Piecewise Convex Function Estimation and Model Selection

Kurt Stewart Riedel
Millennium Partners

arXiv:1803.03903v1 [stat.ME] 11 Mar 2018

Piecewise Convex Function Estimation and Model Selection

Kurt S. Riedel

Abstract. Given noisy data, function estimation is considered when the unknown function is known *a priori* to consist of a small number of regions where the function is either convex or concave. When the regions are known *a priori*, the estimate is reduced to a finite dimensional convex optimization in the dual space. When the number of regions is unknown, the model selection problem is to determine the number of convexity change points. We use a pilot estimator based on the expected number of false inflection points.

§1. Introduction

Our basic tenet is: "Most real world functions are piecewise ℓ-convex with a small number of change points of convexity." Given N measurements of the unknown function, $f(t)$, contaminated with random noise, we seek to estimate $f(t)$ while preserving the geometric fidelity of the estimate, $\hat{f}(t)$, with respect to the true function. In other words, the number and location of the change points of convexity of $\hat{f}(t)$ should approximate those of $f(t)$.

We say that $f(t)$ has k change points of ℓ-convexity with change points $x_1 \leq x_2 \ldots \leq x_k$ if $(-1)^{k-1} f^{(\ell)}(t) \geq 0$ for $x_k \leq t \leq x_{k+1}$. For $\ell = 0$, $f(t)$ is nonnegative and for $\ell = 1$, the function is nondecreasing. In regions where the constraint of ℓ-convexity is active, $f^{(\ell)}(t) = 0$ and $f(t)$ is a polynomial of degree $\ell - 1$. For 1-convexity, $f(t)$ is constant in the active constraint regions and for 2-convexity, the function is linear. Our subjective belief is that most people prefer smoothly varying functions such as quadratic or cubic polynomials even in the active constraint regions. Thus, piecewise 3-convexity or 4-convexity are also reasonable hypotheses. The idea of constraining the function fit to preserve ℓ-convexity properties has been considered by a number of authors. The more difficult problems of determining the number and location of the ℓ-convexity breakpoints will be a focus of this article. We refer to the estimation of the number of change points as the "model selection problem" because it resembles model selection in an infinite family of parametric models.

Approximation Theory VIII
Charles K. Chui and Larry L. Schumaker (eds.), pp. 0—1.

ISBN 0-12-xxxxxx-x

§2 Convex Analysis

In this section, we assume that the change points $\{x_1 \ldots x_k\}$ of ℓ-convexity are given and that the function is in the Sobolev space, $W_{m,p}[0,1]$ with $m \geq \ell$ and $1 < p < \infty$ where

$$W_{m,p} = \{f | f^{(m)} \in L_p[0,1] \text{ and } f, f' \ldots f^{(m-1)}(t) \ \text{ absolutely continuous}\} \ .$$

We decompose $W_{m,p}$ into a direct sum of the space of polynomials of degree $m-1$, P_{m-1} plus the set of functions whose first $m-1$ derivatives vanish at $t=0$ which we denote by $W^0_{m,p}$ [10].

Given change points, $\{x_1, x_2 \ldots x_k\}$, we define the closed convex cone

$$V^{k,\ell}_{m,p}[x_1, \ldots, x_k] = \{f \in W_{m,p} \mid (-1)^{k-1} f^{(\ell)}(t) \geq 0 \ \text{ for } \ x_{k-1} \leq t < x_k\} \ .$$

Let $\mathbf{x}$ denote the k row vector, $(x_1, x_2 \ldots x_k)$. We define the class of functions with at most k change points as

$$V^{k,\ell}_{m,p} \equiv \bigcup_{x_1 \leq x_2 \ldots \leq x_k} \left\{ V^{k,\ell}_{m,p}[x_1, \ldots, x_k] \cup (-V^{k,\ell}_{m,p}[x_1, \ldots, x_k]) \right\} \ .$$

By allowing $x_{k'} = x_{k'+1}$, we have embedded $V^{k,\ell}_{m,p}$ into $V^{k+1,\ell}_{m,p}$. $V^{k,\ell}_{m,p}$ is the union of convex cones, and is closed but not convex. For the case $p = \infty$, similar piecewise ℓ-convex classes are defined in [2]. To decompose $W_{m,p}$ in terms of $V^{k,\ell}_{m,p}$, we require that each function in $W_{m,p}$ has a piecewise continuous ℓ-th derivative. By the Sobolev embedding theorem, this corresponds to the case $m \geq \ell + 1$.

Let $\|f\|^p_{j,p} \equiv \int_0^1 |f^{(j)}(t)|^p dt$. We endow $W_{m,p}$ with the norm:

$$|||f|||^p_{m,p} = \sum_{j=0}^{m-1} |f^{(j)}(t=0)|^p + \|f\|^p_{m,p} \ .$$

The dual space of $W_{m,p}$ is isomorphic to the direct sum of P_{m-1} and $W^0_{m,q}$ with $q = p/(p-1)$ and the duality pairing:

$$\langle\langle g, f \rangle\rangle = \sum_{j=0}^{m-1} b_j a_j + \int_0^1 f^{(m)}(t) g^{(m)}(t) dt \ . \tag{2.1}$$

In (2.1), $f \in W_{m,p}$, $g \in W_{m,q}$, $a_j \equiv f^{(j)}(0)$ and $b_j \equiv g^{(j)}(0)$. We denote the duality pairing by $\langle\langle \cdot \rangle\rangle$ and the L_2 inner product by $\langle \cdot \rangle$. The space $W_{m,p}$ has a reproducing kernel, $R(t,s)$, such that for each t, $f(t) = \langle\langle R_t, f \rangle\rangle$ [10]. A linear operator, $L_i^* \in W^*_{m,p}$ has representations $L_i f = \langle\langle L_i R(\cdot, s), f(s) \rangle\rangle$ and $L_i f = \langle L_i \delta(s - \cdot), f(s) \rangle$.

We are given n measurements of $f(t)$:

$$y_i = L_i f + \epsilon_i = \langle h_i, f \rangle + \epsilon_i = \langle\langle m_i, f \rangle\rangle + \epsilon_i \ , \tag{2.2}$$

where $L_i R(\cdot, s)$ are linear operators in $W^\perp_{m,p}$, and the ϵ_i are independent, normally distributed random variables with variance $\sigma_i^2 > 0$. We represent L_i as $m_i(s) \equiv L_i R(\cdot, s)$ and $h_i(s) \equiv L_i \delta(s - \cdot)$ and assume $h_i \in W^\perp_{\ell,1}$. In the standard case where $y_i = f(t_i) + \epsilon_i$, $m_i(s) = R(t_i, s)$ and $h_i(s) = \delta(s - t_i)$.

A robustified estimate of $f(t)$ given the measurements $\{y_i\}$ is $\hat{f} \equiv$ argmin VP$[f \in V_{m,p}^{k,\ell}[x_1, \ldots, x_k]]$:

$$\mathrm{VP}[f] \equiv \frac{\lambda}{p} \int |f^{(m)}(s)|^p ds + \sum_{i=1}^{N} \psi_i \left(\langle h_i, f \rangle - y_i \right) \ , \qquad (2.3)$$

where the ψ_i are strictly convex, continuous functions. The standard case is $p = 2$ and $\psi_i(y_i - \langle h_i, f \rangle) = |y_i - f(t_i)|^2/n\sigma_i^2$. The set of $\{h_i, i = 1, \ldots, N\}$ separate polynomials of degree $m-1$ means that $\langle h_i, \sum_{k=0}^{m-1} c_k t^k \rangle = 0$, $\forall i$ implies $c_k \equiv 0$.

Theorem 1. *Let $\{h_i\}$ separate polynomials of degree $m-1$, then the minimization problem (2.3) has an unique solution in $V_{m,p}^{k,\ell}[\mathbf{x}]$ and the minimizing function is in $C^{2m-\ell-2}$ and satisfies the differential equation:*

$$(-1)^m d^m [|\hat{f}^{(m)}|^{p-2} \hat{f}^{(m)}(t)] + \sum_{i=1}^{n} \psi_i'(\langle h_i, \hat{f} \rangle - y_i) h_i(t) = 0 \ , \qquad (2.4)$$

in those regions where $|f^{(\ell)}| > 0$ for $1 < p < \infty$.

Proof: The functional (2.3) is strictly convex, lower semicontinuous and coercive, so by Theorem 2.1.2 of Ekeland and Temam, it has a unique minimum, f_0, on any closed convex set. From the generalized calculus of convex analysis, the solution satisfies

$$0 \in (-1)^m d^m [(|f^{(m)}|^{p-2} f^{(m)}(t)] \Sigma \psi_i'(\langle h_i, \hat{f} \rangle - y_i) h_i(t) + \partial N_V(f) \qquad (2.4)$$

where $N_V(f)$ is the normal cone of $V_{m,p}^{k,\ell}[\mathbf{x}]$ at f [1, p. 189]. From [9], each element of $N_V(f)$ is the limit of a discrete sum: $\sum_t a_t \delta^{(\ell)}(\cdot - t)$ where the $t's$ are in the active constraint region. Integrating (2.4) yields

$$|f^{(m)}|^{p-2} f^{(m)}(t) = \sum_{i=1}^{n} \frac{\psi_i'(\langle h_i, \hat{f} \rangle - y_i) \langle h_i(s), (s-t)_+^{m-1} \rangle}{(m-1)!} + \int \frac{(s-t)_+^{m-\ell-1} d\mu(s)}{(m-\ell-1)!} \ , \qquad (2.5)$$

where $d\mu$ corresponds to a particular element of $N_V(f)$. Since $(s-t)_+^{m-\ell-1}$ is $m-\ell-2$ times differentiable, the right hand side of (2.5) is $m-\ell-2$ times differentiable. Integrating (2.5) yields $f \in C^{2m-\ell-2}$. ■

The intervals on which $f^{(\ell)}(t)$ vanishes are unknowns and need to be found as part of the optimization. Using the differential characterization (2.3) loses the convexity properties of the underlying functional. For this reason, extremizing the dual functional is now preferred.

Theorem 2. *The dual variational problem is: Minimize over $\alpha \in \mathbb{R}^n$*

$$\mathrm{VP}^*[\alpha; \mathbf{x}] \equiv \frac{\lambda^{1-q}}{q} \int |[\mathbf{P_x}^* M \alpha]^{(m)}(s)|^q ds + \sum_{i=1}^{n} \psi_i^*(\alpha_i) - \alpha_i y_i \ , \qquad (2.6)$$

where $M\alpha(t) \equiv \sum_i m_i(t)\alpha_i$ *and* ψ_i^* *is the Fenchel/Legendre transform of* ψ_i. *The dual projection* $\mathbf{P_x}^*$ *is defined as*

$$\int |[\mathbf{P_x}^* g]^{(m)}(s)|^q ds \equiv \inf_{\tilde{g}\in V^-} \int_0^1 |g^{(m)} - \tilde{g}^{(m)}(s)|^q \ , \qquad (2.7)$$

where the minimization is over $\tilde{g}$ *in the dual cone subject to* $g^{(j)}(0) = \tilde{g}^{(j)}(0), \quad 0 \le j < m$. *The dual problem is strictly convex and its minimum is the negative of the infimum of (2.3).*

Proof: Let ψ_V be the indicator function of $V^{k,\ell}_{m,p}[\mathbf{x}]$ and define

$$U(f) = \frac{\lambda}{p} \int_0^1 |f^{(m)}(s)|^p ds + \psi_V(f) \ . \qquad (2.8)$$

We claim that the Legendre transform of $U(f)$ is the first term in (2.6). Note that $\psi_V^*(g) = \psi_{V^-}(g)$, the indicator function of the dual cone V^-. Since the Legendre transform of the first term in (2.8) is

$$V_1^*(g) = \frac{\lambda^{1-q}}{q} \int_0^1 |g^{(m)}(s)|^q ds \ \text{ for } \ g \in W^0_{m,q}, \quad \text{and} \ \ \infty \ \ \text{otherwise.}$$

Our claim follows from $[U_1 + U_2]^*(g) = \inf_{g'} \{U_1^*(g - g') + U_2^*(g')\}$. The remainder of the theorem follows from the general duality theorem of Aubin and Ekeland [1, p. 221]. ■

For the case $\ell = m$, the minimization over the dual cone can be done explicitly. For $\ell < m$, Theorem 1 is proven in [9] and Theorem 2 is proven in [6] for the case $p = 2$ and $\psi(y) = y^2$. Equation (2.5) and the corresponding smoothness results appear in [9] for the case $\ell = 1$, $p = 2$ and $L_i = \delta(t - t_i)$.

§3. Change point estimation

When the number of change points is fixed, but the locations are unknown, we can estimate them by minimizing the functional in (2.3) with respect to the change point locations. We now show that there exists a set of minimizing change points.

Theorem 3. *For each* k, *there exist change points* $\{x_j, \ j = 1, \dots k\}$ *that minimize the variational problem (2.3).*

Proof: We use the dual variational problem (2.5) and maximize over $\mathbf{x} \in [0,1]^k$ after minimizing over the $\alpha \in R^N$. The functional (2.5) is jointly continuous in $\alpha, \mathbf{x}$ and convex in α. Theorem 3 follows from the min-max theorem [1,p. 296]. ■

The change point locations need not be unique. The proof requires $\le$ instead of $<$ in the ordering $x_j \le x_{j+1}$ to make the change point space compact. When $x_j = x_{j+1}$, the number of effective change points is less than k. Finding the $\mathbf{x}$ that minimizes VP* is computationally intensive and requires the solution of a convex programming problem at each step. Theorems 1-3 are valid when $\ell \le m$ including $\ell = m$. Restricting to $p = 2$, we have the following theorem from [9]:

Theorem 4. [Utreras] *Let f be in a closed convex cone, $V \subseteq W_{m,2}$, let $\hat{f}_u$ be the unconstrained minimizer of (2.3) given y_i and $\hat{f}_c$ be the constrained minimizer (with $p = 2$ and $\psi_i(y) = |y|^2/\sigma_i^2$). Then $\|f - \hat{f}_c\|_V \leq \|f - \hat{f}_u\|_V$ where $\|f\|_V^2 \equiv \frac{\lambda}{2}\int |f^{(m)}(s)|^2 ds + \sum_{i=1}^N \psi_i(L_i f)$.*

Theorem 4 shows that if one is certain that f is in a particular closed convex cone, the constrained estimate is always better than the unconstrained one. Unfortunately Theorem 4 does not generalize to unions of convex cones and thus does not apply to $V_{m,2}^{k,\ell}$.

§4. Number of false inflection points

We now consider unconstrained estimates of $f(t)$ and examine the number of false ℓ-inflection points. We assume that the noisy measurements of f occur at nearly regularly spaced locations, t_i (with $h_i(t) = \delta(t - t_i)$). Specifically, we assume that $d_n \equiv sup_t\{F_n(t) - F(t)\}$ tends to zero as n^{-b} with $b \geq 0$ where $F_n(t)$ is the empirical distribution of the t_i and $F(t)$ is the limiting distribution. For regularly spaced points, $d_n \sim 1/n$. This nearly regularly spaced assumption allows us to approximate the discrete sums over the t_i by integrals.

A smoothing kernel estimate of $f^{(\ell)}(t)$ is a weighted average of the y_i:

$$\widehat{f^{(\ell)}}(t) = \sum_i y_i \kappa(\frac{t - t_i}{h_n}) \left[\frac{t_{i+1} - t_{i-1}}{2}\right] \ , \tag{4.1}$$

where h_n is the kernel halfwidth and κ is the kernel. κ is required to satisfy the moment conditions: $\int_{-1}^{1} s^j \kappa(s) ds = \ell! \delta_{j,\ell}$, $0 \leq j < \ell + 2$, with $\kappa \in C^2[-1,1]$ and that $\kappa(\pm 1) = \kappa'(\pm 1) = 0$. We call such functions-C^2 extended kernels. When $f \in C^m$, the optimal halfwidth scales as $h_n \sim n^{-1/(2m+1)}$, and the optimal spline smoothing parameter scales as $\lambda_n \sim n^{-2m/(2m+1)}$. In [5], Mammen et al. derive the number of false inflection points for kernel estimation of a probability density. We present the analogous result for regression function estimation. The proofs in our case are easier because we need only show that discrete sums converge to their limits. Our results are for arbitrary ℓ while [4,5] considered $\ell = 1, 2$.

Theorem 5. (Analog of [4,5]) *Let $f(t) \in C^{\ell+1}[a, b]$ have K ℓ-inflection points $\{x_1, \ldots x_k\}$ with $f^{(\ell+1)}(x_j) \neq 0$, $f^{(\ell)}(a) \neq 0$ and $f^{(\ell)}(b) \neq 0$. Consider a sequence of kernel smoother estimates with C^2 extended kernels. Let the sequence of kernel halfwidths, h_n, satisfy $0 < \liminf_n h_n n^{1/(2\ell+3)} \leq \limsup_n h_n n^{1/(2\ell+3)} < \infty$, then the expected number of ℓ-inflection points is*

$$\mathbf{E}[\hat{K}] - K = \ 2\sum_{j=1}^{K} H\left(\sqrt{\frac{nh^{2\ell+3}|f^{(\ell+1)}(x_j)|^2}{\sigma^2\|\kappa^{(\ell+1)}\|F'(x_j)}}\right) \ , \tag{4.2}$$

where $\sigma^2 = \mathbf{Var}[\epsilon_i]$*,* $H(z) \equiv \phi(z)/z + \Phi(z) - 1$ *with* ϕ *and* Φ *being the Gaussian density and distribution provided that* $d_n < n^{-1/2}$*.*

Proof: The proof consists of applying the Cramér-Leadbetter zero-crossing formula to (4.1) and then taking the limit as $n \to \infty$.

Theorem (Cramér-Leadbetter) *Let* N *be the number of zero crossings of a differentiable Gaussian process,* $Z(t)$*, in the time interval [0,T]. Then*

$$\mathbf{E}[N] = \int_0^T \frac{\gamma(s)\rho(s)}{\sigma(s)} \phi\left(\frac{m(s)}{\sigma(s)}\right) G(\eta(s))ds \ , \tag{4.3}$$

where $\sigma^2(s) = \mathbf{Var}[Z(s)]$*,* $\gamma^2(s) = \mathbf{Var}[Z'(s)]$*,* $\mu(s) = \mathbf{Corr}[Z(s)Z'(s)]$*,* $\rho(s)^2 = 1 - \mu(s)^2$*,* $m(s) = \mathbf{E}[Z(s)]$*,* $\eta(s) = \frac{m'(s) - \gamma(s)\mu(s)m(s)/\sigma(s)}{\gamma(s)\rho(s)}$*.*

We claim that for (4.1), $\sigma_n^2(s) \to \sigma^2 \|\kappa^{(\ell)}\|^2 F'(s)/nh^{2\ell+1}$, $\gamma_n^2(s) \to \sigma^2 \ \|\kappa^{(\ell+1)}\|^2 \ F'(s)/nh^{2\ell+3}$, $\mu_n(s) \to \mathcal{O}_R(h_n + 1/nh^{2\ell+1})$, $\rho_n(s)^2 \to 1$, $m_n(s) \to f^{(\ell)}(s) + \mathcal{O}(h_n + 1/nh^{\ell+1})$. To show the convergence of the discrete sums to integrals, we use $\int g(s)ds = \sum_i g(t_i)[t_{i+1} - t_{i-1}]/2 + \mathcal{O}_R(sup_i[t_{i+1} - t_{i-1}]^2/h_n^2)$ where $\mathcal{O}_R$ denotes a relative size of $\mathcal{O}$. More detailed proofs of the convergence of the discrete sums to integrals can be found in [2]. Since the integrand in (4.3) is bounded and converging pointwise, the dominated convergence theorem shows that the sequence of integrals given by (4.3) converges. Equation (4.2) follows by evaluating the integral using the method of steepest descent. ■

Corollary. *Let* $f(t) \in C^{\ell+1}[a,b]$*,* $d_n/h_n^{\ell+1} \to 0$ *and* $nh_n^{2\ell+3} \to \infty$ *with* κ *a* C^2 *extended kernel. The probability that* $\widehat{f^{(\ell+1)}}$ *has a false inflection point outside of a width of* δ *from the actual* $(\ell+1)$*-inflection points is* $\mathcal{O}(\exp(-nh_n^{2\ell+3}))$*.*

For the case $p = 2$, the smoothing spline estimate is a linear estimate of the form $\widehat{f^{(\ell)}}(t) = \sum_i y_i g_{n,\lambda}(t, t_i)$ where $g_{n,\lambda}(t, t_i)$ solves the equation: $(-1)^m \lambda_n g_{n,\lambda}^{(2m)}(t, s) + \sum_i^n g_{n,\lambda}(t_i, s) = \delta(t - s)$, with the boundary conditions, $\partial_t^j g_{n,\lambda}(0, s) = 0 = \partial_t^j g_{n,\lambda}(1, s)$ for $m \le j < 2m$.

Theorem 6. [Silverman] *Let* $\lambda_n^{-1/2m} d_n \to 0$*,* $|F''(t)| < \infty$ *and* $0 < c_1 < F'(t) < c_2$*. The Green's function,* $g_{n,\lambda}(t, t_i)$*, of the smoothing spline converges to a kernel function with the halfwidth,* $h(t) = [\lambda F'(t)]^{1/2m}$*:*

$$h(t) \ \partial_s^j g_{n,\lambda}(t + h(t)s, t) \to \partial_s^j \kappa(s)/F'(t) \ , \quad 0 \le j < m$$

where the equivalent kernel satisfies $(-1)^m \kappa^{(2m)}(t) + \kappa(t) = \delta(t)$ *with decay at infinity boundary conditions. The convergence is uniform for in any closed subdomain,* $t \in [\delta, 1-\delta]$ *and* $t + h(t)s \in [0, 1]$*.*

Although [7] considers only $m = 2$, the proof easily extends to $m > 2$. Using this convergence result, Theorem 5 also holds for smoothing splines:

Theorem 7.*For a sequence of smoothing spline estimates of* f *as given by Thm. 6, Eq. (4.2) holds provided that the smoothing parameters satisfy*

$0 < \mathrm{liminf}_n \lambda_n^{1/2m} n^{1/(2\ell+3)} \leq \mathrm{limsup}_n \lambda_n^{1/2m} n^{1/(2\ell+3)} < \infty$ and $\ell < 2m - 5/2$.

§5. Data-based Pilot Estimators with Geometric Fidelity

We consider two step estimators that begin by estimating $f^{(\ell)}$ and $f^{(\ell+1)}$ using an unconstrained estimate with $h_n \sim log^2(n) n^{1/(2\ell+3)}$. In the second step, we perform a constrained fit, at some locations requiring $\widehat{f^{(\ell)}}$ to be monotone and in other regions requiring $\widehat{f^{(\ell-1)}}$ to be monotone. From the pilot estimate, we determine the number, $\hat{k}$, and approximate locations of the inflection points. At each empirical inflection point, $\hat{x}_j$, we define the α uncertainty interval by $[\hat{x}_j - z_\alpha \hat{\sigma}(\hat{x}_j), \hat{x}_j + z_\alpha \hat{\sigma}(\hat{x}_j)]$, where $\hat{\sigma}^2(\hat{x}_j) = \sigma^2 \|\kappa^{(\ell)}\|^2 F'(s) / |\widehat{f^{(\ell+1)}}(\hat{x}_j)|^2 n h^{2\ell+1}$ and z_α is the two sided α-quantile for a normal distribution.

If an even number of uncertainty intervals overlap, we constrain the fit such that $\widehat{f^{(\ell)}}$ to be positive/negative in each interval. If an odd number of uncertainty intervals overlap, we constrain the fit such that $\widehat{f^{(\ell+1)}}$ to be positive/negative in a subregion of the uncertainty interval which contains an even number of inflection points of $\widehat{f^{(\ell+1)}}$. (The sign of $\widehat{f^{(\ell)}}$ or $\widehat{f^{(\ell+1)}}$ is chosen to match the outer region.) Asymptotically, the uncertainty intervals do not overlap and we constrain the fit such that $\widehat{f^{(\ell+1)}}$ is positive/negative in each uncertainty interval.

Theorem 8. *Consider a two stage estimator that with probability, $1 - \mathcal{O}(p_n)$, correctly chooses a closed convex cone V, with $f \in V$, in the first stage and then performs a constrained regression as in (2.3) with $p = 2$. For $f \in W_{m,2}$, under the restrictions of Thm 4.4 of [9], the estimate, $\hat{f}$, converges as $\mathbf{E}\|\hat{f} - f\|_j^2 \sim \alpha_j \lambda^{(m-j)/m} \|f\|_m^2 + \beta_j \sigma^2 / (n \lambda^{\frac{2k+1}{2m}})$, where n is large enough that $\lambda_n \|f\|_m^2 > p_n (1/n\sigma^2) \sum_i |f(t_i)|^2$ and $p_n n \lambda^{\frac{1}{2m}} \to 0$.*

Proof: If the constraints are correct, Theorem 4 yields the asymptotic error bound [9]. We need to show that misspecified models do not contribute significantly to the error. If the model is misspecified, then $\|\hat{f} - f\|_V \leq \lambda(\|f\|_m + \|\hat{f}\|_m) + (1/n\sigma^2) \sum_i |\hat{f}(t_i) - f(t) - \epsilon_i|^2 + \epsilon_i^2 \leq \lambda \|f\|_m + (1/n\sigma^2) \sum_i (y_i^2 + \epsilon_i^2) \leq \|f\|_V + 1.1 \chi_n^2(p_n)/n$. The expected error is

$$\mathbf{E}\|\hat{f} - f\|_j^2 \leq \mathbf{E}_{f \in V} \|\hat{f} - f\|_j^2 + p_n \mathbf{E}_{f \notin V} \|\hat{f} - f\|_j^2 \ .$$

Note Theorem 4.4 of [9] applies to both pieces. Asymptotically as $p_n \to 0$, $\chi_n(p_n) \leq 1.5n + \mathcal{O}(p_n)$, where $\chi_n^2(p_n)$ is defined by $\int_\chi^\infty dp_{\chi_n} = p_n$. ■

A similar result is given in [3] for the case of constrained least squares. The trick of Theorem 8 is to constrain $\widehat{f^{(\ell+1)}}$ to be positive (or negative) in the uncertainty interval of the estimated inflection points rather than constraining $\widehat{f^{(\ell)}}$ to have a single zero around $\hat{x}_j$.

We recommend choosing the first stage halfwidth, h_n proportional to the halfwidth chosen by generalized crossvalidation (GCV): $h_n = \iota(n) h_{GCV}$ where $\iota(n) = log^2(n) n^{1/(2\ell+3)-1/(2m+1)}$. The second stage smoothing parameter, λ_n, is chosen to be the GCV value $\lambda_n = \lambda_{GCV}$. Other schemes [8] choose the final smoothing parameter to be the smallest value that yields only k inflection points in an unconstrained fit. Since spurious inflection points asymptotically occur only in a neighborhood of an actual inflection point, these earlier schemes oversmooth away from the actual inflection points. In contrast, our second stage use the asymptotically optimal amount of smoothing while preserving geometric fidelity.

Acknowledgments. Work funded by U.S. Dept. of Energy Grant DE-FG02-86ER53223..

References

1. Aubin, J.-P. and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley, New York 1984.
2. Gasser, Th. and Müller, H., Estimating functions and their derivatives by the kernel method, Scand. J. of Stat. **11** (1984), 171–185.
3. Mammen, E., Nonparametric regression under qualitative smoothness assumptions, Ann. Statist., **19** (1991), 741-759.
4. Mammen, E., On qualitative smoothness of kernel density estimates, Univeristy of Heidelberg Report 614.
5. Mammen, E., J.S. Marron and N.J. Fisher, Some asymptotics for multimodal tests based on kernel density estimates, Prob. Th. Rel. Fields, **91** (1992), 115-132.
6. Michelli, C. A., and F. Utreras, Smoothing and interpolation in a convex set of Hilbert space, SIAM J. Stat. Sci. Comp. **9** (1985), 728-746.
7. Silverman, B. W., Spline smoothing: the equivalent variable kernel method, Ann. Stat. **12** (1984), 898-916.
8. Silverman, B. W., Some properties of a test for multimodality based on kernel density estimates, in "Probability, Statistics and Analysis", J. F. C. Kingman and G. E. H. Reuter eds., pp. 248-259. Cambridge University Press, 1983.
9. Utreras, F., Smoothing noisy data under monotonicity constraints - Existence, characterization and convegence rates, Numerische Math. **47** (1985), 611-625.
10. Wahba, G., *Spline Models for Observational Data*, SIAM, Philadelphia, PA 1991.

Kurt S. Riedel
New York University, Courant Institute
251 Mercer St., New York, NY 10012
riedel@cims.nyu.edu