Statistica Sinica **24** (2014), 555-575
doi:http://dx.doi.org/10.5705/ss.2012.007

ASYMPTOTIC LAWS FOR CHANGE POINT ESTIMATION IN INVERSE REGRESSION

Sophie Frick, Thorsten Hohage and Axel Munk

Universität Göttingen

Abstract: We derive rates of convergence and asymptotic normality of 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: Asymptotic normality, change point analysis, confidence bands, dynamic stress moduli, entropy bounds, Hammerstein integral equations, jump detection, penalized least squares estimator, reproducing kernel Hilbert spaces, sparsity, statistical inverse problems.

1. Introduction

We consider the inverse regression model

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

where $X = (x_1, \dots, x_n)$, $n \in \mathbb{N}$ is a (possibly random) vector of design points in a bounded interval $I \subset \mathbb{R}$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ denotes the observation error that 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 $(X, Y) = ((x_1, y_1), \dots, (x_n, y_n))$.

This class of models covers a large variety of 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 good 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 degenerate. This is well known already from much simpler cases, see e.g., Boysen, Bruns, and Munk (2009a).

The particular case in which f_0 has no jumps but kinks is treated in detail. Here 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 those in 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, 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.

When 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: 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) that 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 our asymptotic results are valid for every injective integral operator Φ with certain properties (cf., Assumption **C**).

Our method can even be applied to 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 one to provide estimators and confidence bands for the time relaxation spectra of polymer melts reconstructed from their dynamic modul (see Roths et al. (2000)).

We apply the asymptotic results to the estimation of a step function from the noisy image of an integral operator with convolution kernel (inverse two phase regression) and to the estimation of a piecewise linear kink function from the noisy image of an integral operator with product kernel (inverse multiphase regression). In both cases, we calculate confidence bands of the reconstructed function that give an impression of the reliability of the estimate.

Our results differ substantially from the "truly nonparametric" kink models that have appeared, including Korostelev (1987), Neumann (1997), Raimondo (1998), Goldenshluger, Tsybakov, and Zeevi (2006), Goldenshluger et al. (2008b), Goldenshluger et al. (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 that 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 between discontinuities. Roughly speaking, the situation treated here can be viewed as a limiting case, when the degree of smoothness tends to infinity.

The 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 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. Notation

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 $x_1, \dots, x_n$ 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 take $(\theta_1, \theta_2) := \{\theta \in \mathbb{R}^d \mid \theta = \theta_1 + t(\theta_2 - \theta_1),$ 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 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 $\mathfrak{f} : [a, b] \times \Psi \longrightarrow \mathbb{R}$ satisfy:

(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 $\vartheta \in \Psi$, where $C(I)$ 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.

Example 1 (Constant functions)**.** If $\Psi = [-M, M]$ and $\mathfrak{f}(y, \vartheta) := \vartheta$ we obtain

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

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

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

Definition 2. Let $\mathcal{F} = \{\mathfrak{f}(\cdot, \vartheta) : \vartheta \in \Psi\}$ be a family of continuous parametric functions on the interval $[a, b]$. 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 $a = \tau_0 < \tau_1 < \cdots < \tau_{k+1} = b$ and parameter vectors $\vartheta^1, \ldots, \vartheta^{k+1} \in \Psi$ 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 f is also denoted by $f(\cdot, \vartheta^1, \tau_1, \ldots, \vartheta^k, \tau_k, \vartheta^{k+1})$. We call the elements of the set

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

change points of the function $f \in \mathbf{F}_k$, 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 $\mathbf{F}_k[a, b]$ (or shortly by $\mathbf{F}_k$). With

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

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

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

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 3. 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.

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. Consider the subset of $\tilde{\mathbf{T}}_k \subset \mathbf{T}_k$, with precisely k jumps,

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

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

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

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 3. $\Theta_k \subset \mathbb{R}^d$ denote the parameter domain of a family $\mathbf{F}_k$ of pc functions. 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

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

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 $\tilde{\mathbf{F}}_k := \{f(\cdot, h(\tilde{\theta})) \mid \tilde{\theta} \in \tilde{\Theta}\}$, and the elements $\tilde{\theta}_0 \in \tilde{\Theta}$ are called *reduced parameter vectors* of the functions $f(\cdot, h(\tilde{\theta})) \in \tilde{\mathbf{F}}_k$.

Note that if we consider a class of pc-functions $\mathbf{F}_k$ that is generated by a parametric class $\mathcal{F}$, and if $(y,\vartheta) \mapsto \mathfrak{f}(y,\vartheta)$ is continuously differentiable, then the condition $[f(\cdot,\theta)](\tau) = 0$ often implies local existence of a function h as in Definition 3, by the Implicit Function Theorem. More precisely, if $f(y,\theta_0)$ is a kink function in such a space, the function

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

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 $\tilde{\Theta}$, with $\tilde{\Theta} \subset (\Theta_l)_{l\in I} \subset \mathbb{R}^{d-k}$, where $I \subset \{1,\dots,d\}$ 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 3 are satisfied.

2.3. Assumptions on the model

Assumption A (Assumptions on the error)**.**

A1: the vector $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)$ 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 a 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 $\nu_n := \max_{i=1,\dots,n}|\delta_i| = o_p(1)$. Here $x_{(i)}$ denotes the i-th order statistic of $x_1,\dots,x_n$. Moreover, the design points $x_1,\dots,x_n$ are independent of the error terms $\varepsilon_1,\dots,\varepsilon_n$.

The above assumption covers random designs. If the design points $x_1,\dots,x_n$ are nonrandom, the $o_p(1)$ term above is to be understood as $o(1)$. We do not pursue this situation further, but 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

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

In the following we 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

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

for $\mu \in \mathcal{M}$. 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 $\mathcal{Y}$. 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.

(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 in the consistency proof for the estimator of f_0. Condition (i) especially is 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**.

The results of this paper can also be formulated for $\Phi : L^2([a,b]) \to L^2(I)$, with an interval $I \subset \mathbb{R}$ that need not 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. An estimate of f for given k and $\boldsymbol{F}_k$ is found by taking $\Phi\hat{f}_n$ to minimize 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

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

This estimator depends implicitly on k and $\boldsymbol{F}_k$, but we supress this when no confusion is expected. It then follows from Definition 2 that there exists a parameter vector $\hat{\theta}_n \in \Theta_k$, such that

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

where $\hat{\vartheta}^i$ and $\hat{\tau}_i$ also depend on the index n.

It is easy to see that the minimum at (3.1) is attained, since $\boldsymbol{F}_k$ is closed and compact. It need not be unique. We do not require 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. We give the asymptotic behavior of the 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.

Let $\Lambda : \Theta_k \to L^2([a,b])$ denote the mapping

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

We 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 1. *Suppose that Ass.* ***A1****,* ***B****, and* ***C*** *are satisfied and let* $\hat{f}_n(y) = f(y,\hat{\theta}_n)$ *be the estimator of the true* $f_0 = f(\cdot,\theta_0) \in \boldsymbol{F}_k$ *at* (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) $|\theta_0 - \hat{\theta}_n|_2 = O_P(n^{-1/2})$,

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

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

If f_0 depends on a reduced parameter vector $\tilde{\theta}$ as in Definition 3, 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 and we have the following.

Corollary 1. *Suppose that Ass.* ***A1****,* ***B****, and* ***C*** *are satisfied and that the true* $f_0(y) = f_0(y, h(\tilde{\theta}))$ *can be parameterized by a reduced parameter domain. Then* $V_{\tilde{\theta}}$ *is nonsingular, and the results of Theorem* 1 *are valid with* θ_0 *and* $\hat{\theta}_n$ *substituted by the reduced parameter vector* $\tilde{\theta}_0$ *and its estimator* $\tilde{\theta}_n$*.*

Nonsingularity of the covariance matrix V_{θ_0} is essential for Theorem 1 to hold. We characterize this property in terms of the partial derivatives $\frac{\partial}{\partial\vartheta^i}\mathfrak{f}(y,\vartheta_0^i)$, $i = 1,\dots,k+1$, for the case where $f(\cdot,\theta_0)$ has precisely k jumps.

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

Accordingly, Theorem 1 cannot be applied, if f_0 is a kink function. This case requires restriction to a reduced parameter set $\tilde{\Theta}$. Then 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 $\mathcal{F}$,

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

Corollary 2. *If the conditions of Corollary* 1 *are satisfied and the true* $f_0(y,h(\tilde{\theta}))$ *is a kink function, then the results of Corollary 1 are valid with the improved rate*

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

For example, we obtain rates of order $n^{-1/2}$ if $f\in\tilde{L}_2$. More generally, if $\mathcal{F}$ consists of Hölder continuous functions with exponent $0<\alpha\le 1$, one gets a rate of order $n^{-(1+\alpha)/4}$.

It is straight forward to see that the $\sqrt{n}$ rate in Theorem 1 (i), (ii) is minimax under a normal error for bounded, continuous integral kernels. A similar argument as in the proof of Theorem 1 in Wishart (2011) can be employed to estimate the Kullback Leibler divergence between the distributions of different (Y,X) and apply Theorem 2.2.(iii) in Tsybakov (2009). In a normal model, we claim the asymptotic variance $V_{\theta_0}^{-1}$ in 1 is asymptotically optimal in Le Cam sense, provided the experiment is differentiable in quadratic mean (see van der Vaart (1998)).

The ill posedness of the problem is not reflected in the rate of convergence but rather in the asymptotic variance$V_{\theta_0}^{-1}$, as can be seen from (3.3). The variance is large when 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 1 (iii) and (iv) and in Corollary 2 are minimax but we do not have a proof for this.

3.2. Unknown number of jumps

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

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

where λ_n is some smoothing parameter converging to zero and $\sharp\mathcal{J}(f)$ is taken to be nonzero. 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 2. *Suppose that Ass.* ***A2****,* ***B*** *and* ***C*** *are satisfied. Let* $f_0 \in \boldsymbol{F}_\infty$ *and choose* $\{\lambda_n\}_{n\in\mathbb{N}}$ *such that* $\lambda_n \longrightarrow 0$ *and* $\lambda_n n^{1/(1+\epsilon)} \longrightarrow \infty$ *for some* $\epsilon > 0$*. Then, the minimizer* $\hat{f}_{\lambda_n}$ *of* (3.6) *satisfies* $P(\sharp\mathcal{J}(\hat{f}_{\lambda_n}) = \sharp\mathcal{J}(f_0)) \longrightarrow 1$.

This can be viewed as a model consistency result in that, for n large enough, the correct number of junps/kinks is selected. Viewed as a post model selection estimator, the normal approximation can become unreliable in Theorem 1 (i), since the model selection step may affect the distributional limit from the post model selection estimator (Leeb and Pötscher (2006)). In fact, the convergence in Theorem 2 is nonuniform in the sense that this probabilty will depend on the true underlying function f.

We do not know whether $\lambda_n \sim \log n/n$ would give model consistency as well, the penailzation rate required in Theorem 2 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 do not address this here. 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 Boysen et al. (2009b). In general, a severe computational burden arises in models with many kinks.

3.3. Examples

Example 4 (Hammerstein integral equations)**.** The structure of $\boldsymbol{F}_k[a,b]$ allows extension of the results in Theorem 1, and Corollaries 1 and 2, to a class of nonlinear integral operators of the form

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

known as *Hammerstein integral operators*. We take L to satisfy the following.

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

(2) The 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.

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

It is straightforward to verify that if $\mathcal{L}$ satisfies (1)−(3) and, if $\mathcal{F}$ is a continuous parametric family, 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. That means $\mathcal{L}$ preserves the properties of $f \in \boldsymbol{F}_k[a,b]$ and all results from the preceding section hold for *Hammerstein integral equations* of the first kind, since 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 1, or Corollary 1 or 2, yields an estimator for θ_0, we obtain an estimator for $f(\cdot,\theta_0)$ simultaneously.

Example 5 (Free knot splines)**.** The question of what happens if the true function f in (1.1) is not an element of $\mathbf{F}_k$, has been treated in Boysen, Bruns, and Munk (2009a, Lemma 3.3). In analogy to this, 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.

For the set of piecewise polynomial functions, there is a connection to distributional asymptotics for splines. According to the Curry and Schoenberg Theorem (cf., De Boor (2001, Chapter VIII, (44))), for fixed change points, the set of piecewise polynomials of degree p is the B-spline space of order p with knots in $\{\tau_0,\dots,\tau_{k+1}\}$ with multiplicity p in the case of jumps, and at most $p-1$ in the case of kinks. Here misspecification of the model could be considered as spline approximation of f_0 and this leads to "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 have to deal with *free-knot* splines and these spaces are no longer linear. 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 is in general a challenging task (see e.g., Jupp (1978)). 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). 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 and can be used for cofidence bands (see also Mao and Zhao (2003)). Finally, Theorem 2 gives model selection consistency of knot penalisation in $\mathbf{F}_\infty$, new to our knowledge. Thus from Theorem 2 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 set $\mathbf{F}_\infty$ of free knot splines.

Example 6 (Confidence bands)**.** Theorem 1 (i) implies that the quadratic form

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

is asymptotically χ^2 distributed 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 an approximate $(1-\alpha)$-confidence ellipsoid for $\hat{\theta}_n$ in $\mathbb{R}^d$ is

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

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

We consider product and convolution kernels that assure L^2 injectivity and range inclusions for the corresponding linear integral operator Φ in (1.2) as required by Ass. **C**.

We start with a theorem that 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* in Borwein and Erdélyi (1997, Thm. 6.2):

Lemma 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 3 (product kernels)**.** *Suppose $0 < a < b$ and $0 \leq c < d$ and that $\varphi(x,y) = \phi(xy)$ for some piecewise continuous function $\phi \in L^\infty([ac,bd])$. Then* (i) *and* (ii) *of Ass. **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,* $\Phi \in \mathcal{L}(L^\infty([a,b]), C^{0,1}([c,d]))$ *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* (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 is discussed in Section 5.2. The Gaussian kernel, $\phi(x) = (2\pi\sigma^2)^{-1/2} e^{-(x/\sigma)^2/2}$, is a well known example of a function satisfying the assumptions of Theorem 3.

Theorem 4 (positive definite convolution kernels)**.** *Suppose* $\varphi(x,y) = \phi(x-y)$ *for all* $x,y \in [a,b]$ *for some function* $\phi \in C(\mathbb{R}) \cap L^1(\mathbb{R})$. *Then* (i) *and* (ii) *of Ass.* ***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 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 5 (analytic convolution kernels)**.** *Suppose* $\varphi(x,y) = \phi(x-y)$ *for* $x,y \in [a,b]$ *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 Ass.* ***C***.

5. Simulations and Data Example

5.1. Example: Inverse two phase regression

To evaluate the speed of convergence and quality of the approximation by the asymptotic law given in Theorem 1, we did a simulation study with the true

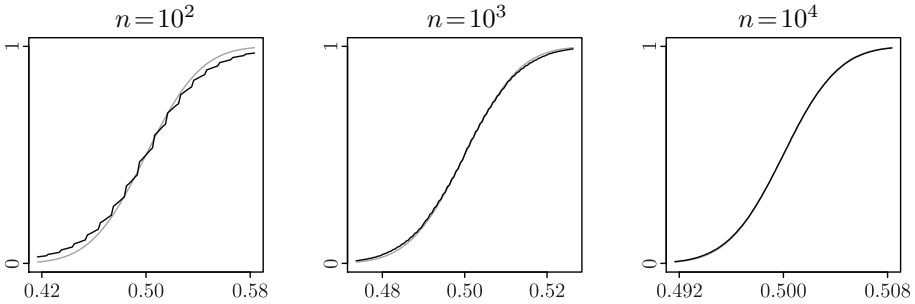

Figure 1. Asymptotic and finite sample size distribution of the jump location for different sample sizes n. 10^5 simulation runs with data generated according to (5.1) were performed. The finite sample size distribution is given by the black line and the asymptotic distribution by the gray line.

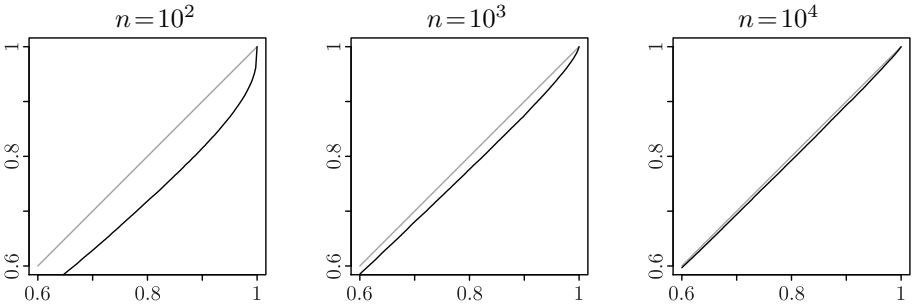

Figure 2. Empirical coverage probability for different sample sizes n of confidence bands for the estimated jump location. 10^5 simulation runs with data generated according to (5.1) were performed. The x-axis shows the nominal and the y-axis the empirical coverage.

function $f_0 \in \tilde{\mathbf{T}}_1$, a step function with one jump given by the parameter vector $\theta_0 = (b_1, \tau, b_2) = (-3, 1/2, 3)$. We generated the observations Y by

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

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 $i = 1, \ldots, n$. Theorem 1 yields

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

where the (non-singular) covariance in (3.3) is

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

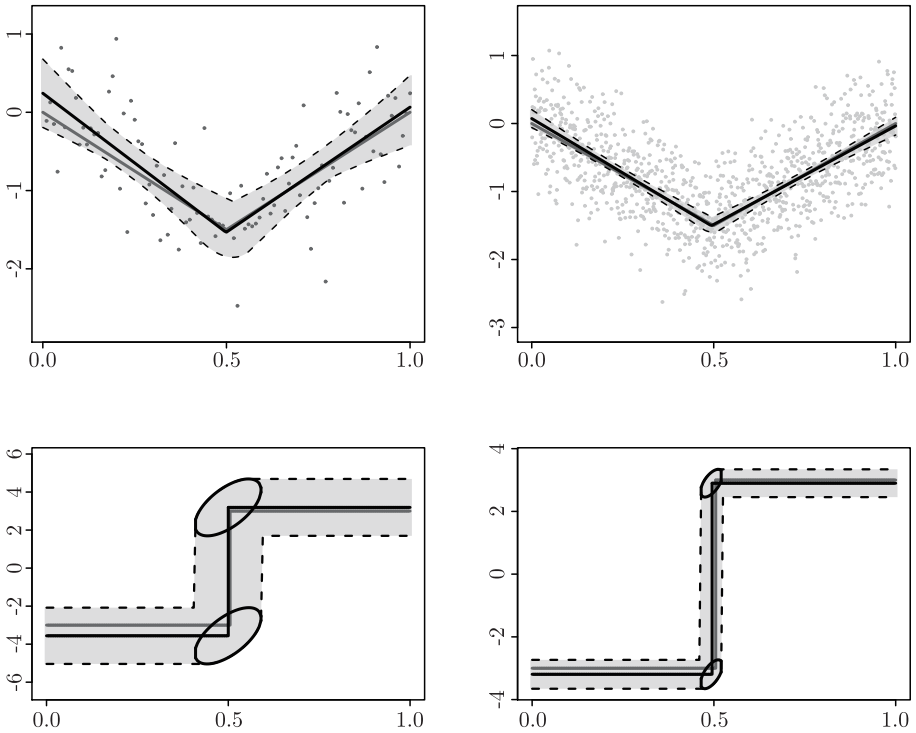

Figure 3. Simulated data examples and confidence bands for the two phase regression with $n = 100$ (left) and n =1,000 (right) observations. The first row displays the observations and the reconstruction in the image space, and the second row shows the estimate for the signal f. The gray line represents 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.

In particular for the jump location we obtain

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

This was used to calculate confidence intervals for $\hat{\tau}$. Figure 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 2.

As described in Example 6 we can calculate confidence bands for the estimated function $\hat{f}_n$ as well as for its image $\Phi\hat{f}_n$. Figure 3 shows two simulated data sets, including their 95%-confidence regions, for $n = 100$ and n =1,000.

5.2. Example: inverse multiphase regression

In this subsection we discuss an application of Corollary 2 to a problem from rheology. The aim here is the determination of the so called *relaxation time spectrum* (see Roths et al. (2000)). This is a characteristic quantity used in rheology that describes the viscoelastic properties of polymer solutions and polymer melts. Given the spectrum, it is very easy to convert one material function into another. 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 (see Roths et al. (2000)).

Definition 4. Let $0 < a < 1 < b < \infty$ and $c \neq 0$. The *relaxation time spectrum transform* is given as

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

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

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

The exponential operator $\mathcal{L}$ satisfies the assumptions claimed in Example 4. Furthermore, the integral operator Φ satisfies Assumption $\boldsymbol{C}$ by virtue of Theorem 3.

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 $\tilde{\theta} = (\vartheta_1^1, \vartheta_2^1, \tau_1, \vartheta_2^2, \tau_2, \vartheta_2^3)$. 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. The true function $f_0(y) = f(y,\theta_0)$ we intend to estimate is an element of the set

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

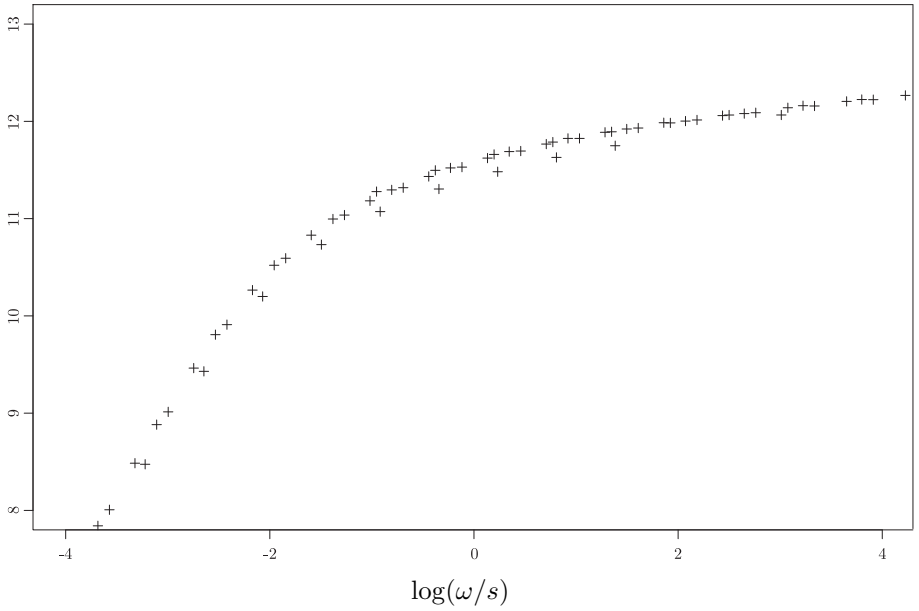

Figure 4. A log − log- plot of the ω frequency of harmonic stress (x-axis) against the dynamic stress moduli of a polymer melt.

which satisfies the conditions of Definition 2. In Roths et al. (2000) it is assumed that the observation model coincides with (1.1) with f_0 substituted by $\mathcal{L}f_0$, namely

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

where Ass. ***A1*** and **B** on error and design are fulfilled. Figure 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 et al. (2000) for details). For estimation we use the estimator at (3.1). Then, application of Corollary 2 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

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

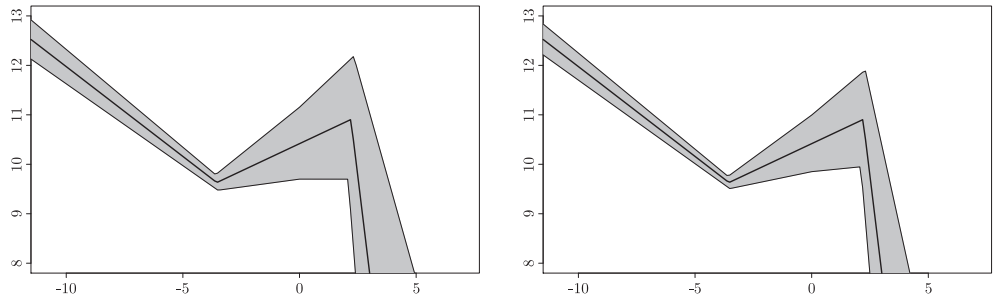

Figure 5. From l.t.r.: 0.95- and 0.80-confidence bands for the estimated kink function $\hat{f}_n$ of the log relaxation time spectrum with two (typical) kinks plotted on a log scale

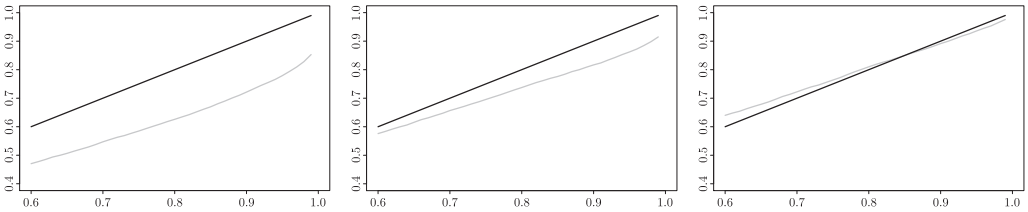

Figure 6. Empirical coverage probability (grey lines) of confidence bands for the estimated kink function for normal observations with $\sigma^2 = 0.01$ for different sample sizes. From the left, $n = 100$, 1,000, 5,000, and 10^4 simulations each. The x-axis shows the nominal and the y-axis the empirical coverage probability. The black line $x = y$ is for comparison, it shows perfect coincidence of empirical and nominal coverage.

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 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.

The results of Theorem 1 hold with the improved rate

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

which for comparison is the square of the rate in (iv) of Theorem 1. Equation (5.4) directly follows from Corollary 2, since linear functions with bounded

slopes are Lipschitz continuous with a 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 shows the estimated kink function for the polymer melt data of the relaxation time spectrum from dynamic moduli (see Roths et al. (2000)), with 95%- and 80%-confidence bands calculated by using a studentized maximum modulus statistic (Miller (1966)).

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 6). Here we used the operator in Definition 4 acting on the space of kink functions with two kinks. A comparison of Figure 2 and 6 illustrates that increasing complexity of the kernel in Subsection 5.2 reduces the finite sample accuracy of the empirical coverage probability.

Acknowledgement

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*. Springer-Verlag, New York.

Borwein, P. and Erdélyi, T. (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*. Springer-Verlag, New York.

Ferry, J. D. (1970). *Viscoelastic Properties of Polymers* 2nd edition. Wiley, New York.

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. Roy. Statist. Soc. Ser. B* **65**, 901-919.

Miller, Jr., R. G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.

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, Reading, Mass.

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, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

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

Wishart, J. (2011). Minimax lower bound for kink location estimators in a nonparametric regression model with long range dependence. *Statist. Prob. Lett.* **81**, 1871-1875.

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

E-mail: sophie.frick@gmx.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

(Received January 2012; accepted February 2013)