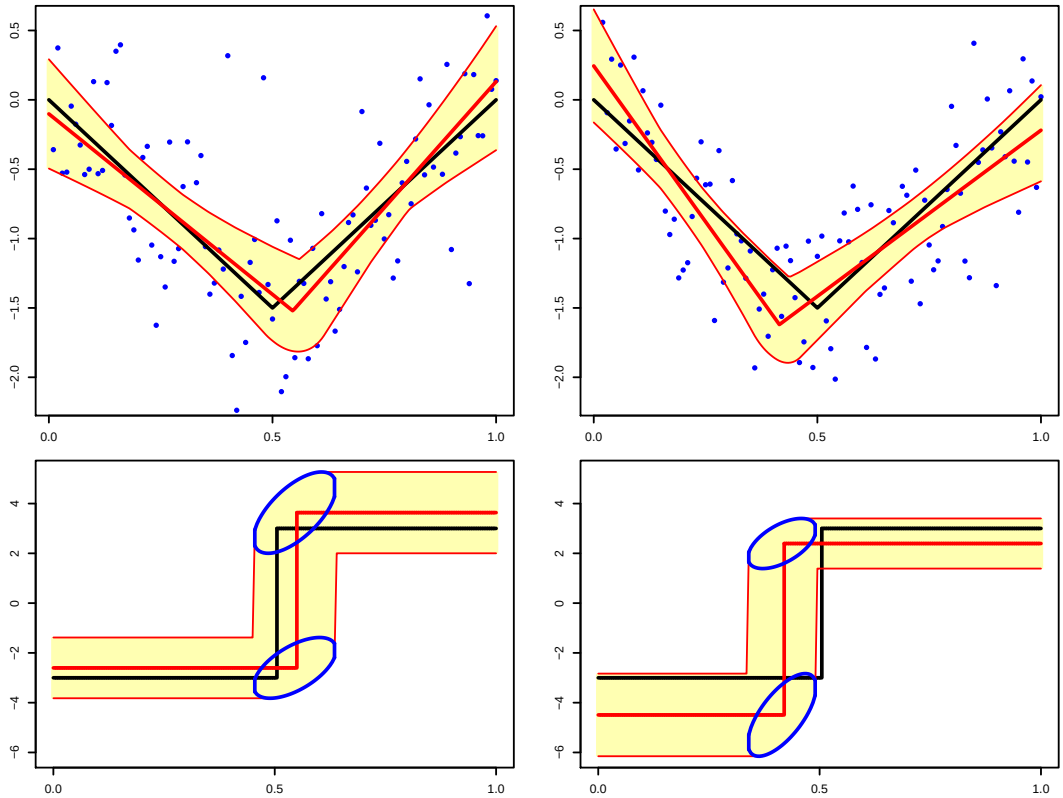

Figure 6.3.: Simulated data examples and confidence bands for the two phase regression with $n = 100$ observations (Test Bed 6.3). The first row displays the observations and the reconstruction in the image space, and the second row shows the estimate for the signal f. The black line represents the true function and the thick red line the estimate. The thin red lines show the confidence bands for the function and the blue dots the observations. The blue ellipses in the second row show the confidence sets for (τ, b_1) and (τ, b_2), respectively.

$$
\begin{aligned}
&= \frac{1}{4}\int_0^\tau (2 - e^{-(\tau-x)/h})^2 dx + \frac{1}{4}\int_\tau^1 e^{-2(x-\tau)/h} dx \\
&= \int_0^\tau \big(1 - e^{-(\tau-x)/h} + \frac{1}{4} e^{-2(\tau-x)/h}\big) dx + \frac{h(1 - e^{-2(1-\tau)/h})}{8} \\
&= h\Big(e^{\frac{-\tau}{h}} - \frac{1}{8}\big(e^{\frac{-2(1-\tau)}{h}} + e^{\frac{-2\tau}{h}}\big) - \frac{3}{4}\Big) + \tau\,.
\end{aligned}
$$

Also,

$$
\begin{aligned}
V_{2,2} &= \int_0^1 \Phi_{L,h}(x-\tau)^2 dx \\
&= \frac{1}{4h^2}\int_0^\tau \exp(2(x-\tau)/h) dx + \frac{1}{4h^2}\int_\tau^1 \exp(-2(x-\tau)/h) dx
\end{aligned}
$$

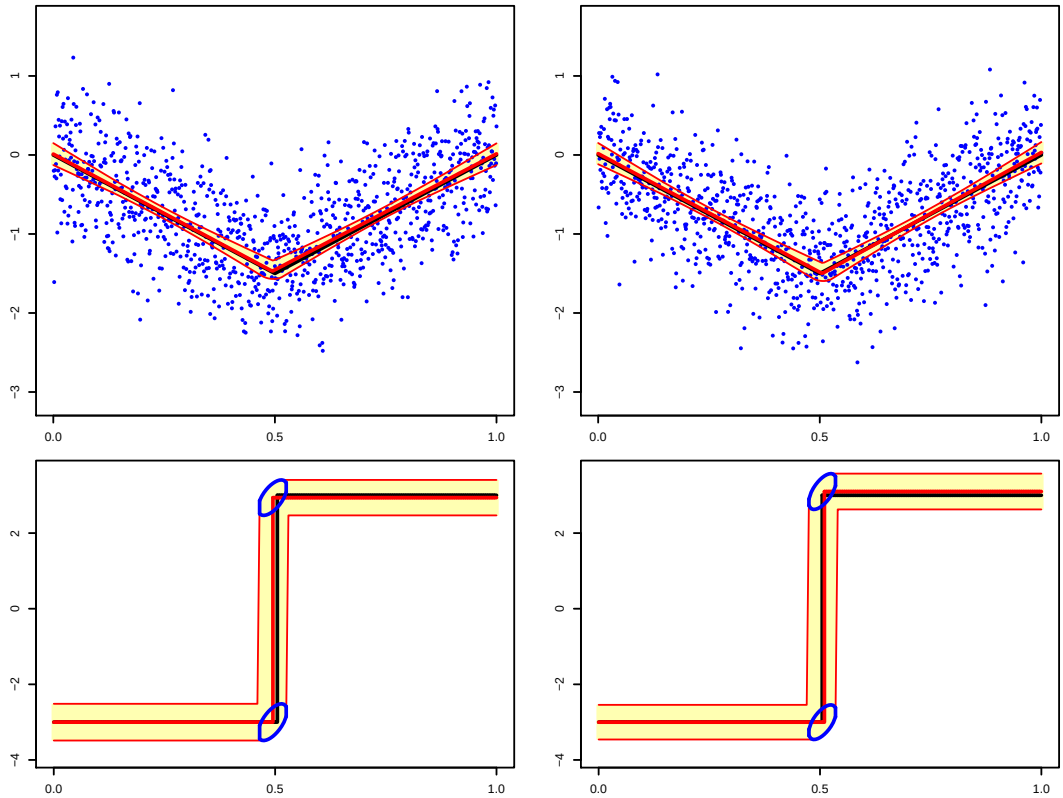

Figure 6.4.: Simulated data examples and confidence bands for two phase regression with $n = 1000$ observations (Test Bed 6.3). See description of Figure 6.3, page 50.

$$= \frac{1}{8h}\left(1 - \exp(-2\tau/h)\right) + \frac{1}{8h}\left(1 - \exp(-2(1-\tau)/h)\right)$$

The calculations for the other entries of V are similar and likewise straightforward. Therefore they are omitted. □

Now turn to the inverse of V. We have

$$\det(V) = V_{1,1}V_{2,2}V_{3,3} + 2V_{1,2}V_{1,3}V_{2,3} - V_{2,3}^2V_{1,1} - V_{1,3}^2V_{2,2} - V_{1,2}^2V_{3,3}\,.$$

Furthermore, V^{-1} is given by

$$\frac{1}{\det(V)}\begin{pmatrix} -V_{2,3}{}^2 + V_{2,2}\,V_{3,3} & V_{1,3}\,V_{2,3} - V_{1,2}\,V_{3,3} & -V_{1,3}\,V_{2,2} + V_{1,2}\,V_{2,3} \\ V_{1,3}\,V_{2,3} - V_{1,2}\,V_{3,3} & -V_{1,3}{}^2 + V_{1,1}\,V_{3,3} & V_{1,2}\,V_{1,3} - V_{1,1}\,V_{2,3} \\ -V_{1,3}\,V_{2,2} + V_{1,2}\,V_{2,3} & V_{1,2}\,V_{1,3} - V_{1,1}\,V_{2,3} & -V_{1,2}{}^2 + V_{1,1}\,V_{2,2} \end{pmatrix}, \qquad (6.5)$$

where the $V_{i,j}$ are given by Lemma 6.4.

Some simulations

As in Section 6.1, we again perform a small simulation study to evaluate the speed of convergence and quality of the approximation by the asymptotic law given in Theorem 4.3.

Test Bed 6.5. *Assume the observations Y are generated by*

$$Y_i = \mathrm{K}(-3 \cdot 1_{[0,0.5)} + 3 \cdot 1_{[0.5,1)})(i/n) + \sigma^2 \varepsilon_i \,, \quad i = 1, \ldots, n \,,$$

where $\mathrm{K}f = \Phi_{L,h} * f$ *and* $\varepsilon_i \sim N(0,1)$ *for* $i = 1, \ldots, n$.

In this setting 10^5 simulation runs are performed. The empirical and asymptotic distribution of $\hat{\tau}$ for different sample sizes n is shown in Figure 6.5.

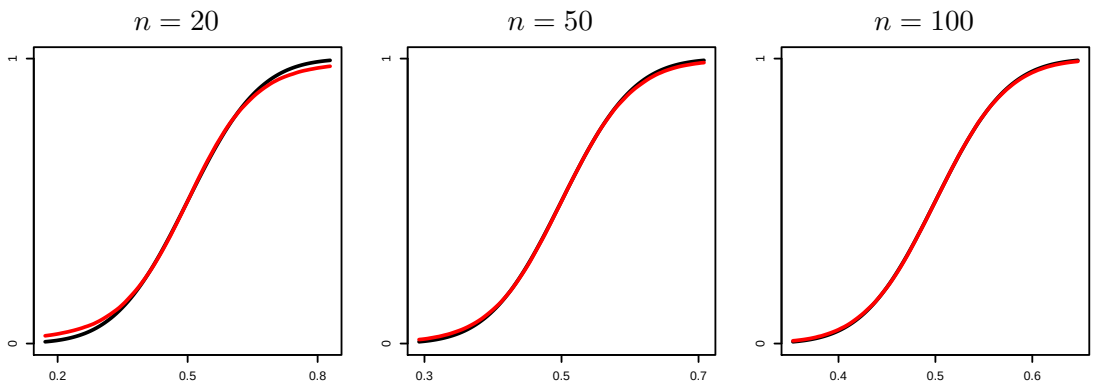

Figure 6.5.: Asymptotic and finite sample size distribution of the estimate of the jump location for Test Bed 6.5 with $\sigma = 0.2$ and $h = 1$ for different sample sizes n and 10^5 simulations. The finite sample size distribution is given by the red line and the asymptotic distribution by the black line.

The approximation by the asymptotic law already is reasonable for a small sample size of $n = 20$. For $n = 100$ the fit is almost perfect. When compared to the results of Section 6.1 the quality of approximation is much better, which is due to fact that the integral kernel $\Phi_{L,h}(x - y)$ is continuous.

As in Section 6.1, the asymptotic coverage of the confidence band for $\hat{\tau}$ is computed. To this end we have to estimate the variance of $\hat{\tau}$. This is again done by replacing (b_1, τ, b_2) in (6.5) with their respective least squares estimates. This gives an estimate $\hat{V}^{-1}$ of V^{-1}. An estimate of $\mathrm{Var}(\hat{\tau})$ is then given by $\hat{\sigma}^2 (\hat{V}^{-1})_{2,2}$, where $\hat{\sigma}^2$ is the mean of the squared residuals of the fitted model.

Figure 6.6 shows that for $n = 20$ the confidence bands for $\hat{\tau}$ are slightly conservative if the nominal coverage is larger than 0.8. For $n = 50$ and $n = 100$ the empirical and the nominal coverage probabilities nearly coincide. Consequently, this procedure gives useful results even if the number of observations is small.

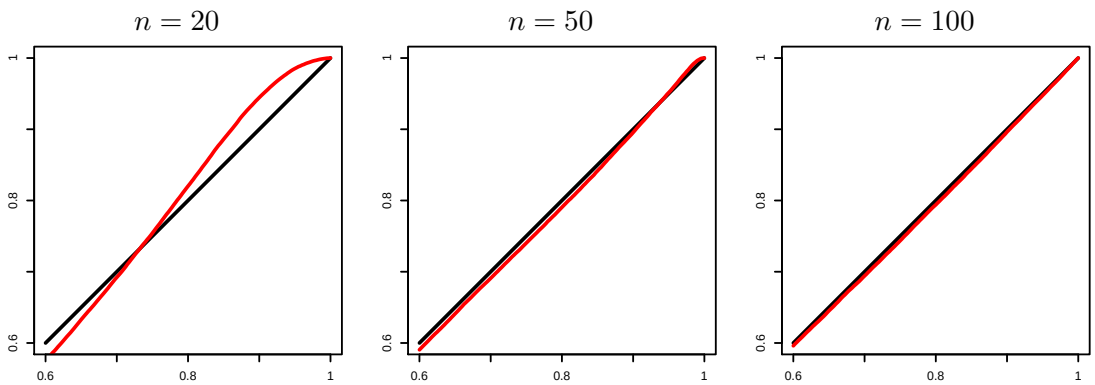

Figure 6.6.: Empirical coverage probability of confidence bands for the estimated jump location for Test Bed 6.5 with $\sigma = 0.2$ and $h = 1$ for different sample sizes n and 10^5 simulations (red lines). The x-axis shows the nominal and the y-axis the empirical coverage probability.

Estimation of the jump location for large and small bandwidths h

The result $|\hat{\tau}-\tau| = O_P(n^{-1/2})$ does not reflect the fact that the variance of $\hat{\tau}$ depends on the ill-posedness of the problem. One possibility of measuring the difficulty of reconstructing the jump location τ is an analysis of the variance for small and large bandwidths h. If h is very small, the convolution operator is close to the identity and one would suspect the variance of $\hat{\tau}$ to be small. On the other hand, if h is very large, the step function is heavily blurred and the estimation of τ is more difficult. Note that, we do not analyze the simultaneous limit of $h \to 0$ (or $h \to \infty$) and $n \to \infty$, but only the limit of the asymptotic variance.
First, consider the case where the bandwidth is small. Observe that

$$\begin{aligned}
V_{1,1} &= \tau - (3/4)h + O(h^2)\,, \\
V_{1,2} &= (b_1 - b_2)/2 + O(h)\,, \\
V_{1,3} &= (3/4)h + O(h^2)\,, \\
hV_{2,2} &= (b_1 - b_2)^2/4 + O(h)\,, \\
V_{2,3} &= (b_1 - b_2)/2 + O(h)\,, \\
V_{3,3} &= (1-\tau) - (3/4)h + O(h^2)\,,
\end{aligned}$$

as $h \to 0$. This gives that

$$h\,\det(V) = \frac{1}{4}(b_1 - b_2)^2(1-\tau)\tau + O(h)\,,$$

and – after some basic calculations – we arrive at

$$V^{-1} = \begin{pmatrix} \tau^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (1-\tau)^{-1} \end{pmatrix} + O(h) \qquad \text{as} \quad h \to 0\,.$$

This means that for small h the variance of $\hat{\tau}$ is negligible when compared to the variances of $\hat{b}_1$ and $\hat{b}_2$. A comparison with the results of Chapter 3 yields, that the limit of V^{-1} for $h \to 0$ corresponds to the variance in the direct case.

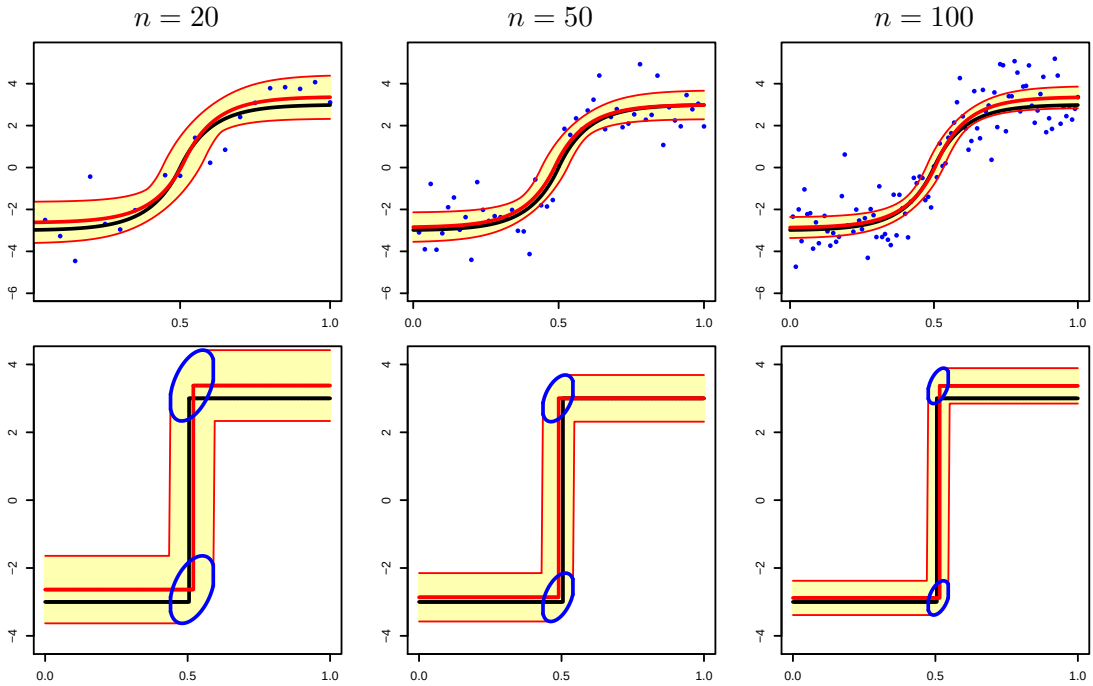

Figure 6.7.: Simulated data examples and confidence bands for Test Bed 6.5 with bandwidth $h = 0.1$ and $\sigma = 1$. See description of Figure 6.3, page 50.

Now consider the case where the bandwidth is large. Extensive calculations give for $h \to \infty$ that

$$\det(V) = \frac{h^{-6}}{144}(b_1 - b_2)^2(1-\tau)^3\tau^3 + o(h^{-6})\,,$$

as well as

$$\det(V)V^{-1} =$$

$$\begin{pmatrix} \frac{3(1+3\tau)}{\tau^3}h^{-4} + o(h^{-4}) & \frac{-3(1+2\tau)}{b(1-\tau)\tau^3}h^{-3} + o(h^{-3}) & \frac{9}{(1-\tau)\tau}h^{-4} + o(h^{-4}) \\ \frac{-3(1+2\tau)}{b(1-\tau)\tau^3}h^{-3} + o(h^{-3}) & \frac{3}{b^2(1-\tau)^3\tau^3}h^{-2} + o(h^{-2}) & \frac{-3(3-2\tau)}{b(1-\tau)^3\tau}h^{-3} + o(h^{-3}) \\ \frac{9}{(1-\tau)\tau}h^{-4} + o(h^{-4}) & \frac{-3(3-2\tau)}{b(1-\tau)^3\tau}h^{-3} + o(h^{-3}) & \frac{3(4-3\tau)}{(1-\tau)^3}h^{-4} + o(h^{-4}) \end{pmatrix},$$

where $b := (b_1 - b_2)$. Consequently, the entries of V^{-1} behave like

$$\begin{pmatrix} h^2 & h^3 & h^2 \\ h^3 & h^4 & h^3 \\ h^2 & h^3 & h^2 \end{pmatrix} \qquad \text{as} \quad h \to \infty\,.$$

This means that for large h the variance of $\hat{\tau}$ dominates the variances of $\hat{b}_1$ and $\hat{b}_2$.

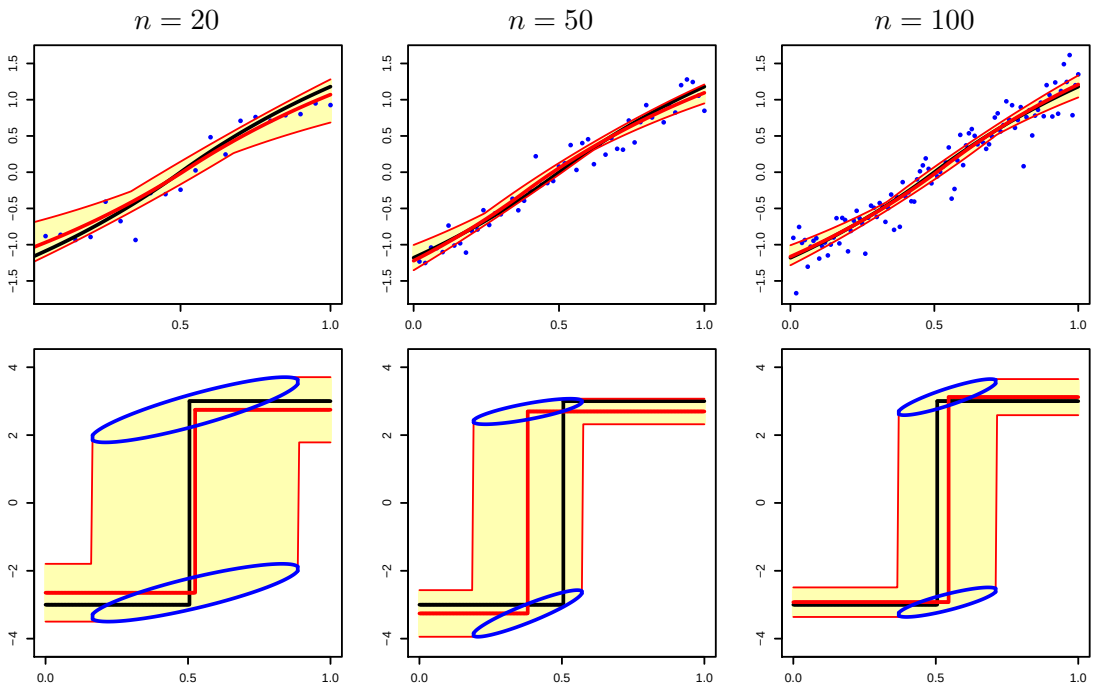

Figure 6.8.: Simulated data examples and confidence bands for Test Bed 6.5 with bandwidth $h = 1$ and $\sigma = 0.2$. See description of Figure 6.3, page 50.

This is also reflected in Figure 6.7 and Figure 6.8, displaying simulated data examples from Test Bed 6.5 for $h = 0.1$ and $h = 1$, respectively. A comparison of the two figures yields that the variance of $\hat{\tau}$ is much larger for $h = 1$ (when compared to the variances of $\hat{b}_1$ and $\hat{b}_2$) than for $h = 0.1$. This is shown by the fact that the confidence ellipses in the lower row of Figure 6.8 spread more in direction of the x-axis (which corresponds to $\hat{\tau}$) than in direction of the y-axis (which corresponds to $\hat{b}_1$ and $\hat{b}_2$). In comparison the confidence ellipses in Figure 6.7 extend more in direction of the y-axis than in direction of the x-axis.

The computation of the confidence bands for Figure 6.7 and Figure 6.8 is similar to the case of two-phase regression as described in Section 6.1. The only difference is that the corresponding quadratic form asymptotically has a χ^2-distribution with three degrees of freedom (compared to four) as we do not have to estimate an additional intercept.

Chapter 7

Proof of main result

This chapter provides the proofs of Theorem 4.3 and Theorem 4.4. The proof of Theorem 4.3 is divided into four parts. After introducing some technical tools in Section 7.1, we start by giving the consistency for the restricted estimate in Section 7.2. This is used in Section 7.3 to show the asymptotic normality of the parameter estimates. In Section 7.4, it is shown that (given some conditions on the smoothing parameter λ_n) the restricted estimate $\hat{f}_n$ and the penalized estimate $\hat{f}_{\lambda_n}$ asymptotically coincide with probability one. Finally, Section 7.5 gives the proof of Theorem 4.4.

To be more precise, the main steps of the proof can be summarized as follows:

1. Compute the L_2-entropy of the space $T_{k,R}(a,b)$ for $-\infty < a < b < \infty$.
2. Use that $\|\mathrm{K}f\|_n$ can be bounded by the $L_2([a,b])$ norm of $\mathrm{K}f$ for suitable chosen $-\infty < a < b < \infty$ to give an upper bound for the empirical entropy of the space $\{\mathrm{K}f : f \in T_{k,R}(\tau_{low}, \tau_{up})\}$.
3. Use the entropy bound and the fact that K^{-1} is continuous on the set of functions $\{\mathrm{K}f : f \in T_{k,R}(\tau_{low}, \tau_{up})\}$ to derive consistency of $\mathrm{K}\hat{f}_n$ and $\hat{f}_n$.
4. Give a local stochastic expansion of the minimized process and use this to derive asymptotic normality.
5. Derive an exponential inequality for
$$\sup_{g \in T_{\infty,R}} \frac{|\langle \mathrm{K}g, \varepsilon \rangle|_n}{\eta(\|\mathrm{K}g\|_n\,,\, \#\mathcal{J}(g)+1)}\,,$$
where $\eta(x,y) = xy(1+\log(y/x))$. Use this exponential inequality to show that the penalized least squares estimate asymptotically coincides with the restricted least squares estimate.
6. Compute the Kullback-Leibler distance of the joint distributions of the observations for two different jump locations and normal error. Use this to derive a lower bound for estimating the jump location.

In Step 4 technical difficulties arise, because the corresponding functions are not differentiable if the integral kernel of the operator K is not continuous. Step 5 is rather technical,

too. The reason for this is that we cannot apply standard techniques from empirical process theory because the used penalty is not a pseudo-norm on the space $T_{\infty,R}(\tau_{low}, \tau_{up})$.

7.1. Technical tools

This section contains technical results necessary for the proofs to come. Various properties of the spaces of step function are given.

Properties of the operator K restricted to the space of step functions

In order to gain some insight into the model, it is useful to have a closer look at the implications of Assumption B for the mapping $f \mapsto \mathrm{K}f$ restricted to the space of step functions. The following lemma collects some properties of this mapping.

Lemma 7.1. *Given Assumption B the following holds true.*

(i) If $\tau_{low} < 0$ and $\tau_{up} > 1$, for all $\epsilon > 0$ with $0 < \epsilon < \max(|\tau_{low}|, |\tau_{up} - 1|)$ there exists a constant $0 < C_0 < \infty$ such that for all $f \in T_\infty(\tau_{low}, \tau_{up})$

$$\|\mathrm{K}f\|_n^2 \leq C_0 \|f|_{[-\epsilon,1+\epsilon]}\|_2^2 \,.$$

Moreover, for all $f \in T_\infty(0,1)$ there exists a constant C_0 such that

$$\|\mathrm{K}f\|_n^2 \leq C_0 \|f|_{[0,1]}\|_2^2 \,.$$

(ii) $\mathrm{K} : T_k(\tau_{low}, \tau_{up}) \to L_2([0,1])$ is one-to-one.

(iii) For all $\epsilon > 0$ with $0 < \epsilon < \max(|\tau_{low}|, |\tau_{up} - 1|)$, the map

$$\mathrm{K} : (\, T_k(\tau_{low}, \tau_{up}), \|\cdot|_{[-\epsilon,1+\epsilon]}\|_2 \,) \to L_2([0,1])$$

is continuous. Moreover, the map $\mathrm{K} : (\, T_k(0,1), \|\cdot|_{[0,1]}\|_2 \,) \to L_2([0,1])$ is continuous.

(iv) The function $(\mathrm{K}f)$ is Lipschitz continuous on $[0,1]$ for all $f \in T_\infty(\tau_{low}, \tau_{up})$.

Proof. By Assumption B we have that $\sup_{x,y} K(x,y) = C < \infty$. This implies that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (\mathrm{K}f)^2(x_i) &= \frac{1}{n}\sum_{i=1}^n \Big(\int K(x_i,y)f(y)dy\Big)^2 \leq \int f(y)^2 \frac{1}{n}\sum_{i=1}^n K(x_i,y)^2 dy \\
&\leq C^2 \int_0^1 f(y)^2 dy + \int_{[\tau_{low},0]\cup[1,\tau_{up}]} f(y)^2 \frac{1}{n}\sum_{i=1}^n K(x_i,y)^2 dy \,,
\end{aligned}$$

for $f \in T_\infty(\tau_{low}, \tau_{up})$ For $\tau_{low} = 0$ and $\tau_{up} = 1$ the second term equals zero, thus this proves the second part of (i). If $\tau_{low} < 0$ and $\tau_{up} > 1$, note that f is constant on $[\tau_{low}, 0)$

and $[1, \tau_{up})$. For $C_K = \sup_{x\in[0,1]} \int_{\tau_{low}}^{\tau_{up}} |K(x,y)|dy$, this gives

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\mathrm{K}f)^2(x_i) &\leq C^2\int_0^1 f(y)^2dy + C_K^2(f|_{[\tau_{low},0)})^2 + C_K^2(f|_{[1,\tau_{up})})^2 \\
&= C^2\int_0^1 f(y)^2dy + \frac{C_K^2}{\epsilon}\Big(\int_{-\epsilon}^0 f(y)^2dy + \int_1^{1+\epsilon} f(y)^2dy\Big) \\
&\leq C_0\int_{-\epsilon}^{1+\epsilon} f(y)^2dy\,,
\end{aligned}$$

for some C_0 depending on K and ϵ only. This proves (i).

Similarly we can show $\|\mathrm{K}f\|_2 \leq C\|f|_{[-\epsilon,1+\epsilon]}\|_2$ for $f \in T_k(\tau_{low}, \tau_{up})$ with $\tau_{low} < 0$ and $\tau_{up} > 1$, and $\|\mathrm{K}f\|_2 \leq C\|f|_{[0,1]}\|_2$ for $f \in T_k(0,1)$ which gives continuity and hence (iii). As argued in section 4.1, (ii) follows from the independency of $\Delta_K(\cdot, \tau_i, \tau_{i+1})$.

To prove (iv), note that if K is of type (B2) we have

$$\begin{aligned}
|(\mathrm{K}1_{[a,b)})(x) - (\mathrm{K}1_{[a,b)})(x+\delta)| &= \Big|\int_{x-b}^{x-a}\Phi(y)dy - \int_{x+\delta-b}^{x+\delta-a}\Phi(y)dy\Big| \\
&\leq 2|\delta|\|\Phi|_{[-\tau_{up},1-\tau_{low}]}\|_\infty\,,
\end{aligned}$$

for $x \in [0,1]$, $x+\delta \in [0,1]$ and $a, b \in [\tau_{low}, \tau_{up}]$. For $f \in T_\infty(\tau_{low}, \tau_{up})$ with $\#\mathcal{J}(g) = k$, this gives

$$|(\mathrm{K}f)(x) - (\mathrm{K}f)(x+\delta)| \leq 2|\delta|k\|f\|_\infty\|\Phi|_{[\tau_{low}-1,\tau_{up}]}\|_\infty\,.$$

This proves Lipschitz continuity of $\mathrm{K}f$ on $[0,1]$. If K is of type (B1) this is clear by continuity of $K(x,y)$. □

Note that the reason, why we look at the interval $[-\epsilon, 1+\epsilon]$ instead of $[0,1]$ at (i) and (iii), is that the object $T_{k,R}(\tau_{low}, \tau_{up})$ is not a linear space. In particular, there is no equivalence of norms and although $\|\mathrm{K}f\|_2 \leq C_K^2\|f|_{[0,1]}\|_\infty$, there is no C such that $\|\mathrm{K}f\|_2 \leq C\|f|_{[0,1]}\|_2$. For $T_k(-\infty,\infty)$ this can be seen by setting $f_\epsilon(x) = 1_{(-\infty,\epsilon)}(x)$. Clearly,

$$\lim_{\epsilon\to 0}\|f_\epsilon|_{[0,1]}\|_2 = 0 \quad\text{and}\quad \lim_{\epsilon\to 0}\|\mathrm{K}f\|_2 = \int_0^1\Big(\int_{-\infty}^0 K(x,y)dy\Big)^2dx\,,$$

which is in general greater than zero.

Implications of the design assumption

As mentioned before, inference on $f - \hat{f}_n$ needs an assumption on the design points. This is necessary, because we need results on the convergence of $\mathrm{K}f - \mathrm{K}\hat{f}_n$ in the L_2-norm instead of the empirical norm to infer on the preimage. The following lemma provides a link of these two norms for design points satisfying Assumption C.

Lemma 7.2. *Suppose Assumption C is satisfied. If f is piecewise Lipschitz continuous on $[0,1]$, i.e. there exist a partition $I_1,\dots,I_k, k<\infty$, with $\bigcup_{i=1}^k I_k = 1$ and $I_j \cap I_r = \emptyset$ for $j \neq r$ such that $f|_{I_j}$ is Lipschitz for all $j=1,\dots,k$, we have that*

$$\int_0^1 f(x)h(x)dx = \frac{1}{n}\sum_{i=1}^n f(x_i) + O_P(n^{-1/2})\,.$$

Proof. Set $H(x) = \int_0^x h(x)dx$, where h is as in Assumption C. Note that H is strictly monotone and the inverse H^{-1} is well defined on $[0,1]$. For $0 \le a \le b \le 1$ we have that

$$b-a = H(H^{-1}(b)) - H(H^{-1}(a)) = \int_{H^{-1}(a)}^{H^{-1}(b)} h(x)dx \ge c_l(H^{-1}(b) - H^{-1}(a))\,.$$

Hence H^{-1} is Lipschitz and so is $(f\circ H^{-1})|_{H^{-1}(I_j)}$ for all $j=1,\dots,k$. By Assumption C we have $H^{-1}(i/n-\delta_i) = x_{(i)}$ with $\nu_n := \max_{i=1,\dots,n}|\delta_i| = O_P(n^{-1/2})$. Assume

$$H^{-1}\Big(\Big[\frac{i-1}{n}\,,\frac{i}{n}\Big]\Big) \subset I_j \qquad \text{and} \qquad H^{-1}\Big(\mathcal{I}(i/n\,,i/n-\delta_i)\Big) \subset I_r \tag{7.1}$$

for some $j,r \in \{1,\dots,k\}$. Here $\mathcal{I}(i/n\,,i/n-\delta_i)$ is the interval spanned by $i/n, i/n-\delta_i$ as defined in Section 2.1. Consequently,

$$\begin{aligned}
&n\int_{(i-1)/n}^{i/n} f(H^{-1}(x))dx \\
&= f(x_{(i)}) + n\int_{(i-1)/n}^{i/n} f(H^{-1}(x)) - f(H^{-1}(i/n))dx \\
&\qquad + n\int_{(i-1)/n}^{i/n} f(H^{-1}(i/n)) - f(H^{-1}(i/n-\delta_i))dx \\
&= f(x_{(i)}) + O(n^{-1}) + O(\nu_n)
\end{aligned}$$

holds by Lipschitz continuity of $f\circ H^{-1}$ on $[(i-1)/n, i/n]$ and $\mathcal{I}(i/n\,,i/n-\delta_i)$. For general i, we get

$$\begin{aligned}
n\int_{(i-1)/n}^{i/n} f(H^{-1}(x))dx &= f(x_{(i)}) + n\int_{(i-1)/n}^{i/n} f(H^{-1}(x)) - f(H^{-1}(i/n+\delta_i))dx \\
&\ge f(x_{(i)}) - 2\|f|_{[0,1]}\|_\infty\,.
\end{aligned}$$

Since f is piecewise Lipschitz continuous, f is bounded in supremum norm on $[0,1]$. Denote the points of discontinuity of f by $\mathcal{J}(f) = \{\vartheta_1,\dots,\vartheta_k\}$. The number of i, which does not satisfy (7.1) is bounded from above by

$$\begin{aligned}
&k + \#\{i : \vartheta_j \in H^{-1}([i/n-\nu_n, i/n+\nu_n]) \text{ for some } j=1,\dots,k\} \\
&\quad = k + \#\{i : H^{-1}(i/n-\nu_n) \le \vartheta_j \le H^{-1}(i/n+\nu_n) \text{ for some } j=1,\dots,k\} \\
&\quad = k + \#\{i : H(\vartheta_j) - \nu_n \le i/n \le H(\vartheta_j) + \nu_n \text{ for some } j=1,\dots,k\} \\
&\quad = k + O(n\nu_n)\,.
\end{aligned}$$

By application of the transformation formula and $k, \|f|_{[0,1]}\|_\infty < \infty$ we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f(x_i) &= \int_0^1 f(H^{-1}(x))dx + O(n^{-1}) + O((k+n\nu_n)n^{-1}\|f|_{[0,1]}\|_\infty) + O(\nu_n)\\ &= \int_0^1 f(x)h(x)dx + O(n^{-1}) + O_P(n^{-1/2})\,,\end{aligned}$$

which proves the claim. □

The facts that h is bounded from below and that $\mathrm{K}f$ is Lipschitz on $[0,1]$ (compare Lemma 7.1, (iv)), together with Lemma 7.2 directly yield the following corollary.

Corollary 7.3. *Suppose that the Assumptions B and C are met. If $f \in T_\infty(\tau_{low}, \tau_{up})$, then*

$$\|(\mathrm{K}f)|_{[0,1]}\|_2^2 = O(\|\mathrm{K}f\|_n^2) + o_P(1)\,.$$

Entropy results

In order to apply the uniform deviation inequalities from empirical process theory, it is necessary to calculate the entropy of the space of interest (compare Appendix A.1 for the uniform deviation inequalities and Section 2.1 for notation concerning entropy numbers and empirical processes). As shown below the entropy $H(\delta, \mathcal{G})$ of the space $\mathcal{G} = T_{k,R}(a,b)$ is polynomial in $\log(\delta^{-1})$. This is typical for small parametric classes. In nonparametric settings the entropy is usually polynomial in δ^{-1} (cf. Devroye and Lugosi, 2001, Chapter 7).

Note that the relevant quantity is the entropy of the space $\mathcal{G} = \{\mathrm{K}f : f \in T_{k,R}(\tau_{low}, \tau_{up})\}$ rather than the entropy of $T_{k,R}(a,b)$. Since the assumptions on K are rather general, it is convenient to first calculate the entropy of $T_{k,R}(a,b)$ and then use Lemma 7.1 to infer on the space $\mathcal{G}$.

Lemma 7.4. *For $-\infty < a < b < \infty$ there exists a constant $C > 0$ independent of δ,k and n, such that*

$$H(\delta, T_{k,R}(a,b)) \le C(k+1)\big(1 + \log\big(\frac{R(k+1)}{\delta}\big)\big)\,.$$

Proof. Define the sets

$$\Delta_K(\delta) = \Big\{-R + mc_2\delta : m = 0, \dots, \lceil 2R(c_2\delta)^{-1}\rceil\Big\}$$

and

$$\Gamma(\delta) = \Big\{a + mc_1\delta^2 : m = 1, \dots, \lfloor (b-a)(c_1\delta^2)^{-1}\rfloor\Big\}\,,$$

where c_1, c_2 will be defined later. Define the function class $\mathcal{H}(\delta)$ as

$$\begin{aligned}\mathcal{H}(\delta) &= \{g : g(x) = \sum_{i=1}^{k+1} b_i 1_{[\gamma_i - 1, \gamma_i)}(x) : b_i \in \Delta_K(\delta), i = 1, \dots, k+1,\\ &\qquad \gamma_0 = a, \gamma_{k+1} = b, \gamma_i \in \Gamma(\delta), \gamma_i < \gamma_{i+1}, i = 1, \dots, k\}\,.\end{aligned}$$

Now for $g_0 \in T_{k,R}(a,b)$ we can choose $g \in \mathcal{H}(\delta)$ such that $\mathrm{d}(\mathcal{J}(g), \mathcal{J}(g_0)) \leq c_1\delta^2/2$, and that for any $x \in [a,b]$ with $\mathrm{d}(x, \mathcal{J}(g)) > c_1\delta^2/2$ we have $(g_0(x) - g(x))^2 \leq c_2^2\delta^2/4$. Since g_0 has k jumps between a and b we get

$$\|g_0 - g\|_2^2 \leq (b-a)c_2^2\,\frac{\delta^2}{4} + k(2R)^2 c_1\,\frac{\delta^2}{2}\,.$$

Choosing $c_1 = (4kR^2)^{-1}$ and $c_2 = (b-a)^{-1/2}$ gives $\|g_0 - g\|_2 \leq \delta$. Hence $\mathcal{H}(\delta)$ is an δ-covering of $T_{k,R}(a,b)$. Since

$$\#\mathcal{H}(\delta) = \left\lceil \frac{2R\sqrt{b-a}}{\delta} \right\rceil^{k+1} \left\lceil \frac{(b-a)4kR^2}{\delta^2} \right\rceil^{k} = O\Big(\Big(\frac{R(k+1)}{\delta}\Big)^{3k+1}\Big)$$

the claim is proved. □

Lemma 7.4 directly gives that $T_{k,R}(a,b)$ is totally bounded for $-\infty < a < b < \infty$, i.e. for each $\epsilon > 0$ there exists a finite subset $\{g_1, \ldots, g_n\} \subset T_{k,R}(a,b)$ such that $T_{k,R}(a,b) \subset \bigcup_{i=1}^n \{f : \|g_i - f\| < \epsilon\}$. Note that $T_{k,R}(a,b)$ also contains functions with less than k jumps and is hence closed. Consequently, it is compact.

Corollary 7.5. *The space $(T_{k,R}(a,b), \|\cdot\|_2)$ is compact for all a, b satisfying $-\infty < a < b < \infty$.*

As previously announced, we will now use the assumptions on the operator K or, to be more precise, Lemma 7.1, to deduce bounds on the entropy of the space $\{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$.

Corollary 7.6. *Assume* K *satisfies Assumption B. For*

$$\mathcal{G}_{k,R}(\mathrm{K}) = \{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$$

there exists a constant C_2 independent of n,k and R such that

$$H(\delta, \mathcal{G}_{k,R}(\mathrm{K}), Q_n) \leq C_2(k+1)\big(1 + \log\big(\frac{R(k+1)}{\delta}\big)\big)\,.$$

Proof. By Lemma 7.1, (i) there exist $-\infty < a < b < \infty$ and $0 < C_0 < \infty$ such that

$$\|\,\mathrm{K}f - \mathrm{K}g\|_n \leq C_0 \|(f-g)|_{[a,b]}\|_2$$

for $f, g \in T_k(\tau_{low}, \tau_{up})$. Assume $\mathcal{H}(\delta)$ is a δ-covering of $T_{k,R}(a,b)$ for every $\delta > 0$. Then $\mathcal{H}(\delta/C_0)$ is a δ-covering of $\mathcal{G}_K(R)$. Consequently, the claim follows from Lemma 7.4. □

Again, this implies that the space $\mathcal{G}_{k,R}(\mathrm{K}) = \{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$ equipped with the empirical norm $\|\cdot\|_n$ is totally bounded. By definition $T_{k,R}(\tau_{low}, \tau_{up})$ contains step functions with k and less jumps. Clearly, a sequence of step functions g_n with $\#\mathcal{J}(g_n) \leq k$ for all n cannot converge to a limit g with $\#\mathcal{J}(g) > k$. Consequently, $\{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$ is closed and hence compact. This means that the functional $\|\cdot - Y\|_n$ has at least one minimizer in this space.

Corollary 7.7. *The functional $\|\cdot - Y\|_n$ has at least one minimizer in the space $\mathcal{G}_{k,R}(\mathrm{K}) = \{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$.*

7.2. Consistency

In this section we will prove consistency of $\hat{f}_n$ in L_2 norm and consistency of the set of jump estimates $\mathcal{J}(\hat{f}_n)$ in the Hausdorff metric.

To deduce consistency of the jump estimates from the L_2 consistency of the function estimator, a result on the dependency of $\mathrm{d}(\mathcal{J}(f), \mathcal{J}(g))$ on the L_2 distance of f and g is needed. This is given by the following lemma.

Lemma 7.8. *Assume $f, g \in T_\infty(\tau_{low}, \tau_{up})$. Then*

$$\mathrm{d}(\mathcal{J}(f), \mathcal{J}(g)) \leq \|(f-g)|_{[0,1]}\|_2^2 \frac{4}{h(f)^2}\,.$$

Proof. Remember that $h(f)$ denotes the minimal jump height of f. Let $\tau \in \mathcal{J}(f)$ and $\gamma \in \mathcal{J}(g)$, such that $|\tau - \gamma| = \mathrm{d}(\mathcal{J}(f), \mathcal{J}(g))$. Then

$$\|(f-g)|_{[0,1]}\|_2^2 \geq (\tau - \gamma)\Big(\frac{h(f)}{2}\Big)^2,$$

which proves the assertion. □

In order to show consistency of $\hat{f}_n$, we start by giving the consistency of $\mathrm{K}\hat{f}_n$. To this end, we use the entropy result of Corollary 7.6 and the inequality given by Lemma A.3.

Lemma 7.9. *Suppose the Assumptions A and B are met. Then*

$$\|\,\mathrm{K}f - \mathrm{K}\hat{f}_n\|_n = o_P(1)\,. \tag{7.2}$$

If additionally Assumption C is met, we have

$$\|(\mathrm{K}f - \mathrm{K}\hat{f}_n)|_{[0,1]}\|_2 = o_P(1)\,. \tag{7.3}$$

Proof. By (4.2)

$$\|\,\mathrm{K}\hat{f}_n - Y\|_n^2 \leq \|\,\mathrm{K}f - Y\|_n^2 + o(n^{-1}).$$

Note that $f - \hat{f}_n \in T_{2k,2R}(\tau_{low}, \tau_{up})$. Use $Y = \mathrm{K}f + \varepsilon$ to obtain

$$\begin{aligned}\|\,\mathrm{K}\hat{f}_n - \mathrm{K}f\|_n &\leq 2\langle \mathrm{K}(\hat{f}_n - f), \varepsilon_n\rangle_n + o(n^{-1}) \\ &\leq 2 \sup_{g \in \mathcal{G}_{2k,2R}(\mathrm{K})} \big|\langle g, \varepsilon_n\rangle_n\big| + o(n^{-1})\,,\end{aligned}$$

where $\mathcal{G}_{k,R}(\mathrm{K}) = \{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$. By Corollary 7.6

$$n^{-1} H(\delta, \mathcal{G}_{2k,2R}(\mathrm{K}), Q_n) \to 0 \quad \text{for all} \quad \delta > 0\,.$$

Hence we can apply Lemma A.3, which gives $\sup_{g \in \mathcal{G}_{2k,2R}(\mathrm{K})} |\langle g, \varepsilon_n\rangle_n| = o_P(1)$.

This proves (7.2) and application of Corollary 7.3 yields (7.3). □

The L_2 consistency of $\mathrm{K}\hat{f}_n$ and the compactness of the spaces of step functions (see Corollary 7.5) allows us to deduce L_2 consistency of $\hat{f}_n$ on each bounded interval I contained in $[\tau_{low}, \tau_{up}]$. For simplicity we give the result only for $I = [0, 1]$.

Lemma 7.10. *Suppose the Assumptions A, B and C are met. Then*

$$\|(f - \hat{f}_n)|_{[0,1]}\|_2 = o_P(1)\,. \tag{7.4}$$

Proof. Note that K is a linear operator, $f - \hat{f} \in T_{2k,2R}(\tau_{low}, \tau_{up})$ and that

$$\{f|_{[-\epsilon,1+\epsilon]} : f \in T_{2k,2R}(\tau_{low}, \tau_{up})\} = T_{2k,2R}(-\epsilon, 1+\epsilon)\,,$$

for any $\epsilon \geq 0$. By Corollary 7.5 the space $(T_{2k,2R}(-\epsilon, 1+\epsilon), \|\cdot\|_2)$ is compact. Lemma 7.1, (ii) and (iii) yield that there exists an $\epsilon \geq 0$ such that the map

$$\mathrm{K} : (\, T_{2k,2R}(-\epsilon, 1+\epsilon), \|\cdot\|_2 \,) \to L_2([0,1])$$

is continuous and one-to-one.

The inverse of a continuous injective mapping f restricted to the image $f(\Omega)$ is continuous if Ω is compact (see e.g. Hohage, 2002, Theorem 2.5). This gives continuity of K^{-1} on $f(\Omega) = \{\mathrm{K}f : f \in T_{2k,2R}(-\epsilon, 1+\epsilon)\}$. Hence, for all $f \in T_{2k,2R}(-\epsilon, 1+\epsilon)$, $\|\mathrm{K}f\|_{L_2([0,1])} \to 0$ implies $\|f\|_{L_2([-\epsilon,1+\epsilon])} = \|\mathrm{K}^{-1}\mathrm{K}f\|_{L_2([-\epsilon,1+\epsilon])} \to 0$. We arrive at

$$\|(f - \hat{f})|_{[0,1]}\|_2 \leq \|(f - \hat{f})|_{[-\epsilon,1+\epsilon]}\|_2 \to 0$$

for $\|(\mathrm{K}f - \mathrm{K}\hat{f})|_{[0,1]}\|_2 \to 0$. Lemma 7.10 gives $\|(\mathrm{K}f - \mathrm{K}\hat{f})|_{[0,1]}\|_2 = o_P(1)$. This proves the claim. □

This allows us to infer the consistency of the parameter estimates.

Corollary 7.11. *Suppose the prerequisites of Lemma 7.10 are met. In this case*

$$\mathrm{d}(\mathcal{J}(f), \mathcal{J}(\hat{f}_n)) = o_P(1)\,,$$

as well as $\#\mathcal{J}(f) = \#\mathcal{J}(\hat{f}_n)$. *Moreover, if* f *is given by (2.1) and* $\hat{f}_n$ *by (4.3), we have for the estimates* $\hat{b}_i$ *of the levels* b_i *that*

$$\max_{i=1,\ldots,k+1} |\hat{b}_i - b_i| = o_p(1)\,.$$

Proof. By Lemma 7.8,

$$\mathrm{d}(\mathcal{J}(f), \mathcal{J}(\hat{f}_n)) = O(\|(f - \hat{f}_n)|_{[0,1]}\|_2^2)\,.$$

Hence, Lemma 7.10 implies consistency of the jump estimates $\mathcal{J}(\hat{f}_n)$. This, together with $\|(f - \hat{f}_n)|_{[0,1]}\|_2 = o_P(1)$, directly gives consistency of the parameter estimates $\hat{b}_i$. □

7.3. Asymptotic normality

A general theorem

To show asymptotic normality for M-Estimators, it is common to assume existence of the derivative of the function which is minimized. However, in the case where the integral kernel is allowed to have discontinuities, a less restrictive result is needed.

As discussed in Chapter 5.3 of van der Vaart (1998) it is sufficient to assume existence of a second order Taylor-type expansion. Following this idea, the next theorem gives the asymptotic normality of the minimizer of a process $Z_n(\theta)$, provided it allows for a certain expansion. It is similar to Theorem 5.23 of van der Vaart (1998), but also covers the case of non i.i.d. random variables, which is required for the fixed design.

Theorem 7.12. *Assume $\Theta \subset \mathbb{R}^d$ is open and $\theta_0 \in \Theta$. Let $(Z_n(\theta))_{\theta\in\Theta}$ be a stochastic process. Assume there exists a sequence of random variables $(W_n)_{n\in\mathbb{N}} \subset \mathbb{R}^d$ and a positive definite matrix $V \in \mathbb{R}^{d\times d}$ such that*

$$Z_n(\theta_0 + \Delta) = Z_n(\theta_0) - 2n^{-1/2}W_n^t\Delta + \Delta^t V\Delta + R_n(\Delta) \tag{7.5}$$

with

$$\sup_{\|\Delta\|\le\delta} \frac{R_n(\Delta)}{\|\Delta\|^2 + n^{-1}} \xrightarrow{p} 0 \qquad \text{as} \quad n \to \infty\,, \delta \to 0\,, \tag{7.6}$$

as well as

$$W_n \xrightarrow{\mathcal{L}} N(0,\Gamma)\,.$$

If $\hat\theta_n$ is a consistent estimator of θ_0 and $\hat\theta_n$ is an approximate minimizer of Z_n, i.e.

$$\|\hat\theta_n - \theta_0\| = o_P(1) \qquad \text{and} \qquad Z_n(\hat\theta_n) \le \inf_{\theta\in\Theta}(Z_n(\theta)) + o_P(n^{-1})\,,$$

then

$$\sqrt{n}(\hat\theta_n - \theta_0) = V^{-1}W_n + o_P(1)\,.$$

Proof. First, we show the $\sqrt{n}$ consistency of $\hat\theta_n$. Set $\Delta_n = (\hat\theta_n - \theta_0)$. Since $\hat\theta_n$ is an approximate minimizer of Z_n,

$$\begin{aligned} Z_n(\theta_0) + o_P(n^{-1}) &\ge Z_n(\hat\theta_n) = Z_n(\theta_0 + (\hat\theta_n - \theta_0)) \\ &= Z_n(\theta_0) - 2n^{-1/2}W_n^t\Delta_n + \Delta_n^t V\Delta_n + R_n(\Delta_n)\,. \end{aligned}$$

Denote by λ_V the smallest eigenvalue of V. The expansion above implies

$$o_P(n^{-1}) \ge \frac{-\|\Delta_n\|}{\sqrt{n}}\frac{2W_n^t\Delta_n}{\|\Delta_n\|} + \lambda_V\|\Delta_n\|^2 + R_n(\Delta_n)\,.$$

Observe that the asymptotic normality of W_n implies $\|\Delta_n\|^{-1}W_n^t\Delta_n = O_P(1)$. Now divide by $\|\Delta_n\|^2 + n^{-1}$ and use condition (7.6) and the consistency of $\hat\theta_n$. This gives

$$o_P\big((n\|\Delta_n\|^2 + 1)^{-1}\big) \ge \frac{O_P(1)}{\sqrt{n}\|\Delta_n\| + (\sqrt{n}\|\Delta_n\|)^{-1}} + \frac{\lambda_V}{1 + (\sqrt{n}\|\Delta_n\|)^{-2}} + o_P(1)\,.$$

Now assume $\sqrt{n}\|\Delta_n\| \xrightarrow{p} \infty$. This leads to

$$o_P(1) \geq o_P(1) + \lambda_V \,,$$

which is a contradiction since $\lambda_V > 0$. This shows

$$\sqrt{n}\|\Delta_n\| = O_P(1)\,.$$

Now we derive the convergence of $\sqrt{n}\Delta_n$ to $V^{-1}W_n$. Observe that $V^{-1}W_n = O_P(1)$. By (7.6)

$$nR_n(n^{-1/2}V^{-1}W_n) = o_P(1) \qquad \text{as well as} \qquad nR_n(\Delta_n) = o_P(1)\,.$$

Together with (7.5) and the minimizing property of $\hat{\theta}_n$ this leads to

$$\begin{aligned}
&o_P(1)\\
&\quad\geq\; n\big(Z_n(\theta_0+\Delta_n) - Z_n(\theta_0 + n^{-1/2}V^{-1}W_n)\big)\\
&\quad=\; 2(V^{-1}W_n - \sqrt{n}\Delta_n)^tW_n + (\sqrt{n}\Delta_n)^tV(\sqrt{n}\Delta_n) - (V^{-1}W_n)^tV(V^{-1}W_n) + o_P(1)\\
&\quad=\; -2(\sqrt{n}\Delta_n)^tV(V^{-1}W_n) + (\sqrt{n}\Delta_n)^tV(\sqrt{n}\Delta_n) + (V^{-1}W_n)^tV(V^{-1}W_n) + o_P(1)\\
&\quad=\; (\sqrt{n}\Delta_n - V^{-1}W_n)^tV(\sqrt{n}\Delta_n - V^{-1}W_n) + o_P(1)\,.
\end{aligned}$$

Since V is positive definite, it follows that

$$\|\sqrt{n}\Delta_n - V^{-1}W_n\|^2 = o_P(1)\,,$$

which proves the claim. □

A second order expansion for the minimized process

To derive an expansion of type (7.5) for the problem in (4.2), let us first introduce some notation. For $b, \tilde{b} \in \mathbb{R}^{k+1}$ and $\tau, \tilde{\tau} \in \Gamma_k(\tau_{low}, \tau_{up})$ set

$$g(x, b, \tau) = \sum_{j=1}^{k+1} b_j \,\mathrm{K}\, 1_{[\tau_{j-1},\tau_j)}(x)\,.$$

and

$$Z_n(\tilde{b}, \tilde{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\Big(g(x_i, b, \tau) + \varepsilon_i - g(x_i, \tilde{b}, \tilde{\tau})\Big)^2\,. \tag{7.7}$$

Assume that f and the estimate $\hat{f}_n$ as defined by (4.2) are given by

$$f(x) = \sum_{i=1}^{k+1} b_i \,\mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x) \qquad \text{and} \qquad \hat{f}_n(x) = \sum_{i=1}^{k+1} \hat{b}_i \,\mathrm{K}\, 1_{[\hat{\tau}_{i-1},\hat{\tau}_i)}(x)\,,$$

respectively. By definition of $Z_n(\tilde{b}, \tilde{\tau})$ it is clear that

$$Z_n(\hat{b}, \hat{\tau}) = \min_{(\tilde{b},\tilde{\tau})\in[-R,R]^{k+1}\times\Gamma_k(\tau_{low},\tau_{up})} Z_n(\tilde{b}, \tilde{\tau}) + o(n^{-1})\,. \tag{7.8}$$

To obtain an expansion for $Z_n(\tilde{b}, \tilde{\tau})$, first examine the difference $g(x, b, \tau) - g(x, \tilde{b}, \tilde{\tau})$.

Lemma 7.13. *Suppose Assumption B is satisfied and $\nu(x)$ is given by (4.5), i.e.*

$$\nu_j(x) = \begin{cases} \mathrm{K}\, 1_{[\tau_{(j+1)/2-1}, \tau_{(j+1)/2})}(x) & j \ \ odd\,, \\ (b_j - b_{j+1}) K(x, \tau_{j/2}) & j \ \ even\,. \end{cases}$$

Define Δ by

$$\Delta = \left(\tilde{b}_1 - b_1, \tilde{\tau}_1 - \tau_1, \tilde{b}_2 - b_2, \tilde{\tau}_2 - \tau_2, \ldots, \tilde{\tau}_k - \tau_k, \tilde{b}_{k+1} - b_{k+1}\right)^t. \tag{7.9}$$

If K *is of type (B1) then*

$$\sum_{j=1}^{k+1} b_j \,\mathrm{K}\, 1_{[\tau_{j-1}, \tau_j]}(x) - \tilde{b}_j \,\mathrm{K}\, 1_{[\tilde{\tau}_{j-1}, \tilde{\tau}_j]}(x) = -\sum_{r=1}^{2k+1} \Delta_r \nu_r(x) + O(\|\Delta\|^2)\,,$$

and if K *is of type (B2), i.e.* $\mathrm{K}f = \Phi * f$*, then*

$$\begin{aligned} &\sum_{j=1}^{k+1} b_j \,\mathrm{K}\, 1_{[\tau_{j-1}, \tau_j]}(x) - \tilde{b}_j \,\mathrm{K}\, 1_{[\tilde{\tau}_{j-1}, \tilde{\tau}_j]}(x) \\ &\quad = \quad -\sum_{r=1}^{2k+1} \Delta_r \nu_r(x) + O(\|\Delta\|^2) + \sum_{i=1}^{k} O(\|\tau - \tilde{\tau}\|) 1_{\mathcal{I}(x-\tau_i, x-\tilde{\tau}_i) \cap \mathcal{J}(\Phi) \neq \emptyset}\,. \end{aligned}$$

Note that $\mathcal{I}(x - \tau_i, x - \tilde{\tau}_i) \cap \mathcal{J}(\Phi) \neq \emptyset$ means that Φ has a discontinuity in the interval with endpoints $x - \tau_i$ and $x - \tilde{\tau}_i$.

Proof of Lemma 7.13. We prove only the more difficult case (B2). By assumption $\mathrm{K}f = \Phi * f$ with $\#\mathcal{J}(\Phi) < \infty$ and $\|\Phi\|_\infty < \infty$.

First assume that $\tilde{\tau}_j \geq \tau_j$ and Φ is continuous on $[x - \tilde{\tau}_j, x - \tau_j]$, i.e. $\mathcal{J}(\Phi) \cap [x - \tilde{\tau}_j, x - \tau_j] = \emptyset$. Then for all $y \in [x - \tilde{\tau}_j, x - \tau_j]$ we have $\Phi(x - y) - \Phi(x - \tau_j) = O(|y - \tau_j|)$. This leads to

$$\begin{aligned} \mathrm{K}\, 1_{[\tau_{j-1}, \tau_j)}(x) - \mathrm{K}\, 1_{[\tau_{j-1}, \tilde{\tau}_j)}(x) &= -\int_{\tau_j}^{\tilde{\tau}_j} \Phi(x - y) dy \\ &= (\tau_j - \tilde{\tau}_j)\Phi(x - \tau_j) - \int_{\tau_j}^{\tilde{\tau}_j} \Phi(x - y) - \Phi(x - \tau_j) dy \\ &= (\tau_j - \tilde{\tau}_j)\Phi(x - \tau_j) - O(1) \int_{\tau_j}^{\tilde{\tau}_j} |y - \tau_j| dy \\ &= (\tau_j - \tilde{\tau}_j)\Phi(x - \tau_j) + O((\tau_j - \tilde{\tau}_j)^2)\,. \end{aligned}$$

If Φ has a discontinuity in $[x - \tilde{\tau}_j, x - \tau_j]$, then

$$\begin{aligned} \mathrm{K}\, 1_{[\tau_{j-1}, \tau_j)}(x) - \mathrm{K}\, 1_{[\tau_{j-1}, \tilde{\tau}_j)}(x) &= (\tau_j - \tilde{\tau}_j)\Phi(x - \tau_j) + \int_{\tau_j}^{\tilde{\tau}_j} O(\|\Phi\|_\infty) dy \\ &= (\tau_j - \tilde{\tau}_j)\Phi(x - \tau_j) + O(|\tau_j - \tilde{\tau}_j|)\,. \end{aligned}$$

The same holds for $\tilde{\tau}_j < \tau_j$. Note that $1_{\mathcal{I}(x-\tau_j,x-\tilde{\tau}_j)\cap\mathcal{J}(\Phi)\neq\emptyset}$ is one if and only if Φ has a discontinuity in $[x-\tilde{\tau}_j, x-\tau_j]$. Consequently,

$$\begin{aligned}\mathrm{K}\,1_{[\tau_{j-1},\tau_j)}(x) - \mathrm{K}\,1_{[\tau_{j-1},\tilde{\tau}_j)}(x) &= (\tau_j-\tilde{\tau}_j)\Phi(x-\tau_j) + O((\tau_j-\tilde{\tau}_j)^2) + \\ &\quad O(|\tau_j-\tilde{\tau}_j|)1_{\mathcal{I}(x-\tau_j,x-\tilde{\tau}_j)\cap\mathcal{J}(\Phi)\neq\emptyset}\,.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{K}\,1_{[\tau_{j-1},\tau_j)}(x) - \mathrm{K}\,1_{[\tilde{\tau}_{j-1},\tau_j)}(x) &= (\tilde{\tau}_{j-1}-\tau_{j-1})\Phi(x-\tau_{j-1}) + O((\tau_{j-1}-\tilde{\tau}_{j-1})^2) + \\ &\quad O(|\tau_{j-1}-\tilde{\tau}_{j-1}|)1_{\mathcal{I}(x-\tau_{j-1},x-\tilde{\tau}_{j-1})\cap\mathcal{J}(\Phi)\neq\emptyset}\,.\end{aligned}$$

Remember $\tau_0 = \tilde{\tau}_0$ and $\tau_{k+1} = \tilde{\tau}_{k+1}$, combine the preceding results and use the notation $\Phi(x-a) = K(x,a)$ to obtain

$$\begin{aligned}&\sum_{j=1}^{k+1}\Big(b_j\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) - \tilde{b}_j\,\mathrm{K}\,1_{[\tilde{\tau}_{j-1},\tilde{\tau}_j]}(x)\Big)\\ &\quad= \sum_{j=1}^{k+1}\Big((b_j-\tilde{b}_j)\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) + \tilde{b}_j\big(\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) - \mathrm{K}\,1_{[\tilde{\tau}_{j-1},\tilde{\tau}_j]}(x)\big)\Big)\\ &\quad= \sum_{j=1}^{k+1}\Big((b_j-\tilde{b}_j)\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) + \tilde{b}_j\big(\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) - \mathrm{K}\,1_{[\tau_{j-1},\tilde{\tau}_j]}(x)\big) + \\ &\qquad\qquad \tilde{b}_j\big(\mathrm{K}\,1_{[\tau_{j-1},\tilde{\tau}_j]}(x) - \mathrm{K}\,1_{[\tilde{\tau}_{j-1},\tilde{\tau}_j]}(x)\big)\Big)\\ &\quad= \sum_{j=1}^{k+1}\Big((b_j-\tilde{b}_j)\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) + \tilde{b}_j(\tau_j-\tilde{\tau}_j)K(x,\tau_j) + O((\tau_j-\tilde{\tau}_j)^2) + \\ &\qquad\qquad O(|\tau_j-\tilde{\tau}_j|)1_{\mathcal{I}(x-\tau_j,x-\tilde{\tau}_j)\cap\mathcal{J}(\Phi)\neq\emptyset} + \tilde{b}_j(\tilde{\tau}_{j-1}-\tau_{j-1})K(x,\tau_{j-1}) + \\ &\qquad\qquad O((\tau_{j-1}-\tilde{\tau}_{j-1})^2) + O(|\tau_{j-1}-\tilde{\tau}_{j-1}|)1_{\mathcal{I}(x-\tau_{j-1},x-\tilde{\tau}_{j-1})\cap\mathcal{J}(\Phi)\neq\emptyset}\Big)\,.\end{aligned}$$

By $\tilde{b}_j(\tau_j-\tilde{\tau}_j) = b_j(\tau_j-\tilde{\tau}_j) + O(\|b-\tilde{b}\|\,\|\tau-\tilde{\tau}\|)$, this gives

$$\begin{aligned}&\sum_{j=1}^{k+1}\Big(b_j\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) - \tilde{b}_j\,\mathrm{K}\,1_{[\tilde{\tau}_{j-1},\tilde{\tau}_j]}(x)\Big)\\ &\quad= \sum_{j=1}^{k+1}(b_j-\tilde{b}_j)\,\mathrm{K}\,1_{[\tau_{j-1},\tau_j]}(x) + \sum_{j=1}^{k}(b_j-b_{j+1})(\tau_j-\tilde{\tau}_j)K(x,\tau_j) + O(\|\tau-\tilde{\tau}\|^2) + \\ &\qquad\qquad O(\|b-\tilde{b}\|\,\|\tau-\tilde{\tau}\|) + \sum_{j=1}^{k}O(\|\tau-\tilde{\tau}\|)1_{\mathcal{I}(x-\tau_i,x-\tilde{\tau}_i)\cap\mathcal{J}(\Phi)\neq\emptyset}\,.\end{aligned}$$

Since $O(\|b-\tilde{b}\|\,\|\tau-\tilde{\tau}\|) = O(\|\Delta\|^2)$ as well as $O(\|\tau-\tilde{\tau}\|) = O(\|\Delta\|)$, this proves the claim. □

Lemma 7.14. *Suppose the Assumptions A, B and C are met. Then the process $Z_n(\tilde{b},\tilde{\tau})$ allows an expansion of type (7.5), namely*

$$Z_n(\tilde{b},\tilde{\tau}) = Z_n(b,\tau) + 2n^{-1/2}W_n^t\Delta + \Delta^t V\Delta + R_n(\Delta)\,,$$

where R_n satisfies condition (7.6), Δ is given by (7.9) and V is the positive definite $(2k+1)\times(2k+1)$ matrix defined by (4.6), i.e.

$$V_{ij} = \int_0^1 \nu_i(x)\nu_j(x)h(x)dx\,,$$

with $\nu(x)$ defined by (4.5). Moreover

$$W_n \xrightarrow{\mathcal{L}} N(0,\mathrm{E}(\varepsilon_1^2)V)\,.$$

Proof. We prove only the more difficult case (B2). By Lemma 7.13,

$$g(x,b,\tau) - g(x,\tilde{b},\tilde{\tau}) = -\sum_{j=1}^{2k+1}\Delta_j\nu_j(x) + O(\|\Delta\|^2) + \sum_{i=1}^{k} O(\|\Delta\|)1_{\mathcal{I}(x-\tau_i,x-\tilde{\tau}_i)\cap\mathcal{J}(\Phi)\neq\emptyset}\,.$$

Expand (7.7) to obtain

$$Z_n(\tilde{b},\tilde{\tau}) = \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\Big(g(x_i,b,\tau) - g(x_i,\tilde{b},\tilde{\tau})\Big) + \frac{1}{n}\sum_{i=1}^{n}\Big(g(x_i,b,\tau) - g(x_i,\tilde{b},\tilde{\tau})\Big)^2 + \|\varepsilon\|_n\,. \quad (7.10)$$

Note that the last term equals $Z_n(b,\tau)$. We will first estimate the second term of (7.10). Denote the points of discontinuity of Φ by $\mathcal{J}(\Phi) = \{\vartheta_1,\ldots,\vartheta_{\#\mathcal{J}(\Phi)}\}$ with $\vartheta_1 < \vartheta_2 < \ldots < \vartheta_{\#\mathcal{J}(\Phi)}$. This means

$$\mathcal{I}(x-\tau_i, x-\tilde{\tau}_i)\cap\mathcal{J}(\Phi)\neq\emptyset \quad\Leftrightarrow\quad \exists s : x\in\mathcal{I}(\vartheta_s-\tau_i,\vartheta_s-\tilde{\tau}_i)\,.$$

By Lemma 4.2,

$$\#\{i : x_i\in\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)\} = O_P(n|\tau_j-\tilde{\tau}_j| + n^{1/2})\,.$$

This gives

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{s=1}^{\#\mathcal{J}(\Phi)} 1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i) &= \frac{\#\mathcal{J}(\Phi)}{n}\sum_{j=1}^{k}O_P(n|\tau_j-\tilde{\tau}_j| + n^{1/2})\\ &= O_P(\|\Delta\| + n^{-1/2})\,.\end{aligned}$$

Note that the functions $\nu_j(x)$ are piecewise Lipschitz continuous by Assumption (B2) and part (iv) of Lemma 7.1. With the help of Lemma 7.2 this gives

$$\frac{1}{n}\sum_{i=1}^{n}\nu_j(x_i)\nu_r(x_j) = \int_0^1 \nu_j(x)\nu_r(x)h(x)dx + o_P(1)\,.$$

Combine the results above to obtain

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(g(x_i,b,\tau)-g(x_i,\tilde{b},\tilde{\tau}))^2\\
&= \sum_{i=1}^{n} n^{-1}\Big(\sum_{j=1}^{2k+1}\Delta_j\nu_j(x_i)+O(\|\Delta\|^2)+O(\|\Delta\|)\sum_{j=1}^{k}\sum_{s=1}^{\#\mathcal{J}(\Phi)}1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i)\Big)^2\\
&= \sum_{i=1}^{n} n^{-1}\Big(\sum_{j=1}^{2k+1}\Delta_j\nu_j(x_i)+\\
&\qquad O(\|\Delta\|^2)\Big)^2+\frac{O(\|\Delta\|^2)}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{s=1}^{\#\mathcal{J}(\Phi)}1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i)\\
&= \sum_{r,j=1}^{2k+1}\Delta_j\Delta_r\Big(\int_0^1\nu_j(x)\nu_r(x)h(x)dx+o_P(1)\Big)+O_P(\|\Delta\|^3+\|\Delta\|^2n^{-1/2})\\
&= \Delta^tV\Delta+O_P(\|\Delta\|^3)+o_P(\|\Delta\|^2)\,,
\end{aligned}
$$

where V is given by (4.6). The remainder terms clearly satisfy condition (7.6).

Next, examine the first term of (7.10). Define W_n by

$$
W_n=\begin{pmatrix} n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\nu_1(x_i)\\ \vdots\\ n^{-1/2}\sum_{i=1}^{n}\varepsilon_i\nu_{2k+1}(x_i)\end{pmatrix},
$$

to derive

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g(x_i,b,\tau)-g(x_i,\tilde{b},\tilde{\tau}))\\
&= \frac{-1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i}{\sqrt{n}}\Big(\sum_{j=1}^{2k+1}\Delta_j\nu_j(x_i)+O(\|\Delta\|^2)+O(\|\Delta\|)\sum_{j=1}^{k}\sum_{s=1}^{\#\mathcal{J}(\Phi)}1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i)\Big)\\
&= -n^{-1/2}\Delta^tW_n+O(\|\Delta\|^2)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i+\frac{O(\|\Delta\|)}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{s=1}^{\#\mathcal{J}(\Phi)}\varepsilon_i1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i)\,.
\end{aligned}
$$

The second term is clearly $o_P(\|\Delta\|^2)$. Obtaining an upper bound for the third term is more involved.

Suppose $\vartheta_s-\tau_j<\vartheta_s-\tilde{\tau}_j$. Set

$$
i_l(s,j)=\min\{i: x_{(i)}\geq\theta_s-\tau_j\}\qquad\text{and}\qquad i_u(s,j)=\max\{i: x_{(i)}<\theta_s-\tilde{\tau}_j\}\,.
$$

Consequently,

$$
\Big|\sum_{i=1}^{n}\varepsilon_i1_{\mathcal{I}(\vartheta_s-\tau_j,\vartheta_s-\tilde{\tau}_j)}(x_i)\Big|=\Big|\sum_{i=i_l(s,j)}^{i_u(s,j)}\varepsilon_i\Big|\,.
$$

By the law of the iterated logarithm for $\varepsilon_1, \varepsilon_2, \ldots$ i.i.d. with $\mathrm{E}(\varepsilon_1) = 0$ and $\mathrm{E}(\varepsilon_1^2) < \infty$ we have

$$\lim_{k_n \to \infty} \max_{j \in \{1,\ldots,k_n\}} (\mathrm{E}(\varepsilon_1^2) k_n \log\log k_n)^{-1/2} \Big| \sum_{i=1}^{j} \varepsilon_i \Big| = 1$$

almost surely. This implies for $\delta_n = i_u(s,j) - i_l(s,j)$ that

$$\max_{j=1,\ldots,\delta_n} \Big| \sum_{i=i_l(s,j)}^{i_u(s,j)} \varepsilon_i \Big| = O_P((\delta_n \log\log \delta_n)^{1/2}) \,.$$

By Lemma 4.2,

$$\delta_n = \#\{i : \vartheta_s - \tau_j \le x_{(i)} < \vartheta_s - \tilde{\tau}_j\} = O_P(n|\tau_j - \tilde{\tau}_j| + n^{1/2}) = O_P(n\|\Delta\| + n^{1/2}) \,.$$

Consequently,

$$\sum_{i=1}^{n} \Big| \varepsilon_i 1_{\mathcal{I}(\vartheta_s - \tau_j, \vartheta_s - \tilde{\tau}_j)}(x_i) \Big| = O_P\Big(\sqrt{(n\|\Delta\| + n^{1/2}) \log\log(n\|\Delta\| + n^{1/2})} \Big) \,.$$

The same can be shown for $\vartheta_j - \tau_j \ge \vartheta_j - \tilde{\tau}_j$. Since $\mathcal{J}(\Phi)$ is a finite set and $k < \infty$, it follows that

$$\begin{aligned} &\frac{O(\|\Delta\|)}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{s=1}^{\#\mathcal{J}(\Phi)} \varepsilon_i 1_{\mathcal{I}(\vartheta_s - \tau_j, \vartheta_s - \tilde{\tau}_j)}(x_i) = \\ &\qquad O(n^{-1}\|\Delta\|) O_P\Big(\sqrt{(n\|\Delta\| + n^{1/2}) \log\log(n\|\Delta\| + n^{1/2})} \Big) \,. \end{aligned} \tag{7.11}$$

To verify condition (7.6) for this term, note that for $\|\Delta\| < n^{-1/2}$,

$$(7.11) = O_P\big(n^{-5/4} \sqrt{\log\log(n^{1/2})}\,\big) = o_P(n^{-1}) \,,$$

and for $\|\Delta\| \ge n^{-1/2}$,

$$(7.11) = O_P\big(\|\Delta\|^{3/2} n^{-1/2} \sqrt{\log\log(n)}\,\big) = O_P\big(\,\|\Delta\|^2 n^{-1/4} \sqrt{\log\log(n)}\,\big) = o_P(\|\Delta\|^2) \,.$$

This gives

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (g(x_i, b, \tau) - g(x_i, \tilde{b}, \tilde{\tau})) = -n^{-1/2} \Delta^t W_n + o_P(\|\Delta\|^2) + o_P(n^{-1}) \,.$$

Next, take a closer look at W_n. For any $a \in \mathbb{R}^{2k+1}$,

$$a^t W_n = \sum_{i=1}^{n} \varepsilon_i \Big(n^{-1/2} \sum_{j=1}^{2k+1} a_j \nu_j(x_i) \Big)$$

and

$$\begin{aligned}\sum_{i=1}^{n}\Big(n^{-1/2}\sum_{j=1}^{2k+1}a_j\nu_j(x_i)\Big)^2 &= \sum_{r,j=1}^{2k+1}a_ja_r\Big(\frac{1}{n}\sum_{i=1}^{n}\nu_j(x_i)\nu_r(x_i)\Big)\\ &= \sum_{r,j=1}^{2k+1}a_ja_r\int\nu_j(x)\nu_r(x)h(x)dx+o_P(1)\,.\end{aligned}$$

By the central limit theorem and the Cramer-Wold device,

$$W_n\xrightarrow{\mathcal{L}}N(0,\sigma^2V)\,,$$

where $\sigma^2=\mathrm{E}(\varepsilon_1^2)$ and V is given by (4.6).

It remains to show that V is positive definite. For any $\beta\in\mathbb{R}^{2k+1}$ calculate that

$$\begin{aligned}\beta^tV\beta &= \sum_{i,j=1}^{2k+1}\beta_i\beta_j\int\nu_i(x)\nu_j(x)h(x)dx=\int\Big(\sum_{i,j=1}^{2k+1}\beta_i\beta_j\nu_i(x)\nu_j(x)\Big)h(x)dx\\ &= \int\Big(\sum_{i=1}^{2k+1}\beta_i\nu_i(x)\Big)^2h(x)dx\geq c_l\int_0^1\Big(\sum_{i=1}^{2k+1}\beta_i\nu_i(x)\Big)^2dx\,.\end{aligned}$$

Where c_l is the lower bound of the design density, given by Assumption C. Observe that

$$\begin{pmatrix}\nu_1(x)\\ \nu_2(x)\\ \nu_3(x)\\ \nu_4(x)\\ \vdots\\ \nu_{2k}(x)\\ \nu_{2k+1}(x)\end{pmatrix}=\begin{pmatrix}\Delta_K(x,\tau_0,\tau_1)\\ (b_1-b_2)\Delta_K(x,\tau_1,\tau_1)\\ \Delta_K(x,\tau_1,\tau_2)\\ (b_2-b_3)\Delta_K(x,\tau_2,\tau_2)\\ \vdots\\ (b_k-b_{k+1})\Delta_K(x,\tau_k,\tau_k)\\ \Delta_K(x,\tau_k,\tau_{k+1})\end{pmatrix},$$

where $b_i-b_{i+1}\neq 0$ for all $i=1,\dots,k$. Hence, by Assumption B, these functions are linearly independent as functions in $L_2([0,1])$. Consequently, for $\beta\neq 0$ we have that

$$\int_0^1\Big(\sum_{i=1}^{2k+1}\beta_i\nu_i(x)\Big)^2dx>0$$

and thus $\beta^tV\beta>0$. □

Asymptotic normality of the estimates

Finally, we are prepared to show the asymptotic normality of the parameter estimates.

Corollary 7.15. *Suppose the Assumptions A, B and C are met. Let f, $\hat{f}_n$ and V be given by (2.1), (4.3) and (4.6), respectively. Set $\theta=(b_1,\tau_1,b_2,\tau_2,\dots,b_k,\tau_k,b_{k+1})$ as the*

parameter vector given by f, and $\hat{\theta}_n$ as the corresponding vector of estimates given by $\hat{f}_n$. Then

$$\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2 V^{-1}),$$

where $\sigma^2 = E(\varepsilon_1)$ and V is given by (4.6).

Proof. Corollary 7.11 implies $\|\theta - \hat{\theta}_n\| = o_P(1)$. By the relation (7.8) and Lemma 7.14 the assumptions of Theorem 7.12 are satisfied. The assertion follows by application of this theorem. □

The result on the asymptotic distribution can be used to derive L_2 rates for $\mathrm{K}\hat{f}_n$ and $\hat{f}_n$.

Corollary 7.16. *Suppose the assumptions of Corollary 7.15 are met. Then*

$$\|(\mathrm{K}f - \mathrm{K}\hat{f}_n)|_{[0,1]}\|_2 = O_P(n^{-1/2})$$

and

$$\|(f - \hat{f}_n)|_{[0,1]}\|_2 = O_P(n^{-1/4}).$$

Proof. By Assumption B the integral kernel of the operator K is bounded in supremum norm. Hence

$$\int_0^1 |\mathrm{K}\, 1_{[a,b)}| dx = \int_0^1 \Big| \int_a^b K(x,y) dy \Big| dx = O(|b-a|).$$

Compute for $\tau_i < \hat{\tau}_i$ and $\tau_{i-1} < \hat{\tau}_{i-1}$ that

$$\begin{aligned}
&\int_0^1 \Big| b_i \mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x) - \hat{b}_i \mathrm{K}\, 1_{[\hat{\tau}_{i-1},\hat{\tau}_i)}(x) \Big| dx \\
\leq\ & |b_i - \hat{b}_i| \int_0^1 |\mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x)| dx + |b_i| \int_0^1 |\mathrm{K}\, 1_{[\tau_{i-1},\hat{\tau}_{i-1})}(x)| dx + \\
& |\hat{b}_i| \int_0^1 |\mathrm{K}\, 1_{[\tau_i,\hat{\tau}_i)}(x)| dx \\
=\ & O(|b_i - \hat{b}_i|) + O(|\tau_{i-1} - \hat{\tau}_{i-1}|) + O(|\tau_i - \hat{\tau}_i|) = O_P(n^{-1/2}).
\end{aligned}$$

Similar calculations yield the same result if $\hat{\tau}_i \leq \tau_i$ or $\hat{\tau}_{i-1} \leq \tau_{i-1}$ and

$$\big\| b_i \mathrm{K}\, 1_{[\tau_{i-1},\tau_i)}(x) - \hat{b}_i \mathrm{K}\, 1_{[\hat{\tau}_{i-1},\hat{\tau}_i)}(x) \big\|^2_{L_2([0,1])} = O_P(n^{-1}).$$

This gives

$$\begin{aligned}
\|\mathrm{K}f - \mathrm{K}\hat{f}_n\|^2_{L_2([0,1])} &= \int_0^1 \Big(\sum_{i=1}^{k+1} b_i \mathrm{K}\, 1_{[\tau_{i-1},\tau_i))}(x) - \hat{b}_i \mathrm{K}\, 1_{[\hat{\tau}_{i-1},\hat{\tau}_i))}(x) \Big)^2 dx \\
&= O_P(n^{-1}).
\end{aligned}$$

To show the second assertion, note that

$$\begin{aligned}
\|f - \hat{f}_n\|_2^2 &= \sum_{i=1}^{k+1} (b_i - \hat{b}_i)^2 \Big((\tau_i \wedge \hat{\tau}_i) - (\tau_{i-1} \vee \hat{\tau}_{i-1}) \Big) + \\
&\quad \sum_{i=1}^{k} \Big(1_{\tau_i \geq \hat{\tau}_i} (b_i - \hat{b}_{i+1})^2 + 1_{\tau_i < \hat{\tau}_i} (b_{i+1} - \hat{b}_i)^2 \Big) |\tau_i - \hat{\tau}_i| \\
&= O_P(n^{-1}) O_P(1) + O_P(1) O_P(n^{-1/2}) = O_P(n^{-1/2}).
\end{aligned}$$

This proves the claim. □

7.3.1. Multi-phase regression with an unknown intercept

The proof of the asymptotic normality in Corollary 7.15 can be easily extended to the setting of Theorem 6.1. As most of the arguments are the same, we only give an outline of the proof.

The first step is to show consistency of the estimates $\hat{f}_n, \hat{b}_0$ as given by (6.2). By Theorem 5.10 the kernel x_+^p satisfies Assumption B with $\tau_{low} = 0$ and $\tau_{up} = 1$. Consequently, we can apply Lemma 7.1 to obtain

$$\max_{g \in T_{k,R}(0,1)} \| \mathrm{K} g\|_n^2 \leq C_0 R^2$$

and

$$\|Y\|_n^2 \leq 3b_0^2 + 3\| \mathrm{K} f\|_n^2 + 3\|\varepsilon\|_n^2 \leq 3b_0^2 + 3C_0R^2 + 3\|\varepsilon\|_n^2 \,.$$

As $\|\varepsilon\|_n^2 \to \sigma^2$ almost surely, we have

$$\max_{g \in T_{k,R}(0,1)} \| \mathrm{K} g - Y\|_n^2 \leq 3b_0^2 + 4C_0R^2 + 4\sigma^2 =: C_U$$

almost surely. This means that

$$\| \mathrm{K} \hat{f}_n + \hat{b}_0 - Y\|_n \leq \min_{g \in T_{k,R}(0,1), b \in \mathbb{R}} \| \mathrm{K} g + b - Y\|_n^2 + o(n^{-1})$$

implies that

$$\| \mathrm{K} \hat{f}_n + \hat{b}_0 - Y\|_n \leq \min_{g \in T_{k,R}(0,1), b \in [-C_U, C_U]} \leq \| \mathrm{K} g + b - Y\|_n^2 + o(n^{-1})$$

holds almost surely. Thus, we can restrict our analysis to the space of functions

$$\mathcal{G}_{k,R,C_U}(K) := \{\mathrm{K} g + b : g \in T_{k,R}(0,1), b \in [-C_U, C_U]\} \,.$$

Note that the δ-entropy of the $\mathcal{G}_{k,R,C_U}(K)$ is bounded from above by the $\delta/2$-entropy of $\{\mathrm{K} g : g \in T_{k,R}(0,1)\}$ times $\log(4C_u/\delta)$. Consequently, consistency of $\hat{f}_n, \hat{b}_0$ can be shown in the same manner as in Section 7.2.

Now we derive the asymptotic normality of the estimates. Set

$$\left(\Delta_0, \Delta_1, \ldots, \Delta_{2k+1}\right) = \left(\tilde{b}_0 - b_0, \tilde{b}_1 - b_1, \tilde{\tau}_1 - \tau_1, \tilde{b}_2 - b_2, \tilde{\tau}_2 - \tau_2, \ldots, \tilde{\tau}_k - \tau_k, \tilde{b}_{k+1} - b_{k+1}\right) .$$

By $\tilde{b}_0 - b_0 = \Delta_0$, Lemma 7.13 directly gives

$$\begin{aligned} & b_0 - \tilde{b}_0 + \sum_{j=1}^{k+1} (b_j \,\mathrm{K}\, 1_{[\tau_{j-1}, \tau_j]}(x) - \tilde{b}_j \,\mathrm{K}\, 1_{[\tilde{\tau}_{j-1}, \tilde{\tau}_j]}(x) \\ = & \; - \sum_{r=0}^{2k+1} \Delta_r \nu_r(x) + O(\|\Delta\|^2) + \sum_{i=1}^{k} O(\|\tau - \tilde{\tau}\|) 1_{\mathcal{I}(x-\tau_i, x-\tilde{\tau}_i) \cap \mathcal{J}(\Phi) \neq \emptyset} \,. \end{aligned}$$

for $\nu_0(x) = 1$. Thus we can derive a local asymptotic expansion in the same manner as in Lemma 7.14. It remains to show that the matrix V is positive definite, where V is defined by its entries (4.6) for $i, j = 0, \dots, 2k+2$ with $\nu_j(x)$ defined by (4.5) for $j = 1, \dots, 2k+1$ and $\nu_0(x) = 1$. To do so, we show that the functions $\nu_0, \nu_1, \dots, \nu_{2k+1}$ are linearly independent in $L_2([0,1])$. This can be done by the same arguments as in Theorem 5.10. Assume

$$\Big\| \sum_{i=0}^{2k+1} \alpha_i \nu_i(\cdot) \Big\|_{L_2([0,1])} = 0 \,.$$

Clearly,

$$\Big(\sum_{i=0}^{2k+1} \alpha_i \nu_i(x) \Big) \Big|_{[0,\tau_1)} = \alpha_0 + \alpha_1 x \,,$$

which directly implies $\alpha_0 = 0$ and $\alpha_1 = 0$. By the same arguments as in the proof of Theorem 5.10 it can be shown that $\alpha_i = 0$ for $i = 2, \dots, 2k+1$. Then the same arguments as in Lemma 7.14 can be used to derive the positive definiteness of V. Theorem 6.1 then follows by the same arguments as in Corollary 7.15.

7.4. Estimator for an unknown number of jumps

In this section we analyze the case where the number of jumps is unknown. We will show that for large n it is possible to estimate the number of jumps correctly with probability one.

In order to reconstruct the number of jumps correctly, it is helpful to use a penalty function which is strictly monotone in the number of jumps. Any penalty term, which depends on the number of jumps only, is not a pseudo-norm on $T_{\infty,R}(\tau_{low}, \tau_{up})$, since $\#\mathcal{J}(\lambda f) = \#\mathcal{J}(f)$ for $\lambda \neq 0$. Hence, the standard results from empirical process theory do not apply. However, it is possible to use similar techniques in the proofs.

Recall the penalized least squares estimate as defined in (4.4). We have

$$\| \mathrm{K}\hat{f}_{\lambda_n} - Y \|_n^2 + \lambda_n(\#\mathcal{J}(\hat{f}_{\lambda_n}) + 1) = \min_{g \in T_{\infty,R}(\tau_{low},\tau_{up})} \| \mathrm{K}g - Y \|_n^2 + \lambda_n(\#\mathcal{J}(g) + 1) + o(n^{-1}) \,,$$

where λ_n is some smoothing parameter. For ease of notation define

$$J_\#(f) := \#\mathcal{J}(f) + 1 \,.$$

The fact that $\hat{f}_{\lambda_n}$ (approximately) minimizes the penalized L_2 functional, implies that for $f \in T_{\infty,R}(\tau_{low}, \tau_{up})$ and $Y_i = \mathrm{K}f(x_i) + \varepsilon_i$, $i = 1, \dots, n$, we get that

$$\| \mathrm{K}\hat{f}_{\lambda_n} - Y \|_n^2 + \lambda_n J_\#(\hat{f}_{\lambda_n}) \leq \| \mathrm{K}f - Y \|_n^2 + \lambda_n J_\#(f) + o(n^{-1}) \,.$$

This gives

$$\| \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f \|_n^2 + 2\langle \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f, -\varepsilon \rangle_n + \|\varepsilon\|_n + \lambda_n J_\#(\hat{f}_{\lambda_n}) \leq \|\varepsilon\|_n + \lambda_n J_\#(f) + o(n^{-1}) \,,$$

which yields the basic inequality

$$\| \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^2 + \lambda_n J_\#(\hat{f}_{\lambda_n}) \leq 2\langle \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f, \varepsilon\rangle_n + \lambda_n J_\#(f) + o(n^{-1})\,. \tag{7.12}$$

This means, a bound for the term $|\langle \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f, \varepsilon\rangle_n|$, would allow immediate conclusions on $\| \mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^2$ as well as $\lambda_n J_\#(\hat{f}_{\lambda_n})$. A bound of this type can be obtained from the following exponential inequality.

Lemma 7.17. *Suppose Assumptions A and B are met and the error additionally satisfies (A1).*

There exist constants $c_1, c_2 > 0$, such that for all $t \geq c_1 n^{-1/2}$ we have

$$\mathrm{P}\Big(\sup_{f\in T_{\infty,R}(\tau_{low},\tau_{up})} \frac{|\langle \varepsilon, \mathrm{K}f\rangle_n|}{\| \mathrm{K}f\|_n J_\#^{1/2}(f)\big(1+\log(J_\#(f)/\| \mathrm{K}f\|_n)_+\big)} \geq t\Big) \leq c_2 \exp\Big(-\frac{nt^2}{c_2^2}\Big)\,.$$

Proof. Set $\mathcal{G}_{k,R}(\mathrm{K}) = \{\mathrm{K}g : g \in T_{k,R}(\tau_{low}, \tau_{up})\}$. By Corollary 7.6 there exists a constant $C > 0$ independent of u,k,R and n such that

$$H\big(u, \mathcal{G}_{k-1,R}(\mathrm{K}), Q_n\big) \leq Ck\big(1+\log\big(\frac{Rk}{u}\big)\big)\,.$$

Compute

$$\begin{aligned}
\int_0^\delta H^{1/2}\big(u, \mathcal{G}_{k-1,R}(\mathrm{K}), Q_n\big)du &\leq \sqrt{C}\sqrt{k}\int_0^\delta \sqrt{\log\big(\frac{\exp(1)Rk}{u}\big)}du \\
&= eRk\sqrt{C}\sqrt{k}\int_0^{\frac{\delta}{eRk}} \sqrt{-\log(u)}du \\
&\leq eRk\sqrt{C}\sqrt{k}\int_0^{\frac{\delta}{eRk}} (-\log(u))du \\
&= eRk\sqrt{C}\sqrt{k}\Big(\frac{\delta}{eRk}\big(1-\log\big(\frac{\delta}{eRk}\big)\big)\Big) \\
&= \delta\sqrt{C}\sqrt{k}(2+\log(R)+\log(k\delta^{-1})) \\
&\leq C_1\delta\sqrt{k}\big(1+\log\big(\frac{k}{\delta}\vee 1\big)\big) = C_1\delta\sqrt{k}\big(1+\log\big(\frac{k}{\delta}\big)_+\big)\,,
\end{aligned}$$

where C_1 is some finite constant independent of k and δ. By Theorem A.2 there exists some constant C_2 depending on the subgaussian error condition (A1) only, such that

$$\sqrt{n}\rho \geq C_2\Big(\int_0^\delta H^{1/2}\big(u, \mathcal{G}_{k-1,R}(\mathrm{K}), Q_n\big)du \vee \delta\Big)$$

implies

$$\mathrm{P}\Big(\sup_{g\in\mathcal{G}_{k-1,R}(\mathrm{K}),\|g\|_n<\delta} |\langle g, \varepsilon\rangle_n| \geq \rho\Big) \leq C_2\exp\Big(-\frac{n\rho^2}{C_2^2\delta^2}\Big)\,.$$

Consequently, for all $t \geq C_2C_1n^{-1/2}$ we have that

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{g\in\mathcal{G}_{k-1,R}(\mathrm{K}),\|g\|_n<\delta} |\langle g, \varepsilon\rangle_n| \geq t\delta\sqrt{k}\big(1+\log\big(\frac{k}{\delta}\big)_+\big)\Big) \\
&\leq C_2\exp\Big(-\frac{nt^2k\big(1+\log\big(\frac{k}{\delta}\big)_+\big)^2}{C_2^2}\Big)\,.
\end{aligned}$$

We arrive at

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{g\in\mathcal{G}_{k-1,R}(\mathrm{K})}\frac{|\langle\varepsilon,g\rangle_n|}{\|g\|_n\sqrt{k}\big(1+\log(k/\|g\|_n)_+\big)}\geq t\Big)\\
&\leq \sum_{s=1}^{\infty}\mathrm{P}\Big(\sup_{g\in\mathcal{G}_{k-1,R}(\mathrm{K}),\|g\|_n\leq 2^{-s+1}R}|\langle\varepsilon,g\rangle_n|\geq t(2^{-s}R)\sqrt{k}\big(1+\big(\log(k2^s/R)\big)_+\big)\Big)\\
&\leq \sum_{s=1}^{\infty}C_2\exp\Big(\frac{-t^2nk(1+(\log(k/R)+s\log(2))_+)}{C_2^2}\Big)\\
&\leq \sum_{s=1}^{\infty}C_2\exp\Big(\frac{-t^2n(1+(s\log(2)-\log(R))_+)}{C_2^2}\Big).
\end{aligned}$$

Splitting this sum at $s=\lceil(1+\log(R))/\log(2)\rceil$ gives

$$\begin{aligned}
&\mathrm{P}\Big(\sup_{g\in\mathcal{G}_{k-1,R}(\mathrm{K})}\frac{|\langle\varepsilon,g\rangle_n|}{\|g\|_n\sqrt{k}\big(1+\log(k/\|g\|_n)_+\big)}\geq t\Big)\\
&\leq C_2\Big\lceil\frac{1+\log(R)}{\log(2)}\Big\rceil\exp\Big(\frac{-t^2n}{C_2^2}\Big)+\sum_{s=\lceil\log(e^1R)/\log(2)\rceil}^{\infty}C_2\exp\Big(\frac{-t^2nC_3(1+s\log(2))}{C_2^2}\Big)\\
&\leq C_5\exp\Big(\frac{-t^2n}{C_2^2}\Big)+\sum_{s=1}^{\infty}C_2\exp\Big(\frac{-t^2nC_4(1+s)}{C_2^2}\Big)\\
&\leq C_5\exp\Big(\frac{-t^2n}{C_2^2}\Big)+\exp\Big(\frac{-t^2nC_4}{C_2^2}\Big)\int_{s=0}^{\infty}C_2\exp\Big(\frac{-t^2nC_4s}{C_2^2}\Big)\\
&\leq C_5\exp\Big(\frac{-t^2n}{C_2^2}\Big)+\frac{C_2^3}{C_4t^2n}\exp\Big(\frac{-t^2nC_4}{C_2^2}\Big)\leq C_6\exp\Big(-\frac{t^2n}{C_4^2}\Big).
\end{aligned}$$

Here C_3,C_4,C_5,C_6 are constants depending on C_1,C_2 and R only. The last inequality holds by $t^2n\geq C_1^2C_2^2$.

Since the constant C_6 does not depend on k, the exponential inequality also holds if we additionally take the supremum over all k. This proves the claim. □

The above lemma yields upper bounds for the rate of $|\langle\mathrm{K}f,\varepsilon\rangle_n|$, which are stated in the subsequent corollary.

Corollary 7.18. *Suppose the prerequisites of Lemma 7.17 are met. Then*

$$\sup_{f\in T_{\infty,R}(\tau_{low},\tau_{up})}|\langle\mathrm{K}f,\varepsilon\rangle_n|=\|\mathrm{K}f\|_n\sqrt{J_\#(f)}\big(1+\log(J_\#(f)/\|\mathrm{K}f\|_n)_+\big)O_P(n^{-1/2}).$$

Moreover, for each $\epsilon>0$ we have

$$\sup_{f\in T_{\infty,R}(\tau_{low},\tau_{up})}|\langle\mathrm{K}f,\varepsilon\rangle_n|=\|\mathrm{K}f\|_n^{1-\epsilon}(J_\#(f))^{(1+2\epsilon)/2}\;O_P(n^{-1/2}).$$

Proof. The first equation follows directly from Lemma 7.17. To show the second equation, observe that $J_\#(f)\geq 1$ and that $\sqrt{x}(1+\log(x))\leq cx^{1/2+\epsilon}$ for $x\geq 1$, $\epsilon>0$ and $c\geq(\epsilon^{-1}\vee 1)$ Moreover, if c is large enough and $x\geq 0$ then $x(1+\log(x^{-1}))\leq cx^{1-\epsilon}$. Combine these observations to derive the second equation from the first. □

Now we are in the position to prove the main theorem of this section, namely that the probability that the penalized estimator $\hat{f}_{\lambda_n}$ correctly estimates the number of jumps tends to one if n tends to infinity (given a proper choice of the penalty term).

Theorem 7.19. *Suppose Assumptions A, B and C met and the error additionally satisfies (A1). If $f \in T_{\infty,R}(\tau_{low}, \tau_{up})$, $\hat{f}_{\lambda_n}$ is given by (4.4) and there exists an $\epsilon > 0$ such that*

$$\lambda_n \overset{n\to\infty}{\longrightarrow} 0 \quad \text{as well as} \quad \lambda_n n^{1/(1+\epsilon)} \overset{n\to\infty}{\longrightarrow} \infty\,, \tag{7.13}$$

then

$$\lim_{n\to\infty} P\big(\#\mathcal{J}(\hat{f}_{\lambda_n}) = \#\mathcal{J}(f)\big) = 1\,.$$

Proof. Application of Corollary 7.18 to (7.12) gives

$$\begin{aligned} &\|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^2 \leq \\ &\|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^{1-\epsilon} J_\#(\hat{f}_{\lambda_n} - f)^{1/2+\epsilon} O_P(n^{-1/2}) + \lambda_n(J_\#(f) - J_\#(\hat{f}_{\lambda_n})) + o(n^{-1})\,, \end{aligned} \tag{7.14}$$

where ϵ is given by (7.13).

First, assume $J_\#(\hat{f}_{\lambda_n}) \leq J_\#(f)$. Then $J_\#(\hat{f}_{\lambda_n} - f)$ is bounded. Then (7.14) implies that either

$$\|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^2 = O(\lambda_n) + o(n^{-1}) \qquad \text{or} \qquad \|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^{1+\epsilon} = O_p(n^{-1/2})\,.$$

Thus, $\|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n = o_P(1)$. By Corollary 7.3, this implies $\|(\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f)|_{[0,1]}\|_2 = o_P(1)$. With the help of Lemma 7.8, it follows $\mathrm{d}(\mathcal{J}(\hat{f}_{\lambda_n}), \mathcal{J}(f)) = o_P(1)$, which in turn implies $J_\#(\hat{f}_{\lambda_n}) \geq J_\#(f)$ eventually.

Now assume $J_\#(\hat{f}_{\lambda_n}) \geq J_\#(f)$. Then (7.14) yields

$$\|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^2 \leq \|\,\mathrm{K}\hat{f}_{\lambda_n} - \mathrm{K}f\|_n^{1-\epsilon} J_\#(\hat{f}_{\lambda_n} - f)^{1/2+\epsilon} O_P(n^{-1/2}) + o(n^{-1})\,.$$

Assume n_k is a subsequence such that $\|\,\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon} \geq c n_k^{-1/2}$ for some $c > 0$. Dividing the last equation by $\|\,\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon}$ gives

$$\begin{aligned} \|\,\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1+\epsilon} &\leq J_\#(\hat{f}_{\lambda_{n_k}} - f)^{1/2+\epsilon} O_P(n_k^{-1/2}) + o(n_k^{-1/2}) \\ &= J_\#(\hat{f}_{\lambda_{n_k}} - f)^{1/2+\epsilon} O_P(n_k^{-1/2})\,. \end{aligned}$$

This yields

$$\|\,\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon} \leq J_\#(\hat{f}_{\lambda_{n_k}} - f)^{(1+\epsilon-2\epsilon^2)/(2+2\epsilon)} O_P(n_k^{-(1-\epsilon)/(2+2\epsilon)})\,.$$

Moreover, by (7.14)

$$\lambda_{n_k}(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) \leq O_P(n_k^{-1/2}) \|\,\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon} J_\#(\hat{f}_{\lambda_{n_k}} - f)^{1/2+\epsilon} + o(n_k^{-1})\,.$$

Combine the last two equations to obtain

$$\lambda_{n_k}(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) \leq O_P(n_k^{-1/(1+\epsilon)}) J_\#(\hat{f}_{\lambda_{n_k}} - f)^{(1+\epsilon-\epsilon^2)/(1+\epsilon)}\,. \tag{7.15}$$

Now assume n_k is a subsequence such that $\|\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon} < cn_k^{-1/2}$ for some $c > 0$. Application of Corollary 7.18 to (7.12) and the observation that $J_\#(g) \geq 1$ for all g gives

$$\begin{aligned}
\lambda_{n_k}(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) &\leq O_P(n_k^{-1/2})\|\mathrm{K}\hat{f}_{\lambda_{n_k}} - \mathrm{K}f\|_{n_k}^{1-\epsilon} J_\#(\hat{f}_{\lambda_{n_k}} - f)^{1/2+\epsilon} + o(n_k^{-1}) \\
&\leq O_P(n_k^{-1}) J_\#(\hat{f}_{\lambda_{n_k}} - f)^{1/2+\epsilon} \\
&\leq O_P(n_k^{-1/(1+\epsilon)}) J_\#(\hat{f}_{\lambda_{n_k}} - f)^{(1+\epsilon-\epsilon^2)/(1+\epsilon)} .
\end{aligned}$$

As each sequence can be decomposed into a subsequence containing only elements smaller than $cn^{-1/2}$ and a subsequence containing only elements greater or equal to $cn^{-1/2}$ for some $c > 0$, we have shown that $J_\#(\hat{f}_{\lambda_n}) \geq J_\#(f)$ implies (7.15).

Now we show that $J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f) \to 0$ in probability. To this end, assume there exists some subsequence n_k such that

$$J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f) \geq c > 0 . \tag{7.16}$$

This implies $J_\#(f) \leq J_\#(f)c^{-1}(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f))$ and

$$\begin{aligned}
J_\#(\hat{f}_{\lambda_{n_k}} - f) &\leq 2J_\#(\hat{f}_{\lambda_{n_k}}) \\
&= 2(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) + 2J_\#(f) \\
&\leq (2 + 2J_\#(f)c^{-1})(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) \\
&= O(1)(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)) .
\end{aligned}$$

Hence

$$J_\#(\hat{f}_{\lambda_{n_k}} - f)^{(1+\epsilon-\epsilon^2)/(1+\epsilon)} = O(1)\big(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)\big)^{(1+\epsilon-\epsilon^2)/(1+\epsilon)} .$$

Together with (7.15), the assumption $\lambda_{n_k} n_k^{1/(1+\epsilon)} \to \infty$ and (7.16), this gives

$$0 < c^{\epsilon^2/(1+\epsilon)} \leq \big(J_\#(\hat{f}_{\lambda_{n_k}}) - J_\#(f)\big)^{\epsilon^2/(1+\epsilon)} = O_P(\lambda_{n_k}^{-1} n_k^{-1/(1+\epsilon)}) = o_P(1) ,$$

which is a contradiction and implies $J_\#(\hat{f}_n) - J_\#(f) \to 0$ in probability. Since $J_\#(f)$ and $J_\#(\hat{f}_n)$ are integers, this yields

$$\mathrm{P}\left(J_\#(\hat{f}_n) = J_\#(f)\right) \to 1 ,$$

for $n \to \infty$. This proves the claim. □

7.5. A lower bound for estimating the jump locations

In this section we show that the obtained rate $\mathrm{d}(\mathcal{J}(\hat{f}_n), \mathcal{J}(f)) = O_P(n^{-1/2})$ is optimal in a minimax sense. To do so, we construct functions $f_0, f_{1,n}, f_{2,n}$ with

$$\mathrm{d}(\mathcal{J}(f_0), \mathcal{J}(f_{i,n})) = cn^{-1/2}$$

for $i = 1, 2$ and some $c > 0$ to be chosen later. Given the observations

$$Y_i = g(x_i) + \varepsilon_i \quad i = 1, \dots, n$$

for $g \in \{\mathrm{K}f_0, \mathrm{K}f_{1,n}, \mathrm{K}f_{2,n}\}$ and $\varepsilon_1, \dots, \varepsilon_n$ independent and identically distributed according to $N(0, \sigma^2)$ with $\sigma^2 > 0$, we show that for any estimator, the probability to choose the true function is strictly smaller than one. Obviously it is sufficient to consider the case of a single jump with a fixed jump height.

Lemma 7.20. *Suppose Assumption B is met, $x_1, \dots, x_n \in [0, 1]$ are arbitrary fixed design points. Moreover, assume that $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed according to $N(0, \sigma^2)$ with $\sigma^2 > 0$. Set $g_\tau = \mathrm{K}\, 1_{[\tau,\infty)}$ for $\tau \in (\tau_{low}, \tau_{up})$. Given observations*

$$Y_i = g_\tau(x_i) + \varepsilon_i \quad i = 1, \dots, n$$

denote the corresponding probability measure by P_τ. There exists some $c, c_1 > 0$ such that

$$\inf_{\hat{\tau}} \sup_{\tau \in (\tau_{low}, \tau_{up})} P_\tau(|\tau - \hat{\tau}| \geq cn^{-1/2}) \geq c_1 > 0\,.$$

Proof. We wish to apply Theorem A.4.

Recall that for product measures $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$ we have

$$d_K(P, Q) = \sum_{i=1}^n d_K(P_i, Q_i)\,.$$

Note that $Y_i \sim N(g_\tau(x_i), \sigma^2)$ and denote the corresponding measures with P_τ^i. By independency of the ε_i the joint measure P_τ of $Y_1, \dots, Y_n$ is given by $P_\tau = \otimes_{i=1}^n P_\tau^i$. Application of Lemma A.5 (which gives the Kullback-Leibler distance for normal measures) yields

$$d_K(P_{\tau_1}^i, P_{\tau_2}^i) = (2\sigma^2)^{-1}(g_{\tau_1}(x_i) - g_{\tau_2}(x_i))^2$$

and

$$d_K(P_{\tau_1}, P_{\tau_2}) = (2\sigma^2)^{-1} \sum_{i=1}^n (g_{\tau_1}(x_i) - g_{\tau_2}(x_i))^2\,.$$

Note that by Assumption B the integral kernel $K(x, y)$ is bounded in supremum norm. Set

$$K_\infty = \sup_{x \in [0,1]\,,\, y \in (\tau_{low}, \tau_{up})} |K(x, y)|\,,$$

and calculate

$$(g_{\tau_1}(x_i) - g_{\tau_2}(x_i))^2 = \Big(\int_{\min(\tau_1,\tau_2)}^{\max(\tau_1,\tau_2)} K(x_i, y) dy \Big)^2 \leq (\tau_1 - \tau_2)^2 K_\infty^2\,.$$

Consequently,

$$d_K(P_{\tau_1}, P_{\tau_2}) \leq (2\sigma^2)^{-1} n (\tau_1 - \tau_2)^2 K_\infty^2\,.$$

Now choose some $0 < \alpha < 1/10$, set $c = (2\alpha\sigma^2/K_\infty^2)^{1/2}$ and choose

$$\tau_0 \in (\tau_{low} + cn^{-1/2}, \tau_{up} - cn^{-1/2}) .$$

Set

$$\tau_1 = \tau_0 + cn^{-1/2} \qquad \text{and} \qquad \tau_2 = \tau_0 - cn^{-1/2} .$$

This gives

$$\frac{1}{2}\sum_{j=1}^{2} d_K(P_{\tau_j}, P_{\tau_0}) \leq (4\sigma^2)^{-1} n \sum_{j=1}^{2} (\tau_0 - \tau_j)^2 K_\infty^2 = \alpha .$$

Consequently, the assumptions of Theorem A.4 are satisfied for $s = c/2n^{-1/2}$ and $d(\tau, \tau') = |\tau - \tau'|$. Application of this theorem gives

$$\inf_{\hat{\tau}} \sup_{\tau \in (\tau_{low}, \tau_{up})} P_\tau(|\tau - \hat{\tau}| \geq 2^{-1} cn^{-1/2}) \geq \frac{\sqrt{2}}{1+\sqrt{2}} \Big(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log 2}}\Big) > 0 .$$

This proves the claim. □

Note that in the proof we used the absolute integrability and the boundedness in supremum norm of the integral kernel $K(x, y)$ only.

Proof of Theorem 4.4. Lemma 7.20 directly implies that the jump estimator attains the minimax rate. By Lemma 7.8 the L_2-norm of $\hat{f}_n - f$ is bounded from below by

$$C\, \mathrm{d}(\mathcal{J}(\hat{f}_n), \mathcal{J}(f))^{1/2}$$

for some $C > 0$. Consequently, f cannot be estimated at a faster rate than $n^{-1/4}$.

If f is a step function with known jump locations and unknown level heights b_i, the inverse regression model (2.2) reduces to a standard linear regression model. It is well known that in this setting the levels b_i cannot be estimated at a rate faster than $O_P(n^{-1/2})$. Consequently, this also holds for the case of unknown jump locations. This proves Theorem 4.4. □

Chapter 8

Extensions

This chapter contains some auxiliary results. Section 8.1 gives an exponential inequality in empirical norm of the image space under weaker assumptions on the operator than Assumption B, but the strong subgaussian assumption on the error. Afterwards, in Section 8.2 we discuss one case, when the operator K is not bounded and may be arbitrary close to the identity. In this case faster rates than $n^{-1/2}$ are obtained.

8.1. An exponential inequality in the image space

In Section 7.2 we used the relation

$$\|\mathrm{K}f\|_n^2 \leq C_0 \|f|_{[-\epsilon,1+\epsilon]}\|_2^2 \quad \text{for some} \quad C_0, \epsilon > 0 \quad \text{and all} \quad f \in T_k(\tau_{low}, \tau_{up}),$$

to prove consistency of the estimator. Being a bit more general, we will assume in the following that

$$\|\mathrm{K}f\|_n \leq C_0 \|f|_{[-\epsilon,1+\epsilon]}\|_2^c \quad \text{for some} \quad C_0, c > 0, \epsilon \geq 0 \quad \text{and all} \quad f \in T_k(\tau_{low}, \tau_{up}). \tag{8.1}$$

Note that the constants may depend on k. Taking this as a starting point, it is possible to prove an exponential inequality for $\mathrm{K}\hat{f}_n$ in empirical norm, which might be interesting on its own. This is the case, if one is interested in weakening either the design assumption or the assumption on the operator. The inequality will also be required in Section 8.2.

In Section 7.2 the entropy of the space of step functions was only used to show consistency of the estimate $\mathrm{K}\hat{f}_n$. If we are interested in the rate of convergence of this estimator, the entropy of the whole space is of minor importance. Once we have established consistency, it is sufficient to study the entropy of δ-balls around the objective function. The entropy of such balls is also called *local entropy*.

Lemma 8.1. *For $f \in T_k(a,b)$ with $\#\mathcal{J}(f) = k$, $-\infty < a < b < \infty$ and*

$$\mathcal{G}_f(\delta) = \{g \in T_k(a,b) : \|g - f\|_2 \leq \delta\},$$

there exist constants $C, \delta_0 > 0$ independent of δ and n, such that for $\delta < \delta_0$

$$H(u, \mathcal{G}_f(\delta)) \leq C(1 + \log(\delta) - \log(u)).$$

First let us introduce some notation for the proof. For $g \in T_\infty(a,b)$ with $\#\mathcal{J}(g) = r$ define $\tau_0(g) := a\,, \tau_{r+1}(g) := b$ and $\tau_i(g)$ as the i-th jump of g for $i = 1, \dots, r$. Note that $g|_{[\tau_{i-1}(g),\tau_i(g))}$ is constant for all $i = 1, \dots, r+1$. Moreover, define

$$\mathrm{m}_J(g) = \min_{i=1,\dots,r+1} |\tau_{i-1}(g) - \tau_i(g)|\,, \tag{8.2}$$

as the minimal distance of two jump locations of g.

Proof. Fix $g_0 \in \mathcal{G}_f(\delta)$. Lemma 7.8 implies that

$$h(f)^2 \ \mathrm{d}(\mathcal{J}(g_0), \mathcal{J}(f)) \le 4\|(f - g_0)|_{[a,b)}\|_2^2 \le 4\delta^2\,,$$

where $h(f)$ denotes the minimal jump height of f. Choose $\delta_0 = 4^{-1}h(f)\sqrt{\mathrm{m}_J(f)}$. Then we have for $\delta \le \delta_0$ that

$$\mathrm{d}(\mathcal{J}(g_0), \mathcal{J}(f)) \le \frac{\mathrm{m}_J(f)}{4}\,.$$

Consequently, $\#\mathcal{J}(g_0) = \#\mathcal{J}(f)$ and

$$\mu\big([\tau_{i-1}(f), \tau_i(f)) \cap [\tau_{i-1}(g_0), \tau_i(g_0))\big) \ge \frac{\mathrm{m}_J(f)}{2} \quad \forall\, i = 1, \dots, k+1\,,$$

where μ denotes the Lebesgue-measure. This implies

$$\frac{\mathrm{m}_J(f)}{2} \max_{i=1,\dots,k+1} \Big(f|_{[\tau_{i-1}(f),\tau_i(f))} - g_0|_{[\tau_{i-1}(g_0),\tau_i(g_0))}\Big)^2 \le \|f - g_0\|_2^2 \le \delta^2\,. \tag{8.3}$$

Consequently,

$$\begin{aligned} h(g_0) &\le h(f) + \max_{i=1,\dots,k+1} \Big|f|_{[\tau_{i-1}(f),\tau_i(f))} - g_0|_{[\tau_{i-1}(g_0),\tau_i(g_0))}\Big| \\ &\le h(f) + \frac{\delta_0\sqrt{2}}{\sqrt{\mathrm{m}_J(f)}} = h(f) + \frac{h(f)\sqrt{2}}{4} \le 2h(f)\,. \end{aligned}$$

For $i \in \{1, \dots, k\}$ define the set

$$\Gamma_i(u,f) := \left\{\tau_i(f) + \frac{u^2 m}{c_1} : m = -\left\lceil \frac{4\delta^2 c_1}{u^2 h(f)^2}\right\rceil, -\left\lceil \frac{4\delta^2 c_1}{u^2 h(f)^2}\right\rceil + 1, \dots, \left\lceil \frac{4\delta^2 c_1}{u^2 h(f)^2}\right\rceil\right\}$$

and for $i \in \{1, \dots, k+1\}$ the set

$$\begin{aligned} \Delta_i(u,f) \ = \Bigg\{ \ & f|_{[\tau_{i-1}(f),\tau_i(f))} + \frac{um}{c_2} : \\ & m = -\left\lceil \frac{\delta c_2\sqrt{2}}{u\sqrt{\mathrm{m}_J(f)}}\right\rceil, -\left\lceil \frac{\delta c_2\sqrt{2}}{u\sqrt{\mathrm{m}_J(f)}}\right\rceil + 1, \dots, \left\lceil \frac{\delta c_2\sqrt{2}}{u\sqrt{\mathrm{m}_J(f)}}\right\rceil \ \Bigg\}\,. \end{aligned}$$

Now define the function class $\mathcal{H}_f(u)$ by

$$\begin{aligned} \mathcal{H}_f(u) \ = \{ \ & b_i 1_{[\gamma_i - 1, \gamma_i)}(x) : b_i \in \Delta_i(u,f), i = 1, \dots, k+1, \\ & \gamma_0 = a, \gamma_{k+1} = b, \gamma_i \in \Gamma_i(u,f), i = 1, \dots, k \ \}\,. \end{aligned}$$

For $g_0 \in \mathcal{G}_f(\delta)$ with $\delta < \delta_0$ equation (8.3) implies that

$$g_0|_{[\tau_{i-1}(g_0),\tau_i(g_0))} \in \big[\min\{x : x \in \Delta_i(u,f)\}, \max\{x : x \in \Delta_i(u,f)\}\big]$$

holds for all $i = 1, \dots, k+1$. This gives

$$\min_{b_i \in \Delta_i(u,f)} \big|b_i - g_0|_{[\tau_{i-1}(g_0),\tau_i(g_0))}\big| \le \frac{u}{c_2}\,.$$

Consequently, we can choose $g \in \mathcal{H}(u)$ with $\mathrm{d}(\mathcal{J}(g), \mathcal{J}(g_0)) \le u^2/(2c_1)$ and

$$\max_{i=1,\dots,k+1} \big|(g - g_0)|_{[\tau_{i-1}(g),\tau_i(g)) \cap [\tau_{i-1}(g_0),\tau_i(g_0))}\big| \le \frac{u}{c_2}\,.$$

Moreover, we have

$$\max_{i=1,\dots,k+1} \big|g_0|_{[\tau_i(g_0),\tau_{i+1}(g_0))} - g|_{[\tau_i(g),\tau_{i+1}(g))}\big| \le h(g_0) + \frac{u}{c_2} \le 2h(f) + \frac{u}{c_2}\,.$$

Since g_0 has k jumps between a and b we arrive at

$$\|g_0 - g\|_2^2 \le (b-a)\,\frac{u^2}{4c_2^2} + k\big(h(f) + \frac{u}{c_2}\big)^2\,\frac{u^2}{2c_1}\,.$$

Consequently, we can choose constants c_1 and c_2 depending on k, f and $b-a$ only, such that $\|g_0 - g_n\|_2 \le u$. Hence $\mathcal{H}(u)$ is an u-covering of $\mathcal{G}_f(\delta)$. Since

$$\#\mathcal{H}(u) \le \Big(2\Big\lceil\frac{4\delta^2 c_1}{u^2 h(f)^2}\Big\rceil + 1\Big)^{k+1}\Big(2\Big\lceil\frac{2\delta c_2}{u\sqrt{\mathrm{m}_J(f)}}\Big\rceil + 1\Big)^{k} = O\Big(\Big(\frac{\delta}{u}\Big)^{3k+1}\Big)$$

the claim is proved. □

Similarly to Corollary 7.6, this result can be used to bound the entropy of the space of interest.

Corollary 8.2. *Assume* K *satisfies (8.1). For*

$$\mathcal{G}_k(\mathrm{K}, \delta) := \{\mathrm{K}g : g \in T_k(\tau_{low}, \tau_{up}) \text{ and } \|\,\mathrm{K}f - \mathrm{K}g\|_n \le \delta\}$$

there exist constants $0 < C_1, \delta_0 < \infty$ *independent of* δ *and* n*, such that for all* $\delta < \delta_0$

$$H(u, \mathcal{G}_{n,K}(\delta), Q_n) \le C_1(1 + \log(\delta) - \log(u))\,.$$

Proof. By (8.1) there exists a finite $\epsilon \ge 0$, a finite $c > 0$ and a finite $C_0 > 0$ such that

$$\|\,\mathrm{K}f - \mathrm{K}g\|_n \le C_0 \|(f-g)|_{[-\epsilon,1+\epsilon]}\|_2^c$$

for $f, g \in T_k(\tau_{low}, \tau_{up})$. Assume $\mathcal{H}(u)$ is a u-covering of $\{g \in T_k(-\epsilon, 1+\epsilon) : \|g - f\|_2 \le \delta\}$ for every $u > 0$. Then $\mathcal{H}(u^{1/c}/C_0)$ is a u-covering of $\mathcal{G}_{n,K}(\delta)$. This means the claim follows directly from Lemma 8.1. □

Now we use this bound on the local entropy and Theorem A.1 to prove an exponential inequality for $\| \mathrm{K} f - \mathrm{K} \hat{f}_n \|_n$.

Theorem 8.3. *Suppose the Assumption A is met and the error satisfies (A1). Moreover, suppose the operator* K *satisfies (8.1). If* Y *is given by (2.2),* $f \in T_k(\tau_{low}, \tau_{up})$ *and* $\hat{f}_n$ *is defined by*

$$\hat{f}_n = \underset{g \in T_k(\tau_{low}, \tau_{up})}{\operatorname{argmin}} \| \mathrm{K} g - Y \|_n \,,$$

there exist $c, c_0 > 0$ *such that for* $\delta \geq c_0 n^{-1/2}$ *we have*

$$\mathrm{P}(\| \mathrm{K} f - \mathrm{K} \hat{f}_n \|_n \geq \delta) \leq c \exp(-n \delta^2 c^{-2}) \,.$$

Proof. By Corollary 8.2 we have for $u < \delta_1 < \delta_0$ that

$$H(u, \mathcal{G}_n(\delta_1), Q_n) \leq C(1 + \log(\delta_1) - \log(u))$$

for some $0 < C < \infty$. Compute

$$\begin{aligned} J(\delta_1, \mathcal{G}_n(\delta_1), Q_n) &= \int_0^{\delta_1} H^{1/2}(u, \mathcal{G}_n(\delta_1), Q_n) du \\ &\leq C \delta_1 \int_0^1 \sqrt{\log(u^{-1} \exp(1))} du =: \delta_1 c_1 \,. \end{aligned}$$

Choosing $\Psi(\delta) = c_1 \delta$, there exists some $c_0 > 0$, such that $\delta_n = c_0 n^{-1/2}$ satisfies $\delta_n < \delta_0$ as well as condition (A.1). The claim follows directly by application of Theorem A.1. □

8.2. Faster rates: Abel-type kernels

The intention of this section is to examine, what happens "between" the case where K is the identity and the case where K is an integral operator with bounded integral kernel. We focus on giving an idea which rates of convergence can be obtained, and do not strive to derive the results in the most general setting. We will use a simplified model and analyze a special integral kernel, namely an Abel-type kernel (cf. Hall et al., 2003). For $0 < \alpha < 1$ define $\mathrm{K}_\alpha f$ as convolution of f with $\Phi_\alpha(x) = (1 - \alpha) x_+^{-\alpha}$, i.e.

$$(\mathrm{K}_\alpha f)(x) = (1 - \alpha) \int_{-\infty}^x (x - y)^{-\alpha} f(y) dy \,.$$

The choice of α is restricted to $(0, 1)$ since this assures that Φ_α is integrable on bounded intervals, which in turn assures that $f \in L_1(\mathbb{R})$ implies $\mathrm{K}_\alpha f \in L_1(\mathbb{R})$. Note that Φ_α is square integrable on bounded intervals if and only if $\alpha < 1/2$.

Throughout this section assume that

$$Y_i = \mathrm{K}_\alpha f_\tau(x_i) + \varepsilon_i \qquad i = 1, \ldots, n \tag{8.4}$$

for $x_i = i/n, i = 1, \ldots, n$, $\varepsilon_1, \ldots, \varepsilon_n$ independent identically distributed according to $N(0, 1)$ and $f(x) = 1_{[\tau, 1)}(x)$ with $\tau \in (0, 1)$ unknown. This simple model is well suited

to give an impression what happens in the case of an unbounded kernel, but keeps the notation as simple as possible. An estimator for τ is given by

$$\hat{\tau} = \operatorname*{argmin}_{\gamma \in (0,1)} \|Y - \mathrm{K}_\alpha 1_{[\gamma,1)}\|_n^2 . \tag{8.5}$$

Accordingly, denote $f_{\hat{\tau}} = 1_{[\hat{\tau},1)}$.

Technical tools

We start by calculating an explicit expression and an upper bound for $(\mathrm{K}_\alpha 1_{[a,b)})(x)$.

Lemma 8.4. *If $a < b$, then*

$$(\mathrm{K}_\alpha 1_{[a,b)})(x) = (x-a)^{1-\alpha} 1_{[a,b)}(x) + \big((x-a)^{1-\alpha} - (x-b)^{1-\alpha}\big) 1_{[b,\infty)}(x) ,$$

and

$$(\mathrm{K}_\alpha 1_{[a,b)})(x) \le (b-a)^{1-\alpha} 1_{[a,b+(b-a))}(x) + (1-\alpha)(b-a)(x-b)^{-\alpha} 1_{[2b-a,\infty)}(x) .$$

Proof. To prove the first claim, calculate

$$\begin{aligned}
(\mathrm{K}_\alpha 1_{[a,b)})(x) &= (1-\alpha) \int_a^b (x-y)^{-\alpha} 1_{[0,\infty)}(x-y) dx \\
&= (1-\alpha) \int_{x-b}^{x-a} (y)^{-\alpha} 1_{[0,\infty)}(y) dx \\
&= (x-a)^{1-\alpha} 1_{[a,b)}(x) + \big((x-a)^{1-\alpha} - (x-b)^{1-\alpha}\big) 1_{[b,\infty)}(x)
\end{aligned}$$

To prove the second claim, first assume $x \in [a,b)$. Note that $(\mathrm{K}_\alpha 1_{[a,b)})(x)$ is increasing on this interval. Therefore,

$$(\mathrm{K}_\alpha 1_{[a,b)})(x) = (x-a)^{1-\alpha} \le (b-a)^{1-\alpha} .$$

The same inequality holds for $x \in [b, b+(b-a))$, since $(\mathrm{K}_\alpha 1_{[a,b)})(x)$ is decreasing on this interval.

Now assume $x \ge b + (b-a)$. By Taylor's formula for $0 < \alpha < 1$ and $a < b$ there exists some $\xi \in (x-b, x-a)$ such that

$$(x-a)^{1-\alpha} = (x-b)^{1-\alpha} + (1-\alpha)(b-a)(x-b)^{-\alpha} + (b-a)^2(-\alpha)(1-\alpha)\xi^{-1-\alpha} .$$

Since the last term is smaller than zero

$$(x-a)^{1-\alpha} - (x-b)^{1-\alpha} \le (1-\alpha)(b-a)(x-b)^{-\alpha} .$$

This gives

$$\mathrm{K}\, 1_{[a,b)}(x) \le \begin{cases} (b-a)^{1-\alpha} & x \in [a, b+(b-a)) , \\ (1-\alpha)(b-a)(x-b)^{-\alpha} & x \ge 2b-a , \end{cases}$$

which proves the second claim. □

The first step on the way to obtain the rate of the least squares estimator of τ is to give a rate for $\mathrm{K}_\alpha f_{\hat{\tau}}$. To this end, we use the results of Section 8.1. The corresponding theorem requires that the empirical norm of $\mathrm{K}f$ is bounded by the L_2 norm of f.

Lemma 8.5. *Suppose $x_i = i/n$ for $i = 1, \dots, n$, $0 \le a \le b \le 1$. For every $0 < \alpha < 1$ there exists $C_\alpha > 0$ such that*

$$\| \mathrm{K}_\alpha 1_{[a,b)} \|_n^2 = \frac{1}{n} \sum_{i=1}^n (\mathrm{K}_\alpha 1_{[a,b)}(x_i))^2 \le C_\alpha (b-a)^{(3-2\alpha)\wedge 2} .$$

Proof. Apply Lemma 8.4 to obtain

$$\begin{aligned}
&\| \mathrm{K}\, 1_{[a,b)} \|_n^2 \\
&\le \frac{1}{n} \sum_{i=1}^n \Big((b-a)^{2-2\alpha} 1_{[a,2b-a)}(x_i) + (1-\alpha)^2 (b-a)^2 (x_i - b)^{-2\alpha} 1_{[2b-a,\infty)}(x_i) \Big) \\
&\le \frac{n+2}{n} 2(b-a)(b-a)^{2-2\alpha} + \frac{(1-\alpha)^2 (b-a)^2}{n} \sum_{i:n(2b-a)+1 \le i \le n} \Big(\frac{i}{n} - b \Big)^{-2\alpha} .
\end{aligned}$$

Compute that

$$\begin{aligned}
\sum_{i:n(2b-a)+1 \le i \le n} n^{-1} (i/n - b)^{-2\alpha} &\le \sum_{i:n(2b-a) \le i+1 \le n} \int_{(i-1)/n}^{i/n} (x-b)^{-2\alpha} dx \\
&\le \int_{(2b-a)}^1 (x-b)^{-2\alpha} dx ,
\end{aligned}$$

where the first inequality holds since $(x-b)^{-2\alpha}$ is decreasing on $[2b-a, 1]$. This gives

$$\| \mathrm{K}\, 1_{[a,b)} \|_n^2 \le 6(b-a)^{3-2\alpha} + (1-\alpha)^2 (b-a)^2 \int_{(2b-a)}^1 (x-b)^{-2\alpha} dx .$$

For $\alpha \ne 1/2$ we have

$$\begin{aligned}
\int_{(2b-a)}^1 (x-b)^{-2\alpha} dx &= (1-2\alpha)((1-b)^{1-2\alpha} - (b-a)^{1-2\alpha}) \\
&\le \begin{cases} (1-2\alpha)(1-b)^{1-2\alpha} & \alpha < 1/2, \\ (2\alpha-1)(b-a)^{1-2\alpha} & \alpha > 1/2. \end{cases}
\end{aligned}$$

By $(b-a) < 1$ we get $(b-a)^{3-2\alpha} \le (b-a)^2$ for $\alpha < 1/2$. Thus, there exists $C_\alpha > 0$ such that for $0 < \alpha < 1/2$,

$$\| \mathrm{K}\, 1_{[a,b)} \|_n^2 \le 6(b-a)^2 + (1-\alpha)^2 (1-2\alpha)(1-b)^{1-2\alpha} (b-a)^2 \le C_\alpha (b-a)^2 .$$

Similarly there exists $C_\alpha > 0$ such that for $1/2 < \alpha < 1$,

$$\| \mathrm{K}\, 1_{[a,b)} \|_n^2 \le 6(b-a)^{3-2\alpha} + (1-\alpha)^2 (2\alpha-1)(1-b)^{1-2\alpha} (b-a)^2 \le C_\alpha (b-a)^{3-2\alpha} .$$

For $\alpha = 1/2$ and $0 \le x \le 1$ note that

$$\begin{aligned}(x-a)^{1/2} - (x-b)^{1/2} &= \big((x-a)^{1/4} - (x-b)^{1/4}\big)\big((x-a)^{1/4} + (x-b)^{1/4}\big) \\ &\le 2\big((x-a)^{1/4} - (x-b)^{1/4}\big)\,.\end{aligned}$$

By Lemma 8.4 this is smaller or equal to $3/2(b-a)(x-b)^{-1/4}$ for $2b-a \le x \le 1$. Consequently,

$$\| \mathrm{K}_{1/2}\, 1_{[a,b)} \|_n^2 \le \frac{1}{n} \sum_{i=1}^n \Big((b-a) 1_{[a,2b-a)}(x_i) + \frac{9}{4}(b-a)^2 (x_i - b)^{-1/2} 1_{[2b-a,1]}(x_i) \Big)\,.$$

By the same arguments as for $\alpha = 1/4$, there exists some $C_{1/2}$ such that the above expression is smaller or equal to $C_{1/2}(b-a)^2$. This proves the claim. □

If we want to infer the rate of $\hat{\tau}$ from the rate of $\mathrm{K} f_{\hat{\tau}}$, we need a lower bound in addition to the upper bound of Lemma 8.5.

Lemma 8.6. *For any $0 < \alpha < 1$ there exists some $c_\alpha > 0$ such that for $0 \le a \le b \le 1$ we have*

$$\| \mathrm{K}_\alpha\, 1_{[a,b)} \|_n^2 \ge c_\alpha (b-a)^{(3-2\alpha)\wedge 2} + O(n^{-1})\,.$$

Proof. First assume $\alpha \ge 1/2$. Compute

$$\begin{aligned}\| \mathrm{K}_\alpha\, 1_{[a,b)} \|_n^2 &\ge \sum_{i: a+1/n \le i/n \le b} n^{-1} (i/n - a)^{2-2\alpha} \\ &\ge \sum_{i: a+1/n \le i/n \le b} \int_{(i-1)/n}^{i/n} (x-a)^{2-2\alpha} dx \\ &= \int_a^b (x-a)^{2-2\alpha} dx - \int_a^{\lceil na \rceil/n} (x-a)^{2-2\alpha} dx - \int_{\lfloor nb \rfloor/n}^b (x-a)^{2-2\alpha} dx \\ &= (3-2\alpha)^{-1} (b-a)^{3-2\alpha} + O(n^{-1})\,,\end{aligned}$$

where the second inequality holds since $(x-a)^{2-2\alpha}$ is increasing. A similar argument gives for $0 < \alpha < 1/2$ that

$$\| \mathrm{K}_\alpha\, 1_{[a,b)} \|_n^2 \ge \int_b^1 ((x-a)^{1-\alpha} - (x-b)^{1-\alpha})^2 dx + O(n^{-1})\,.$$

By Taylor's formula there exists some $\xi \in (x-b, x-a)$ such that

$$(x-b)^{1-\alpha} = (x-a)^{1-\alpha} + (a-b)(1-\alpha)(x-a)^{-\alpha} + (-\alpha)(1-\alpha)\xi^{-1-\alpha}\,.$$

Since the last term is smaller than zero we get

$$(b-a)(1-\alpha)(x-a)^{-\alpha} \le (x-a)^{1-\alpha} - (x-b)^{1-\alpha}\,.$$

This gives

$$\begin{aligned}\int_b^1 ((x-a)^{1-\alpha} - (x-b)^{1-\alpha})^2 dx &\ge (b-a)^2 (1-\alpha)^2 \int_b^1 (x-b)^{-\alpha} dx \\ &= (b-a)^2 (1-\alpha)(1-b)^{1-\alpha}\,,\end{aligned}$$

which proves the claim. □

Consistency and minimax rates

As mentioned before, Lemma 8.5 can be used to derive rates for the least squares estimate of $\mathrm{K}_\alpha f_\tau$.

Lemma 8.7. *Given the model defined at the beginning of Section 8.2, we have*

$$\| \mathrm{K}_\alpha f - \mathrm{K}_\alpha f_{\hat\tau}\|_n = O_P(n^{-1/2}) \,.$$

Proof. Lemma 8.1 gives the local entropy of $T_k(0,1)$. As we do not allow for general jump heights, the relevant space is the subspace of $T_k(0,1)$ that contains only functions f with $f(x) \in \{-1,0,1\}$ for all $x \in [0,1]$. The entropy of this subspace is clearly smaller than the entropy of the whole space.

Suppose f is in this subspace, and is given by $f(x) = \sum_{j=1}^{k+1} b_j 1_{[\tau_{j-1},\tau_j)}(x)$ with $b_j \in \{-1,0,1\}$. Lemma 8.5 and $\tau_j - \tau_{j-1} < 1$ for $j = 1,\dots,k+1$ imply

$$\begin{aligned}
\| \mathrm{K} f\|_n^2 &\le \sum_{j=1}^{k+1} (k+1) b_j^2 \| \mathrm{K}\, 1_{[\tau_j - \tau_{j-1})}\|_n^2 \\
&\le (k+1) \sum_{j=1}^{k+1} b_j^2 (\tau_j - \tau_{j-1})^{(3-2\alpha)\wedge 2} \\
&\le (k+1) \sum_{j=1}^{k+1} b_j^2 (\tau_j - \tau_{j-1}) = (k+1) \| f|_{[0,1]}\|_2 \,.
\end{aligned}$$

Thus by the same arguments as in Section 8.1 the exponential inequality of Theorem 8.3 holds for $\| \mathrm{K}_\alpha f - \mathrm{K}_\alpha f_{\hat\tau}\|_n$. This implies the claimed stochastic $O_P(n^{-1/2})$ rate. □

Usage of Lemma 8.6 directly gives that $|\hat\tau - \tau| = O_P(n^{-1/(2\wedge(3-2\alpha))})$. Similar arguments as in Section 7.5 show that this rate is optimal in a minimax sense.

Lemma 8.8. *Suppose $0 < \alpha < 1$. Given observations*

$$Y_i = \mathrm{K}_\alpha 1_{[\tau,1)}(x_i) + \varepsilon_i \quad i = 1,\dots,n$$

for $x_i = i/n, i = 1,\dots,n$, $\varepsilon_1,\dots,\varepsilon_n$ independent identically distributed according to $N(0,1)$, denote the corresponding probability measure by P_τ.

There exists some $c > 0$ such that

$$\inf_{\hat\tau} \sup_{\tau\in(0,1)} P_\tau(|\tau - \hat\tau| \ge c n^{-1/(2\wedge(3-2\alpha))}) > 0 \,.$$

Proof. The proof is very similar to the proof of Lemma 7.20. Set $g_{\tau_i} = \mathrm{K}_\alpha 1_{[\tau_i,1)}$ for $i = 1,2$. By Lemma 8.5,

$$\sum_{i=1} (g_{\tau_1}(x_i) - g_{\tau_2}(x_i))^2 \le n C_\alpha (\tau_1 - \tau_2)^{(3-2\alpha)\wedge 2} \,.$$

This gives

$$d_K(P_{\tau_1}, P_{\tau_2}) \le (2\sigma^2)^{-1} n C_\alpha (\tau_1 - \tau_2)^{(3-2\alpha)\wedge 2} \,.$$

The rest of the proof is done exactly the same way as the proof of Lemma 7.20. □

The results of this section are summarized in the following theorem.

Theorem 8.9. *Suppose Y is given by (8.4) and the assumptions of Lemma 8.8 on the design points and the error are satisfied. Then for*

$$\hat{\tau} = \operatorname*{argmin}_{\gamma \in (0,1)} \|Y - \mathrm{K}_\alpha 1_{[\gamma,1)}\|_n^2$$

we have that

$$|\hat{\tau} - \tau| = O_P(n^{-1/(2\wedge(3-2\alpha))}),$$

and this rate is minimax.

Note that the "elbow" in the rates of convergence occurs at $\alpha = 1/2$, and that the $n^{-1/2}$ rate holds for the case where Φ_α is square integrable on bounded intervals.

Neumann (1997) and Goldenshluger et al. (2006) also observe an elbow in the rate of convergence of recovering a change point in an inverse problem. To compare their results to the rate given by Theorem 8.9, we first calculate the Fourier transform of Φ_α.

Denote by $\Gamma(a) = \int_0^\infty x^{a-1}\exp(-x)dx$ the Γ-function. For $0 < \alpha < 1$ one can show

$$\int_0^\infty \sin(x)x^{-\alpha}dx = \cos(\alpha\pi/2)\int_0^\infty x^{-\alpha}\exp(-x)dx = \cos(\alpha\pi/2)\Gamma(1-\alpha)$$

and

$$\int_0^\infty \cos(x)x^{-\alpha}dx = \sin(\alpha\pi/2)\int_0^\infty x^{-\alpha}\exp(-x)dx = \sin(\alpha\pi/2)\Gamma(1-\alpha).$$

Basic calculations yield

$$\widehat{\Phi}_\alpha(x) = |x|^{-1+\alpha}\Gamma(1-\alpha)\big(\sin(\alpha\pi/2) + i\,\mathrm{sign}(x)\cos(\alpha\pi/2)\big).$$

Consequently,

$$\left|\widehat{\Phi}_\alpha(x)\right| = |x|^{-1+\alpha}\Gamma(1-\alpha).$$

Goldenshluger et al. (2006) assume that the regression function is given by $\Phi * f$ where f is a sum of a step function and a function with bounded m-th derivative, and the Fourier transform of Φ satisfies

$$C_1|1+x|^{-\beta} \geq |\widehat{\Phi}(x)| \geq C_2|1+x|^{-\beta} \tag{8.6}$$

for some $0 < C_2 < C_1 \leq \infty$. In this setting they show (in a white noise model) that the minimax rate for estimating the jumps of f is

$$\begin{aligned} n^{-1/(2\beta+1)} &: \quad \beta < 1/2, \\ n^{-(m+1)/(2\beta+2m+1)} &: \quad \beta \geq 1/2. \end{aligned}$$

Neumann (1997) obtains a similar result in a deconvolution setting for $m = 1$.

Though $|\widehat{\Phi}_\alpha(x)|$ is not bounded for $x \to 0$ and thus formally does not satisfy condition (8.6), the behavior for $|x| \to \infty$ corresponds to $\beta = 1 - \alpha$ in this condition. A comparison of the rates gives that the rates of Neumann (1997) and Goldenshluger et al. (2006) coincide with the rates of Theorem 8.9 for $\beta \leq 1/2$ and $\alpha \geq 1/2$, respectively.

Some simulated data

Estimation in a setting with an Abel-type kernel is a good example, that the additional information that f is a step function significantly improves the reconstruction of f. For illustrational purposes, consider the following simulated data example.

Test Bed 8.10. *Assume the observations Y are generated by*

$$Y_i = \mathrm{K}_\alpha(1_{[0,0.3)} + 2 \cdot 1_{[0.3,1)})(i/n) + 0.5\,\varepsilon_i\,, \quad i = 1, \ldots, n\,,$$

where $n = 200$, $\alpha = 0.6$ and $\varepsilon_i \sim N(0,1)$ for $i = 1, \ldots, n$.

The generated data is displayed in Figure 8.1.

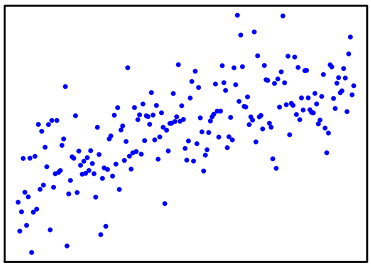

Figure 8.1.: Observations according to the Test Bed 8.10.

There are several ways to reconstruct the signal f from the observations. The approach, which is perhaps most natural from a statisticians point of view, is to compute an estimate for $\mathrm{K}f$ and then invert this estimate. Mathematical properties of this approach were examined by Bissantz et al. (2004). In the following we use Tikhonov regularization with a small regularization parameter to compute the estimates for f. For regularization methods in inverse problems compare Engl et al. (1996).

A first glance at the observations in Figure 8.1 reveals that the data look quite linear. Thus it might be tempting to fit a linear regression to the data and then estimate f by applying the inverse of K to this regression estimate. As Figure 8.2 shows, the corresponding estimate is quite far from the true signal f.

As the assumption of a linear regression model is rather restrictive, a more realistic approach is to use a local linear regression estimate. This, however, includes a choice of bandwidth. For practical purposes, it is helpful to use several bandwidths and then to visually examine the corresponding estimates, to identify different features of the data (cf. Chaudhuri and Marron, 2000). This is done in Figure 8.3 for three different bandwidths. Visual inspection of the estimates of the signal f might lead to the guess that the true signal has a jump, but it is rather unclear, whether one or more discontinuities are present.

Note that given the information that the true function is a step function, it is possible to fit a step function to the estimate of f given by first using a local polynomial fit and than applying the (regularized) inverse of K. The corresponding estimates are represented by the dashed line in Figure 8.3. It can be seen that this procedure leads to rather reasonable estimates of f.

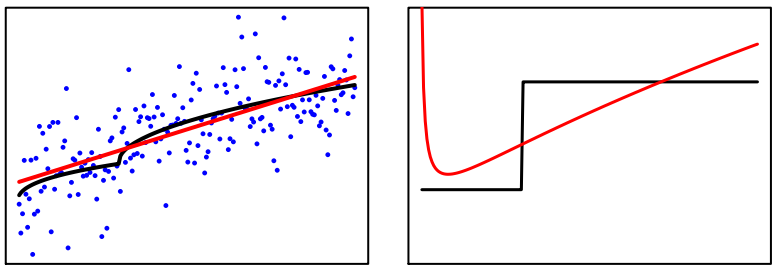

Figure 8.2.: Linear regression estimate of $\mathrm{K}f$ and the corresponding estimate for f. The red line shows the estimate and the black line the signal. The blue dots represent the observations.

Figure 8.4 shows the least squares estimate $\mathrm{K}\hat{f}$ of $\mathrm{K}f$, where $\hat{f}$ is restricted to $T_1(0,1)$. The estimate is close to what one would suspect, when examining Figure 8.3.

The main message of this small study is that in inverse problems additional information, such as knowledge that the signal is a step function, can be used to construct significantly improved estimates of the signal. This is of course also true for direct problems, but as the pictures above show, the error for reconstructing $\mathrm{K}f$ without prior knowledge is small, when compared to the error for reconstructing f without prior knowledge. Here, the method of reconstructing the signal is of minor importance, if the given prior information is used in a clever way.

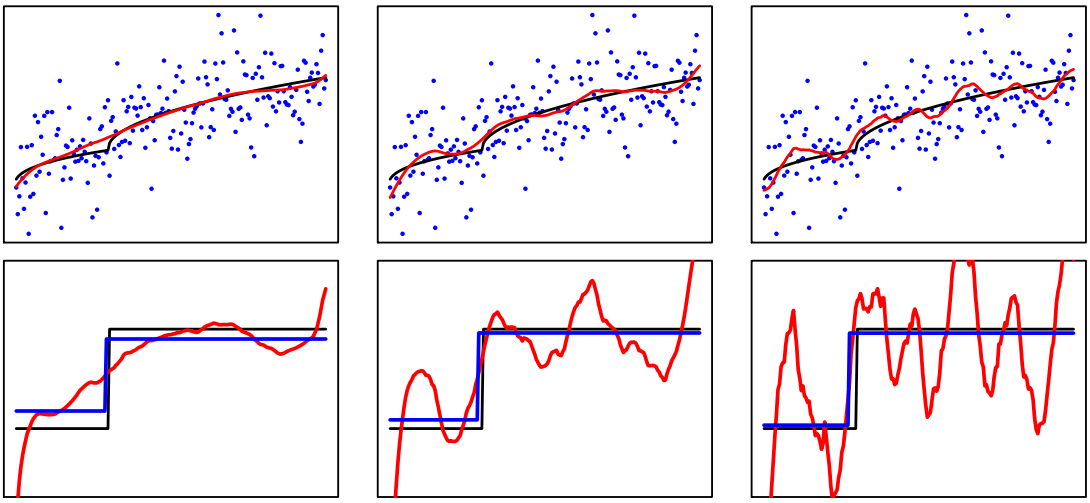

Figure 8.3.: Local linear regression estimator of $\mathrm{K}f$ with different bandwidths (first row) and the corresponding estimates for f (second row). The red line shows the estimate, and the black line the signal. The blue dots represent the observations. The blue line (lower row) is the least squares fit of a step function to the estimate.

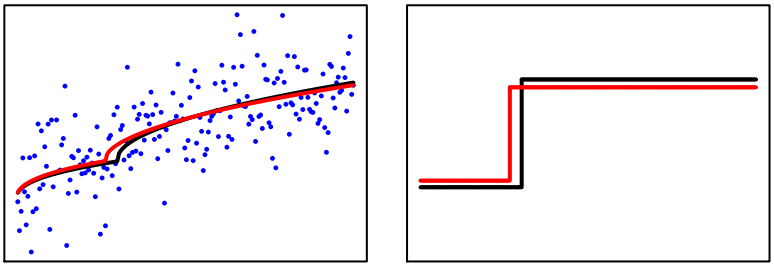

Figure 8.4.: Least squares of $\mathrm{K}f$ for $f \in T_1(0,1)$ and the corresponding estimate for f. The red line shows the estimate and the black line the signal. The blue dots represent the observations.

Chapter 9

Discussion

Computational Feasibility

Unlike the direct setting, the inverse regression model does not allow for a division of the problem into smaller independent subproblems. This makes finding a solution $f \in T_k(\tau_{low}, \tau_{up})$ of the minimization problem

$$\| \mathrm{K} f - Y \|_n \longrightarrow \min!$$

difficult. For any given vector $\tau \in \Gamma_k(\tau_{low}, \tau_{up})$ of jump points, the corresponding model reduces to the simple linear regression model

$$\| \sum_{i=1}^{k+1} b_i \, \mathrm{K} \, 1_{[\tau_{i-1}, \tau_i)}(\cdot) - Y \|_n \longrightarrow \min! \, .$$

The solutions $b_1, \ldots, b_{k+1}$ of this minimization problem can be efficiently calculated using standard software. Hence, the corresponding loss can be seen as function of τ which can be numerically minimized. If the number of jump points is small (say no larger than four or five) this is computationally feasible, but the computation time grows exponentially in the number of jump locations. This means our estimate is not very useful if step functions with a large number of jumps shall be reconstructed. In this case, it would be better to fit some nonparametric regression estimate to the data, use some regularized inversion method to calculate an initial estimate of f and then fit a step function to this initial estimate (compare end of Section 8.2). However, it is much more involved to derive an asymptotic distribution of the resulting estimates of the jump locations.

Unknown k: Choice of the smoothing parameter

Theorem 4.3 assures that if the smoothing parameter λ_n of the penalized least squares estimate tends to zero slower than $n^{-(1+\epsilon)}$ for some $\epsilon > 0$ the number of jump locations is asymptotically correctly estimated with probability one. However, the practical use of this statement is rather limited. For finite sample sizes, the statistician is still confronted with the delicate task of choosing λ_n. In the setting of this thesis, this is equivalent to choosing the number of jumps of the reconstruction f.

There is a huge amount of literature on the topic of model and parameter selection. It is beyond the scope of this thesis to examine, which methods work best in this particular setting.

However, analysis of any model selection procedure is limited by the computational feasibility of the underlying models. As argued before, this limits the analysis to models with very few jump locations. Nonetheless it can be interesting to construct procedures to test a model with one or two jumps against against a model without jumps. In the multi-phase regression setting this was done by Quandt (1960), Feder (1975a) and, more recently, by Horváth (1995) and Hušková (2000).

A closely related question is how the restricted least squares estimate behaves if the true function has less jumps than specified. As the entropy of the corresponding space is small, one can assume that the estimate $\mathrm{K}\hat{f}_n$ converges to the best approximating function of $\mathrm{K}f$, which is of course $\mathrm{K}f$ itself. By the same arguments used in the proof of Lemma 7.10, this implies consistency of $\hat{f}_n$. However, a construction of confidence intervals in this case may lead to completely wrong conclusions.

Closure of the spaces of step function

This thesis does not examine the case where the model is misspecified, i.e. the true function f is not a step function. If this is the case, the number of jumps of the penalized least squares estimate $\hat{f}_{\lambda_n}$ converges to ∞ for $n \to \infty$. For the direct problem, the properties of $\hat{f}_{\lambda_n}$ were examined by Boysen et al. (2005). It turns out, that the estimate achieves the optimal rates, if the true function f is of bounded total variation. Thus, it is natural to assume that this also holds for more general operators K. However, it is known that in the general inverse regression model this rate of convergence depends on the ill-posedness of the problem. Consequently, different methods of analysis would be necessary to derive rates of convergence. Moreover, as argued before, the computation time of the estimator grows exponentially in the number of jumps. Therefore such an estimate would be of theoretical interest only.

If one is interested in reconstructing the function f in an inverse regression model by a step function, it is probably more reasonable to use a three step procedure as described at the end of the discussion of the computational feasibility.

Singular integral kernels

Section 8.2 provides an example of an integral kernel which is not bounded. However, only the special case of an Abel-type kernel and one jump is discussed. It would be interesting to expand these results, and give convergence rates for a more general class of unbounded kernels.

Moreover, it is quite unclear, what the asymptotic distribution of the jump estimates is in such a setting. An indication is given by the results of Müller (1992), who derived a normal distribution and faster rates than $n^{-1/2}$ for a kernel based estimate of a jump in the p-th derivative of the regression function.

Appendix A

Tools from mathematical statistics

A.1. Empirical process theory

In this section we introduce some uniform deviation inequalities from empirical process theory. There is a large amount of literature on inequalities of this type. Good references are Pollard (1984), van der Vaart and Wellner (1996), van der Vaart (1998), van de Geer (2000) and Devroye and Lugosi (2001) to mention just a few.

The results cited below, are taken from van de Geer (2000). The error condition used in that book is that there exist $0 < C_0, \sigma_0 < \infty$ such that

$$\lim_{n\to\infty} \max_{i=1\ldots n} C_0^2 \,\mathrm{E}(\exp(\varepsilon_i^2/C_0^2)) \leq \sigma_0^2 \,.$$

This is obviously weaker than Assumption A with errors satisfying (A1).

The first theorem gives an exponential inequality for the least squares estimate depending on the local entropy of some function space $\mathcal{G}$.

Theorem A.1. *Suppose Assumption A is met, the error satisfies (A1) and $\hat{g}_n$ is given by*

$$\hat{g}_n := \operatorname*{argmin}_{g\in\mathcal{G}} \|g_0 + \varepsilon - g\|_n \,.$$

For

$$\mathcal{G}_n(\delta) := \{g \in \mathcal{G} : \|g - g_0\|_n \leq \delta\}$$

take $\Psi(\delta) \geq J(\delta, \mathcal{G}_n(\delta), Q_n)$ in such a way that $\Psi(\delta)/\delta^2$ is a nonincreasing function of δ. Then for a constant c depending only on Assumption (A1) and for

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n) \tag{A.1}$$

we have for all $\delta \geq \delta_n$

$$\mathrm{P}(\|\hat{g}_n - g_0\|_n \geq \delta) \leq c\exp(-n\delta^2 c^{-2}) \,.$$

Proof. See Theorem 9.1, p. 151 in van de Geer (2000). □

The next theorem gives a uniform deviation inequality depending on the entropy of the function space.

Theorem A.2. *Suppose Assumption A is met and the error satisfies (A1). Moreover, assume* $\sup_{g\in\mathcal{G}} \|g\|_n \le R$. *There exists a constant* C *depending only on Assumption (A1), such that for all* $\delta > 0$ *satisfying*

$$\sqrt{n}\delta \ge C\Big(\int_0^R H^{1/2}(u,\mathcal{G},Q_n)du \vee R\Big) \tag{A.2}$$

we have that

$$\mathrm{P}\Big(\sup_{g\in\mathcal{G}}\Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_{n,i}g(x_i)\Big| \ge \delta\Big) \le C\exp\Big(-\frac{n\delta^2}{C^2R^2}\Big). \tag{A.3}$$

Proof. See Lemma 3.2, page 29 in van de Geer (2000). □

The last result gives stochastic convergence to zero of $\sup_{g\in\mathcal{G}_n(R)} |\langle\varepsilon, g\rangle_n|$ under quite general conditions on the entropy of $\mathcal{G}_n(R)$.

Lemma A.3. *Assume* $\varepsilon_1,\dots,\varepsilon_n$ *are i.i.d. with mean zero and* $\mathrm{E}(\varepsilon_1^2) = \sigma^2 < \infty$. *Set* $\mathcal{G}_n(R) = \{g \in \mathcal{G} : \|g\|_n \le R\}$ *and suppose that*

$$\frac{1}{n}\,H(\delta,\mathcal{G}_n(R),Q_n) \to 0 \quad \textit{for all} \quad \delta > 0, R > 0.$$

Then

$$\sup_{g\in\mathcal{G}_n(R)} \big|\langle\varepsilon,g\rangle_n\big| = \sup_{g\in\mathcal{G}_n(R)} \big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\big| = o_P(1)$$

for every $R > 0$.

Proof. This follows directly from the proof of Theorem 4.8, page 56 in van de Geer (2000). □

A.2. Minimax estimation

In this section we introduce some tools from Tsybakov (2004) useful to obtain lower bounds for estimation.

Assume $\theta \in \Theta$ is some parameter, $\{P_\theta, \theta \in \Theta\}$ is some parametric family of probability measures and we are given observations distributed according P_θ. The following theorem gives a lower bound on the probability to identify $\theta = \theta_0$ given $\theta \in \{\theta_0,\theta_1,\dots,\theta_M\}$ in dependence of the Kullback-Leibler distance $d_K(P_{\theta_0}, P_{\theta_j})$ for $j = 1,\dots,M$.

Theorem A.4. *Suppose* $M \ge 2$ *and that* Θ *contains elements* $\theta_0,\theta_1,\dots,\theta_M$ *with*

$$d(\theta_j,\theta_k) \ge 2s > 0, \qquad \forall\, 0 \le j < k \le M\,,$$

$P_{\theta_j} \ll P_{\theta_0}$ *for all* $j = 1,\dots,M$ *and*

$$\frac{1}{M}\sum_{j=1}^M d_K(P_{\theta_j},P_{\theta_0}) \le \alpha\log M\,,$$

with $0 < \alpha < 1/10$. *Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \Big(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log M}}\Big) > 0\,.$$

Proof. See Theorem 2.5, page 85 in Tsybakov (2004). Note that, Tsybakov gives $0 < \alpha < 1/8$, which is to be a typo, since $1 - 2\alpha - 2\sqrt{\alpha/\log M} < 0$ for $\alpha = 1/8$, $M = 2$ if the natural logarithm is used. □

Application of Theorem A.4 requires the knowledge of the Kullback-Leibler distance of two measures. The following lemma provides this distance for two normal measures with the same variance. Though this result is standard, it is given for the sake of completeness.

Lemma A.5. *Assume P and Q are measures belonging to normal distributions with variances σ^2 and means μ_1 and μ_2. Then*

$$d_K(P, Q) = (2\sigma^2)^{-1}(\mu_1 - \mu_2)^2\,.$$

Proof. Denote by φ_{μ,σ^2} the density of a normal distribution with mean μ and variance σ^2. Compute

$$\begin{aligned} d_K(P, Q) &= \int \varphi_{\mu_1,\sigma^2}(x)(\log(\varphi_{\mu_1,\sigma^2}(x)) - \log(\varphi_{\mu_2,\sigma^2}(x)))dx \\ &= (2\sigma^2)^{-1} \int \varphi_{\mu_1,\sigma^2}(x)(-(x-\mu_1)^2 + (x-\mu_2)^2)dx \\ &= (2\sigma^2)^{-1} \int \varphi_{\mu_1,\sigma^2}(x)(2x(\mu_1-\mu_2) + \mu_2^2 - \mu_1^2)dx \\ &= (2\sigma^2)^{-1}(\mu_1-\mu_2)^2\,. \end{aligned}$$

This proves the claim. □

Appendix B

Tools from approximation theory

B.1. Positive definite functions

In the literature the definition below is mostly given for complex-valued functions. If we consider real-valued functions only, we have to restrict the definition to symmetric functions to be consistent with the usual notion.

Definition B.1. *A symmetric, continuous function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is called positive definite if for all $n \in \mathbb{N}$, all sets of pairwise distinct points $X = x_1, \ldots, x_n \subset \mathbb{R}^d$ and all $\alpha \in \mathbb{R}^n \backslash \{0\}$ we have*

$$\sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k \Phi(x_i - x_k) > 0\,. \tag{B.1}$$

If in (B.1) "$\geq$" holds instead of "$>$", the corresponding function is called positive semi-definite.

Some elementary properties of positive semi-definite functions are $\Phi(0) \geq 0$ and

$$|\Phi(x)| \leq \Phi(0) \quad \text{for all} \quad x \in \mathbb{R}^d\,.$$

Positive semi-definite functions can be characterized in terms of Fourier transforms. The most well-known result in this direction is probably Bochner's theorem, which states that positive semi-definite functions are those functions, which are Fourier transform of finite nonnegative Borel measures.

The following theorem gives an easily verifiable condition for positive definite functions.

Theorem B.2. *Suppose that $\Phi \in L_1(\mathbb{R}^d)$ is continuous. Then Φ is positive definite if and only if Φ is bounded and its Fourier transform is nonnegative and nonvanishing.*

Proof. See Theorem 6.11 in Wendland (2005). □

One example for a class of positive definite function are the truncated power functions.

Theorem B.3. *The function $(1 - |x|)_+^n$ is positive definite on $\mathbb{R}^d$ for $n \geq \lfloor d/2 \rfloor + 1$.*

Proof. See Theorem 6.20 in Wendland (2005). □

B.2. Native spaces and reproducing kernel Hilbert spaces

Reproducing kernel Hilbert spaces are well known and have been thoroughly studied in numerical analysis. A classical reference is Meschkowski (1962). Results on the properties of the reproducing kernel Hilbert space given the reproducing kernel, can be found in Madych and Nelson (1988, 1990) as well as in the overview articles Schaback (1999, 2000). Another good reference is the recent book of Wendland (2005).

This section gives the main notions and some useful results. We start with the definition of a reproducing kernel.

Definition B.4. *Let $\mathcal{H}$ be a real Hilbert space of functions $f : \Omega \to \mathbb{R}$. A function $\Phi : \Omega \times \Omega \to \mathbb{R}$ is called a reproducing kernel for $\mathcal{H}$ if*

$$\Phi(y, \cdot) \in \mathcal{H} \quad \textit{for all} \quad y \in \Omega$$

and

$$f(y) = \langle f, \Phi(y, \cdot)\rangle_{\mathcal{H}} \quad \textit{for all} \quad f \in \mathcal{H} \quad \textit{and all} \quad y \in \Omega\,.$$

Closely linked to the notion of a reproducing kernel is that of a native space.

Definition B.5. *If a symmetric positive definite function $\Phi : \Omega\times\Omega \to \mathbb{R}$ is the reproducing kernel of a real Hilbert space $\mathcal{H}$ of real functions on Ω then $\mathcal{H}$ is called the native space for Φ.*

Existence and uniqueness of the native space given some positive definite function Φ is given by the following theorem.

Theorem B.6. *Any positive definite function Φ on some domain Ω has a unique native space $\mathcal{N}_\Phi(\Omega)$. It is the closure of the space*

$$\mathcal{F}_\Phi(\Omega) := \Big\{ \sum_{i=1}^{M} \lambda_i \Phi(x_i, \cdot) \,:\, \lambda_i \in \mathbb{R},\, x_i \in \Omega,\, M \in \mathbb{N} \Big\}\,,$$

under the inner product

$$\Big\langle \sum_{i=1}^{M} \lambda_i \Phi(x_i, \cdot) \sum_{j=1}^{M'} \mu_j \Phi(y_j, \cdot) \Big\rangle_\Phi = \sum_{i=1}^{M} \sum_{j=1}^{M'} \lambda_i \mu_j \Phi(x_i, y_j)\,.$$

The elements of $\mathcal{N}_\Phi(\Omega)$ can be interpreted as functions via

$$f(x) = \big\langle f, \Phi(x, \cdot)\big\rangle_\Phi\,.$$

Proof. See Schaback (1999), Theorem 8. □

As in Theorem B.6, for any given positive definite function Φ we will denote the corresponding native space by $\mathcal{N}_\Phi(\Omega)$.

The next theorem gives a characterization of the native space in the case $\Omega = \mathbb{R}^d$ by Fourier transforms. It shows that the native space consists of smooth functions.

Theorem B.7. *Suppose $\Phi \in C(\mathbb{R}^d) \cap L_1(\mathbb{R}^d)$ is a real-valued positive definite function. Define*

$$\mathcal{G} := \left\{ f \in C(\mathbb{R}^d) \cap L_2(\mathbb{R}^d) : \widehat{f} \Big/ \sqrt{\widehat{\Phi}} \in L_2(\mathbb{R}^d) \right\}$$

and equip this space with the bilinear form

$$\langle f, g \rangle_{\mathcal{G}} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\widehat{f}(x)\overline{\widehat{g}(x)}}{\widehat{\Phi}(x)} dx \,.$$

Then $\mathcal{G}$ is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ and reproducing kernel Φ, where Φ is interpreted as kernel via $\Phi(x, y) = \Phi(x - y)$. Hence $\mathcal{G}$ is the native space of the kernel Φ on $\mathbb{R}^d$, i.e. $\mathcal{N}_\Phi(\mathbb{R}^d) = \mathcal{G}$.

Proof. See Theorem 10.12 in Wendland (2005). □

The next results gives the dependency of $\mathcal{N}_\Phi(\Omega)$ on Ω.

Theorem B.8. *Assume that $\Omega \subset \mathbb{R}^d$ and $f \in \mathcal{N}_\Phi(\mathbb{R}^d)$. Then the restriction $g := f|_\Omega$ is contained in $\mathcal{N}_\Phi(\Omega)$ and*

$$\langle g, g \rangle_{\mathcal{N}_\Phi(\Omega)} \leq \langle f, f \rangle_{\mathcal{N}_\Phi(\mathbb{R}^d)} \,.$$

Proof. See Theorem 10.47 in Wendland (2005). □

List of symbols

$\xrightarrow{\mathcal{L}}$	Convergence in distribution, page 16
$\xrightarrow{p}$	Convergence in probability, page 16
$\|\cdot\|_n$	Empirical L_2 norm, page 15
$\langle\cdot,\cdot\rangle_n$	Empirical scalar product, page 15
$\lceil a \rceil$	Integer larger than or equal to a, page 17
$\lfloor a \rfloor$	Integer smaller than or equal to a, page 17
$d_K(P,Q)$	Kullback-Leibler distance of P and Q, page 16
$P \ll Q$	P is absolutely continuous with respect to Q, page 16
$\widehat{\Phi}$	Fourier transform of Φ, page 33
$1_{(cond)}$	Indicator for some boolean expression $(cond)$, page 17
$a \vee b$	$\max(a,b)$, page 17
$a \wedge b$	$\min(a,b)$, page 17
β^t	Transpose of β, page 17
$\mathrm{d}(A,B)$	Hausdorff distance of A, B, page 15
$\Delta_K(x,a,b)$	Modified version of $\mathrm{K}\,1_{[a,b)}(x)$, page 26
$\hat{f}_{\lambda_n}$	Penalized least squares estimate, page 29
$\hat{f}_n$	Restricted least squares estimate, page 28
$f^{(m)}$	m-th derivative of f, page 34
$f(x)_+$	Positive part of f, page 17
$\Gamma_k(\tau_{low},\tau_{up})$	Set of ordered points $\subset [\tau_{low},\tau_{up}]^{k+2}$, page 14
$H(\delta,\mathcal{G})$	Entropy for Lebesgue measure, page 16
$H(\delta,\mathcal{G},Q)$	Entropy for measure Q, page 16
$h(g)$	Minimal jump height of g, page 15
$\mathcal{I}(a,b)$	Interval $[a \wedge b, a \vee b]$, page 17
$\mathcal{J}(g)$	Set of jumps of g, page 15
$J(\delta,\mathcal{G},Q)$	Entropy integral, page 16
K	Operator, page 26
K	Integral kernel, page 26
$\mathrm{m}_J(g)$	Minimal distance of two jumps of g, page 84
$N(\delta,\mathcal{G})$	Covering number for Lebesgue measure, page 16
$N(\delta,\mathcal{G},Q)$	Covering number, page 16
$\mathcal{N}_\Phi(\Omega)$	Native space, page 102

$O_P(a_n)$	Stochastic order symbol, page 16
$o_P(a_n)$	Stochastic order symbol, page 16
Q_n	Empirical measure, page 16
$\mathcal{S}(\mathbb{R})$	Schwartz space, page 34
$\mathrm{supp}(f)$	Support of f, page 17
$T_\infty(\tau_{low}, \tau_{up})$	Set of step functions on $[\tau_{low}, \tau_{up}]$, page 14
$T_{\infty,R}(\tau_{low}, \tau_{up})$	Set of bounded step functions on $[\tau_{low}, \tau_{up}]$, page 14
$T_k(\tau_{low}, \tau_{up})$	Set of step functions with k jumps on $[\tau_{low}, \tau_{up}]$, page 14
$T_{k,R}(\tau_{low}, \tau_{up})$	Set of bounded step functions with k jumps on $[\tau_{low}, \tau_{up}]$, page 14

Bibliography

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85** 115–129.

Achieser, N. I. (1992). *Theory of approximation.* Dover Publications Inc., New York. Translated from the Russian and with a preface by Charles J. Hyman, Reprint of the 1956 English translation.

Bissantz, N., Hohage, T. and Munk, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems* **20** 1773–1789.

Boysen, L., Kempe, A., Liebscher, V., Munk, A. and Wittich, O. (2005). Consistencies and rates of convergence of jump-penalized least squares estimators. Preprint.

Braess, D. (1986). *Nonlinear approximation theory*, vol. 7 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin.

Braun, J. V., Braun, R. K. and Müller, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314.

Carlstein, E. and Müller, H.-G. (eds.) (1994). *Change-point problems.* Institute of Mathematical Statistics Lecture Notes—Monograph Series, 23, Institute of Mathematical Statistics, Hayward, CA. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28** 408–428.

Christensen, J. and Rudemo, M. (1996). Multiple change-point analysis of disease incidence rates. *Prev. Vet. Med.* 54–76.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation.* Springer Series in Statistics, Springer-Verlag, New York.

Dümbgen, L. and Johns, R. B. (2004). Confidence bands for isotonic median curves using sign tests. *J. Comput. Graph. Statist.* **13** 519–533.

Engl, H. W., Hanke, M. and Neubauer, A. (1996). *Regularization of inverse problems*, vol. 375 of *Mathematics and its Applications.* Kluwer Academic Publishers Group, Dordrecht.

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272.

FEDER, P. I. (1975a). The log likelihood ratio in segmented regression. *Ann. Statist.* **3** 84–97.

FEDER, P. I. (1975b). On asymptotic distribution theory in segmented regression problems–identified case. *Ann. Statist.* **3** 49–83.

FREDKIN, D. and RICE, J. (1992). Baysian restoration and single-channel patch clamp recordings. *Biometrics* **48** 427–428.

GOLDENSHLUGER, A., TSYBAKOV, A. and ZEEVI, A. (2006). Optimal change-point estimation from indirect observations. *Ann. Statist.* To appear.

HALL, P., PAIGE, R. and RUYMGAART, F. H. (2003). Using wavelet methods to solve noisy Abel-type equations with discontinuous inputs. *J. Multivariate Anal.* **86** 72–96.

HINKLEY, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56** 495–504.

HINKLEY, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17.

HOHAGE, T. (2002). *Lecture Notes on Inverse Problems.*

HORVÁTH, L. (1995). Detecting changes in linear regressions. *Statistics* **26** 189–208.

HUŠKOVÁ, M. (2000). Some invariant test procedures for detection of structural changes. *Kybernetika (Prague)* **36** 401–414.

ISING, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z. Physik* **31** 253.

KARLIN, S. (1968). *Total positivity. Vol. I.* Stanford University Press, Stanford, Calif.

KARLIN, S. and STUDDEN, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics.* Pure and Applied Mathematics, Vol. XV, Interscience Publishers John Wiley & Sons, New York-London-Sydney.

KEMPE, A. (2004). *Statistical Analysis of Discontinuous Phenomena with Potts Functionals.* Ph.D. thesis, Institut für Biomathematik und Biometrie an der Gesellschaft für Umwelt und Gesundheit, München-Neuherberg.

KOROSTELËV, A. P. (1987). Minimax estimation of a discontinuous signal. *Teor. Veroyatnost. i Primenen.* **32** 796–799.

KOUL, H. L., QIAN, L. and SURGAILIS, D. (2003). Asymptotics of M-estimators in two-phase linear regression models. *Stochastic Process. Appl.* **103** 123–154.

MADYCH, W. R. and NELSON, S. A. (1988). Multivariate interpolation and conditionally positive definite functions. *Approx. Theory Appl.* **4** 77–89.

MADYCH, W. R. and NELSON, S. A. (1990). Multivariate interpolation and conditionally positive definite functions. II. *Math. Comp.* **54** 211–230.

MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283.

MESCHKOWSKI, H. (1962). *Hilbertsche Räume mit Kernfunktion.* Die Grundlehren der mathematischen Wissenschaften, Bd. 113, Springer-Verlag, Berlin.

MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761.

MUNK, A. (2002). Testing the goodness of fit of parametric regression models with random Toeplitz forms. *Scand. J. Statist.* **29** 501–533.

NEUMANN, M. H. (1997). Optimal change-point estimation in inverse problems. *Scand. J. Statist.* **24** 503–521.

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer Series in Statistics, Springer-Verlag, New York.

POTTS, R. (1952). Some generalized order-disorder transitions. *Proc. Camb. Phil. Soc.* **48** 106–109.

QUANDT, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53** 873–880.

QUANDT, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Amer. Statist. Assoc.* **55** 324–330.

RAIMONDO, M. (1998). Minimax estimation of sharp change points. *Ann. Statist.* **26** 1379–1397.

SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors. III. *Ann. Math. Statist.* **41** 2057–2074.

SCHABACK, R. (1999). Native Hilbert spaces for radial basis functions. I. In *New developments in approximation theory (Dortmund, 1998)*, vol. 132 of *Internat. Ser. Numer. Math.* Birkhäuser, Basel, 255–282.

SCHABACK, R. (2000). A unified theory of radial basis functions. Native Hilbert spaces for radial basis functions. II. *J. Comput. Appl. Math.* **121** 165–177. Numerical analysis in the 20th century, Vol. I, Approximation theory.

SPRENT, P. (1961). Some hypotheses concerning two phase regression lines. *Biometrics* **17** 634–645.

TSYBAKOV, A. B. (2004). *Introduction à l'estimation non-paramétrique*, vol. 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin.

VAN DE GEER, S. A. (2000). *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

VAN DER VAART, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics, Springer-Verlag, New York.

WENDLAND, H. (2005). *Scattered data approximation*, vol. 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge.

WERNER, D. (2000). *Funktionalanalysis*. extended ed. Springer-Verlag, Berlin.

WINKLER, G. and LIEBSCHER, V. (2002). Smoothers for discontinuous signals. *J. Nonparametr. Stat.* **14** 203–222.

YAKIR, B., KRIEGER, A. M. and POLLAK, M. (1999). Detecting a change in regression: first-order optimality. *Ann. Statist.* **27** 1896–1913.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189.

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381.

Curriculum Vitae

Leif Boysen

November 26, 1976	born in Bremen, Germany
1983 – 1987	(elementary school), Achim bei Bremen
1987 – 1989	Orientierungstufe am Markt, Achim bei Bremen
1989 – 1996	Gymnasium Achim (grammar school)
June 1996	Abitur (graduation from grammar school)
1996 – 1997	Civilian cervice
1997 – 2002	Diploma studies in mathematics, minor computer science, Georg-August-Universität Göttingen
November 2002	Graduation ("Diplom") in mathematics title of the diploma thesis: "Analyse von intra-individuellen Effekten bei longitudinalen Daten"
2003 – 2006	Ph.D. studies at the institute of Mathematical Stochastics, Georg-August-Universität Göttingen
April, 2003 – May, 2006	Member of the Ph.D.-Program "Applied Statistics and Empirical Methods"
April, 2003 – March, 2006	"Georg-Lichtenberg"-Scholarship
July, 2004 – May, 2006	Associated member of the graduate school "Identifikation in mathematischen Modellen"
April, 2006 – May, 2006	Scholarship of the graduate school "Identifikation in mathematischen Modellen"