The Annals of Statistics
1997, Vol. 25, No. 6, 2592–2606

PIECEWISE CONVEX FUNCTION ESTIMATION: PILOT ESTIMATORS[1]

By Kurt S. Riedel

Courant Institute of Mathematical Sciences

Given noisy data, function estimation is considered when the unknown function is known a priori to be either convex or concave on each of a small number of regions where the function. When the number of regions is unknown, the model selection problem is to determine the number of convexity change points. For kernel estimates in Gaussian noise, the number of false change points is evaluated as a function of the smoothing parameter. To insure that the number of false convexity change points tends to zero, the smoothing level must be larger than is generically optimal for minimizing the mean integrated square error (MISE). A two-stage estimator is proposed and shown to achieve the optimal rate of convergence of the MISE. In the first stage, the number and location of the change points is estimated using strong smoothing. In the second stage, a constrained smoothing spline fit is performed with the smoothing level chosen to minimize the MISE. The imposed constraint is that a single change point occur in a region about each empirical change point from the first-stage estimate. This constraint is equivalent to the requirement that the third derivative of the second-stage estimate has a single sign in a small neighborhood about each first-stage change point. The change points from the second stage are first-stage change points, but need not be at the identical locations.

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 $x_1 \leq x_2 \leq \cdots \leq x_K$ if $(-1)^{k-1} f^{(l)}(t) \geq 0$ or ≤ 0 for $x_k \leq t \leq x_{k+1}$.

The idea of constraining the function fit to preserve *prescribed* l-convexity properties has been considered by a number of authors [6, 10, 18, 19]. The more difficult problems of determining the number and location of the l-convexity breakpoints will be a focus of this article. Historical perspectives on the problem may be found in [5, 7]. "Bump hunting" dates back at least to [3]. Silverman [16], Mammen [8] and Mammen, Marron and Fisher [9] formulated the problem as a sequential hypothesis testing problem. We refer to the estimation of the number of change points as the "model selection prob-

Received June 1995; revised October 1996.

[1] Supported by U.S. Department of Energy Grant DE-FG02-86ER53223.

AMS 1991 *subject classifications*. 62G07, 65D10, 65D07, 60G35.

Key words and phrases. Pilot estimators, shape constraints, nonparametric estimation, convexity.

lem” because it resembles model selection in an infinite family of parametric models.

An interesting result of [8, 9, 15] is that kernel smoothers will often produce too many inflection points or wiggles. If the amount of smoothing is chosen to minimize the mean integrated square error (MISE), then with nonvanishing asymptotic probability the estimate will have multiple inflection points in a neighborhood of an actual one. For many applications, estimating the correct shape is more important than minimizing the MISE.

In this article, we propose a class of two-stage estimators which estimate the l-change points in the first stage and then perform a constrained regression in the second-stage. In the first stage, the function is strongly smoothed while the smoothing in the constrained second stage is optimized for the minimal mean square error. When the change points are correctly specified, the constrained spline estimate has a smaller square error (as measured in a particular norm) than the unconstrained estimate.

Our second-stage estimate achieves the asymptotically optimal MISE convergence rate while suppressing artificial change points that can occur with the unconstrained method. Our proof does not exclude the possibility that the second-stage estimate has spurious inflection points far from the first-stage inflection points. We believe that our estimator has the same *relative* convergence rate as standard nonparametric methods. Thus our estimator suppresses artificial wiggles at nontrivial computational costs but no loss of MISE.

In Section 2, we show that the constrained smoothing spline estimate achieves the optimal rate of convergence for the expected square error even when the constraints are occasionally misspecified, provided that the misspecification rate is sufficiently small.

In Section 3, we evaluate the expected number of false (extra) empirical change points [9] for kernel smoothers when the errors are Gaussian. By adjusting the smoothing parameter, we can guarantee an asymptotically small probability of an error in our imposed constraints.

In Section 4, we propose two-stage estimators which estimate the number and location of the l-change points in the first stage. In the second stage, we impose that $\hat{f}^{(l+1)}$ be positive or negative in a small region about each of the empirical l-change points. The main advantage of the two-stage procedure is that less smoothing is required in the second stage than in the first stage while preserving the geometric fidelity.

Section 5 discusses potential extensions of the method. Section 6 describes a global shape optimization that is *heuristically* designed to be efficient in the amount of smoothing subject to determining the number of change points consistently.

2. Expected error under inexact convexity constraints.

We are given N measurements of the unknown function $f(t)$:

$$y_i = f(t_i) + \varepsilon_i. \tag{1}$$

The mean integrated squared error (MISE) for the estimate $\hat{f}^{(j)}$ of the jth derivative is defined as $\mathbf{E}[\int |\hat{f}^{(j)}_{N,\lambda}(t) - f^{(j)}(t)|^2]$. We consider the MISE for constrained estimation as the number of measurements N tends to infinity. In describing the large N asymptotics, we consider a sequence of measurement problems. For each N, the measurements occur at $\{t_i^N, i = 1, \dots, N\}$, We suppress the superscript N on the measurement locations $t_i \equiv t_i^N$. We define the empirical distribution of measurements, $F_N(t) = \sum_{t_i \le t} 1/N$, and let $F(t)$ be its limiting distribution.

Assumption A. Consider the sequence of estimation problems: $y_i^N = f(t_i^N) + \varepsilon_i^N$, where the ε_i^N are zero mean random variables and $\mathbf{Cov}[\varepsilon_i^N, \varepsilon_j^N] = \sigma^2\delta_{i,j}$. Assume that the distribution of measurement locations converges in the sup norm: $|F_N(t) - F(t)| \to 0$, where $F(t)$ is $C^\infty[0,1]$ and $0 < c_F < F'(t) < C_F$.

We measure the convergence of a set of measurement locations to the continuum limit in terms of the discrepancy of the point set:

Definition. The star discrepancy of $\{t_1 \dots t_N\}$ with respect to the continuous distribution $F(t)$ is $D_N^* \equiv \sup_t \{F_N(t) - F(t)\}$.

Equivalently, $D_N^* \sim 1/2N + \max_{1 \le i \le N} |F(t_i) - (i - 1/2)/N|$. For regularly spaced points, $D_N^* \sim 1/N$, while for randomly spaced points $D_N^* \sim \sqrt{\ln[\ln[N]]/N}$ by the Glivenko–Cantelli theorem.

A popular linear estimator is the smoothing spline [20]: $\hat{f} = \arg\min \mathrm{VP}[f]$, where

$$\mathrm{VP}[f] \equiv \frac{\lambda}{2} \int |f^{(m)}(s)|^2 \, ds + \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - f(t_i)|^2}{\sigma^2}. \tag{2}$$

We denote the standard Sobolev space of functions with square integrable derivatives by $W_{m,2}$ [20]. In general, the smoothing parameter will be decreased as N increases: $\lambda_N \to 0$. For smoothing splines, we add the following stronger requirements:

Assumption A* [1]. Let Assumption A hold with $f \in W_{m,2}$ and $m > 3/2$. Consider the sequence of smoothing spline minimizers of (2). Let the smoothing parameter λ_N satisfy $\lambda_N \to 0$ and $D_N^* \lambda_N^{-5/(4m)} \to 0$ as $N \to \infty$.

The constraint that $D_N^* \lambda_N^{-5/(4m)} \to 0$ is very weak since the optimal value of λ_N satisfies $\lambda_N \sim N^{-2m/(2m+1)}$. The assumption that $F(t)$ is $C^\infty[0,1]$ can be weakened.

For unconstrained smoothing spline estimates, the MISE has the following upper bound:

Theorem 2.1 [13, 1]. *Let Assumption* A* *hold, and denote the smoothing spline estimate from* (2) *by* $\hat{f}_{N,\lambda}(t)$. *As* $N \to \infty$,

$$\mathbf{E}\left[\int |\hat{f}^{(j)}_{N,\lambda}(t) - f^{(j)}(t)|^2\right] \le \alpha_j \lambda^{(m-j)/m} \|f\|_m^2 + \frac{\beta_j \sigma^2}{N\lambda^{(2j+1)/(2m)}}, \qquad j \le m, \tag{3}$$

where α_j and β_j are positive constants. The MISE is minimized by $\lambda = O(N^{-2m/(2m+1)})$ and

$$E\left[\int |\hat{f}^{(j)}_{N,\lambda}(t) - f^{(j)}(t)|^2\,dt\right] = O\big(N^{-2(m-j)/(2m+1)}\big). \tag{4}$$

For uniformly spaced measurement points, this result is in [13], while Cox [1] generalizes the result to these more general conditions on the measurement points. Cox imposes the condition that $j < m$, while his proof applies to $j = m$ as well.

Given change points, $\{x_1, x_2, \dots, x_K\}$, we define the closed convex cone:

$$V^{K,l}_{m,2}[x_1, \dots, x_K] = \{f \in W_{m,p} | (-1)^{k-1} f^{(l)}(t) \geq 0 \text{ for } x_{k-1} \leq t \leq x_k\}, \tag{5}$$

where $x_0 \equiv 0$ and $x_{K+1} \equiv 1$. When the change points are unknown, we need to consider $V^{K,l}_{m,2} = \bigcup_{\mathbf{x}\in[0,1]^K} V^{K,l}_{m,2}[\mathbf{x}] \cup (-V^{K,l}_{m,2}[\mathbf{x}])$, where $\mathbf{x} \equiv (x_1, x_2, \dots, x_K)$.

If we know the change point locations $\mathbf{x}$, a natural estimator is $\hat{f} = \arg\min \mathrm{VP}[f]$ subject to the convex constraints that $\hat{f} \in V^{K,l}_{m,2}[\mathbf{x}]$. This constrained spline estimate is the basis of our analysis. Detailed representation and duality results are in [14]. If we correctly impose the constraint that $f \in V^{K,l}_{m,2}[\mathbf{x}]$, the constrained spline fit always outperforms the nonconstrained fit as the following theorem indicates:

THEOREM 2.2 [18]. *Let f be in a closed convex cone $C \subseteq W_{m,2}$. Let $\hat{f}_u$ be the unconstrained minimizer of* (2) *given y_i, and let $\hat{f}_c$ be the constrained minimizer. Then $\|f - \hat{f}_c\|_V \leq \|f - \hat{f}_u\|_V$, where*

$$\|f\|^2_V \equiv \frac{\lambda}{2}\int |f^{(m)}(s)|^2\,ds + \frac{1}{N\sigma^2}\sum_{i=1}^{N} |f(t_i)|^2. \tag{6}$$

Theorem 2.2 applies to any set of y_i and does not use $y_i = f(t_i) + \varepsilon_i$. Theorem 2.2 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 2.2 does not generalize to unions of convex cones, and thus does not apply to $V^{K,l}_{m,2}$.

THEOREM 2.3. *Consider $f \in W_{m,2}$ and let Assumption* A* *hold. Consider a sequence of estimates $\hat{f}_{u,N}$ which satisfy the error bound given by* (3), *and a second sequence of estimates $\hat{f}_{c,N}$, which satisfy the error bound*: $\|f - \hat{f}_c\|_V \leq \|f - \hat{f}_u\|_V$. *The asymptotic error bound given by* (3) *holds for $\hat{f}_{c,N}$ with different constants, α'_j and β'_j.*

PROOF. For uniform sampling, this theorem is proved in [18] using interpolation inequalities. To generalize Utreras's result to our sampling hypotheses, we replace his Lemma 4.3 with (15) and (16) in Appendix A (with $\delta = 0.05$) and substitute $D_N^{*0.45}$ everywhere $1/n$ appears. □

This generalization of Utreras's result does not require the ratio of $\overline{\Delta}_N \equiv \sup_{i<N}(t_{i+1}-t_i)$ to $\underline{\Delta}_N \equiv \inf_{i<N}(t_{i+1}-t_i)$ to be bounded. In practice, we choose our constraints empirically and sometimes impose an incorrect constraint. We now show that occasionally imposing the wrong constraint does not degrade the asymptotic rate of convergence provided that the probability of an incorrect constraint is sufficiently small. The following theorem is a basis for our data adaptive estimators in Section 4.

THEOREM 2.4 (Occasional misspecification). *Consider a two-stage estimator that with probability* $1 - \mathscr{O}(p_N)$ *correctly chooses a closed convex cone* C, *with* $f \in C$, *in the first stage and then performs a constrained regression as in* (2). *Under Assumption* A*, *the estimate* $\hat{f}^{(j)}$ *satisfies the asymptotic bound* (3) (*with different constants*, α'_j *and* β'_j) *provided that*, *as* $N \to \infty$, p_N *vanishes rapidly enough*: $p_N/\lambda_N \to 0$ *and* $p_N N \lambda^{1/(2m)} \to 0$.

The proof is given in Appendix B.

3. Convergence of kernel smoothers.

In this section, we examine the expected number of zeros of a kernel smoother as a function of the halfwidth parameter. The results in this section are proven in [15]. We begin by presenting convergence results for kernel estimators $\hat{f}_N^{(l)}(t)$ of $f^{(l)}(t)$ as $N \to \infty$. We define

$$\sigma_N^2(t) = \mathbf{Var}[\hat{f}_N^{(l)}(t)], \qquad \xi_N^2(t) = \mathbf{Var}[\hat{f}_N^{(l+1)}(t)],$$
$$\mu_N(t) = \mathbf{Corr}[\hat{f}_N^{(l)}(t), \hat{f}_N^{(l+1)}(t)].$$

We use the notation $\mathscr{O}_R(\cdot)$ to denote a size of $\mathscr{O}(\cdot)$ relative to the main term: $\mathscr{O}_R(\cdot) = \times[1+\mathscr{O}(\cdot)]$ and $o_R = \times[1+o(\cdot)]$. We define $\|f\|_{bv}$ to be the sum of the L_∞ and total variation norms of f.

LEMMA 3.1 (Generalized Gasser–Müller [2, 15]). *Let* $f(t) \in C^{l+1}[0,1] \cap TV[-1,1]$ *and consider a sequence of estimation problems satisfying Assumption* A. *Let* $\hat{f}_N^{(l)}(t)$ *be a kernel smoother estimate of the form*

$$\hat{f}^{(l)}(t) = \frac{1}{N h_N^{l+1}} \sum_i^N \frac{y_i w_i}{F'(t_i)} \kappa^{(l)}\left(\frac{t-t_i}{h_N}\right), \tag{7}$$

where h_N *is the kernel halfwidth and the weights* w_i *satisfy* $|w_i - 1| \sim \mathscr{O}(D_N^*/h_N)$. *Let the kernel* $\kappa^{(l+1)} \in TV[-1,1] \cap C[-1,1]$ *satisfy moment condition* $\int_{-1}^1 \kappa(s)\,ds = 1$ *and boundary conditions* $\kappa^{(j)}(-1) = \kappa^{(j)}(1) = 0$ *for* $0 \le j \le l$. *Choose the kernel halfwidths such that* $h_N \to 0$ *and* $D_N^*/h_N^{l+2} \to 0$.

Then

$$\begin{aligned}\mathbf{E}[\hat{f}_N^{(l)}](t) &\to f^{(l)}(t) + \mathscr{O}_R(h_N + D_N^*/h_N^{l+1}),\\ \mathbf{E}[\hat{f}_N^{(l+1)}](t) &= \int_{-1}^{1} f^{(l)}(t+hs)\kappa(-s)\,ds + \mathscr{O}(\|f\kappa^{(l+1)}\|_{bv} D_N^*/h_N^{l+2}),\\ \sigma_N^2(t) &\to \sigma^2\|\kappa^{(l)}\|^2/(NF'(t)h_N^{2l+1}) + \mathscr{O}_R(h_N + D_N^*/h_N),\\ \xi_N^2(s) &\to \sigma^2\|\kappa^{(l+1)}\|^2/(NF'(s)h_N^{2l+3}) + \mathscr{O}_R(h_N + D_N^*/h_N)\end{aligned}$$

and

$$\mu_N^2(t) \to \mathscr{O}(h_N + D_N^*/h_N) \textit{ uniformly in the interval } [h_N, 1-h_N].$$

Lemma 3.1 applies to all of the common kernel smoother weightings [2] such as Priestley–Chao and Gasser–Müller. This result is slightly stronger than previous theorems on kernel smoothers [2]. Our hypotheses are stated in terms of the star discrepancy while previous convergence theorems [2] place restrictions on both $\bar{\Delta}_N \equiv \sup_{i<N}\{t_{i+1} - t_{i-1}\}$ and $\varepsilon_N \equiv \sup_{i<N}\{|1 - (t_{i+1} - t_i)NF'(t_i)\}$.

We now evaluate the expected number of false l-change points for a sequence of kernel estimates of $f^{(l)}(t)$. We restrict to independent *Gaussian* errors: $\varepsilon_i \sim N(0, \sigma^2)$. Thus, $\hat{f}_N^{(l)}(t)$ is a Gaussian process. The following assumption rules out nongeneric cases:

ASSUMPTION B. Let $f(t) \in C^{l+1}[0,1]$ have K l-change points $\{x_1, \dots, x_K\}$ with $f^{(l+1)}(x_k) \neq 0$, $f^{(l)}(0) \neq 0$ and $f^{(l)}(1) \neq 0$. Consider a sequence of estimation problems with independent, normally distributed measurement errors ε_i^N with variance σ^2. Let $\hat{f}_N^{(l)}(t)$ be a sequence of kernel estimates of $f^{(l)}$ on the sequence of intervals $[\delta_N, 1-\delta_N]$.

To neglect boundary effects, we take $\delta_N = h_N$ for kernel estimators and $\delta_N = \delta$ for splines. For each change point x_k, we define the change point variance [2]: $\sigma_{if}^2(x_k) \equiv \mathbf{Var}[\hat{f}_N^{(l)}(x_k)]/|f^{(l+1)}(x_k)|^2$. The following theorem bounds the probability of a false estimate of a change point far away from a true change point.

THEOREM 3.2. *Let Assumption* B *hold and consider a sequence of kernel estimators $\hat{f}_N^{(l)}(t)$ that satisfy the hypotheses of Lemma* 3.1. *Choose kernel halfwidths h_N and uncertainty intervals w_N such that $h_N/w_N \to 0$, $w_N \to 0$, $w_{N,k}^2 N h_N^{2l+1} \geq 1$. The probability $p_N(w_N)$ that $\hat{f}_N^{(l)}$ has a false change point outside of a width of w_N from the actual $(l+1)$-change points satisfies*

$$p_N(w_N) \leq \sum_{k=1}^{K} \mathscr{O}\left(\frac{\sigma_{if}(x_k)}{h_N}\exp\left(\frac{-w_N^2}{2\sigma_{if}^2(x_k)}\right)\right), \tag{8}$$

where $\sigma_{if}^2(x_k) \to \sigma^2\|\kappa^{(l)}\|^2/|f^{(l+1)}(x_k)|^2 NF'(x_k)h_N^{2l+1}$ on the interval $[h_N, 1-h_N]$.

In [8, 9], Mammen and co-workers derive the number of false change points for kernel estimation of a probability density. We present the analogous result for regression function estimation.

THEOREM 3.3 (Analog of [8, 9]). *Let Assumption* B *hold. Consider a sequence of kernel smoother estimates* $\hat{f}_N$ *which satisfy the hypotheses of Lemma* 3.1 *with* $\int_{-1}^{1} s\kappa(s)\,ds = 0$. *Let the sequence of kernel halfwidths* h_N *satisfy* $D_N^* N^{1/2} h_N^{1/2} \to 0$ *and* $0 < \liminf_N h_N N^{1/(2l+3)} \leq \limsup_N h_N N^{1/(2l+3)} < \infty$. *The expected number of l-change points of* $\hat{f}_N$ *in the estimation region* $[h_N, 1 - h_N]$ *is asymptotically*

$$\mathbf{E}[\hat{K}] - K = 2 \sum_{k=1}^{K} H\left(\sqrt{\frac{|f^{(l+1)}(x_k)|^2 N F'(x_k) h^{2l+3}}{\sigma^2 \|\kappa^{(l+1)}\|^2}}\right) + o_{\mathscr{P}}(1) \tag{9}$$

where $H(z) \equiv \phi(z)/z + \Phi(z) - 1$ *with* ϕ *and* Φ *being the Gaussian density and cdf respectively. If* $f^{(l+1)}(t)$ *has Hölder smoothness of order* ν *for some* $0 < \nu < 1$, *and* $h_N N^{1/(2l+3)} \to 0$, *then* (9) *remains valid provided that* $h_N N^{1/(2l+3+2\nu)} \to 0$.

In [8, 9], the correction in (9) is shown to be $o(1)$ when $\limsup_N h_N N^{1/(2l+3)} < \infty$. We strengthen this result by showing that (9) continues to represent the leading order asymptotics even when $h_N N^{1/(2l+3)} \to \infty$.

For each change point, x_k, we define the α uncertainty interval by

$$[x_k - z_\alpha \sigma_{if}(x_k), x_k + z_\alpha \sigma_{if}(x_k)],$$

where z_α is the two-sided $\alpha[1 + 2H(h_N \|\kappa^{(l)}\| / (\sigma_{if}(x_k) \|\kappa^{(l+1)}\|))]$-critical value for a normal distribution. The probability that an empirical change point is more than $z_\alpha \sigma_{if}(x_k)$ away from the kth actual change point is less than α. We consider two change points well resolved if the two uncertainty intervals do not overlap.

A similar variance for change point estimation is given in [11], where Müller shows that the leftmost change point is asymptotically normally distributed with variance $\sigma_{if}^2(x_k)$. For his result, Müller imposes stricter requirements on $f^{(l+1)}$ and $\hat{f}^{(l+1)}$, and does not obtain results pertaining to the expected number of false change points.

When $f \in C^m$, the halfwidth which minimizes the MISE scales as $h_N \sim N^{-1/(2m+1)}$. Other schemes for piecewise convex fitting [5] choose the kernel halfwidth or smoothing parameter to be the smallest value $h_{cr,K}$ that yields only K change points in an unconstrained fit. Theorem 3.3 shows that $h_{cr,K}$ is asymptotically larger than the halfwidth which minimizes the MISE for $l = m$, $m - 1$. As a result, these schemes oversmooth.

4. Data-based pilot estimators with geometric fidelity. We consider two-stage estimators that begin by estimating $f^{(l)}$ and $f^{(l+1)}$ using an unconstrained estimate with $h_N \gtrsim \log(N) N^{1/(2l+3)}$. From the pilot estimate, we determine the number, $\hat{K}$, and approximate locations of the change points. In the second stage, we perform a constrained fit, requiring that $\hat{f}^{(l)}$ be monotone in small regions about each empirical change point. Since spurious change points asymptotically occur only in a neighborhood of an actual change point, the second-stage fit need only be constrained in a vanishingly small portion of the domain asymptotically.

THEOREM 4.1 (Asymptotic MISE for pilot estimation). *Let $f(t)$ satisfy Assumption* B *and consider a sequence of two-stage estimators. In the first stage, let the hypotheses of Theorem* 3.2 *be fulfilled. From the first-stage estimate, denote the empirical l-change points by $\hat{x}_k$, $k = 1, \dots, \hat{K}$. Choose widths $w_{N,k}$ such that $w_{N,k} \to 0$, $h_N/w_{N,k} \to 0$ and $w_{N,k}^2 N h_N^{2l+1}/\ln(N) \to \infty$, where h_N is the first-stage halfwidth. In the second stage, perform a constrained regression as in* (2), *where the second-stage smoothing parameter λ_N satisfies $\lambda_N \to 0$ and $D_N^* \lambda_N^{-5/4m} \to 0$. In the second-stage regression, impose the constraints that the second-stage $\hat{f}^{(l+1)}$ has a single sign in the regions $[\hat{x}_k - w_{N,k}, \hat{x}_k + w_{N,k}]$ [which matches the sign of $\hat{f}^{(l)}(\hat{x}_k + w_{N,k} - \hat{f}^{(l)}(\hat{x}_k - w_{N,k})]$. For $f \in W_{m,2}$, the second-stage estimate $\hat{f}$ satisfies the expected error bounds of* (3) (*with different constants*).

PROOF. Theorem 3.2 shows that $\hat{x}_k$ lie within $w_N/2$ of the x_k with probability $1 - p_N$, where $p_N = \mathscr{O}(\exp(-c_0^2 w_N^2 N h_N^{2l+1}/8\sigma^2))$ and $c_o = \inf_k\{|f^{(l+1)}(x_k)|\}$. In the remainder of the proof, we implicitly neglect this set of measure p_N and use arguments that are valid for large N. By Assumption B, there are no zeros of $f^{(l+1)}(x_k)$ in $[\hat{x}_k - w_{N,k}, \hat{x}_k + w_{N,k}]$ and thus $|f^{(l)}(\hat{x}_k + w_{N,k}) - f^{(l)}(\hat{x}_k - w_{N,k})| > c_o w_{N,k}$. Note that $\hat{f}^{(l)}(\hat{x}_k + w_{N,k}) - \hat{f}^{(l)}(\hat{x}_k - w_{N,k})$ has a Gaussian distribution with variance $2\sigma_N^2(t)$ and a bias error bounded by $\mathscr{O}(h_N)$ [2]. Thus, the sign of $\hat{f}^{(l)}(\hat{x}_k + w_{N,k}) - \hat{f}^{(l)}(\hat{x}_k - w_{N,k})$ is determined correctly with a probability of $1 - p_N$. The result follows from Theorem 2.4. □

The trick of Theorem 4.1 is to constrain $\hat{f}^{(l+1)}$ to be positive (or negative) in the uncertainty interval of the estimated change points (linear constraints) rather than constraining $\hat{f}^{(l)}$ to have a single zero around $\hat{x}_j$ (nonlinear constraints). Theorem 4.1 implies that the second-stage estimate has no false l change points within $\pm w_{N,k}$ of $\hat{x}_k$ with high probability. It does not exclude the possibility that false change points occur outside of $[\hat{x}_k - w_{N,k}, \hat{x}_k + w_{N,k}]$, but we believe that such false change points seldom occur in practice. We believe that it is adequate to eliminate false inflection points in the regions where they occur in the unconstrained nonparametric estimates.

Asymptotically, the zeros of $\hat{f}^{(l+1)}$ will occur in clusters with an odd number of zeros. If a cluster with an even number of zeros occurs in the first stage, it is spurious (with high asymptotic probability). We recommend imposing the constraint that the second-stage $\hat{f}^{(l)}$ (not $\hat{f}^{(l+1)}$) has a single sign in each neighborhood where an even number of change points of the first stage occur.

For data adaptive methods, we modify Theorem 4.1 slightly:

COROLLARY 4.2. *The hypotheses of Theorem* 4.1 *on the first-stage estimate* [*such as $D_N^*/h_N^{l+2} \to 0$, $w_{N,k}^2 N h_N^{2l+1}/\ln(N) \to \infty$, $w_{N,k} \to 0$ and $h_N/w_{N,k} \to$*

0] *need only be true with probability* $(1 - p_N)$ *for the conclusion of Theorem* 4.1 *to be valid, where* p_N *satisfies* $p_N/\lambda_N \to 0$ *and* $p_N N \lambda^{1/(2m)} \to 0$.

Let $h_{\rm GCV}$ denote the smoothing bandwidth chosen by generalized cross-validation. Under certain conditions [4], it can be shown that $h_{\rm GCV}$ has an asymptotically normal distribution with mean $cN^{-\beta}$, where c and β depend on f and the kernel shape. For the first-stage halfwidth, we propose using $h_N = \iota(N)h_{\rm GCV}$, where $\iota(N)$ is chosen such that $\iota(N)h_{\rm GCV}$ satisfies Corollary 4.2. If $h_{\rm GCV} \to cN^{-1/(2l+1)}$, we recommend choosing $\iota(N) \approx \log(N)N^{\alpha}$ with $\alpha = 1/(2l+1) - 1/(2l+3)$. This scaling corresponds to a uniformly consistent estimate of $f^{(l+1)}$ [2]. The overall moral is: the smoothing level chosen by GCV is asymptotically optimal for estimating functions, but shape estimation requires more smoothing.

Numerical implementations [19] of constrained least squares usually apply the active set method of quadratic programming. The constrained smoothing spline regression reduces to a finite dimensional minimization when the constraints are on the mth derivative. Since we constrain $\hat{f}_N^{(l+1)}(t)$ in each neighborhood of an estimated l-change point, our algorithm is most readily implemented for $l = m - 1$ using the duality result of [14].

To reduce the number of constraints, we seek to minimize the length of the constraint intervals, $w_{N,k}$. When the hypotheses of Theorem 3.3 are satisfied with $h_N N^{1/(2l+3)} \to \infty$, we can estimate the uncertainty interval for inflection points by $\hat{\sigma}^2_{if}(\hat{x}_k) \equiv \sigma^2 \|\kappa^{(l)}\|^2 / [|\hat{f}^{(l+1)}(\hat{x}_k)|^2 N F'(\hat{x}_k) h_N^{2l+1}]$. We recommend choosing the constraint width such that $w_{N,k} \gg \hat{\sigma}^2_{if}(\hat{x}_k)$. (For smaller constraint widths, Theorem 4.1 is true, but false inflection points can still occur near the actual inflection point.)

5. Potential extensions.

1. At present, we cannot exclude the posibility of false change points arising in the second-stage estimate away from the constraint intervals. We believe that this is a technical gap in our analysis and not a practical difficulty. The risk of false change points can be reduced by choosing larger intervals to impose the constraints on $\hat{f}^{(l+1)}$. Let $\{\hat{u}_j\}$ denote the zeros of $\hat{f}^{(l+1)}$ in the first stage. Let u_j and u_{j+1} be the two closest zeros of $\hat{f}^{(l+1)}$ to x_k with $u_j < x_k < u_{j+1}$. A judicious choice of constraint intervals is $[(x_k - u_j)/2, (x_k + u_{j+1})/2]$, which gives large constraint intervals with only a small chance of imposing an incorrect constraint on the second-stage $\hat{f}^{(l+1)}$.
2. The pilot method suppresses false zeros of $f^{(l)}(t)$, but does not suppress false zeros of $f^{(l)}(t) - c$ where c, is a nonzero real number. It may be more desirable to apply the pilot estimator to $f^{(l)}(t) - q(t)$, where $q(t)$ is a prescribed function possibly involving a small number of empirically estimated free parameters. The constraints in the second stage will virtually never need to be imposed if $f^{(l)}(t)$ is always positive or negative. Thus, we suggest centering $f^{(l)}(t)$ about zero by subtracting off a polynomial fit of order

l [and thereby centering $f^{(l)}(t)$ about zero] prior to applying our two-stage estimator.

3. Asymptotically, smoothing splines are equivalent to kernel smoothers [1, 17]. Using this convergence, results analogous to Theorems 3.2, 3.3 and 4.1 can be proved for the case where smoothing splines are used in the first-stage estimate.
4. It is tempting to try the pilot estimation procedure using local polynomial regression (LPR) in the second stage. Unfortunately, there is a difficulty with shape constrained LPR. Let $\hat{f}(x) \sim a_0(t)+a_1(t)[x-t]+a_2(t)[x-t]^2/2$ for $|x-t| < h$. If $a_2(t)$ is constrained to be nonnegative, $a_0(t)$ need not be convex because LPR does not require $a_0''(t) = a_2(t)$.
5. Our data adaptive convergence results in Sections 3 and 4 are for quadratic estimation with Gaussian errors. The results should be extendable to non-Gaussian errors using the central limit theorem and Brownian bridges as in [9].

6. Piecewise convex information criterion. Instead of the two-stage pilot estimator, we now propose a second class of estimators which penalize both smoothness and the number of change points. Information or discrepancy criteria are used to measure whether the improvement in the goodness of fit is sufficient to justify using additional free parameters. Both the number of free parameters and their values are optimized with respect to the discrepancy criterion $d(\hat{f}, \{y_i\})$. Let $\hat{\sigma}^2$ be a measure of the average residual error:

$$\hat{\sigma}^2(\hat{f}, \{y_i\}) = \frac{1}{N\sigma^2}\sum_{i=1}^{N}[y_i - \hat{f}(t_i)]^2,$$

or its L_1-analog. Typical discrepancy functions are

$$d^I(\hat{f}, \{y_i\}) = \hat{\sigma}^2/[1-(\gamma_1 p/N)]^2 \tag{10}$$

and

$$d^B(\hat{f}, \{y_i\}) = \hat{\sigma}^2[1+(\gamma_2 p\ln(N)/N)], \tag{11}$$

where p is the effective number of free parameters in the smoothing spline fit. For $\gamma_1 = 1$, $d^I(\hat{f}, \{y_i\})$ is generalized cross-validation (GCV) which has the same asymptotic behavior as the Akaike information criterion. For $\gamma_2 = 1$, d^B is the Bayesian or Schwartz information criterion. For a nested family of models, $\gamma_2 = 1$ is appropriate while $\gamma_2 = 2$ corresponds to a nonnested family with $2\binom{N}{K}$ candidate models at the kth level. In very specialized settings in regression theory and time series, it has been shown that functions like d^I are asymptotically efficient while those like d^B are asymptotically consistent. In other words, using d^I-like criteria will asymptotically minimize the expected error at the cost of not always yielding the correct model. In contrast, the Bayesian criteria will asymptotically yield the correct model at the cost of having a larger expected error.

Our goal is to select consistently the number of convexity change points and efficiently estimate the model subject to the change point restrictions. Therefore, we propose the following *new* discrepancy criterion:

$$\text{PCIC} = \sigma^2(\hat{f}, \{y_i\})\left[\frac{1+\gamma_2 K \ln(N)/N}{(1-\gamma_1 p/N)^2}\right], \tag{12}$$

where K is the number of convexity change points and p is the number of free parameters; PCIC stands for piecewise convex information criterion. In selecting the positions of the K change points, there are essentially $2\binom{N}{K}$ possible combinations of change point locations if we categorize the change points by the nearest measurement location. Thus, our default values are $\gamma_1 = 1$ and $\gamma_2 = 2$.

We motivate PCIC: to add a change point requires an improvement in the residual square error of $O(\sigma^2 \ln(N))$, which corresponds to an asymptotically consistent estimate. If the additional knot does not increase the number of change points, it will be added if the residual error decreases by $\gamma_1 \sigma^2$. Presently, PCIC is purely a heuristic principle. We conjecture that it consistently selects the number of change points and is asymptotically efficient within the class of methods that are asymptotically consistent with regard to convexity change points.

7. Summary. Theorem 3.3 shows that, for $l = m$ and $l = m - 1$, the amount of smoothing necessary for geometric fidelity is larger than the optimal value for minimizing the mean integrated square error. Therefore, we have considered two-stage estimators which estimate the l-change points and their uncertainty intervals in the first stage. In the second stage, a constrained smoothing spline fit is applied using a data adaptive estimate of the smoothing parameter.

Our main result is that such two-stage schemes achieve the same asymptotic rate of convergence as standard methods such as GCV that do not guarantee geometric fidelity. We prove this result incrementally. Theorem 2.4 evaluates an acceptable rate of failure for imposing the wrong constraints. Theorem 4.1 proves the asymptotic MISE result when the change points are estimated in the first stage while the widths of the constraint intervals and the smoothing parameter in the first stage satisfy certain scaling bounds. The second-stage estimates have no false change points in the regions where the unconstrained estimators have all of their false change points with probability approaching unity.

Linear constraints are necessary only in small neighborhoods about each l-change point asymptotically. This suggests that the ratio of the MISE from our two stage estimate to that of kernel smoothers or spline tends to 1. Our estimators should be useful in situations where obtaining the correct shape is important and computational costs are not an issue. Piecewise convex fitting may offer larger potential gains in the MISE in small sample situations because the *a priori* knowledge that there are only a small number of inflection points should be of more value when less data are available. Numerical simulations are underway and will be discussed elsewhere.

APPENDIX A

Interpolation inequalities. We measure the distance from an arbitrary set of measurement times to an equispaced set of points in terms of the *discrepancy* as defined in Section 2. The discrepancy is useful because it describes how closely a discrete sum over an arbitrarily placed set of points approximates an integral. In this appendix, we summarize these results and present a new interpolation identity for discrete sums. A useful condition is the following:

ASSUMPTION 0. Assume that the limiting distribution of the measurement locations $F(t)$ is $C^1[0,1]$ and $0 < c_F < F'(t) < C_F$.

We denote the set of functions of bounded variation by $TV[0,1]$ and the corresponding norm by $\|\cdot\|_{TV}$.

THEOREM A.1 (Generalized Koksma [15]). *Let g be a bounded function of bounded variation $\|g\|_{TV}$ on $[0,1]$: $g \in TV[0,1] \cap L_\infty[0,1]$. Let the star discrepancy be measured by a distribution $F(t)$ which satisfies Assumption 0. If the discrete sum weights $\{w_i, i = 1, \dots, N\}$ satisfy $|w_i - 1| \le CD_N^*$, then*

$$\left|\int_0^1 g(t)\,dF(t) - \frac{1}{N}\sum_{i=1}^N g(t_i)w_i\right| \le \left[\|g\|_{TV} + C\|g\|_\infty\right] D_N^*. \tag{13}$$

In our version of Koksma's theorem, we have added two new effects: a nonuniform weighting $\{w_i, i = 1, \dots, N\}$, and a nonuniform distribution of points dF. The total variation of $g(t(F))$ with respect to dF is equal to the total variation of $g(t)$ with respect to dt. Theorem A.1 follows from Koksma's theorem by a change of variables.

In the continuous case, the following Sobolev interpolation result [13] is well known:

LEMMA A.2. *There exist constants c_j depending only on m such that, for all $g \in W_{m,2}[0,1]$ and $\theta \in [0,1]$ $0 \le j \le m$,*

$$\theta^{2j}\int_0^1 |g^{(j)}(s)|^2\,ds \le c_j\left[\int_0^1 |g(s)|^2\,ds + \theta^{2m}\int_0^1 |g^{(m)}(s)|^2\,ds\right]. \tag{14}$$

Using Koksma's theorem and Lemma A.2, we can arrive at the following inequalities:

COROLLARY A.3. *Let g be in $W_{m,2}[0,1]$ and assume the star discrepancy satisfies Assumption 0 with $m < N$. The following interpolation bounds hold:*

$$\frac{1}{N}\sum_{i=1}^N g^2(t_i) \le C_1\int_0^1 g^2(t)\,dt + c_1 D_N^{*m}\int_0^1 |g^{(m)}(s)|^2\,ds, \tag{15}$$

where $C_1 = C_F + c_1 + D_N^*$, *and*

$$[c_F - c_1 D_N^{*^\delta} - D_N^*] \int_0^1 |g(s)|^2 \, ds \\ \leq \frac{1}{N} \sum_{i=1}^N g(t_i)^2 + c_1 D_N^{*^{m(1-2\delta)}} \int_0^1 |g^{(m)}(s)|^2 \, ds, \tag{16}$$

for all δ *in* $(0, 1/2)$ *such that* $c_F > c_1 D_N^{*^\delta} + D_N^*$.

PROOF. For $g \in W_{1,2}$, Koksma's theorem implies

$$\left| \int_0^1 g^2(t) \, dF(t) - \frac{1}{N} \sum_{i=1}^N g^2(t_i) \right| \leq \|g^2\|_{TV} D_N^* \leq (\|g\|_{0,2}^2 + \|g\|_{1,2}^2) D_N^*.$$

We then apply (14) to $\|g\|_{1,2}^2$ with $\theta = |D_N^*|^{1/2+\delta}$ for arbitrarily small δ,

$$\left| \int_0^1 g^2(t) \, dF(t) - \frac{1}{N} \sum_{i=1}^N g(t_i)^2 \right| \leq c_1 D_N^{*^\delta} [\|g\|_{0,2}^2 + D_N^{*^{m(1-2\delta)}} \|g\|_{m,2}^2] + \|g\|_{0,2}^2 D_N^*,$$

yielding the bound (16). □

APPENDIX B

PROOF OF THEOREM 2.4. If the constraints are correct, Theorem 2.3 yields the asymptotic error bound. We need to show that misspecified models do not contribute significantly to the error. For any realization of the $\{y_i\}$, we have the bound

$$\|\hat{f} - f\|_V^2 \leq \lambda(\|f\|_m^2 + \|\hat{f}\|_m^2) + \frac{1}{N\sigma^2} \sum_i^N |\hat{f}(t_i) - f(t) - \varepsilon_i|^2 + \varepsilon_i^2 \\ \leq \lambda \|f\|_m^2 + \frac{1}{N\sigma^2} \sum_i (y_i^2 + \varepsilon_i^2) \leq \|f\|_V^2 + \frac{1}{N} \sum_i \frac{\varepsilon_i^2}{\sigma^2}. \tag{17}$$

This paragraph is devoted to bounding the expectation of $\sum_i \varepsilon_i^2$ over the worst possible set with probability p_N. Note $\sum_i \varepsilon_i^2$ has a χ_N^2 distribution with density $p_{\chi_N^2}(w) = w^{N/2-1} \exp(-w/2)/2^{N/2}\Gamma(N/2)$. We seek to bound $I_1 \equiv \int_{\chi_0^2}^\infty w \, dp_{\chi_N^2}(w)$, where $\chi_0^2(p_N)$ is defined by $\int_{\chi_0^2}^\infty dp_{\chi_N^2}(w) = p_N$. We claim that $I_1 \leq 2Np_N$. We assume that $\chi_0^2(p_N) > 1.5N$. [If $\chi_0^2(p_N) < 1.5N$, we split the integral into $w \leq 1.5N$ and $w > 1.5N$.] We define $\tilde{w} = w/N$ and use Sterling's formula to find $p_{\chi_N^2}(N\tilde{w}) \approx \tilde{w}^{N/2-1} \exp(-N(\tilde{w} - 1)/2)/\sqrt{4\pi N}$. Evaluating the integrals by Laplace's method under the assumption that $p_N \ll 1$ and $\chi_0^2 > N$ yields $p_N \approx 2p_{\chi_N^2}(\chi_0^2)\chi_0^2/(\chi_0^2 - N + 1/2)$ and $I_1 \approx 2Np_{\chi_N^2}(\chi_0^2)\chi_0^2/(\chi_0^2 - N)$. For $\chi_0^2 > 1.5N$, we have $I_1 \approx p_N N(\chi_0^2 - N + 1/2)/(\chi_0^2 - N) \ll 0.5Np_N$.

Taking the expectation conditional on $f \notin V$ yields $\mathbf{E}_{f \notin V} \|\hat{f} - f\|_V^2 \leq \|f\|_V^2 + 2$. Since $(1/N)\sum_i |f(t_i)|^2 \to \int_0^1 f(s)^2 \, dF(s)$, we select N large enough that

$\lambda_N\|f\|_m^2 > (p_N/(N\sigma^2))\sum_i |f(t_i)|^2$. We now bound the contribution to $\mathbf{E}[\|\hat{f} - f\|_m^2]$ and $\mathbf{E}[\|\hat{f} - f\|_0^2]$ from $f \notin V$:

$$\begin{aligned} p_N\mathbf{E}_{f\notin V}\|\hat{f} - f\|_m^2 &\le \frac{p_N}{\lambda_N}\biggl(2 + \lambda_N\|f\|_m^2 + \frac{1}{N\sigma^2}\sum_i^N f(t_i)^2\biggr) \\ &\le \frac{2\sigma^2}{N\lambda^{(2m+1)/(2m)}} + [1 + \mathscr{O}(p_N)]\|f\|_m^2, \end{aligned} \tag{18}$$

by assumption on p_N. To bound $\mathbf{E}_{f\notin V}\|\hat{f} - f\|_0^2$, we apply Lemma A.2 with $\delta = 0.05$ in (16) and D_N^* large enough that $1/(c_F - c_1 D_N^{*^{\delta}} + D_N^*)$ is bounded by a constant, γ:

$$\begin{aligned} p_N\mathbf{E}_{f\notin V}\|\hat{f} - f\|_0^2 &\le \gamma p_N\mathbf{E}_{f\notin V}\Biggl[\frac{1}{N}\sum_{i=1}^N |\hat{f}(t_i) - f(t_i)|^2 + c_1 D_N^{*\,0.9m}\|\hat{f} - f\|_m^2\Biggr] \\ &\le \gamma p_N\biggl(\sigma^2 + \frac{D_N^{*\,0.9m}}{\lambda}\biggr)\mathbf{E}_{f\notin V}[\|\hat{f} - f\|_V^2] \\ &\le \gamma p_N\biggl(\sigma^2 + \frac{D_N^{*\,0.9m}}{\lambda}\biggr)\biggl(2 + \lambda_N\|f\|_m^2 + \frac{1}{N\sigma^2}\sum_{i=1}^N f(t_i)^2\biggr) \\ &\le \gamma(\sigma^2 + 1)\biggl(\frac{\sigma^2}{N\lambda_N^{1/(2m)}} + \lambda_N[1 + \mathscr{O}(p_N)]\|f\|_m^2\biggr). \end{aligned} \tag{19}$$

Equations (18) and (19) are in the form required by Theorem 4.4 of [18]. Using Lemma 14 and duplicating the proof of Theorem 4.6 of [18] yields the result. □

Acknowledgments. We thank the referee and the Editors for their help.

REFERENCES

[1] COX, D. D. (1984). Multivariate smoothing splines functions. *SIAM J. Numer. Anal.* **21** 789–813.

[2] GASSER, TH. and MÜLLER, H. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

[3] GOOD, I. J. and GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering data. *J. Amer. Statist. Assoc.* **75** 43–73.

[4] HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–101.

[5] LI, W., NAIK, D. and SWETITS, J. (1996). A data smoothing technique for piecewise convex/concave curves. *SIAM J. Sci. Comput.* **17** 517–537.

[6] MÄCHLER, M. (1995). Variational solution of penalized likelihood problems and smooth curve estimation. *Ann. Statist.* **23** 1496–1517.

[7] MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

[8] MAMMEN, E. (1995). On qualitative smoothness of kernel density estimates. *Statistics* **26** 253–267.

[9] MAMMEN, E., MARRON, J. S. and FISHER, N. J. (1992). Some asymptotics for multimodal tests based on kernel density estimates. *Probab. Theory Related Fields* **91** 115–132.

[10] MICHELLI, C. A. and UTRERAS, F. (1985). Smoothing and interpolation in a convex set of Hilbert space. *SIAM J. Statist. Sci. Comput.* **9** 728–746.

[11] MÜLLER, H. G. (1985). Kernel estimators of zeros and of the location and size of extrema of regression functions. *Scand. J. Statist.* **12** 221–232.

[12] NIEDERRIETER, H. (1992). *Random Number Generators and Quasi-Monte Carlo Methods*. SIAM, Philadelphia.

[13] RAGOZIN, D. L. (1983). Error bounds for derivative estimation based on spline smoothing of exact or noisy data. *J. Approx. Theory.* **37** 335–355.

[14] RIEDEL, K. S. (1995). Piecewise convex function estimation and model selection. In *Proceedings of Approximation Theory VIII* (C. K. Chui and L. L. Schumaker, eds.) 467–475. World Scientific, Singapore.

[15] RIEDEL, K. S. (1997). Improved asymptotics for zeros of kernel estimates via a reformulation of the Leadbetter–Cryer integral. *Statist. Probab. Lett.* **32** 351–356.

[16] SILVERMAN, B. W. (1983). 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.) 248–259. Cambridge Univ. Press.

[17] SILVERMAN, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

[18] UTRERAS, F. (1985). Smoothing noisy data under monotonicity constraints—existence, characterization and convergence rates. *Numer. Math.* **47** 611–625.

[19] VILLALOBOS, M. and WAHBA, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82** 239–248.

[20] WAHBA, G. (1991). *Spline Models for Observational Data*. SIAM, Philadelphia.

COURANT INSTITUTE OF MATHEMATICAL SCIENCES
NEW YORK UNIVERSITY
NEW YORK, NEW YORK 10012-1185
E-MAIL: riedel@cims.nyu.edu