ASYMPTOTIC LAWS FOR CHANGE POINT ESTIMATION IN INVERSE REGRESSION

Sophie Frick[1], Thorsten Hohage[2], Axel Munk[1]

[1] *Institute for Mathematical Stochastics,*
[2] *Institute for Numerical and Applied Mathematics,*
Georgia Augusta Universität Göttingen

Abstract: We derive rates of convergence and asymptotic normality for the least squares estimator for a large class of parametric inverse regression models $Y = (\Phi f)(X) + \varepsilon$. Our theory provides a unified asymptotic tretament for estimation of f with discontinuities of certain order, including piecewise polynomials and piecewise kink functions. Our results cover several classical and new examples, including splines with free knots or the estimation of piecewise linear functions with indirect observations under a nonlinear Hammerstein integral operator. Furthermore, we show that ℓ_0-penalisation leads to a consistent model selection, using techniques from empirical process theory. The asymptotic normality is used to provide confidence bands for f. Simulation studies and a data example from rheology illustrate the results.

Key words and phrases: Statistical inverse problems, jump detection, asymptotic normality, change point analysis, penalized least squares estimator, sparsity, entropy bounds, confidence bands, Hammerstein integral equations, reproducing kernel Hilbert spaces, dynamic stress moduli.

1 Introduction

We consider the inverse regression model

$$y_i = (\Phi f_0)(x_i) + \varepsilon_i \text{ for } i = 1, ..., n, \tag{1.1}$$

where $X = (x_1, ..., x_n)$, $n \in \mathbb{N}$ is a (possibly random) vector of design points in a bounded interval $I \subset \mathbb{R}$ and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ denotes the observation error, which

is assumed to be independent of X, with mean zero. Further, Φ denotes some integral operator $\Phi : L^2([a,b]) \longrightarrow L^2(I)$,

$$(\Phi f)(x) := \int_a^b \varphi(x,y) f(y) \mathrm{d}y, \tag{1.2}$$

acting on a piecewise continuous function $f(y) = f(y,\theta)$, which is determined by a parameter vector $\theta \in \Theta_k \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$. Here k describes the number of (unknown) discontinuities of f. The aim is to reconstruct the true function $f_0(y) = f(y,\theta_0)$ from the observations $(X,Y) = ((x_1,y_1),...,(x_n,y_n))$.

This class of models covers a large variety of important applications, ranging from multiphase regression to piecewise polynomial splines. Model (1.1) has been introduced in Boysen, Bruns, and Munk (2009a) for piecewise constant functions f, where the integral kernels φ was restricted to the class of piecewise Lipschitz continuous convolution kernels $\varphi(x,y) = \phi(x-y)$.

Integral equations as in (1.2) are well known to generate *ill posed* problems, that is, small perturbations on the right hand side of (1.1) induce large errors in the solution. Therefore, reconstruction of f_0 from (1.1) requires appropriate regularization. In this paper we show that this can be achieved in quite generality by an ℓ_0 penalized least squares estimator restricted to suitable compact function classes,indexed in Θ_k. To this end we extend the model of Boysen, Bruns, and Munk (2009a) for piecewise constant functions with respect to the considered classes of objective functions as well as with respect to the integral kernels φ. We show $n^{-1/4}$ convergence rates of the least squares estimator $f(y,\hat{\theta}_n)$ of a piecewise continuous parametric function $f(y,\theta_0)$ with known number of change points. Furthermore we obtain $n^{-1/2}$ rates for the convergence of the respective parameter estimate $\hat{\theta}_n$ of the true parameter θ_0 and show that it is asymptotically multivariate normally distributed. However, we mention that the obtained asymptotic normality together with "model consistency" in general is not uniform in these models, as the kinks or jumps may degeneraty. This is well known already from much simpler cases, see e.g. the Introduction in Boysen, Bruns, and Munk (2009a).

The particular case, when f_0 is additionally known to be continuous, i.e. f_0 has no jumps but kinks, is treated in detail. In this case the continuity assumption on f_0 improves the convergence rate of the least squares estimate $f(y,\hat{\theta}_n)$. The

improvement depends directly on the smoothness of the pieces between the kinks. For instance, for piecewise linear kink functions, we obtain $n^{-1/2}$-consistency of $\hat{f}_n := f(y, \hat{\theta}_n)$.

In order to obtain our results, we require techniques that are substantially different from Boysen, Bruns, and Munk (2009a). The extension of the class of objective functions from step functions to general piecewise continuous parametric functions requires existence and uniform L^2 boundedness of the first derivative of the pieces of the objective function $\theta \mapsto f(y, \theta)$ for almost every $y \in [a, b]$. This differentiability allows for a general estimate of the entropy of the class of piecewise continuous parametric functions, which is a main ingredient in the proof of consistency. Moreover, we will see, that exactly this property implies continuous differentiability of the mapping $\theta \mapsto (\Phi f)(y, \theta)$. This differentiability in turn paves the way to the second order expansion of the expectation of the score function, required for the proof of asymptotic normality. This is more straightforward and in particular more general, than the elementary expansion in Boysen, Bruns, and Munk (2009a). Remarkably, this approach abandons the assumption of Lipschitz continuity of $y \mapsto f(y, \theta)$ and $(x, y) \mapsto \varphi(x, y)$. The generality of the applied techniques furthermore covers the case of dependencies between the parameter components of θ, as in the case of kinks functions.

In the case, where the number of change points of the objective function in (1.1) is not known, we show that under the additional assumption of subgaussian tails of the error distribution, the number of change points can be asymptotically estimated correctly with probability one.

A key ingredient of our consistency proof is the injectivity of the integral operator Φ in (1.2). Two main classes are discussed in detail, namely product kernels $\varphi(x, y) = \phi(xy)$ and convolution kernels $\varphi(x, y) = \phi(x - y)$. For the asymptotic normality to hold injectivity of the corresponding integral operator plays an important role. To this end we introduce an injectivity condition for general symmetric and positive definite kernels (not restricted to one of the above classes), which is based on the theory of native Hilbert spaces and on the so called *full Müntz Theorem* Borwein and Erdélyi (1995). We mention, however, that the asymptotic results of this paper are not restricted to this selection, i.e. they are valid for every injective integral operator Φ with certain properties (cf.

Assumption **C**).

We further want to emphasize that the extension to piecewise continuous functions, instead of piecewise constant functions covers several interesting examples, including splines. Moreover, we show that our method can even be applied to specific *nonlinear* integral operators, namely the *Hammerstein integral operators* (see e.g. Hammerstein (1930))

$$f \longmapsto \int_a^b \varphi(\cdot, y)\mathrm{L}(f(y), y)\mathrm{d}y,$$

where the additional operator $\mathcal{L}f(y) := \mathrm{L}(f(y), y)$ is injective and satisfies certain smoothness conditions to preserve essential properties of f as e.g. the differentiability for $\mathcal{L}f$. This allows, to provide estimators and confidence bands for the time relaxation spectra of polymer melts reconstructed from their dynamic modul (see Roths, Maier, Friedrich, Marth, and Honerkamp (2000)).

Finally, we apply the asymptotic results in two examples: the estimation of a step function from the noisy image of an integral operator with convolution kernel, i.e. inverse two phase regression, and the estimation of a piecewise linear kink function from the noisy image of an integral operator with product kernel, i.e. inverse multiphase regression. In both cases, we calculate confidence bands of the reconstructed function, which give an impression of the reliability of the estimate.

We stress, that our results in this paper substantially differ from "truly nonparametric" kink models which have been the topic in a series of paper in the last two decades, including Korostelev (1987), Neumann (1997), Raimondo (1998), Goldenshluger, Tsybakov, and Zeevi (2006), Goldenshluger, Juditsky, Tsybakov, and Zeevi (2008b), Goldenshluger, Juditsky, Tsybakov, and Zeevi (2008a) for independent error, and recently Wishart (2010, 2011) for long range dependent error. In the present paper f is modeled as a piecewise "parametric" function, which is $\sqrt{n}$ estimable between kinks, leading to asymptotic normality and a parametric rate of convergence. It is easily seen that this rate is minimax for *bounded kernels* φ in (1.2), and can be even improved for singular kernels (see Boysen, Bruns, and Munk (2009a)). This is in contrast to the afore mentioned papers, where piecewise (nonparametric) smooth functions are treated which requires a different estimation technique and analysis. This also leads to different rates of convergence which are additionally deteriorated by the smoothness be-

tween discontinuities. Roughly speaking, the situation treated here can be viewed as a limiting case, when the degree of smoothness tends to infinity.

This paper is structured as follows. Section 2 gives some basic notation and the main assumptions. The estimator and its asymptotic properties are given in Section 3. Section 4 discusses injectivity of the considered integral operators. In Section 5 we show how the asymptotic normality can be used for the construction of confidence bands for the case of jump and kink functions, respectively. The finite sample performance of the asymptotic distribution is briefly investigated in a simulation study. The proofs of the asymptotic results from Section 3 and the injectivity statements from Section 4 are given in a supplement to this paper.

2 Definitions and assumptions

2.1 Notations

For functions $g, f : I \to \mathbb{R}$, we denote by $\|f\|_{L^2(I)}$ the L^2-norm and by $\langle f, g\rangle_{L^2(I)}$ the corresponding inner product. The essential supremum is denoted by $\|f\|_\infty$, the empirical norm and the empirical inner product by

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2 \quad \text{and} \quad \langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i),$$

where $x_1, ..., x_n$ are given design points. Accordingly, the empirical measure is $P_n := n^{-1}\sum_{i=1}^n \delta_{x_i}$. For vectors $\theta, \theta_1, \theta_2 \in \mathbb{R}^d$, we use the Euclidean norm $|\theta|_2$ and the maximum norm $|\theta|_\infty$ and by $(\theta_1, \theta_2) \subset \mathbb{R}^d$ we denote the segment between θ_1 and θ_2, that is $(\theta_1, \theta_2) := \{\theta \in \mathbb{R}^d \mid \theta = \theta_1 + t(\theta_2 - \theta_1), \text{ for } t \in (0, 1)\}$.

2.2 Piecewise continuous parametric functions

We start by introducing the class of functions f to be estimated in model (1.1). Throughout this paper we assume that $a, b \in \mathbb{R}$, $a < b$ and $r, k \in \mathbb{N} \setminus \{0\}$.

Definition 2.1. *Assume that $\Psi \subset \mathbb{R}^r$ is convex and compact and choose $M > 0$ such that $|\vartheta|_\infty \le M$ for all $\vartheta \in \Psi$. Let*

$$\mathfrak{f} : [a, b] \times \Psi \quad \longrightarrow \quad \mathbb{R}$$

be a function $\mathfrak{f}(y, \vartheta)$ satisfying the following conditions:

i) $\mathfrak{f}$ *is continuous and continuously differentiable with respect to* ϑ*.*

ii) For all open subintervals $I \subset [a,b]$ *the mapping* $\mathfrak{F}_I : \Psi \to C(I)$*,* $\mathfrak{F}_I(\vartheta) := \mathfrak{f}(\cdot,\vartheta)|_I$ *is injective, and its derivative* $\mathfrak{F}_I'[\vartheta] : \mathbb{R}^r \to C(I)$ *is also injective for all* $\vartheta \in \Psi$*. Here* $C(I)$ *denotes the set of continuous functions on* I*.*

Then, $\mathcal{F} := \{\mathfrak{f}(\cdot,\vartheta) \mid \vartheta \in \Psi\}$ *is called a family of* ***continuous parametric functions*** *with parameter domain* Ψ*.*

For example, this includes the following two families.

Example 2.2 (Constant functions)**.** If $\Psi = [-M, M]$ and $\mathfrak{f}(y,\vartheta) := \vartheta$ we obtain

$$\mathcal{F}_T := \{\vartheta \mathbf{1}_{[a,b]} \mid |\vartheta| \le M\},$$

Example 2.3 (Linear functions)**.** If $\Psi = [-M, M]^2$ and $\mathfrak{f}(y,\vartheta) := \vartheta_1 + y\vartheta_2$ we obtain

$$\mathcal{F}_L := \{\vartheta_1 + \vartheta_2 \bullet \mid |\vartheta_1|, |\vartheta_2| \le M\}.$$

These functions from a known family $\mathcal{F}$ will now constitute the building blocks for the class of functions in the next definition.

Definition 2.4. *Let* $\mathcal{F} = \{\mathfrak{f}(\cdot,\vartheta) : \vartheta \in \Psi\}$ *be a family of continuous parametric functions on the interval* $[a,b]$ *in the sense of Definition 2.1. A function* $f \in L^\infty([a,b])$ *is called a* ***parametric piecewise continuous function*** *(pc-function) generated by* $\mathcal{F}$ *if there exists a partition* $a = \tau_0 < \tau_1 < ... < \tau_{k+1} = b$ *and parameter vectors* $\vartheta^1, .., \vartheta^{k+1} \in \Psi$ *such that*

$$f = \sum_{j=1}^{k+1} \mathfrak{f}(\cdot,\vartheta^j)\, \mathbf{1}_{[\tau_{j-1},\tau_j)}\,. \tag{2.1}$$

The function above will also be denoted by $f(\cdot, \vartheta^1, \tau_1, ..., \vartheta^k, \tau_k, \vartheta^{k+1})$*. We call the elements of the set*

$$\mathcal{J}(f) := \{\tau_i \mid i \in \{1, ..., k\} \text{ such that } \vartheta^i \ne \vartheta^{i+1} \text{ and } \tau_i < \tau_{i+1}\}$$

change points of the function $f \in \boldsymbol{F}_k$ *and denote its cardinality by* $\sharp\mathcal{J}(f)$*. The set of all parametric piecewise continuous functions with at most* k *change points generated by* $\mathcal{F}$ *is denoted by* $\boldsymbol{F}_k[a,b]$ *(or shortly by* $\boldsymbol{F}_k$*). Using the notation*

$$[f](\tau) \;:=\; \lim_{\epsilon \searrow 0} (f(\tau+\epsilon) - f(\tau-\epsilon)),$$

we say that f has a jump at τ if $[f(\cdot,\theta)](\tau) \neq 0$, and that f has a kink at τ if τ is a change point and $[f(\cdot,\theta)](\tau) = 0$. Moreover, we say that f is a kink function (or jump function) if it has kinks (or jumps) at all change points.

Note that $\theta := (\vartheta^1, \tau_1, .., \vartheta^k, \tau_k, \vartheta^{k+1})$ lies in the convex and compact parameter set $\Theta_k \subset \mathbb{R}^d$, $d = (r+1)k + r$ where

$$\Theta_k = \{(\vartheta^1, \tau_1, .., \vartheta^k, \tau_k, \vartheta^{k+1}) \in (\Psi \times [a,b])^k \times \Psi \mid a \leq \tau_1 \leq ... \leq \tau_k \leq b\}\,. \quad (2.2)$$

Thus $\mathbf{F}_k = \{f(\cdot,\theta) \mid \theta \in \Theta_k\}$. Accordingly we define

$$\boldsymbol{F}_\infty[a,b] = \bigcup_{k=1}^{\infty} \boldsymbol{F}_k[a,b].$$

Example 2.5. Continuing Examples 2.2 and 2.3 above, the families $\mathcal{F}_T$ and $\mathcal{F}_L$ generate sets of step functions $\mathbf{T}_k$ and piecewise linear functions $\mathbf{L}_k$, respectively, in the sense of Definition 2.4.

Note that for functions $f \in \mathbf{F}_k$ with less than k change points there is more than one parameter vector in Θ_k generating f. In other words, the implication $f(\cdot,\theta) = f(\cdot,\theta_0) \Rightarrow \theta = \theta_0$ is true if and only if $\sharp\mathcal{J}(f) = k$. If uniqueness of the parameter vector is required, we have to confine ourselves to functions in $\mathbf{F}_k$ with precisely k change points. To illustrate, again we continue Example 2.2 and 2.3 and consider the subset of $\tilde{\mathbf{T}}_k \subset \mathbf{T}_k$, with precisely k jumps, i.e.

$$\tilde{\mathbf{T}}_k := \{f = f(\cdot,\theta) \in \mathbf{T}_k \mid [f](\tau_i) \neq 0, \tau_{i-1} < \tau_i,\, i = 1, \cdots, k+1\} \quad (2.3)$$

and the subset $\tilde{\mathbf{L}}_k \subset \mathbf{L}_k$ of piecewise linear functions with precisely k kinks, i.e.

$$\tilde{\mathbf{L}}_k := \{f \in \mathbf{L}_k \mid \vartheta_1^i = \vartheta_1^{i-1} - (\vartheta_2^{i-1} - \vartheta_2^i)\tau_{i-1}, \text{ and } \vartheta_2^{i-1} \neq \vartheta_2^i, \tau_{i-1} < \tau_i\, i = 2, ..., k+1\}. \quad (2.4)$$

As in the case of kinks there may occur dependencies among the parameter components such that actually the number of parameters which determine $f(\cdot,\theta)$ is smaller than the dimension of θ. Therefore we define a so called *reduced parameter domain.*

Definition 2.6. *$\Theta_k \subset \mathbb{R}^d$ denote the parameter domain of a family $\boldsymbol{F}_k$ of pc functions in the sense of Definition 2.4. If $\tilde{\Theta} \subset \mathbb{R}^{\tilde{d}}$ is convex and compact and*

if there exists a continuously differentiable function $h : \tilde{\Theta} \to \Theta_k$ *such that the mapping*

$$\tilde{\Theta} \to \boldsymbol{F}_k, \qquad \tilde{\theta} \mapsto f(\cdot, h(\tilde{\theta}))$$

and its derivative $\delta\tilde{\theta} \mapsto \frac{\partial f}{\partial \theta}(\cdot, h(\tilde{\theta}))\delta\tilde{\theta}$ *are injective, then* $\tilde{\Theta}$ *is called a* ***reduced parameter domain*** *of* $\tilde{\boldsymbol{F}}_k := \{f(\cdot, h(\tilde{\theta})) \mid \tilde{\theta} \in \tilde{\Theta}\}$*, and the elements* $\tilde{\theta}_0 \in \tilde{\Theta}$ *are called* ***reduced parameter vectors*** *of the functions* $f(\cdot, h(\tilde{\theta})) \in \tilde{\boldsymbol{F}}_k$*.*

Note that if we consider a class of pc-functions $\mathbf{F}_k$ as in Definition 2.4, which is generated by a parametric class $\mathcal{F}$ as in Definition 2.1 and if moreover $(y, \vartheta) \mapsto \mathfrak{f}(y, \vartheta)$ is continuously differentiable, then the condition $[f(\cdot, \theta)](\tau) = 0$ often implies local existence of a function h as in Definition 2.6 by the implicit function theorem. More precisely, if $f(y, \theta_0)$ is a kink function in such a space, the function

$$\begin{aligned} F : \Theta_k &\longrightarrow \mathbb{R}^k \\ \theta &\longmapsto F(\theta) := \Big(\mathfrak{f}(\tau_1, \vartheta^1) - \mathfrak{f}(\tau_1, \vartheta^2), ..., \mathfrak{f}(\tau_k, \vartheta^k) - \mathfrak{f}(\tau_k, \vartheta^{k+1})\Big)^\top \end{aligned}$$

vanishes in θ_0. Due to the differentiability of the map $\theta \mapsto F(\theta)$, the implicit function theorem implies that there exists a function h and a reduced parameter domain $\tilde{\Theta}$ as in Definition 2.6, with $\tilde{\Theta} \subset (\Theta_l)_{l \in I} \subset \mathbb{R}^{d-k}$, where $I \subset \{1, ..., d\}$ if the Jacobian $\partial / (\partial \theta_l)_{l \notin I} F(\theta_0)$ is invertible.

Consider for example the set $\tilde{\mathbf{L}}_1$ in (2.4). There we have $\vartheta_1^2 = \vartheta_1^1 + (\vartheta_2^1 - \vartheta_2^2)\tau_1$ and choosing the reduced parameter vector $\tilde{\theta} = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_2^2)$ and the function $h(\tilde{\theta}) = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_1^1 + (\vartheta_2^1 - \vartheta_2^2)\tau_1, \vartheta_2^2)$ the conditions of Definition 2.6 are satisfied.

2.3 Assumptions on the model

Assumption A (Assumptions on the error)**.** Throughout this paper we assume that

A1: the vector $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ consists of independent identically distributed random variables with mean zero for every n and $E(\varepsilon_1^2) = \sigma^2 < \infty$.

In some situations, the error is additionally needed to satisfy the following sub-gaussian condition.

A2: ε satisfies **A1** and there exists some $\alpha > 0$ such that $E(e^{\varepsilon_1^2/\alpha}) < \infty$.

Assumption B (Assumptions on the design)**.** There exists a function $s : I \to [s_u, s_l]$ with $0 < s_u < s_l < \infty$ and $\int_a^b s(x)\mathrm{d}x = 1$, such that

$$\frac{i}{n} = \int_a^{x_{(i)}} s(x)\mathrm{d}x + \delta_i$$

with $\nu_n := \max_{i=1,...,n} |\delta_i| = o_p(1)$. Here $x_{(i)}$ denotes the i-th order statistic of $x_1, ..., x_n$. Moreover, the design points $x_1, ..., x_n$ are independent of the error terms $\varepsilon_1, ..., \varepsilon_n$.

The above assumption covers random designs. If the design points $x_1, ..., x_n$ are nonrandom, the $o_p(1)$ term above is to be understood as $o(1)$. We will not pursue this situation further, however, we mention that with a slight change of technicalities all subsequent results hold analoguously.

2.4 Integral operator

The integral operator Φ in (1.2) acts on $\boldsymbol{F}_k \subset L^2([a,b])$, hence it can be considered as a map, acting on the parameter space Θ_k, for $x \in [a,b]$, by

$$\theta \quad \longmapsto \quad \Phi f(\cdot, \theta) := \int_a^b \varphi(\cdot, y) f(y, \theta)\mathrm{d}y. \tag{2.5}$$

In the following we will require the Frechet differentiability of Φ to ensure identifiability of the parametrization in (2.5). To this end we introduce the space $\mathcal{M}([a,b])$ of all signed Borel measures μ on $[a,b]$ of the form $\mu = f + \sum_{j=1}^n \gamma_j \delta_{x_j}$ with $f \in L^1([a,b])$, $n \in \mathbb{N}$, $x_j \in [a,b]$ and $\gamma_j \in \mathbb{R}$ and define

$$(\Phi\mu)(x) := \int_a^b \varphi(x,y)\,\mathrm{d}\mu(y) = \int_a^b \varphi(x,y) f(y)\,\mathrm{d}y + \sum_{j=1}^n \gamma_j \varphi(x, x_j), \qquad x \in I \tag{2.6}$$

for $\mu \in \mathcal{M}$ as above. In the following we denote by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the space of bounded linear operators of a normed space $\mathcal{X}$ into a normed space $\mathcal{Y}$. Moreover, we denote by $C^{0,1}(I)$ the space of uniformly Lipschitz continuous functions with norm $\|f\|_{C^{0,1}} := \|f\|_\infty + \sup_{x \neq y} |f(x) - f(y)|/|x - y|$.

Assumption C (Assumptions on the integral operator)**.** The operator Φ in (1.2) satisfies the following conditions:

i): $\Phi \in \mathcal{L}\left(L^\infty([a,b]), C^{0,1}(I)\right)$ and $\Phi \in \mathcal{L}\left(L^1([a,b]), L^\infty(I)\right)$.

ii): The mapping $[a,b] \to L^2(I)$, $y \mapsto \varphi(\cdot, y)$ is continuous, so in particular Φ is well defined on $\mathcal{M}([a,b])$ by (2.6). Moreover, $\Phi : \mathcal{M}([a,b]) \longrightarrow L^2(I)$ is injective.

Conditions *ii*) is essential for the consistency proof for the estimator of f_0 in the following chapters. Condition *i*) especially will be needed to estimate the L^2-norm of Φf by means of the empirical norm. In Section 4 we introduce some special classes of operators satisfying Assumption **C**.

Moreover, we want to mention that the results of this paper can also be formulated for $\Phi : L^2([a,b]) \to L^2(I)$, with an interval $I \subset \mathbb{R}$ which does not need to coincide with the interval $[a,b]$, but for ease of notation we only discuss the case $I = [a,b]$.

3 Estimate and asymptotic results

3.1 Known number of jumps

Estimate. Estimating f for given k and $\boldsymbol{F}_k$ can be performed using the least squares estimator $\hat{f}_n$, which is defined such that $\Phi\hat{f}_n$ minimizes the empirical distance to the observations Y in (1.1) with respect to the space $\boldsymbol{F}_k$. That is, $\hat{f}_n \in \mathbf{F}_k$ and

$$\|\Phi\hat{f}_n - Y\|_n^2 \leq \min_{f \in \boldsymbol{F}_k} \|\Phi f - Y\|_n^2 + o_P(n^{-1}). \tag{3.1}$$

Note, that this estimator implicitely depends on k and $\boldsymbol{F}_k$, of course. However, we supress this dependence in the notation whenever no confusion is expected. It then follows from Definition 2.4 that there exists a parameter vector $\hat{\theta}_n \in \Theta_k$, such that

$$\hat{f}_n(y) = f(y, \hat{\theta}_n) = \sum_{i=1}^{k+1} \mathfrak{f}(y, \hat{\vartheta}^i)\mathbf{1}_{[\hat{\tau}_{i-1}, \hat{\tau}_i)}.$$

Note further, that the parameters $\hat{\vartheta}^i$ and $\hat{\tau}_i$ also depend on the index n.

It is easy to see that the minimum on the right hand side of (3.1) always is attained, because the candidate set $\boldsymbol{F}_k$ is closed and compact. Note, that the minimum does not need to be unique. Furthermore, we mention that in (3.1) it is not required that $\hat{f}_n$ minimizes the functional $\|\Phi f - Y\|_n^2$ exactly, but only

up to a term of order $o_P(n^{-1})$. This allows for numerical approximation of the minimizer and gives an intuition of the required precision for the asymptotic results to be valid.

Consistency and asymptotic results. Now we present the main results of this paper, i.e. the asymptotic behavior of the least squares estimator in (3.1), for the case, where the true function $f_0 \in \mathbf{F}_k$ has precisely k change points, that is $\sharp\mathcal{J}(f_0) = k$ and for the case where the number of change points is not known.

We require some more notation. Let $\Lambda : \Theta_k \to L^2([a,b])$ denote the mapping

$$\Lambda\theta := \Phi f(\cdot, \theta)\,. \tag{3.2}$$

We will show in the supplement that Λ is differentiable and denote by $\Lambda'[\theta] \in L^2([a,b])^d$ its gradient at θ. With this, we define the $d \times d$ matrix V_θ by

$$(V_\theta) = \int_a^b \Lambda'[\theta](\Lambda'[\theta])^t s(x)\mathrm{d}x, \tag{3.3}$$

where s is as in Assumption **B**.

Theorem 3.1. *Suppose that Assumptions **A1**, **B** and **C** are satisfied and let $\hat{f}_n(y) = f(y, \hat{\theta}_n)$ be the least squares estimator of the true function $f_0 = f(\cdot, \theta_0) \in \boldsymbol{F}_k$ as in (3.1), with $\sharp\mathcal{J}(f_0) = k$. If the matrix V_{θ_0} is nonsingular, then*

(i) $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{\theta_0}^{-1})$,

(ii) $|\theta_0 - \hat{\theta}_n|_2 = O_P(n^{-\frac{1}{2}})$,

(iii) $\|f_0 - \hat{f}_n\|_{L^p([a,b])} = O_P(n^{-\frac{1}{2p}})$ *for any* $p \in [1, \infty)$ *and,*

(iv) $\|\Phi f_0 - \Phi \hat{f}_n\|_{L^\infty([a,b])} = O_P(n^{-\frac{1}{2}})$.

If f_0 depends on a reduced parameter vector $\tilde{\theta}$ as in Definition 2.6, the derivative of $\tilde{\theta} \mapsto \Lambda(h(\tilde{\theta}))$ can be calculated by the chain rule, due to the differentiability of the function h (cf. Definition 2.6) and we have the following corollary.

Corollary 3.2. *Suppose that Assumptions **A1**, **B** and **C** as in Theorem 3.1 are satisfied and that the true function $f_0(y) = f_0(y, h(\tilde{\theta}))$ can be parameterized by a reduced parameter domain as in Definition 2.6. Then $V_{\tilde{\theta}}$ is nonsingular, and the results of Theorem 3.1 are valid with θ_0 and $\hat{\theta}_n$ substituted by the reduced parameter vector $\tilde{\theta}_0$ and its estimator $\tilde{\theta}_n$.*

Non singularity of the covariance matrix V_{θ_0} is essential for Theorem 3.1 to hold. The next proposition gives a characterization of this property in terms of the partial derivatives $\frac{\partial}{\partial \vartheta^i}\mathfrak{f}(y,\vartheta_0^i)$, $i=1,...,k+1$, for the case, where $f(\cdot,\theta_0)$ has precisely k jumps. Moreover, it states that V_{θ_0} is always singular if $f(y,\theta_0)$ has a kink in some change point.

Proposition 3.3. *Suppose that Assumptions* $\boldsymbol{B}$ *and* $\boldsymbol{C}$ *are satisfied and* $f(\cdot,\theta)=\sum_{i=1}^{k+1}\mathfrak{f}(\cdot,\vartheta^i)\mathbf{1}_{[\tau_{i-1},\tau_i)} \in \boldsymbol{F}_k$ *(cf. Definition 2.4) has* k *change points. Then the matrix* V_θ *as defined in (3.3) is nonsingular if and only if* $f(\cdot,\theta)$ *has jumps in all change points.*

Hence, if the true function f_0 is known to be a kink function, Proposition 3.3 implies that Theorem 3.1 can not be applied, since the condition of non singularity of V_θ is violated. So this case requires restriction to a reduced parameter set $\tilde{\Theta}$ as in Definition 2.6. In this case, it is even possible to improve the rate of convergence of $\hat{f}_n$, which depends on the *modulus of continuity* of the considered function class $\mathcal{F}$ defined as

$$\nu(\mathcal{F},\delta) := \sup_{f\in\mathcal{F}}\sup_{|y_1-y_2|\le\delta}|f(y_1)-f(y_2)|. \tag{3.4}$$

Corollary 3.4. *Assume that the conditions of Corollary 3.2 are satisfied, but the true function* $f_0(y,h(\tilde{\theta}))$ *with* h *as in Definition 2.6 is a kink function and let* ν *be defined as in* (3.4)*. Then the results of Corollary 3.2 are valid, with the improved rate*

$$\|f_0-\hat{f}_n\|_{L^p([a,b])} = O_P(n^{-\frac{1}{2}}+n^{-\frac{1}{2p}}\nu(\mathcal{F},n^{-\frac{1}{2}})) \qquad \text{for } p\in[1,\infty). \tag{3.5}$$

For example, in Subsection 5.3 we obtain rates of order $n^{-1/2}$ if $f\in\tilde{L}_2$ as in (2.4). More generally, if $\mathcal{F}$ consists of Hölder continuous functions with exponent $0<\alpha\le 1$, Equation (3.5) yields rates of order $n^{-(1+\alpha)/4}$.

We finally comment on the asymptotic optimality of Theorem 3.1 and Corollary 3.4. As mentioned in the Introduction it is straight forward to see that the $\sqrt{n}$ rate in Theorem 3.1 (i), (ii) is minimax under a normal error for bounded, continuous integral kernels as considered Assumtion **C**. To this end a similar argument as in the proof of Theorem 1 in Wishart (2011) can be employed by means of estimating the Kullback Leibler divergence between the distributions

of different (Y, X) and applying Theorem 2.2.(iii) in Tsybakov (2009). Again, in a normal model, we claim the the asymptotic variance $V_{\theta_0}^{-1}$ appearing in 3.1 (and analogue $V_{\tilde{\theta}}^{-1}$ in Corollary 3.4) is asymptotically optimal in Le Cam sense, provided the experiment is differentiable in quadratic mean (see van der Vaart (1998)).

Albeit the ill posedness of the problem is not reflected in the rate of convergence it is reflected in the asymptotic variance$V_{\theta_0}^{-1}$ as can be seen from (3.3). The variance becomes large when $\Lambda'(\theta)$ becomes small, i.e. the gradient of $\Phi f(\cdot, \theta_0)$ is flat. Loosely speaking, this happens when kinks or jumps in the signal are only weakly propagated through the operator Φ and hence hard to detect.

We finally mention that we believe that the rates in Theorem 3.1 (iii) and (iv) and in Corollary 3.4 are minimax but we do not have a proof for this which might be interesting for further investigations.

3.2 Unknown number of jumps

If we do not know the number of change points of the objective function, we can use the above introduced least squares estimator $\hat{f}_n$ penalized by the number of change points $\sharp\mathcal{J}(\hat{f}_n)$. More precisely, we consider the ℓ_0-minimizer $\hat{f}_{\lambda_n}$:

$$\|\Phi\hat{f}_{\lambda_n} - Y\|_n^2 + \lambda_n \sharp\mathcal{J}(f_{\lambda_n}) \leq \min_{f\in\mathbf{F}_\infty} \|\Phi f - Y\|_n^2 + \lambda_n \sharp\mathcal{J}(f) + o_P(n^{-1}) \qquad (3.6)$$

where λ_n is some smoothing parameter converging to zero and $\sharp\mathcal{J}(f)$ is assumed to be nonzero. Otherwise take $\sharp\mathcal{J}(f) + 1$ instead for technical reasons. In the following result we show that for a large range of parameters $(\lambda_n)_{n\in\mathbb{N}}$, the correct number of change points is estimated with probability tending to one. That means, for large enough n, the estimators $\hat{f}_n$ in (3.1) and $\hat{f}_{\lambda_n}$ in (3.6) coincide.

Theorem 3.5. *Suppose that Assumptions **A2**, **B** and **C** are satisfied. Let $f_0 \in \boldsymbol{F}_\infty$ and choose $\{\lambda_n\}_{n\in\mathbb{N}}$ such that*

$$\lambda_n \longrightarrow 0 \quad \textit{and} \quad \lambda_n n^{\frac{1}{1+\epsilon}} \longrightarrow \infty,$$

for some $\epsilon > 0$. Then, the minimizer $\hat{f}_{\lambda_n}$ of (3.6) satisfies

$$P\left(\sharp\mathcal{J}(\hat{f}_{\lambda_n}) = \sharp\mathcal{J}(f_0)\right) \longrightarrow 1.$$

Thel last theorem can be viewed as a model consistency result, i.e. eventually for n large enough the correct number of junps/kinks is selected. In this sense, the resulting nonlinear least squares estimator is a post model selection estimator. It is well known that the model selection step may affect the distributional limit from the post model selection estimator in general and in accordance to this in Theorem 3.1 (i) the the normal approximation can become unreliable (Leeb and Pötscher (2006). In fact, the convergence in 3.5 is nonuniform in the sense that this probabilty will depend on the true underlying function f.

We did not succeed to prove or disprove whether $\lambda_n \sim \log n/n$ would give model consistency as well, the penailzation rate required in Theorem 3.5 is stronger. The choice of $\lambda_n \sim \log n/n$ would correspond to the classical BIC criterion. The practical choice of λ_n in the last theorem is a subtle task and we will not address this here in detail. In general, (generalized) cross validation methods could be employed (see e.g. Mao and Zhao (2003) in the context of splines) or residual based multiresolution techniques following the lines in Boysen, Kempe, Liebscher, Munk, and Wittich (2009b). We finally mention, that in general a severe computational burden is given in models with many kinks because the computation of the estimator $\hat{f}_n$ often leads to a difficult nonlinear optimization problem, a well known problem in nonlinear regresison. We will not pursue this issue further.

3.3 Examples

Example 3.6 (Hammerstein integral equations)**.** The structure of the function set $\boldsymbol{F}_k[a,b]$ allows extension of the results in Theorem 3.1 and Corollaries 3.2 and 3.4 to a prominent class of nonlinear integral operators of the form

$$Hf(x) = \int_a^b \varphi(x,y)\mathrm{L}(f(y),y)\mathrm{d}y\,, \tag{3.7}$$

which are known as *Hammerstein integral operators*. To be more precise, we assume that L has the following properties:

1.) L is continuously differentiable with respect to the first variable and continuous with respect to the second variable.

2.) The following operator $\mathcal{L} : L^2([a,b]) \to L^2([a,b])$ is injective:

$$(\mathcal{L}f)(y) := L(f(y), y), \qquad y \in [a,b]$$

3.) For any $f \in C([a,b])$ the derivative $\mathcal{L}'[f] : L^2([a,b]) \to L^2([a,b])$ is injective, i.e. the function $\frac{\partial L}{\partial f}(f, \cdot) \in C([a,b])$ does not vanish on any open subsets.

For a specific application from rheology we refer to Subsection 5.3.

It is straightforward to verify that if $\mathcal{L}$ satisfies conditions 1.) - 3.) and if $\mathcal{F}$ is a continuous parametric family in the sense of Definition 2.1, then the image $\mathcal{L}(\mathcal{F})$ is a continuous parametric family with $\mathfrak{f}$ replaced by $\mathfrak{f}_L(y, \vartheta) := L(\mathfrak{f}(y, \vartheta), y)$. Moreover, $\mathcal{L}(\boldsymbol{F}_k[a,b])$ is again a set of pc-functions in the sense of Definition 2.4. That means $\mathcal{L}$ preserves the properties of $f \in \boldsymbol{F}_k[a,b]$ and all results from the preceding section also hold for *Hammerstein integral equations* of the first kind. This is due to the fact that for $f \in \boldsymbol{F}_k[a,b]$ we can consider $Hf = \Phi\tilde{f}$ as a linear operator, where $\tilde{f}$ is an element of the transformed function space $\mathcal{L}(\boldsymbol{F}_k[a,b])$. Since estimating a function $\tilde{f}(\cdot, \theta_0) \in \mathcal{L}(\boldsymbol{F}_k[a,b])$, under the conditions of Theorem 3.1, or Corollary 3.2 or 3.4 respectively, yields an estimator for θ_0, we obtain an estimator for $f(\cdot, \theta_0)$ simultaneously.

Example 3.7 (Free knot splines)**.** Misspecification of the model, i.e. the question, what happens if the true function f in (1.1) is not an element of $\mathbf{F}_k$, has been treated already in Boysen, Bruns, and Munk (2009a, Lemma 3.3). In analogy to this lemma, under certain conditions on the design, the minimizer of (3.1) converges to a pc function $\bar{f} \in \mathbf{F}_k$ such that $\Phi\bar{f}$ is the best approximation of Φf.

If $\mathbf{F}_k$ is the set of piecewise polynomial functions, this offers an interesting connection to distributional asymptotics for splines. According to the *Curry and Schoenberg Theorem* (cf. De Boor (2001, Chapter VIII, (44))), for fixed change points, we have that the set of piecewise polynomials of degree p is equal to the B-spline space of order p with knots in $\{\tau_0, ..., \tau_{k+1}\}$ with multiplicity p, in the case of jumps and at most $p-1$ in the case of kinks. Thus, in this case, misspecification of the model, could be considered as spline approximation of f_0 and thus leads to the well known "spline-regularization". Results concerning spline-regularization with *fixed* knots and its relationship to inverse problems as in (1.1) is a classical topic and can be found e.g. in Cardot (2002). Here, we

actually have to deal with *free-knot* splines, i.e. the knots are free parameters and not known in advance. This is a serious difference to fixed knot splines, such as B-splines, starting with the fact that those spaces are no longer linear. It has long been known that approximation of a function by splines improves dramatically if the knots are free (Rice (1969), Burchard (1973/74)), although stable and effective computation of optimal knots in general is a challenging task (see e.g. Jupp (1978)). For a general discussion on free knot spline spaces in the context of approximation theory, we refer to De Boor (2001, Chapter XII §4). In the context of regression the optimal knot number and the optimal density for the knot distribution minimizing the asymptotic IMSE has been characterized by Agarwal and Studden (1980). Note that our results do not only yield an asymptotic expression for the variance of the estimated parameters including knot locations (which yields the MSE and can be optimized following the lines of Agarwal and Studden (1980)), but also show that they are asymptotically multivariate normally distributed, which can be used for cofidence bands (see also Mao and Zhao (2003)). Finally, Theorem 3.5 gives model selection consistency of knot penalisation in $\mathbf{F}_\infty$. This has never been shown before to our knowledge, e.g. in Mao and Zhao (2003) a GCV criterion is suggested for selection of λ_n without giving a proof of model selection consistency. In other words, from Theorem 3.5 it follows that for a large range of regularization parameters λ_n (which should converge to zero slower than $O(n^{-1})$) penalization with the number of knots picks asymptotically the right number of knots eventually in the entire set of free knot splines, i.e. the space $\mathbf{F}_\infty$.

Example 3.8. (Confidence bands) Statement (i) in Theorem 3.1 implies that the quadratic form

$$n\sigma^{-2}(\hat{\theta}_n - \theta_0)V_{\theta_0}(\hat{\theta}_n - \theta_0)^\top$$

is asymptotically distributed according to a χ^2-distribution with d degrees of freedom. This is still true if σ and V_{θ_0} are replaced by consistent estimators $\hat{\sigma}_n$ and $V_{\hat{\theta}_n}$, respectively. Hence we are now able to determine a $(1-\alpha)$-confidence ellipsoid for $\hat{\theta}_n$ in $\mathbb{R}^d$ by

$$n(\hat{\sigma}_n)^{-2}(\hat{\theta}_n - \theta_0)(V_{\hat{\theta}_n})(\hat{\theta}_n - \theta_0)^\top \leq \chi^2_d(1-\alpha). \tag{3.8}$$

Here $\chi^2_d(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with d degrees

of freedom. By maximizing and minimizing $f(y,\theta)$ for θ inside this confidence ellipsoid, we obtain simultaneous confidence bands for $\hat{f}_n$. Of course, any of the common methods for approximate confidence sets, namely Bonferroni, Scheffée or studentized maximum modulus statistics (for details see e.g. Miller (1966)) can be applied as well. In fact, some simulation studies show (not presented) that for functions with discontinuities including jump functions as treated in this paper, the studentized statistic is the least conservative of them, even for a small number of parameters as long as these are less than the number of observations. Moreover, if we consider the surface area of the respective bands as a further criterion, simulations show that for increasing number of parameters the bands corresponding to the studentized statistic outperform in terms of smaller surface area even the exact bands obtained from the elliptic confidence set. Therefore, we confine ourselves in Section 5.2 to the maximum modulus statistics. Note, that this extends the pointwise confidence intervals for free knot splines constructed in Mao and Zhao (2003) (see the previous example) in a simple way to bands.

4 Injectivity and mapping properties for some classes of integral operators

The following theorems give conditions on two classes of kernels φ, namely product and convolution kernels, that assure L^2 injectivity and range inclusions for the corresponding linear integral operator Φ in (1.2) as required in Assumption **C**.

We start with a theorem, which establishes a connection between injectivity of an integral operator with product kernel $\varphi(x,y) = \phi(xy)$ and the expansion of ϕ. The main argument in the proof is given by the *Full Müntz Theorem* proven by Borwein & Erdélyi Borwein and Erdélyi (1997, Thm 6.2):

Lemma 4.1 (Full Müntz-Theorem)**.** *Suppose that $J \subset \mathbb{N}$ and that $0 < a < b$. Then, span$(\{y^j : j \in J\})$, is dense in $C([a,b])$ with respect to the maximum norm if and only if*

$$\sum_{j\in J} j^{-1} = \infty. \tag{4.1}$$

Theorem 4.2 (product kernels)**.** *Assume that $0 < a < b$ and $0 \le c < d$ and that $\varphi(x,y) = \phi(xy)$ for some piecewise continuous function $\phi \in L^\infty([ac,bd])$. Then part i) and ii) of Assumption **C** are satisfied under the following conditions:*

***C**i): We have $\Phi \in \mathcal{L}\left(L^1([a,b]), L^\infty([c,d])\right)$. Moreover, $\Phi \in \mathcal{L}\left(L^\infty([a,b]), C^{0,1}([c,d])\right)$ if $\phi \in BV([ac,bd])$, the space of functions of bounded variation on $[ac,bd]$.*

***C**ii): Suppose there exists an interval $[\rho_1,\rho_2] \subset [ac,bd]$ with $\frac{\rho_1}{a} < \frac{\rho_2}{b}$, such that ϕ has an absolutely convergent expansion*

$$\phi(z) = \sum_{j=0}^{\infty} \alpha_j z^j \quad \text{with } \alpha_j \in \mathbb{R} \quad \text{for all } j \in \mathbb{N},\ z \in [\rho_1,\rho_2] \tag{4.2}$$

and the set $J := \{j \in \mathbb{N} : \alpha_j \neq 0\}$ satisfies the Müntz-condition (4.1). Then $\Phi : \mathcal{B}([a,b]) \longrightarrow L^2([a,b])$ is injective on the space $\mathcal{B}([a,b])$ of signed Borel measures on $[a,b]$. If $\rho_1 = ac$ and $\rho_2 = bd$, then (4.1) is also necessary for injectivity of Φ on $\mathcal{B}([a,b])$.

One example of such a kernel occurs in the example from rheology, which will be discussed in detail in Section 5.2. The Gaussian kernel $\phi(x) = (2\pi\sigma^2)^{-1/2} e^{-(x/\sigma)^2/2}$, mentioned above, is another well known example for a function satisfying the assumptions of Theorem 4.2.

Theorem 4.3 (positive definite convolution kernels)**.** *Assume that $\varphi(x,y) = \phi(x-y)$ for all $x,y \in [a,b]$ for some function $\phi \in C(\mathbb{R}) \cap L^1(\mathbb{R})$. Then part i) and ii) of Assumption **C** are satisfied under the following conditions:*

***C**i): If $\phi \in BV([a-b,b-a])$, then $\Phi \in \mathcal{L}\left(L^\infty([a,b]), C^{0,1}([a,b])\right)$ and $\Phi \in \mathcal{L}\left(L^1([a,b]), L^\infty([a,b])\right)$.*

***C**ii): If the Fourier transform $\widehat{\phi}$ is integrable and strictly positive a.e. on $\mathbb{R}$, then $\Phi : \mathcal{M}([a,b]) \to L^2([a,b])$ is injective.*

Examples of kernels satisfying the assumptions of Theorem 4.3 include the Laplace kernel $\phi(x) = \frac{1}{2} e^{-|x|}$ and kernels of the type $\phi(x) = \max(1-|x|,0))^p$ for $p = 2,3,\ldots$.

Theorem 4.4 (analytic convolution kernels)**.** *Assume that $\varphi(x,y) = \phi(x-y)$ for all $x,y \in [a,b]$ for some analytic function $\phi \in L^2(\mathbb{R})$ and that the Fourier*

transform $\widehat{\phi}$ *vanishes at most on a set of Lebesgue measure* 0. *Then the operator* Φ *satisfies Assumption* ***C****.*

5 Simulations and Data Example

5.1 Example: Inverse two phase regression

In order to evaluate the speed of convergence and quality of the approximation by the asymptotic law given in Theorem 3.1, we perform a simulation study for the case where the true function is a step function with one jump, that is $f_0 \in \tilde{\mathbf{T}}_1$ (cf. Example 2.2 and (2.3)), given by the parameter vector $\theta_0 = (b_1, \tau, b_2) = (-3, \frac{1}{2}, 3)$. Therefore we generate the observations Y by

$$Y_i = \Phi\left(-3 \cdot \mathbf{1}_{[0,\frac{1}{2})} + 3 \cdot \mathbf{1}_{[\frac{1}{2},1]}\left(\tfrac{i}{n}\right)\right) + \frac{1}{2}\varepsilon_i, \quad i = 1, ..., n, \tag{5.1}$$

where $(\Phi f)(x) = \int_0^1 \mathbf{1}_{[0,\infty]}(x-y) f(y)\mathrm{d}y$ and $\varepsilon \sim N(0,1)$ for $i = 1, ..., n$. Theorem 3.1 yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{\theta_0}^{-1})$$

where the covariance in (3.3) in calculated as

$$\sigma^{-2} V_{\theta_0} = \begin{pmatrix} \frac{12}{\tau^3} & \frac{-6}{(b_1-b_2)\tau^2} & 0 \\ \frac{-6}{(b_1-b_2)\tau^2} & \frac{4}{(b_1-b_2)^2(1-\tau)\tau} & \frac{-6}{(b_1-b_2)(1-\tau)^2} \\ 0 & \frac{-6}{(b_1-b_2)(1-\tau)^2} & \frac{12}{(1-\tau)^3} \end{pmatrix},$$

with $\det V_{\theta_0} = 16 > 0$, which implies existence of the inverse $V_{\theta_0}^{-1}$. In particular for the jump location we obtain

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{\mathcal{D}} N\left(0, \frac{4\sigma^2}{(b_1 - b_2)^2(1-\tau)\tau}\right).$$

This has been used to calculate confidence intervals for $\hat{\tau}$. Figure 5.1 shows the empirical and the asymptotic distribution of $\hat{\tau}$ for different sample sizes n.

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 5.2. As described in Subsection 3.8 we can also calculate confidence bands for the estimated function $\hat{f}_n$ as well as for its image $\Phi\hat{f}_n$. Figure 5.3 shows two simulated data sets, including their 95%-confidence regions for $n = 100$ and $n = 1000$.

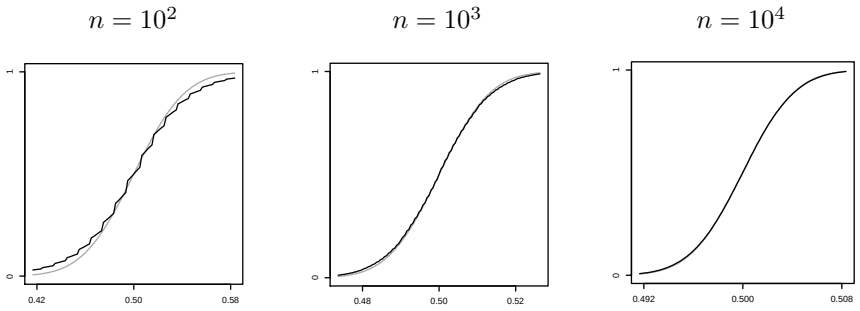

Figure 5.1: Asymptotic and finite sample size distribution of the jump location for different sample sizes n. 10^5 simulation runs with data generated according to (5.1) were performed. The finite sample size distribution is given by the black line and the asymptotic distribution by the gray line.

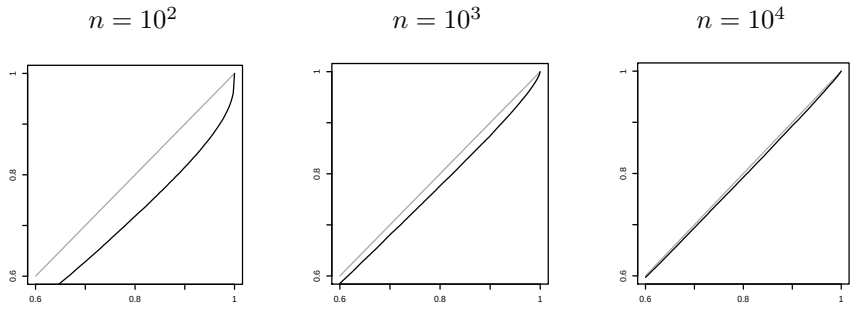

Figure 5.2: Empirical coverage probability for different sample sizes n of confidence bands for the estimated jump location. 10^5 simulation runs with data generated according to (5.1) were performed. The x-axis shows the nominal and the y-axis the empirical coverage.

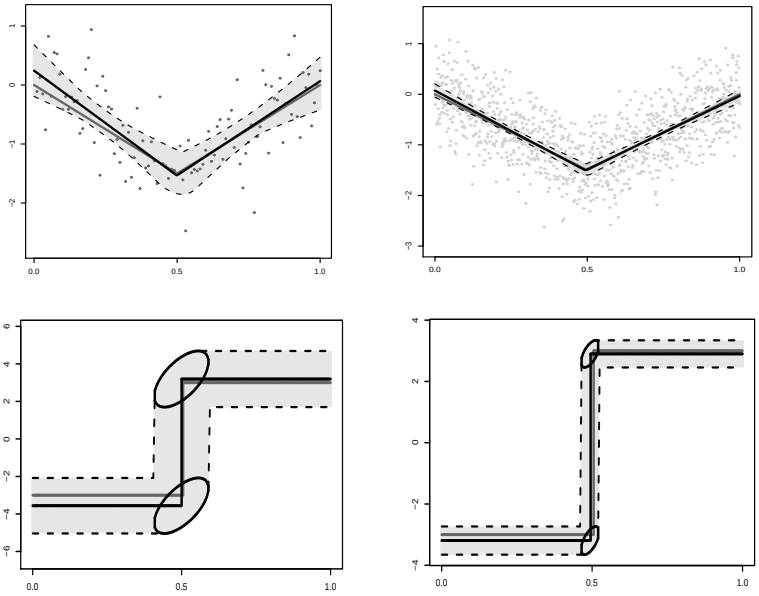

Figure 5.3: Simulated data examples and confidence bands for the two phase regression with $n = 100$ (left) and $n = 1000$ (right) observations. The first row displays the observations and the reconstruction in the image space, and the second row shows the estimate for the signal f. The gray line represents the true function and the solid black line the estimate. The dashed lines show the confidence bands for the function and the gray dots the observations. The ellipses in the second row show the confidence sets for (τ, b_1) and (τ, b_2), respectively.

5.2 Example: Inverse multiphase regression

In this subsection we are going to discuss an application of Corollary 3.4 to a problem from rheology. The aim here is the determination of the so called *relaxation time spectrum* (see Roths, Maier, Friedrich, Marth, and Honerkamp (2000)). The relaxation time spectrum is a characteristic quantity used in rheology which describes the viscoelastic properties of polymer solutions and polymer melts. Given this spectrum, it is very easy to convert one material function into another one. Additionally, many theories are based on the spectrum or provide predictions about its character (see for example Ferry (1970)). The relaxation time spectrum is not directly accessible by experiment and has to be inferred from dynamic stress moudli. It is common to assume that these are observed (with gaussian noise) under a nonlinear integral transform as follows (see Roths, Maier, Friedrich, Marth, and Honerkamp (2000)).

Definition 5.1. *Let* $0 < a < 1 < b < \infty$ *and* $c \neq 0$. *The* **relaxation time spectrum transform** *is given as*

$$\begin{aligned} H : L^\infty([a,b]) &\longrightarrow L^2([a,b]) \\ f &\longmapsto Hf(x) := \int_a^b \frac{x^2 y}{1+x^2y^2} e^{cf(y)} \mathrm{d}y. \end{aligned}$$

Note that this is a Hammerstein integral $H = \Phi \circ \mathcal{L}$, where $\mathcal{L} : L^\infty([a,b]) \to L^2([a,b])$ and $\Phi : L^2([a,b]) \to L^2([a,b])$ are defined by

$$\begin{aligned} (\mathcal{L}f)(y) &:= y^{-1} e^{cf(y)} \\ (\Phi g)(y) &:= \int_a^b \frac{x^2y^2}{1+x^2y^2} g(y) \mathrm{d}y. \end{aligned}$$

Note that the exponential operator $\mathcal{L}$ satisfies the assumptions claimed in Example 3.6. Furthermore, the integral operator Φ satisfies Assumption $\boldsymbol{C}$ by virtue of Theorem 4.2.

The function f describing the relaxation time spectrum is known to have the interpretation $f(\cdot,\theta) = \tilde{f}(\log(\cdot),\theta)$ such that $\tilde{f}(\cdot,\theta)$ is continuous and piecewise linear with two kinks (see Prince (1953)). This means that $\tilde{f}$ is an element of $\tilde{\mathbf{L}}_2$ as defined in (2.4) with reduced parameter vector $\tilde{\theta} = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_2^2, \tau_2, \vartheta_2^3)$ (cf. Definition 2.6). For simplicity we rename $\tilde{\theta}$ as $\theta = (b_0, b_1, \tau_1, b_2, \tau_2, b_3)$. Then we

have

$$\tilde{\mathbf{L}}_2 = \{\tilde{f} \in L^2([\log(a), \log(b)]) \mid \tilde{f}(y,\theta) = b_0 + b_1 y + b_2(y-\tau_1)_+ + b_3(y-\tau_2)_+,\ \theta \in \Theta_2\},$$

where Θ_2 is assumed to be compact. Then, the true function $f_0(y) = f(y, \theta_0)$, we intend to estimate, is an element of the set

$$\mathbf{L}_{\log} := \{f(y,\theta) = \tilde{f}(\log(y), \theta) \mid \tilde{f} \in \tilde{\mathbf{L}}_2\},$$

which obviously satisfies the conditions of Definition 2.4. In Roths, Maier, Friedrich, Marth, and Honerkamp (2000) it is assumed that the observation model coincides with (1.1) with f_0 substituted by $\mathcal{L}f_0$, namely

$$y_i = Hf(x_i, \theta_0) + \varepsilon = \Phi\mathcal{L}f(x_i, \theta_0) + \varepsilon_i \text{ for } i = 1, ..., n,$$

where Assumptions ***A1*** and **B** on error and design are fulfilled. Figure 5.4 shows a sample of stress moduli measurements on a log-scale performed at the Center of Material Sciences at Freiburg for a certain polymer melt (see Roths, Maier, Friedrich, Marth, and Honerkamp (2000) for details). For estimation we use the least squares estimator defined in (3.1). Then, application of Corollary 3.4 yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 V_{\theta_0}^{-1}), \tag{5.2}$$

where $\sigma^2 = \mathbf{E}(\varepsilon^2)$ if V_{θ_0} is regular. By the chain rule the derivative of the mapping $\Lambda : \mathbb{R}^6 \to L^2([a,b])$, $\Lambda\theta := \Phi\mathcal{L}f(\cdot, \theta_0)$ is given by

$$(\Lambda'[\theta_0]h)(x) = \Phi\left(\frac{\partial}{\partial\theta}\left[\mathcal{L}f(\cdot,\theta_0)\right]h\right)(x) = c\int_a^b \frac{x^2 y}{1+x^2y^2} e^{cf(y,\theta_0)}\left(h^\top df(y,\theta_0)\right)\mathrm{d}y, \tag{5.3}$$

where

$$df(y,\theta) = \begin{pmatrix} 1 \\ \log(y) \\ -b_2\mathbf{1}_{[e^{\tau_1},b]} \\ (\log(y)-\tau_1)\mathbf{1}_{[e^{\tau_1},b]} \\ -b_3\mathbf{1}_{[e^{\tau_2},b]} \\ (\log(y)-\tau_2)\mathbf{1}_{[e^{\tau_2},b]} \end{pmatrix}.$$

Remembering that $b_2 \neq 0 \neq b_3$ and $\tau_1 < \tau_2$ by Definition 2.4, it is easy to see that the components of $df(\cdot, \theta_0)$ are linearly independent. Together with the

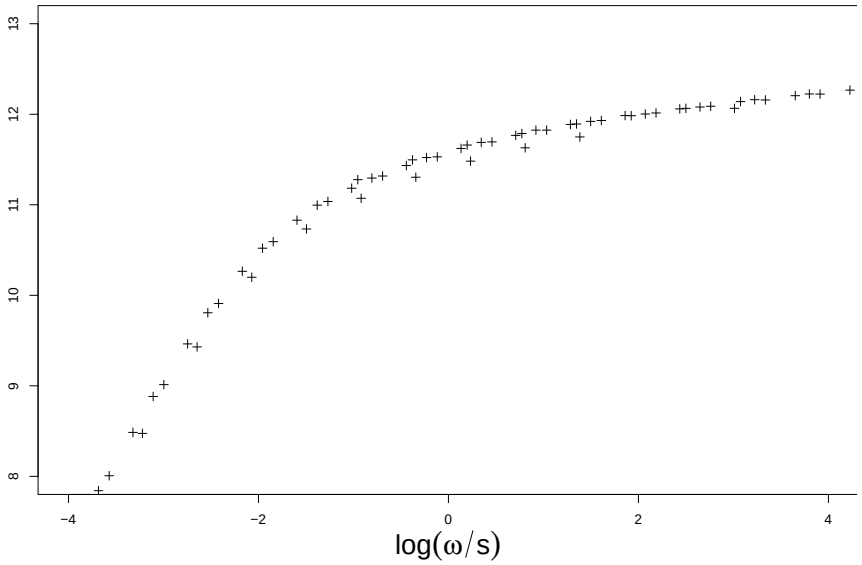

Figure 5.4: A $\log - \log$- plot of the ω frequency of harmonic stress (x-axis) against the dynamic stress moduli of a polymer melt.

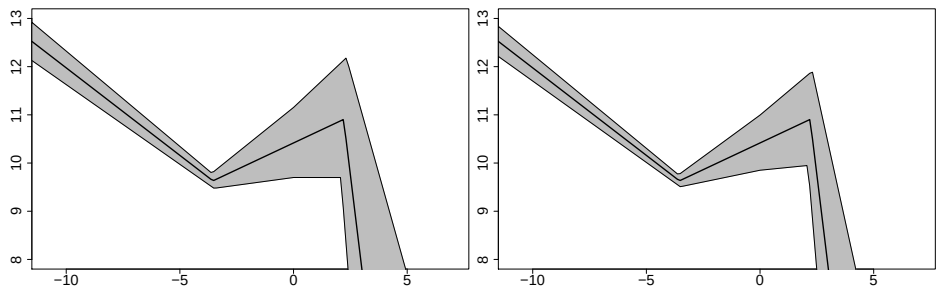

Figure 5.5: From l.t.r.: 0.95- and 0.80-confidence bands for the estimated kink function $\hat{f}_n$ of the log relaxation time spectrum with two (typical) kinks plotted on a log scale

injectivity of Φ, it follows that $\Lambda'[\theta_0]$ is injective, and hence $V_{\theta_0} \in \mathbb{R}^{6\times 6}$ defined as in (3.3) is nonsingular.

Moreover, the results of Theorem 3.1 hold with the improved rate

$$\|f_0 - \hat{f}_n\|_{L^2[a,b]} = O_P(n^{-\frac{1}{2}}), \tag{5.4}$$

which for comparison is the square of the rate in (iv) in Theorem 3.1. Note that Equation (5.4) directly follows from Corollary 3.4, since linear functions with bounded slopes i.e. functions in $\mathcal{F}_L$ (c.f. (2.3)), are Lipschitz continuous, with uniform Lipschitz constant. Hence, for the modulus of continuity it holds that $\nu(\mathcal{F}_L, n^{-1/2}) = O(n^{-1/2})$.

Figure 5.5 shows the estimated kink function for the polymer melt data of the relaxation time spectrum from dynamic moduli (see Roths, Maier, Friedrich, Marth, and Honerkamp (2000)) with 95%- and 80%-confidence bands, calculated by using a studentized maximum modulus statistic as discussed in Subsection 3.8 (for details see also pp.70 in Miller (1966)). Finally, as in Subsection 5.1, we evaluated the accuracy of the normal approximation from (5.2) in this special example, by performing a simulation study (see Figure 5.6). Here we used the operator in (5.1) acting on the space of kink functions with two kinks. A comparison of Figure 5.2 and 5.6 illustrates that increasing complexity of the kernel in Subsection 5.2 reduces the finite sample accuracy of the empirical coverage probability.

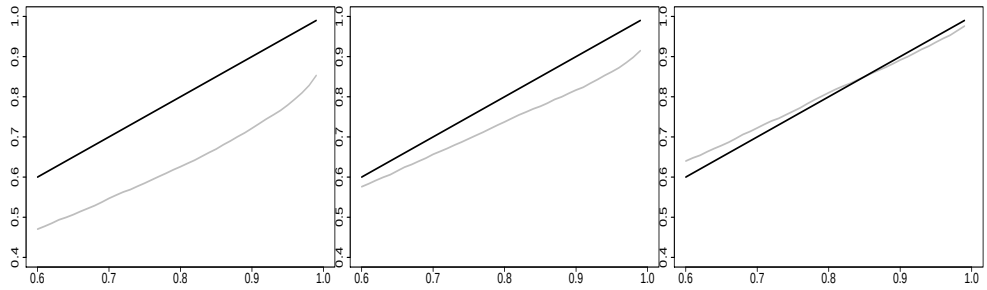

Figure 5.6: Empirical coverage probability (grey lines) of confidence bands for the estimated kink function for normal observations with $\sigma^2 = 0.01$ for different sample sizes. From l.t.r.: $n = 100, 1000, 5000$, and 10^4 simulations each. The x-axis shows the nominal and the y-axis the empirical coverage probability. The black line $x = y$ is for comparison, it shows perfect coincidence of empirical and nominal coverage.

Acknowledgment

Part of this paper is content of the PhD thesis of S. Frick, who acknowledges support of DFG, RTN 1023. T. Hohage and A. Munk acknowledge support of DFG FOR 916 and CRC803. We thank N. Bissantz for providing us the data of Example 5.3. We gratefully acknowledge helpful comments of L. Dümbgen, K. Frick and J. Schmidt-Hieber.

References

Agarwal, G. G. and Studden, W. J. (1980), "Asymptotic integrated mean square error using least squares and bias minimizing splines," *Ann. Statist.*, 8, 1307–1325.

Borwein, P. and Erdélyi, T. (1995), *Polynomials and polynomial inequalities*, vol. 161 of *Graduate Texts in Mathematics*, New York: Springer-Verlag.

— (1997), "Generalizations of Müntz's theorem via a Remez-type inequality for Müntz spaces," *J. Amer. Math. Soc.*, 10, 327–349.

Boysen, L., Bruns, S., and Munk, A. (2009a), "Jump estimation in inverse regression," *Electron. J. Statist.*, 3, 1322–1359.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009b), "Consistencies and rates of convergence of jump-penalized least squares estimators," *Ann. Statist.*, 37, 157–183.

Burchard, H. G. (1973/74), "Splines (with optimal knots) are better," *Applicable Anal.*, 3, 309–319.

Cardot, H. (2002), "Spatially adaptive splines for statistical linear inverse problems," *J. Multivariate Anal.*, 81, 100–119.

De Boor, C. (2001), *A practical guide to splines*, vol. 27 of *Applied Mathematical Sciences*, New York: Springer-Verlag, revised ed.

Ferry, J. D. (1970), *Viscoelastic properties of polymers [by] John D. Ferry*, Wiley New York,, 2nd ed.

Goldenshluger, A., Juditsky, A., Tsybakov, A., and Zeevi, A. (2008a), "Change-point estimation from indirect observations. II. Adaptation," *Ann. Inst. Henri Poincaré Probab. Stat.*, 44, 819–836.

Goldenshluger, A., Juditsky, A., Tsybakov, A. B., and Zeevi, A. (2008b), "Change-point estimation from indirect observations. I. Minimax complexity," *Ann. Inst. Henri Poincaré Probab. Stat.*, 44, 787–818.

Goldenshluger, A., Tsybakov, A., and Zeevi, A. (2006), "Optimal Change-Point Estimation from Indirect Observations," *Ann. Statist.*, 34, 350–372.

Hammerstein, A. (1930), "Nichtlineare Integralgleichungen nebst Anwendungen," *Acta Math.*, 54, 117–176.

Jupp, D. L. B. (1978), "Approximation to data by splines with free knots," *SIAM J. Numer. Anal.*, 15, 328–343.

Korostelev, A. P. (1987), "Minimax estimation of a discontinuous signal," *Teoryia Veroyatnostei iee Primeneniya*, 32, 796–799.

Leeb, H. and Pötscher, B. M. (2006), "Can one estimate the conditional distribution of post-model-selection estimators?" *Ann. Stat.*, 34, 2554–2591.

Mao, W. and Zhao, L. (2003), "Free-knot polynomial splines with confidence intervals," *J. R. Statist. Soc. Ser. B*, 65, 901–919.

Miller, Jr., R. G. (1966), *Simultaneous statistical inference*, New York: McGraw-Hill Book Co.

Neumann, M. H. (1997), "Optimal change-point estimation in inverse problems," *Scand. J. Statist.*, 24, 503–521.

Prince, E. and. Rouse, A. (1953), "A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers," *The Journal of Chemical Physics*, 21, 1272–1280.

Raimondo, M. (1998), "Minimax estimation of sharp change points," *Ann. Statist.*, 26, 13979–1397.

Rice, J. R. (1969), *The approximation of functions. Vol. 2: Nonlinear and multivariate theory*, Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

Roths, T., Maier, D., Friedrich, C., Marth, M., and Honerkamp, J. (2000), "Determination of the relaxation time spectrum from dynamic moduli using an edge preserving regularization method," *Rheologica Acta*, 39, 163–173.

Tsybakov, A. (2009), *Introduction to nonparametric estimation*, Springer Series in Statistics, New York, Berlin, Tokyo: Springer Publishing Company Incorporated.

van der Vaart, A. W. (1998), *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press.

Wishart, J. (2010), "Kink estimation in stochatsic regression with dependent errors and predictors," *Electron. J. Statist.*, 4, 875–913.

— (2011), "Minimax lower bound for kink location estimators in a nonparametric regression model with long range dependence," *Statist. Prob. Letters*, 81, 1871–1875.

Institut für Mathematische Stochastik, Universität Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

E-mail: sbruns@math.uni-goettingen.de

Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestrasse 16-18, 37083 Göttingen, Germany

E-mail: hohage@math.uni-goettingen.de

Institut für Mathematische Stochastik, Universität Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany

E-mail: munk@math.uni-goettingen.de

Statistica Sinica: Supplement

ASYMPTOTIC LAWS FOR CHANGE POINT ESTIMATION IN INVERSE REGRESSION

Sophie Frick[1], Thorsten Hohage[2], Axel Munk[3]

[1] *Institute for Mathematical Stochastics,*
[2] *Institute for Numerical and Applied Mathematics,*
Georgia Augusta Universität Göttingen

Supplementary Material

In this supplement the proofs of the results of the paper are collected.

S1 Proofs of Theorem 3.1, Proposition 3.3, and Corollaries 3.2 and 3.4

This section is dedicated to the proof of Theorem 3.1 and Corollaries 3.2 and 3.4 in Section 3. The proof is separated in four parts. We start with some technical results, then we calculate the entropy numbers of the considered function spaces, which yield the basic arguments to show consistency of the estimator in (3.1). Finally, we give the proof of the asymptotic results given in Theorem 3.1 and Corollaries 3.2 and 3.4. This section is dedicated to the proof of Theorem 3.1 and Corollaries 3.2 and 3.4 in Section 3. The proof is separated in four parts. We start with some technical results, then we calculate the entropy numbers of the considered function spaces, which yield the basic arguments to show consistency of the estimator in (3.1). Finally, we give the proof of the asymptotic results given in Theorem 3.1 and Corollaries 3.2 and 3.4.

S1.1 Some technical lemmata

Before we give the proofs of the main results in Section 3, we need some technical lemmata.

Lemma S1.1. *For $p \in [1, \infty)$ the mapping $\mathfrak{F} : (\Theta_k, |\cdot|_\infty) \to (\mathbf{F}_k, \|\cdot\|_{L^p})$, $\theta \mapsto f(\cdot, \theta)$ is Hölder continuous with index $1/p$, and there exists a constant L independent of k such*

that for all $\theta^{(1)}, \theta^{(2)} \in \Theta$ the following estimate holds true:

$$\left\|\mathfrak{F}(\theta^{(2)}) - \mathfrak{F}(\theta^{(1)})\right\|_{L^p} \le \left(L(k+1)\left|\theta^{(2)} - \theta^{(1)}\right|_\infty\right)^{1/p}.$$

Proof. Let $\theta^{(1)} = (\vartheta_1^{(1)}, \tau_1^{(1)}, \dots, \vartheta_k^{(1)}), \theta^{(2)} = (\vartheta_1^{(2)}, \tau_1^{(2)}, \dots, \vartheta_k^{(2)}) \in \Theta_k$ and consider the mapping $\lambda : [0,1] \to L^1([a,b])$,

$$\lambda(t) := f(\cdot, \theta^{(1)} + t(\theta^{(2)} - \theta^{(1)})).$$

Setting $\vartheta_i(t) := \vartheta_i^{(1)} + t(\vartheta_i^{(2)} - \vartheta_i^{(1)})$ for $i = 1, \dots, k+1$ and and $\tau_i(t) := \tau_i^{(1)} + t(\tau_1^{(2)} - \tau_i^{(1)})$ for $i = 0, \dots, k+1$ we get from the integral form of the mean value theorem that

$$\begin{aligned}\left\|\mathfrak{F}(\theta^{(2)}) - \mathfrak{F}(\theta^{(1)})\right\|_{L^p}^p = \|\lambda(1) - \lambda(0)\|_{L^p}^p &\le \sum_{i=1}^{k+1} \int_{\tau_{i-1}(t)}^{\tau_i(t)} \int_0^1 \left|\left(\vartheta_i^{(2)} - \vartheta_i^{(1)}\right)^\top \frac{\partial \mathfrak{f}}{\partial \vartheta}(y, \vartheta_i(t))\right|^p \mathrm{d}t\mathrm{d}y \\ &\quad + \sum_{i=1}^{k} \int_0^1 \left|[f(\tau_i(t), \theta(t))]\right|^p dt \left|\tau_i^{(2)} - \tau_i^{(1)}\right| \\ &\le (1+k)\, C \left|\theta^{(2)} - \theta^{(1)}\right|_\infty\end{aligned}$$

with the constant $C := \sup_{\vartheta \in \Psi, y \in [a,b]} \max((2|\mathfrak{f}(y,\vartheta)|)^p, (b-a)\mathrm{diam}_\infty(\Psi)^{p-1} |\frac{\partial \mathfrak{f}}{\partial \vartheta}(y,\vartheta)|_1^p)$, which is finite since Ψ is compact. □

Lemma S1.2. *Suppose that Assumption **C** holds true. Then $\Lambda : \Theta_k \to L^2(I)$ is continuously differentiable and the derivative is given by*

$$(\Lambda'[\theta]e_i)(x) = \begin{cases} \displaystyle\int_a^b \varphi(x,y) \tfrac{\partial}{\partial \theta_i} f(y,\theta) \mathrm{d}y & i \neq 0 \bmod (r+1), \\ \varphi(x, \tau_{\frac{i}{r+1}}) \left[f(\cdot, \theta)\right](\tau_{\frac{i}{r+1}}) & i = 0 \bmod (r+1). \end{cases} \tag{S1.1}$$

Proof. We show that the mapping $\Lambda_0 : \Psi \times [a,b]^2 \to L^2(I)$

$$\Lambda_0(\vartheta, \tau_1, \tau_2) := \int_{\tau_1}^{\tau_2} \varphi(\cdot), y) \mathfrak{f}(y, \vartheta)\, \mathrm{d}y$$

is continuously differentiable with derivative

$$\begin{aligned}\Lambda_0'[\vartheta, \tau_1, \tau_2](\delta\vartheta, \delta\tau_1, \delta\tau_2) = &\int_{\tau_1}^{\tau_2} \varphi(\cdot), y) \frac{\partial \mathfrak{f}}{\partial \vartheta}(y, \vartheta) \delta\vartheta \, \mathrm{d}y \\ &- \varphi(\cdot, \tau_1) \mathfrak{f}(\vartheta)(\tau_1, \vartheta)\, \delta\tau_1 + \varphi(\cdot, \tau_2) \mathfrak{f}(\vartheta)(\tau_2, \vartheta)\, \delta\tau_2\end{aligned} \tag{S1.2}$$

from which the assertion follows immediately. We write $\Lambda_0 = \Phi_0 \circ \mathfrak{F}$ as the composition of the mappinng $\mathfrak{F} : \Psi \times [a,b]^2 \to C([a,b]) \times [a,b]^2$, $\mathfrak{F}(\vartheta, \tau_1, \tau_2) := (\mathfrak{f}(\cdot, \vartheta), \tau_1, \tau_2)^\top$, which is continuously differentiable with derivate $\mathfrak{F}'[\vartheta, \tau_1, \tau_2](\delta\vartheta, \delta\tau_1, \delta\tau_2) = (\frac{\partial \mathfrak{f}}{\partial \vartheta}(\cdot, \vartheta)\delta\vartheta, \delta\tau_1, \delta\tau_2)^\top$ by the first property in Definition 2.1, and the integral operator $\Phi_0 : C([a,b]) \times [a,b]^2 \to$

$L^2(I)$, $(\Phi_0(g,\tau_1,\tau_2))(x) := \int_{\tau_1}^{\tau_2} \varphi(x,y)g(y)\,\mathrm{d}y$ which is continuously differentiable with derivative $\Phi_0'[g,\tau_1,\tau_2](\delta g,\delta\tau_1,\delta\tau_2) = \Phi_0(\delta g,\tau_1,\tau_2) - \varphi(\cdot,\tau_1)g(\tau_1)\delta\tau_1 + \varphi(\cdot,\tau_2)g(\tau_2)\delta\tau_2$ by the fundamental theorem of calculus and Assumption **C**. Now (S1.2) follows from the chain rule $\Lambda_0'[\bar\theta](\delta\bar\theta) = \Phi_0'[\mathfrak{F}(\bar\theta)]\mathfrak{F}'[\bar\theta](\delta\bar\theta)$ with $\bar\theta := (\vartheta,\tau_1,\tau_2)$. □

Corollary S1.3. *Suppose that Assumptions **B** and **C** are met. Then, uniformly for all $f \in F_k([a,b])$, it holds*

$$o_P(1) + s_l\|\Phi f\|_n^2 \le \|\Phi f\|^2_{L^2([a,b)]} \le s_u\|\Phi f\|_n^2 + o_P(1)$$

*with constants s_l, s_u depending on the design density (cf. Assumption **B**), only.*

Proof. The claim follows from Boysen, Bruns, and Munk (2009, Lemma 4.3) together with Assumption **C**i). □

S1.2 Entropy results

In order to show consistency of the least squares estimator $\hat f_n$ in (3.1), we apply uniform deviation inequalities from empirical process theory. To this end, it is necessary to calculate the *entropy* of the space of interest, which is defined in the following way.

Definition S1.4. *Given a subset $\mathcal{G}$ of a linear space G, a semi-norm $\|\cdot\| : G \to [0,\infty)$, and a real number $\delta > 0$, the δ-**covering number** $\boldsymbol{N}(\delta,\mathcal{G},\|\cdot\|)$ is defined as the smallest value of N such that there are functions $g_1,\dots,g_N$ with*

$$\min_{1\le j\le N}\|g - g_j\| \le \delta \quad \text{for all } g \in \mathcal{G}.$$

*Moreover, the δ-**entropy** $\boldsymbol{H}$ and the **entropy integral** $\boldsymbol{J}$ of $\mathcal{G}$ are defined as*

$$\boldsymbol{H}(\delta,\mathcal{G},\|\cdot\|) = \log \boldsymbol{N}(\delta,\mathcal{G},\|\cdot\|) \text{ and}$$

$$\boldsymbol{J}(\delta,\mathcal{G},\|\cdot\|) := \max\left(\delta, \int_0^\delta \boldsymbol{H}^{1/2}(u,\mathcal{G},\|\cdot\|)\mathrm{d}u\right),$$

respectively.

We are interested in the entropy of the set

$$\mathbf{G}_k := \{\Phi f \in L^2(I) \mid f \in \boldsymbol{F}_k[a,b]\}, \tag{S1.3}$$

where Φ is a known integral operator with kernel φ as defined in (1.2). In order to deduce consistency of $\hat f_n$, additionally we have to know the entropy of the set $\boldsymbol{F}_k$. By definition, all functions $f \in \boldsymbol{F}_k$ are determined by a parameter vector θ. Thus the core of the problem reduces to determination of the entropy of the parameter set Θ_k.

Lemma S1.5. *Let $\boldsymbol{F}_k$ and $d = (k+1)r + k$ be as in Definition 2.4. Then there exists a constant $T_{\mathcal{F}}, \tilde{T}_{\mathcal{F}} > 0$ depending only on the considered function class $\mathcal{F}$ in Definition 2.1, such that*

$$\boldsymbol{H}(\delta, \boldsymbol{F}_k, \|\cdot\|_{L^1}) \le d \log\left(\frac{(k+1)T_{\mathcal{F}} + \delta}{\delta}\right), \tag{S1.4}$$

$$\boldsymbol{H}(\delta, \boldsymbol{G}_k, \|\cdot\|_n) \le d \log\left(\frac{(k+1)\tilde{T}_{\mathcal{F}} + \delta}{\delta}\right). \tag{S1.5}$$

Proof. Note that the diameter of Θ_k with respect to the maximum norm is bounded by a constant M independent of k and recall that $\mathfrak{F} : (\Theta_k, |\cdot|_\infty) \to (\mathbf{F}_k, \|\cdot\|_{L^1})$, $\theta \mapsto f(\cdot, \theta)$ is Lipschitz continuous with constant $L(k+1)$ (cf. Lemma S1.1). Hence, (S1.4) with $T_{\mathcal{F}} := 2ML$ follows from the fact that the number of balls with radius $\delta/(L(k+1))$ which are needed to cover a subset of $\mathbb{R}^d$ with diameter bounded by M can be estimated by $(2ML(k+1)+\delta)^d/\delta^d$ (cf. del Barrio, Deheuvels, and van de Geer (2007, Lem. 2.5)). Analogously, we obtain (S1.5) with $\tilde{T}_{\mathcal{F}} := 2ML\|\Phi\|_{L^1 \to L^\infty}$. □

S1.3 Consistency

Theorem S1.6. *Let Φ be an operator satisfying Assumption **C** and $f_0 = f(\cdot, \theta_0) \in F_k$. Furthermore, assume that Assumption **A1** and **B** are met. Then, for $\hat{f}_n = f(\cdot, \hat{\theta}_n)$, the least squares estimator in (3.1), it holds that*

$$\|\Phi \hat{f}_n - \Phi f_0\|_n = o_P(1).$$

Proof. Due to Inequality (3.1) we have

$$\|\Phi \hat{f}_n - Y\|_n^2 \le \|\Phi f_0 - Y\|_n^2 + o_p(n^{-1}).$$

Inserting $Y = \Phi f_0 + \varepsilon$ leads to

$$\|\Phi \hat{f}_n - \Phi f_0\|_n^2 - 2\langle \Phi \hat{f}_n - \Phi f_0, \varepsilon\rangle_n + \|\varepsilon\|_n^2 \le \|\varepsilon\|_n^2 + o_p(n^{-1})$$

which implies

$$\begin{aligned} \|\Phi \hat{f}_n - \Phi f_0\|_n^2 &\le 2\langle \Phi \hat{f}_n - \Phi f_0, \varepsilon\rangle_n + o_p(n^{-1}) \\ &= 2(\langle \Phi \hat{f}_n, \varepsilon\rangle_n - \langle \Phi f_0, \varepsilon\rangle_n) + o_p(n^{-1}) \\ &\le 4 \sup_{g \in \mathbf{G}_k} |\langle g, \varepsilon\rangle_n| + o_p(n^{-1}). \end{aligned}$$

Lemma S1.5 gives boundedness of the entropy $\boldsymbol{H}(\delta, \mathbf{G}_k, P_n)$ uniformly in n, for all $\delta > 0$ and so $n^{-1}\boldsymbol{H}(\delta, \mathbf{G}_k, P_n) \to 0$ as $n \to \infty$. With this result it follows directly from van de Geer (2000, Theorem 4.8) that $\sup_{g \in \mathbf{G}_k} |\langle g, \varepsilon\rangle_n| = o_P(1)$. □

Corollary S1.7. *Under the assumptions of Theorem S1.6 one has*

$$\|\Phi\hat{f}_n - \Phi f_0\|_{L^2(I)} = o_P(1).$$

Proof. Since the design in Definition (1.1) is assumed to satisfy Assumption **B** the claim follows directly from Theorem S1.6 and Corollary S1.3. □

Lemma S1.8. *Under the assumptions of Theorem S1.6 it holds that*

$$\|\Phi\hat{f}_n - \Phi f_0\|_{L^2(I)} = o_p(1) \quad \textit{implies} \quad \|f(\cdot,\theta_0) - f(\cdot,\hat{\theta}_n)\|_{L^2([a,b])} = o_p(1).$$

Proof. The operator $\Phi : (\mathbf{F}_k, \|\cdot\|_{L^2([a,b])}) \longrightarrow (L^2(I), \|\cdot\|_{L^2(I)})$ is linear and bounded and hence continuous. According to Assumption **C**ii) it is injective and it follows from Lemma S1.5 that the set $(\mathbf{F}_k, \|\cdot\|_{L^2([a,b])})$ is totally bounded. Since it also contains functions with less than k change points, it is additionally closed and therefore compact. Hence $\Phi : \mathbf{F}_k \longrightarrow \{\Phi f \in L^2(I) : f \in \mathbf{F}_k\}$ is a bijective continuous mapping from a compact set to a Hausdorff space, hence a homeomorphism (see tom Dieck (2008, Prop 1.5.3)). □

Lemma S1.9. *Assume that $f_0 = f(\cdot,\theta_0) \in \boldsymbol{F}_k$ with $\sharp\mathcal{J}(f_0) = k$ and let $\{f(\cdot,\theta_n)\}_{n\in\mathbb{N}}$ be a sequence in $\boldsymbol{F}_k$. Then*

$$\|f(\cdot,\theta_0) - f(\cdot,\theta_n)\|_{L^2([a,b])} = o(1) \quad \textit{implies} \quad |\theta_0 - \theta_n|_\infty = o(1).$$

Proof. Due to the definition of $\mathcal{J}(\cdot)$ in Subsection 2.2, the assumption $\sharp\mathcal{J}(f_0) = k$ implies that $f(\cdot,\theta_0)$ has precisely k change points. That means, $f(\cdot,\theta_0) \equiv f(\cdot,\theta)$ implies $\theta = \theta_0$, i.e. for all $\theta_0 \neq \theta \in \Theta_k$ we have $\|f(\cdot,\theta_0) - f(\cdot,\theta)\|_{L^2([a,b])} > 0$. Now assume that $\|f(\cdot,\theta_0) - f(\cdot,\theta_n)\|_{L^2([a,b])} = o(1)$ but that there exist a subsequence $\{\theta_{k_n}\}_{n\in\mathbb{N}}$ and a constant $c_1 > 0$, such that $|\theta_0 - \theta_{k_n}|_\infty > c_1$ for all $n \in \mathbb{N}$. Since Θ_k is compact, we can choose a further subsequence of this subsequence, which converges to some $\hat{\theta} \in \Theta_k$. W.l.o.g we assume $\lim_{n\to\infty} |\hat{\theta} - \theta_{k_n}|_\infty = 0$. By construction $|\theta_0 - \hat{\theta}|_\infty > c_1$ and so uniqueness of θ_0 implies $\|f(\cdot,\theta_0) - f(\cdot,\hat{\theta})\|_{L^2([a,b])} > c_2 > 0$ for some constant c_2. Since the mapping $\theta \mapsto \|f(\cdot,\theta) - f(\cdot,\theta_0)\|_{L^2([a,b])}$ is continuous by Lemma S1.1, there exists some $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$ we have

$$\|f(\cdot,\theta_0) - f(\cdot,\theta_{k_n})\|_{L^2([a,b])} > \frac{1}{2}c_2 > 0.$$

This is a contradiction to $\|f(\cdot,\theta_0) - f(\cdot,\theta_n)\|_{L^2([a,b])} = o(1)$ and the claim follows. □

Corollary S1.10. *Under the assumptions of Theorem S1.6 it holds that*

$$\|f(\cdot,\theta_0) - f(\cdot,\hat{\theta}_n)\|_{L^2([a,b])} = o_P(1).$$

Moreover, if the true function f_0 has exactly k change points it also holds that

$$|\theta_0 - \hat{\theta}_n|_\infty = o_P(1).$$

Proof. This follows from Theorem S1.6 by application of Lemma S1.8 and S1.9. □

S1.4 Asymptotic normality

In this subsection we show asymptotic normality of the least squares estimator $\hat{\theta}_n$ in (3.1). Therefore, we focus on the stochastic process $\|Y-\Phi\hat{f}_n\|_n^2 = n^{-1}\sum_{i=1}^n (y_i-\Phi\hat{f}_n(x_i))^2$ for the random observations (Y,X) as in (1.1), which henceforth we write as the empirical expectation

$$\mathbb{E}_n m(\cdot,\cdot,\theta) := n^{-1}\sum_{i=1}^n m(x_i,y_i,\theta),$$

with m defined as

$$m(x,y,\theta) := (y-(\Lambda(\theta))(x))^2. \tag{S1.6}$$

Hence, $\hat{\theta}_n$ the least squares estimator is the minimizer of $\theta \longmapsto \mathbb{E}_n m(\cdot,\cdot,\theta)$. Let $\mathbf{E}\varepsilon_1 = 0$ and $\mathbf{E}\varepsilon_1^2 = \sigma^2$ then expectation of $m(\cdot,\cdot,\theta)$ can be calculated as

$$\begin{aligned}\mathbf{E}m(\cdot,\cdot,\theta) &= \mathbf{E}(\Phi f(\cdot,\theta_0)-\Phi f(\cdot,\theta))^2+\sigma^2 \\ &= \mathbf{E}(\Phi f(\cdot,\theta_0)-\Phi f(\cdot,\theta))^2+\mathbf{E}m(\cdot,\cdot,\theta_0).\end{aligned} \tag{S1.7}$$

By Lemma S1.2, the function $\theta\mapsto m(\cdot,y,\theta)$ is differentiable with derivative $\partial/\partial\theta\, m(\cdot,y,\theta) = 2(\Lambda(\theta)-y)\Lambda'[\theta]$ such that for all $h_1,h_2\in\mathbb{R}^d$

$$\mathbf{E}\left(\frac{\partial m}{\partial\theta}(\cdot,\cdot,\theta_0)h_1\right)\left(\frac{\partial m}{\partial\theta}(\cdot,\cdot,\theta_0)h_2\right) = 4\sigma^2\mathbf{E}(\Lambda'[\theta_0]h_1)(\Lambda'[\theta_0]h_2) = 4\sigma^2 h_1^\top V_{\theta_0}h_2. \tag{S1.8}$$

Classical conditions for asymptotic normality of $\hat{\theta}_n$ require that the function $\theta\mapsto m(x,y,\theta)$ is twice differentiable, which is not the case on our situation. Therefore, we follow a different route according to Theorem 5.23 (Chapter 5.3) in van der Vaart (1998) where a second order expansion of the expectation $\theta\mapsto \mathbf{E}m(\cdot,\cdot,\theta)$ instead of the function m itself is sufficient to obtain the desired normality.

Theorem S1.11. *For each θ in an open subset of Euclidean space let $(x,y)\mapsto m(x,y,\theta)$ be a measurable function such that $\theta\mapsto m(x,y,\theta)$ is differentiable at θ_0 for $\boldsymbol{P}$-almost every (x,y) and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $\boldsymbol{E}\dot{m}^2<\infty$*

$$|m(x,y,\theta_1)-m(x,y,\theta_2)| \le \dot{m}(x,y)|\theta_1-\theta_2|_\infty. \tag{S1.9}$$

Furthermore, assume that the map $\theta\mapsto \boldsymbol{E}m(\cdot,\cdot,\theta)$ has the asymptotic behavior

$$\boldsymbol{E}m(\cdot,\cdot,\theta) = \boldsymbol{E}m(\cdot,\cdot,\theta_0)+\frac{1}{2}(\theta-\theta_0)^\top V(\theta-\theta_0)+o(|\theta_0-\theta|_\infty^2), \qquad \text{as } |\theta_0-\theta|_\infty\to 0 \tag{S1.10}$$

at a point of minimum θ_0 with a nonsingular symmetric matrix V. If $\mathbb{E}_n m(\cdot,\cdot,\hat{\theta}_n) \le \inf_\theta \mathbb{E}_n m(\cdot,\cdot,\theta)+o_P(n^{-1})$ and $\hat{\theta}_n\xrightarrow{P}\theta_0$, then

$$\sqrt{n}(\hat{\theta}_n-\theta_0) = -V^{-1}\frac{1}{\sqrt{n}}\sum_{1=1}^n \frac{\partial m}{\partial\theta}(x_i,y_i,\theta_0)+o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n-\theta_0)$ is asymptotically normal with mean zero and covariance matrix $V^{-1}\boldsymbol{E}\frac{\partial m}{\partial\theta}(\cdot,\cdot,\theta_0)\frac{\partial m}{\partial\theta}(\cdot,\cdot,\theta_0)^\top V^{-1}$.

Proof. Along the lines of the proof of van der Vaart (1998, Thm 2.23). □

Proof. (of Theorem 3.1) We show that the assumptions of Theorem S1.11 are satisfied: It follows from Lemma S1.1 and the assumed boundedness of $\Phi : L^1([a,b]) \to L^\infty(I)$ that $\Lambda = \Phi \cdot \mathfrak{F} : (\Theta, |\cdot|_\infty) \to (L^\infty, \|\cdot\|_{L^\infty})$ is Lipschitz continuous, which implies condition (S1.9) is satisfied with constant $\dot{m}$. Moreover, (S1.10) with $V = V_{\theta_0}$ follows from Lemma S1.2. According to this theorem, together with (S1.8), the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $\sigma^2 V_{\theta_0}^{-1}$, which proves (i). Part (ii) follows from van der Vaart (1998, Cor. 5.53). Part (iii) is now a consequence of part (ii) and Lemma S1.1. Finally, part (iv) follows from part (iii) with $p = 1$ and the boundedness of $\Phi : L^1([a,b]) \to L^\infty(I)$. □

Proof. (of Corollary 3.2) Due to the differentiability of h in Definition 2.6 Lemma S1.1 and S1.2 hold analogously for the reduced parameter domain by the chain rule. Moreover, the mapping $\delta\tilde{\theta} \mapsto \Lambda'[h(\tilde{\theta})h'[\tilde{\theta}]\delta\tilde{\theta}$ is injective by Assumption **C** and the injectivity assumption in Definition 2.6, and hence $V_{\tilde{\theta}}$ is nonsingular. Therefore, the proof of the corollary is completely analogous to the proof of Theorem 3.1. □

Proof. (of Corollary 3.4) Statements (i) - (iv) from Theorem 3.1 are valid for the reduced parameter vectors $\tilde{\theta}_0$ and $\tilde{\theta}_n$ by Corollary 3.2. In order to show (3.6), we skip the dependencies of the parameter components, for the sake of simplicity and consider the pieces $\mathfrak{f}(y, \vartheta^i)$ instead of $\mathfrak{f}(y, \vartheta^i(\tilde{\theta}))$ for all $i = 1, ..., k+1$, keeping in mind that for all occurring derivatives we actually need to apply the chain rule.

Now f has a kink in τ_i for all $i = 1, ..., k$. W.l.o.g. we assume that $\tau_i > \hat{\tau}_i$, then we have

$$\int_{\tau_i}^{\hat{\tau}_i} \left(\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(y, \hat{\vartheta}^i)\right)^p \mathrm{d}y \leq \int_{\tau_i}^{\hat{\tau}_i} \Big(|\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^{i+1})| + |\mathfrak{f}(\tau_i, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^i)| + |\mathfrak{f}(\tau_i, \vartheta^i) - \mathfrak{f}(\tau_i, \hat{\vartheta}^i)| + |\mathfrak{f}(\tau_i, \hat{\vartheta}^i) - \mathfrak{f}(y, \hat{\vartheta}^i)|\Big)^p \mathrm{d}y.$$

By the mean value theorem we have $|\mathfrak{f}(\tau_i, \vartheta^i) - \mathfrak{f}(\tau_i, \hat{\vartheta}^i)| = O(|\vartheta^i - \hat{\vartheta}^i|)$. The term $|\mathfrak{f}(\tau_i, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^i)|$ vanishes because there is a kink at τ_i. Finally, remembering the definition of the modulus of continuity ν in (3.5), we get

$$\sup_{y \in [\tau_i, \hat{\tau}_i]} (|\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(\tau_i, \vartheta^{i+1})|, |\mathfrak{f}(\tau_i, \hat{\vartheta}^i) - \mathfrak{f}(y, \hat{\vartheta}^i)|) = \nu(\mathcal{F}, |\tau_i - \hat{\tau}_i|)\,.$$

Hence, it follows from (ii) that

$$\begin{aligned} \int_{\tau_i}^{\hat{\tau}_i} \left(\mathfrak{f}(y, \vartheta^{i+1}) - \mathfrak{f}(y, \hat{\vartheta}^i)\right)^p \mathrm{d}y &= O(|\tau_i - \hat{\tau}_i|)(\nu(\mathcal{F}, |\tau_i - \hat{\tau}_i|) + |\vartheta^i - \hat{\vartheta}^i|)^2 \\ &= O_P(n^{-\frac{1}{2}}(\nu(\mathcal{F}, n^{-\frac{1}{2}})^p + n^{-p/2})). \end{aligned}$$

Since this holds for all $i = 1, ..., k$, this proves (3.6). □

S1.5 Proof of Proposition 3.3

Proof. (Proposition 3.3) It is obvious from the definition (3.4) that the matrix V_θ is symmetric and positive semi-definite, so we have to study under which conditions it is positive definite. Since $h^\top V_\theta h = \int_a^b |\Lambda'[\theta_0]h|^2 s\,dy$ for $h \in \mathbb{R}^d$, it follows from Assumption **B** on s that that $h^\top V_\theta h = 0$ is equivalent to $\Lambda'[\theta]h = 0$. Hence, V_{θ_0} is non-singular if and only if $\Lambda'[\theta]$ is injective. It follows from Lemma S1.2 that $\Lambda'[\theta] = \Phi \circ \mathfrak{F}'[\theta]$ is the composition of the integral operator $\Phi : \mathcal{M} \to L^2(I)$, which is injective by Assumption **C**ii) and the formal derivative $\mathfrak{F}'[\theta_0] : \Theta_k \to \mathcal{M}$, of the mapping $\mathfrak{F}\theta := f(\cdot,\theta)$ given by

$$\begin{aligned}&\mathfrak{F}'[\vartheta_1,\tau_1,\vartheta_2,\dots,\tau_k,\vartheta_{k+1}](\delta\vartheta_1,\delta\tau_1,\delta\vartheta_2\dots,\delta\tau_k,\delta\vartheta_{k+1})\\ &\quad:=\sum_{j=1}^{k+1}(\delta\vartheta_j)^\top\frac{\partial \mathfrak{f}}{\partial\vartheta_j}(\cdot,\vartheta_j)\mathbf{1}_{[\tau_{j-1},\tau_j)}+\sum_{j=1}^{k}\left[f(\cdot,\theta)\right](\tau_j)\delta\tau_j\,.\end{aligned}$$

Since the mappings $\mathfrak{F}'_{[\tau_{j-1},\tau_j)}[\vartheta_j]\delta\vartheta_j = (\delta\vartheta_j)^\top\frac{\partial \mathfrak{f}}{\partial\vartheta_j}(\cdot,\vartheta_j)$ are assumed to be injective (see Definition 2.1) $\mathfrak{F}$ is injective if and only if $[f(\cdot,\theta)](\tau_j) \neq 0$ for $j = 1,\dots,k$, i.e. if and only if $f(\cdot,\theta)$ has jumps at all change points. □

S2 Proof of Theorem 3.5

From Inequality (3.2) we obtain the basic inequality

$$\|\Phi\hat f_{\lambda_n} - \Phi f_0\|_n^2 \le 2\langle \Phi f_0 - \Phi\hat f_{\lambda_n},\varepsilon\rangle_n + \lambda_n(\sharp J(f_0) - \sharp J(\hat f_{\lambda_n})) + o(n^{-1}). \tag{S2.11}$$

Again we have to consider the behavior of the empirical process $\langle \Phi f_0 - \Phi\hat f_{\lambda_n},\varepsilon\rangle_n$, and therefore the entropy of the respective function space to gain a bound for this process. We use the results from Boysen, Bruns, and Munk (2009).

Lemma S2.1. *Suppose that Assumptions A and A*1 *are satisfied. Then, for all* $\Phi f \in \boldsymbol{G}_\infty = \{\Phi f \in L^2(I) \mid f \in \boldsymbol{F}_\infty\}$*, we have*

$$|\langle \Phi f,\varepsilon\rangle_n| = O_P(n^{-\frac12})\|\Phi f\|_n^{1-\epsilon}(\sharp J(f))^{\frac12(1+2\epsilon)},$$

for any $\epsilon > 0$.

Proof. For fixed number of jumps k, we find from Lemma S1.5 that

$$H(\delta, G_k, P_n) \le d\log\left(\frac{\tilde T_{\mathcal{F}}\sqrt{k+1}+\delta}{\delta}\right),$$

with $d = (k+1)r + k$ and a constant $\tilde T_{\mathcal{F}}$, which is independent of k. Using this entropy bound, it follows along the lines of the proof of Lemma 4.18 in Boysen, Bruns, and Munk (2009) that

$$\sup_{g\in G_k,\,\|g\|_n\le\delta}\frac{|\langle g,\varepsilon\rangle_n|}{\sqrt{k}\delta\left(1+\log\left(\frac{\tilde T_{\mathcal{F}}\sqrt{k}+\delta}{\delta}\right)\right)} = O_P(n^{-\frac12}),$$

holds uniformly for all k. For all $\Phi f \in \mathbf{G}_\infty$ this implies

$$\frac{|\langle \Phi f, \varepsilon\rangle_n|}{\sqrt{\sharp J(f)}\|\Phi f\|_n\left(1+\log\left(\frac{\tilde{T}_{\mathcal{F}}\sqrt{\sharp J(f)}+\|\Phi f\|_n}{\|\Phi f\|_n}\right)\right)} \leq \sup_{\substack{g\in G_{\sharp J(f)},\\ \|g\|_n\leq\|\Phi f\|_n}} \frac{|\langle g,\varepsilon\rangle_n|}{\sqrt{\sharp J(f)}\|\Phi f\|_n\left(1+\log\left(\frac{\tilde{T}_{\mathcal{F}}\sqrt{\sharp J(f)}+\|\Phi f\|_n}{\|\Phi f\|_n}\right)\right)} = O_P(n^{-\frac{1}{2}}).$$

Analogously to the proof of Corollary 4.19 in Boysen, Bruns, and Munk (2009), this directly yields the claim. □

Lemma S2.2. *Let $f_0 \in \boldsymbol{F}_\infty$ and $\{f_n\}_{n\in\mathbb{N}}$ a sequence in $\boldsymbol{F}_{\sharp J(f_0),D}$, with*

$$\|f_0 - f_n\|_{L^2([a,b])} = o(1).$$

Then, there exists an $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$

$$\sharp J(f_0) = \sharp J(f_n).$$

Proof. W.l.o.g let $\sharp J(f_0) = 1$. Now we assume that there exists a subsequence f_{k_n} with no jumps, i.e. $f_{k_n} \in \boldsymbol{F}_{0,D}$ for all n. Furthermore f_{k_n} is a subsequence of a converging sequence, and thus converges to the same limit function f_0. As shown in the proof of Lemma S1.8, the set $\boldsymbol{F}_{0,D}$ is compact thus the limit function of f_{k_n} has to be contained in $\boldsymbol{F}_{0,D}$, which leads the contradiction

$$f_0 \in \boldsymbol{F}_{0,D}.$$

□

Now we are prepared for the proof of Theorem 3.5.

Proof. (of Theorem 3.5) Throughout the proof w.l.o.g we assume that $\epsilon \leq 1$. From Lemma S2.1 and (S2.11), it follows that

$$\begin{aligned}\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^2 &\leq O_P(n^{-\frac{1}{2}})\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon}(\sharp\mathcal{J}(\hat{f}_{\lambda_n} - f_0))^{\frac{1}{2}(1+\epsilon)} \\ &\quad +\lambda_n(\sharp\mathcal{J}(f_0) - \sharp\mathcal{J}(\hat{f}_{\lambda_n})) + o(n^{-1}) \\ &\leq O_P(n^{-\frac{1}{2}})\|\Phi\hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon}\sharp\mathcal{J}(\hat{f}_{\lambda_n})^{\frac{1}{2}(1+\epsilon)} - \lambda_n\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n\sharp\mathcal{J}(f_0),\end{aligned}$$

where we took into account that λ_n is assumed to converge slower than n^{-1} and that we have $\sharp\mathcal{J}(f_0) < \infty$, which implies that $\sharp\mathcal{J}(\hat{f}_{\lambda_n} - f_0) = O_P(\sharp\mathcal{J}(\hat{f}_{\lambda n}))$.

Choosing $f \equiv 0$ on the right hand side of Equation (3.2) implies $\lambda_n\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \|Y\|_n^2 = O_P(1)$ and hence, we have

$$\sharp\mathcal{J}(\hat{f}_{\lambda_n}) = O_P(\lambda_n^{-1}). \tag{S2.12}$$

We assumed that $\lambda_n^{-1} n^{-1/(1+\epsilon)} \to 0$, for $n \to \infty$, which gives

$$n^{-1} = o(\lambda_n^{1+\epsilon}). \tag{S2.13}$$

By compactness of Ψ we have that $\sup_{f \in \mathbf{F}_\infty} \|f\|_\infty \leq R$ and thus

$$\sup_{f \in \mathbf{F}_\infty} \|\Phi f\|_n \leq \|\varphi\|_\infty R < \infty \tag{S2.14}$$

Inserting (S2.14), (S2.12) and (S2.13) into (S2.12), we obtain

$$\begin{aligned} \|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^2 &\leq o_P(\lambda_n^{\frac{1+\epsilon}{2}}) O_P(\lambda_n^{\frac{1-\epsilon}{2}}) \|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^{1-\frac{1}{2}\epsilon} \sharp\mathcal{J}(\hat{f}_n) - \lambda_n \sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n \sharp\mathcal{J}(f_0) \\ &= (o_p(\lambda_n) - \lambda_n)\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n \sharp\mathcal{J}(f_0). \end{aligned} \tag{S2.15}$$

Since $o_p(\lambda_n) - \lambda_n$ becomes negative for increasing n, this implies

$$\|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_n^2 = O_P(\lambda_n)$$

and with Corollary S1.3,

$$\|\Phi \hat{f}_{\lambda_n} - \Phi f_0\|_{L^2(I)}^2 = O_P(\lambda_n) + o_P(1) = o_P(1). \tag{S2.16}$$

Again considering Equation (S2.15) we find that this is equivalent to

$$0 \leq (o_p(\lambda_n) - \lambda_n)\sharp\mathcal{J}(\hat{f}_{\lambda_n}) + \lambda_n \sharp\mathcal{J}(f_0),$$

which means

$$(1 - o_P(1))\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0).$$

Because $\sharp\mathcal{J}(f_0)$ and $\sharp\mathcal{J}(\hat{f}_{\lambda_n})$ are integers, this implies $P(\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0)) \to 1$. For $\sharp\mathcal{J}(\hat{f}_{\lambda_n}) \leq \sharp\mathcal{J}(f_0)$ in turn, it holds that $f_0, \hat{f}_{\lambda_n} \in \boldsymbol{F}_{\sharp\mathcal{J}(f_0)}$ and Lemma S1.8 together with (S2.16), yields

$$\|f_0 - \hat{f}_{\lambda_n}\|_{L^2([a,b])} = o_P(1).$$

Using Lemma S2.2 this implies that $\lim_{n\to\infty} P\left(\sharp\mathcal{J}(f_0) = \sharp\mathcal{J}(\hat{f}_{\lambda_n})\right) = 1$, which is the claim. □

S3 Proofs of the Theorems 4.2, 4.3 and 4.4

Proof. (Theorem 4.2) ad **C**i): It is straightforward to show that $\|\Phi f\|_{L^\infty} \leq \|\phi\|_{L^\infty} \|f\|_{L^1}$, so $\phi \in \mathcal{L}\left(L^1([a,b]), L^\infty([c,d])\right)$. Since $\phi \in BV([ac,bd])$ there exist monotonically increasing and bounded functions ϕ_1, ϕ_2 such that $\phi = \phi_1 - \phi_2$. Setting $\varphi_i(x,y) := \phi_i(xy)$ for $i = 1,2$ we obtain for $x, x+\delta \in [a,b]$ with $\delta > 0$

$$\begin{aligned} |(\Phi f)(x) - (\Phi f)(x+\delta)| &= \left| \int_a^b (\varphi_1(x,y) - \varphi_1(x+\delta,y) - \varphi_2(x,y) + \varphi_2(x+\delta,y)) f(y) \mathrm{d}y \right| \\ &\leq \|f\|_\infty \left[\int_a^b |\varphi_1(x+\delta,y) - \varphi_1(x,y)| \,\mathrm{d}y + \int_a^b |\varphi_2(x+\delta,y) - \varphi_2(x+\delta,y)| \,\mathrm{d}y \right] \\ &= \|f\|_\infty \left[\int_a^b (\varphi_1(x+\delta,y) - \varphi_1(x,y)) \,\mathrm{d}y + \int_a^b (\varphi_2(x+\delta,y) - \varphi_2(\delta,y)) \,\mathrm{d}y \right] \end{aligned} \tag{S3.17}$$

using the monotonicity of ϕ_i in the last line. The integrals on the left hand side can be estimated by

$$\begin{aligned}\int_a^b [\phi_i((x+\delta)y) - \phi_i(xy)]\,\mathrm{d}y &= \frac{1}{x+\delta}\int_{a(x+\delta)}^{b(x+\delta)} \phi_i(u)\mathrm{d}u - \frac{1}{x}\int_{ax}^{bx}\phi_i(u)\mathrm{d}u \\ &= \left(\frac{1}{x+\delta} - \frac{1}{x}\right)\int_{ax}^{bx}\phi_i(u)\mathrm{d}u + \frac{1}{x+\delta}\left(\int_{bx}^{b(x+\delta)}\phi_i(u)\mathrm{d}u - \int_{ax}^{a(x+\delta)}\phi_i(u)\mathrm{d}u\right) \\ &\le \left(\frac{b-a}{x+\delta}\delta + \frac{\delta b - \delta a}{x+\delta}\right)\|\phi_i\|_\infty \le \frac{b-a}{a}2\delta\|\phi_i\|_\infty\,,\end{aligned}$$

so $|(\Phi f)(x) - (\Phi f)(x+\delta)| \le \frac{b-a}{a}2\delta(\|\phi_1\|_\infty + \|\phi_2\|_\infty)\,\|f\|_\infty$.

ad **C**ii): Assume that the Müntz condition (4.1) holds true and that

$$(\Phi\mu)|_{\left[\frac{\rho_1}{a},\frac{\rho_2}{b}\right]} \equiv 0$$

for some Borel measure $\mu \in \mathcal{B}([a,b])$. Since $xy \in [\rho_1,\rho_2]$ if $x \in [\frac{\rho_1}{a},\frac{\rho_2}{b}]$ and $y \in [a,b]$ and since the series expansion of ϕ converges absolutely and hence uniformly on the compact interval $[\rho_1,\rho_2]$, integration and summation may be interchanged, and we obtain

$$(\Phi\mu)(x) = \int_a^b \phi(xy)\mathrm{d}\mu(y) = \sum_{j=0}^\infty x^j \int_a^b \alpha_j y^j \mathrm{d}\mu(y) = \sum_{j=0}^\infty c_j x^j, \qquad x \in \left[\frac{\rho_1}{a},\frac{\rho_2}{b}\right]$$

with $c_j := \alpha_j \int_a^b y^j \mathrm{d}\mu(y)$. In order to see that the power series $\sum_{j=0}^\infty c_j x^j$ converges absolutely and uniformly for $x \in [\frac{\rho_1}{a},\frac{\rho_2}{b}]$, note that $|c_j| \le |\mu|([a,b])|\alpha_j|b^j$, so

$$\sum_{j=0}^\infty \left|c_j x^j\right| \le |\mu|([a,b])\sum_{j=0}^\infty |\alpha_j|\rho_2^j < \infty, \qquad x \in \left[\frac{\rho_1}{a},\frac{\rho_2}{b}\right].$$

Since a power series with positive radius of convergence vanishes identically if and only if all its coefficients vanish, we obtain

$$\int_a^b y^j \mathrm{d}\mu(y) = 0 \qquad \text{for all } j \in J.$$

By the *Müntz-Theorem* 4.1 this implies that $\int_a^b g(y)\mathrm{d}\mu(y) = 0$ for all $g \in C([a,b])$, so $\mu \equiv 0$, i.e. $\Phi : \mathcal{B}([a,b]) \to L^2([a,b])$ is injective.

If the Müntz condition (4.1) is violated, then the converse implication of the full Müntz Theorem 4.1 entails that the closure of $\mathrm{span}(\{y^j : j \in J\})$ does not coincide with $C([a,b])$, and as a consequence of the Hahn-Banach theorem (cf. Yosida (1995, §IV.6)) there exists a functional $\overline{\mu}_0 \neq 0$ in the dual space $C([a,b])'$ which vanishes on $\mathrm{span}(\{y^j : j \in J\})$. By the Riesz representation theorem (cf. Rudin (1987, Thm 6.19)) $\overline{\mu}_0$ can be expressed by a (signed) Borel measure $\mu_0 \in \mathcal{B}([a,b])$ via $\overline{\mu}_0(g) = \int_a^b g d\mu_0$, and our previous computations show that $\Phi\mu_0 = 0$. □

Proof. (Theorem 4.3) ad **C**i): Obviously, $\|\Phi f\|_{L^\infty} \le \|\phi\|_{L^\infty}\|f\|_{L^1}$, so $\phi \in \mathcal{L}\left(L^1([a,b]), L^\infty([a,b])\right)$. As in the proof of Theorem 4.2, we can write $\phi = \phi_1 - \phi_2$ with bounded, monotonically increasing functions ϕ_1, ϕ_2, and define $\varphi_i(x,y) := \phi_i(x-y)$ such that eq. (S3.17) holds true. Here the integrals on the left hand side of (S3.17) can be estimated by

$$\int_a^b (\phi_i(x+\delta-y) - \phi_i(x-y))\,\mathrm{d}y \quad = \quad -\int_{x-b}^{x-b+\delta} \phi_i(u)\mathrm{d}u + \int_{x-a}^{x-a+\delta} \phi_i(u)\mathrm{d}u \le 2|\delta|\|\phi_i\|_\infty,$$

and we obtain $|(\Phi f)(x) - (\Phi f)(x+\delta)| \le 2\delta(\|\phi_1\|_\infty + \|\phi_2\|_\infty)\|f\|_\infty$.

ad **C**ii): Take $\mu = f + \sum_{j=1}\gamma_j\delta_{x_j} \in \mathcal{M}([a,b])$ and assume that

$$(\Phi\mu)(x) = \int_a^b \phi(x-y)f(y)\,\mathrm{d}y + \sum_{j=1}^n \gamma_j\phi(x-x_j) = 0, \qquad \text{for all } x \in [a,b].$$

Extending f by 0 on $\mathbb{R}\setminus[a,b]$, it follows from the Plancherel theorem and the Fourier convolution theorem that

$$\begin{aligned}
0 &= \int_{-\infty}^\infty f(x)(\Phi\mu)(x)\,\mathrm{d}x + \sum_{k=1}^n \gamma_k(\Phi\mu)(x_k) \\
&= \int_{-\infty}^\infty f(x)\int_{-\infty}^\infty \phi(x-y)f(y)\,\mathrm{d}y\,\mathrm{d}x + \sum_{j=1}^n \gamma_j\int_{-\infty}^\infty f(x)\phi(x-x_j)\,\mathrm{d}x \\
&\quad + \sum_{k=1}^n \gamma_k\int_{-\infty}^\infty \phi(x_k-y)f(y)\,\mathrm{d}y + \sum_{k=1}^n\sum_{j=1}^n \gamma_k\gamma_j\phi(x_j-x_k) \\
&= \int_{-\infty}^\infty \left|\widehat{f}(\xi) + \sum_{j=1}^n \gamma_j e^{-2\pi i\xi x_j}\right|^2 \widehat{\phi}(\xi)\,\mathrm{d}\xi\,.
\end{aligned}$$

Using the assumption $\widehat{\phi} > 0$ a.e., we find that $\widehat{f}(\xi) + \sum_{j=1}^n \gamma_j e^{-2\pi i\xi x_j} = 0$ for a.e. $\xi \in \mathbb{R}$. Since $\lim_{|\xi|\to 0}\widehat{f}(\xi) = 0$ by the Riemann-Lebesgue lemma, this implies $\widehat{f} = 0$ and $\gamma_1 = \cdots = \gamma_n = 0$, so $\mu = 0$. □

Proof. (Theorem 4.4) ad **C**ii): This follows from the first part of Theorem 4.3 since analytic functions are of bounded variation.

ad **C**ii): Assume that $\Phi\mu = 0$ for $\mu = f + \sum_{j=1}\gamma_j\delta_{x_j} \in \mathcal{M}([a,b])$. Since ϕ is analytic, it has a holomorphic extension to a neighborhood $\mathcal{U}$ of $\mathbb{R}$ in $\mathbb{C}$. By a compactness argument $\mathcal{U}_0 := \bigcap_{y\in[a,b]}\mathcal{U} - y$ is also a neighborhood of $\mathbb{R}$ in $\mathbb{C}$. Define

$$g(z) := \int_a^b \phi(z-y)f(y)\,dy + \sum_{j=1}^n \gamma_j\phi(z-x_j), \qquad z \in \mathcal{U}_0.$$

Interchanging differentiation and integration, it follows that g is holomorphic, and $g(x) =$

$(\Phi\mu)(x) = 0$ for $x \in [a, b]$. Hence, g vanishes identically. Therefore,

$$0 = \int_{-\infty}^{\infty} e^{-2\pi i\xi x} g(x)\, dx = \widehat{\phi}(\xi) \left(\widehat{f}(\xi) + \sum_{j=1}^{n} \gamma_j e^{2\pi i \xi x_j} \right), \qquad \xi \in \mathbb{R}.$$

Since we have assumed that $\widehat{\phi} \neq 0$ a.e., it follows that the term in parenthesis vanishes a.e., and hence $\mu = 0$. □

References

Boysen, L., Bruns, S., and Munk, A. (2009), "Jump estimation in inverse regression," *Electron. J. Statist.*, 3, 1322–1359.

del Barrio, E., Deheuvels, P., and van de Geer, S. (2007), *Lectures on empirical processes*, EMS Series of Lectures in Mathematics, European Mathematical Society (EMS), Zürich, theory and statistical applications, With a preface by Juan A. Cuesta Albertos and Carlos Matrán.

Rudin, W. (1987), *Real and complex analysis*, New York: McGraw-Hill Book Co., 3rd ed.

tom Dieck, T. (2008), *Algebraic Topology*, EMS Textbooks in Mathematics, European Mathematical Society.

van de Geer, S. A. (2000), *Applications of empirical process theory*, vol. 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press.

van der Vaart, A. W. (1998), *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press.

Yosida, K. (1995), *Functional analysis*, Classics in Mathematics, Berlin: Springer-Verlag, reprint of the sixth (1980) edition.