The Annals of Statistics
2003, Vol. 31, No. 3, 921–941

SEQUENTIAL METHODS FOR DESIGN-ADAPTIVE ESTIMATION OF DISCONTINUITIES IN REGRESSION CURVES AND SURFACES[1]

BY PETER HALL AND ILYA MOLCHANOV

Australian National University and Universität Bern

In fault-line estimation in spatial problems it is sometimes possible to choose design points sequentially, by working one's way gradually through the "response plane," rather than distributing design points across the plane prior to conducting statistical analysis. For example, when estimating a change line in the concentration of resources on or under the sea bed, individual measurements can be particularly expensive to make. In such cases, sequential, design-adaptive methods are attractive. Appropriate methodology is largely lacking, however, and the potential advantages of taking a sequential approach are unclear. In the present paper we address both these problems. We suggest a methodology based on "sequential refinement with reassessment" that relies upon assessing the correctness of each sequential result, and reappraising previous results if significance tests show that there is reason for concern. We focus part of our attention on univariate problems, and we show how methods for the spatial case can be constructed from univariate ones.

1. Introduction. Consider the problem of estimating a fault line in a response surface by sampling the surface sequentially. For example, the surface might represent the concentration of a mineral at a given depth in the earth's crust, or the level of a nutrient on the ocean floor. Each sampling operation incurs a cost, which is reduced by minimizing the number of samples drawn for a given order of accuracy. We shall show that sequential sampling offers an opportunity for making large savings. In particular, if the fault line is estimated using a second-order method, requiring two derivatives, then the number of sampling operations needed in order to achieve $O(\delta)$ accuracy, as $\delta \to 0$, is reduced from $O(\delta^{-3/2})$, if the points are scattered across the plane prior to estimation, to $O(\delta^{-1/2})$, multiplied by a logarithmic factor, when the points are placed sequentially into the plane. Relative expense is reduced by an even greater amount if the alternative is a predetermined gridded design, which gives particularly poor performance per sample point. The rate $O(\delta^{-1/2})$ is optimal, although the logarithmic factor may depend on the nature of the error distribution (in particular, whether it is heavy-tailed) or the method used.

Received June 2001; revised March 2002.

[1]Supported in part by Visiting Fellowship Grant, UK EPSRC.

AMS 2000 subject classifications. Primary 62L12; secondary 62G20, 62H11.

Key words and phrases. Changepoint, fault line, hypothesis test, nonparametric estimation, recursive, search methods, spatial statistics.

Sequential sampling for changepoint estimation on the line is a closely related problem. Indeed, in many circumstances a solution to the spatial problem would involve repeated application of methods in the univariate case, and so we address the latter problem first. There, the expense of achieving $O(\delta)$ accuracy can be reduced from $O(\delta^{-1})$, if design points are placed in predetermined positions, to little more than $O(|\log \delta|)$ if they are chosen sequentially.

These results are closely linked to optimal convergence rates in more familiar, deterministic problems. Consider, for example, the problem of estimating the location θ of a jump discontinuity in an otherwise-smooth univariate function f, defined on the line and which we may observe without error. Make the task relatively simple by supposing f takes constant, known, unequal values a and b to the left and right, respectively, of θ, and consider θ to be a random variable that is uniformly distributed in a unit interval. Then the search algorithm that minimizes the expected length (with respect to the uniform distribution of θ) of an interval that is known to contain θ involves observing the value of f at the midpoint of the previously computed interval. Thus, after n steps the value of θ is narrowed to an interval of length 2^{-n} within which it lies with probability 1.

In the following sense, the algorithm suggested in Section 2 attains this optimal convergence rate arbitrarily closely, in the context of functions observed with noise. Suppose only that the noise distribution has zero mean and finite variance; assume only that f is smooth, rather than strictly constant, away from the jump; and take $\rho = \rho(n)$ to be any positive sequence converging to 0. Then we can produce, after n sequential sampling operations, a confidence interval of width $e^{-\rho n}$ (rather than $e^{-n \log 2}$ in the algorithm of the previous paragraph) within which the true value of θ lies with probability converging to 1 as $n \to \infty$.

If the error distribution is known then our algorithm can be modified so that ρ is kept fixed at a strictly positive value. On the other hand, assuming only that the error distribution has a finite moment generating function, and taking ρ to converge to 0 at rate $(\log n)^{-\gamma}$ for some $\gamma > 2$, we may ensure that the confidence interval for θ has coverage $1 - O(n^{-C})$ for all $C > 0$. That is, our point estimator $\hat{\theta}$ of θ satisfies

$$P\big[|\hat{\theta} - \theta| \le \exp\{-n(\log n)^{-\gamma}\}\big] = O(n^{-C})$$

for all $C > 0$. Of course, since we may take $C > 1$ then strong convergence also obtains: $|\hat{\theta} - \theta| \le \exp\{-n(\log n)^{-\gamma}\}$ with probability 1.

It follows that convergence rates attainable using sequential algorithms are much faster than those available using traditional methods based on predetermined design points. In particular, if the n points at which f is observed are equally spaced across the interval then the rate at which θ is estimated cannot be improved beyond $O(n^{-1})$, with or without stochastic error in observations of f. See, for example, Loader (1996), Müller and Song (1997) and Gijbels, Hall and Kneïp (1999). These results imply the improvements claimed earlier for sequential

sampling. While the gains are theoretical, they are so great that their practical implications too can be expected to be significant; see the numerical results in Section 5.

The algorithm that we suggest involves sequential refinement of confidence intervals for the unknown changepoint and makes a reappraisal of the accuracy of the interval after each sequential step. If the reappraisal suggests that an error may have been committed then the next step (perhaps the next few steps) will involve checking current and previous decisions rather than refining the current confidence interval. One can obtain a simpler procedure by ignoring the reappraisal step, but from a theoretical viewpoint this is suboptimal, and in numerical practice it does not enjoy as good performance as the method introduced in Section 2.

There is a particularly extensive literature on estimation of jump points in otherwise-smooth functions of a single variable. In addition to the work cited above, recent wavelet-based methods [e.g., Wang (1995) and Raimondo (1996)] should be mentioned. Wang (1995) gives a particularly good literature survey, which we shall not repeat here except to note that a conference proceedings edited by Carlstein, Müller and Siegmund (1994) discusses an extensive variety of changepoint estimation problems in univariate cases.

In the spatial context there is a large, multidisciplinary literature on boundary detection, although seldom involving sequential methods. Techniques for global search [e.g., Zhigljavsky (1991) and Pronzato, Wynn and Zhigljavsky (2000)] are exceptions. However, while they frequently involve random aspects of design, they are seldom constructed to accommodate stochastic errors in observations of response functions. Optimal convergence rates and methods, for estimating boundaries using predetermined (i.e., nonsequential) design, have been discussed by Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995), for example. A likelihood-based approach has been suggested by Rudemo and Stryhn (1994) and alternative procedures have been proposed by Qiu and Yandell (1997), Qiu (1998) and Hall and Rau (2000). Particular properties of boundary estimation problems when design points are restricted to a regular lattice have been addressed by Hall and Raimondo (1997, 1998). The connections that exist between methods for image analysis and statistical techniques based on smoothing have been elucidated and developed by Titterington (1985a, b) and Cressie [(1993), pages 528–530].

The problem of sequentially inverting or minimizing a function observed with error, which is at the heart of a particularly extensive literature on stochastic approximation and sequential inference, is also related to that of sequential estimation of a changepoint. For the former, see, for example, Ruppert (1991) and Chapter 15 of Ghosh, Mukhopadhyay and Sen (1997). However, the nature of the results there is very different, not least in terms of the convergence rate. Moreover, the sequential sampling considered in the present paper is in batches, rather than individual data.

2. One-dimensional problem.

2.1. *Overview of problem and methodology.* Assume the function f is defined on an interval $\mathcal{I}$, and has a jump discontinuity at a point θ in the interior of $\mathcal{I}$. Specifically, we ask that, for differentiable functions g_1 and g_2,

$$f(x) = f(x|\theta) = g_1(x) + g_2(x) I(x > \theta) \tag{2.1}$$

where

$$\sup_{x \in \mathcal{I}} \max\{|g_1'(x)|, |g_2'(x)|\} < \infty, \qquad \gamma \equiv g_2(\theta_0) \neq 0,$$

and θ_0 denotes the true value of θ. We shall observe f at points $x = x_i \in \mathcal{I}$, subject to error: $Y_i = f(x_i) + \varepsilon_i$, where the design points x_i are open to sequential choice and the errors ε_i are independent and identically distributed with zero mean. The case where there is more than one changepoint and the number of changepoints is known would be treated very similarly. It has virtually identical theoretical properties and is omitted here only in order to simplify our discussion.

The case where the number of changepoints is unknown is more difficult. From a theoretical viewpoint it can be resolved satisfactorily as long as the number is known to be finite. There, the number can be determined empirically, to such accuracy that the probability of error converges to zero faster than the inverse of any polynomial in sample size.

Section 2.2 will introduce our recursive method for estimating θ. In practice this technique would be applied only after a "pilot" estimator, $\tilde{\theta}$, had been constructed using a portion of the permitted sample size, n. (A likelihood ratio approach is one technique for constructing $\tilde{\theta}$. We use this approach in the simulation study in Section 5.) This would lead to a preliminary interval $\mathcal{I}_1$, a strict subset of $\mathcal{I}$, in which the first estimator in the recursion would be constructed, using m design points $x_1 < \cdots < x_m$ equally spaced on $\mathcal{I}_1$. (Here and below, saying that $x_1, \ldots, x_m$ are "equally spaced" on $[c, d]$ means that, if we define $x_0 = c$ and $x_{m+1} = d$, then the values of $x_i - x_{i-1}$, $1 \le i \le m+1$, are all equal.) For notational simplicity, in Section 2.2 we shall take the permitted sample size for the recursive part of the algorithm to be n, although in our theoretical account in Section 4 we shall reduce this by the number of data that are used to construct $\tilde{\theta}$.

The interval $\mathcal{I}_1$ is the first of a sequence of confidence sets for the true value of θ. At the kth stage of the algorithm we shall determine $\mathcal{I}_k$. Assume $n = \ell m$, where ℓ, m are positive integers. Each sequential sample will be of size m, and there will be ℓ stages in the algorithm. In the first stage, distribute m equally spaced points on the first interval $\mathcal{I}_1$ and sample f at those places. Under the temporary assumption that the data are Normally distributed with known variance, compute the statistic $T(\theta)$ associated with a likelihood ratio test of the null hypothesis that f is constant on $\mathcal{I}_1$, against the alternative that f takes different but constant values on either side of θ. Take $\hat{\theta}_1$ to be that value of θ, chosen from among the m design points, that gives an extremum for the test.

2.2. *Sequential refinement with reassessment.* Let $\lambda > 0$. Assume that at the kth stage of the method, for $1 \le k \le \ell - 1$, an estimator $\hat\theta_k$ was obtained. Distribute m equally spaced points on $\mathcal{I}_k = [\hat\theta_k - (m^{-1}\lambda)^k, \hat\theta_k + (m^{-1}\lambda)^k]$ and construct the likelihood ratio test restricted to the new data on $\mathcal{I}_k$. The test leads to one of two possible conclusions. Either the maximum of the test statistic, over values of θ equal to the design points, exceeds a certain critical point c_{crit}, in which case we define $\hat\theta_{k+1}$ to be the value at which the maximum is attained, and pass to the next stage; or the maximum of the test statistic does not exceed c_{crit}, in which case we reassess our position.

We conduct the reassessment by considering again the interval $\mathcal{I}_k$, distributing m equally spaced points there, and constructing the likelihood-ratio test statistic for these new design points. (The data drawn at each step of the reassessment are completely independent of those used at any previous stages or steps.) If the test statistic computed on the latest occasion exceeds c_{crit}, then we deem the $(k+1)$st stage to have terminated and take $\hat\theta_{k+1}$ to equal the value of the design point in $\mathcal{I}_k$ at which the most recently computed test statistic achieved its maximum. On the other hand, if the most recently computed maximum does not exceed c_{crit} then the reassessment should continue. In this event we go back to the previous interval $\mathcal{I}_{k-1}$, distribute m new points there, compute the test statistic for these points, and compare it with the value obtained earlier for the previous dataset on $\mathcal{I}_{k-1}$. This makes it possible to correct estimation errors that would otherwise perpetuate, resulting from a wrong decision being taken at some stage. See Sections 2.5 and 5.5 for variants of this sequential refinement with reassessment (SRR) method.

This sequence of operations, in the reassessment part of the $(k+1)$st stage, continues until one of the following occurs: (a) in the next sampling step we would exceed the total number of data, n, that we are permitted to draw; or (b) we get back to $\mathcal{I}_1$ without having obtained a significant value (i.e., a value exceeding c_{crit}) of the test statistic; or (c) neither (a) nor (b) occurs before we obtain a significant value of the statistic. In case (c) we take $\hat\theta_{k+1}$ to be the design point, in the most recent sample, at which the most recently computed test statistic achieved its maximum value. If, at this time, we have used up all the n permitted sampling operations, then we take the final estimator $\hat\theta_{\text{SRR}}$ to equal $\hat\theta_{k+1}$. If we still have data remaining, however, then the sequential procedure continues to the next stage. In case (a) we take $\hat\theta_{\text{SRR}} = \hat\theta_k$. In case (b) we continue drawing new samples of size m, with design points equally spaced on $\mathcal{I}_1$, until either we reach the end of our allowance of n data (in which case we take $\hat\theta_{\text{SRR}} = \hat\theta_k$) or we obtain a test statistic whose value exceeds c_{crit} (in which case $\hat\theta_{k+1}$ is taken to be the point at which the most recently computed test statistic achieved its maximum value, and we pass to the next stage).

This algorithm can be represented graphically in at least two ways: first, as a tree diagram, in which all but one of the branches of the tree denote false starts that terminated as the result of a reassessment cycle; and, second, as a sequence

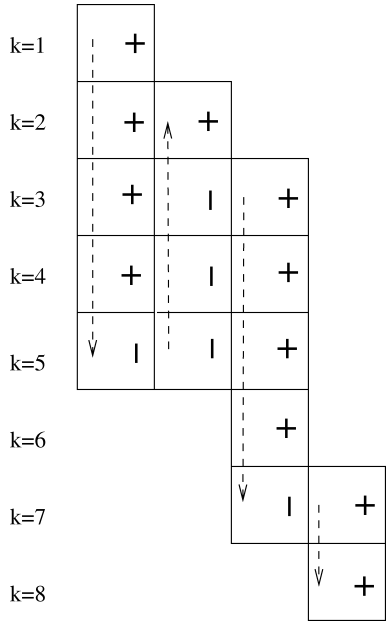

FIG. 1. *Schematic representation of the SRR method. Depth, k, in the algorithm is represented by the number of units on the vertical axis; pluses represent steps where the test statistic exceeds the critical value, and minuses represent the opposite outcome.*

of directed parallel lines, in which lines from left to right denote sequences of consecutive steps in which the value of the test statistic exceeded c_{crit}, and lines from right to left denote consecutive steps where the test statistic was less than c_{crit} and that step was used to reassess the step indicated immediately above it. Figure 1 illustrates the latter representation. The process starts from the top left corner and the vertical positions of the boxes represent depth, k, in the algorithm. The pluses represent steps where the test statistic exceeded the critical value, and the minuses represent the complementary situation.

2.3. *Main features of sequential refinement with reassessment.* For a general sequential method constructed along lines similar to those suggested in Section 2.2, the final estimators of θ would nominally have an accuracy equal to the width of the interval $\mathcal{I}_k$ at which the sequential construction terminated. If the interval at termination is $\mathcal{I}_\ell$ then its width will be proportional to $(m^{-1}\lambda)^\ell = (m^{-1}\lambda)^{n/m}$. However, without the reassessment step the estimator may stray from the true value of θ well before the end of the sequence of ℓ stages, so that later stages will be unreliable. In this case more data need to be used to guard against incorrect decisions at successive stages. The reassessment step in the SRR algorithm renders this unnecessary, however. As a result, more data can be used to estimate the changepoint itself.

For the SRR method, while the number of stages is random, with probability tending to 1 it exceeds $\delta\ell$ for some fixed $\delta \in (0, 1)$. Consequently, with high probability the width of the interval on termination will be no greater than $(m^{-1}\lambda)^{\delta n/m}$. And because of reassessment, the probability that this interval actually contains θ will also be high.

2.4. *Likelihood-ratio test.* Assume that at a given stage of the algorithm, data Y_i (where $1 \le i \le m$) are generated by the model $Y_i = f(x_i) + \varepsilon_i$, where the ε_i's are independent and identically distributed errors with zero mean and finite variance σ^2, and the design points $x_1 < \cdots < x_m$ are equally spaced on an interval $\mathcal{I}_k$. Assuming that σ^2 is known, a likelihood ratio test of the null hypothesis that f is constant on the interval, against the alternative that it takes different but constant values on either side of a changepoint θ, is to reject the null hypothesis if the quantity

$$\begin{aligned} T(\theta) &\equiv m_1(\theta)\bar{Y}_1(\theta)^2 + m_2(\theta)\bar{Y}_2(\theta)^2 - m\bar{Y}^2 \\ &= m^{-1}m_1(\theta)m_2(\theta)\{\bar{Y}_1(\theta) - \bar{Y}_2(\theta)\}^2 \end{aligned} \tag{2.2}$$

exceeds a critical point. Here, $\bar{Y}$, $\bar{Y}_1$ and $\bar{Y}_2$ denote the average values of Y_i over all indices i, over i such that $x_i \le \theta$ and over i such that $x_i > \theta$, respectively, and m, $m_1(\theta)$ and $m_2(\theta)$ are the respective numbers of terms in these averages.

Although $T(\theta)$ is motivated under the assumption that f is piecewise constant, and that the errors are Normally distributed, it is applicable in a wide range of other cases. Our theory will bear this out. The method could be refined by, for example, using a piecewise linear (rather than simply piecewise constant) approximation to f, and estimating the slopes of f to the left- and right-hand sides of a putative value of θ.

If the interval $\mathcal{I}_k$ on which the test statistic is constructed is short, if m is large, and if the true value θ_0 of θ divides the interval $\mathcal{I}_k$ into the proportion $p : (1 - p)$, then the maximum value attained by $T(\theta)$ will equal approximately $mp(1 - p)\gamma^2$, and the value at which it is achieved will be near to θ_0. [We defined γ at (2.1).] These heuristic considerations suggest taking the critical point c_{crit} for a test based on $T(\theta)$ to equal $m\xi(1 - \xi)$, where $0 < \xi < p$. This we do; see Section 4. In our asymptotic treatment, other aspects of the size of c_{crit} are unimportant.

2.5. *Refinements.* The SRR method suggested in Section 2.2 is only an example of a range of sequential techniques. In particular, one does not need to reassess at each step; reassessment at an appropriate proportion of steps is adequate. It is not essential to retrace one's path as soon as a reassessment contradicts a previous decision; one can wait until a number of consecutive contradictions are obtained. And one can reuse, perhaps in a weighted form, values obtained in the same interval in previous steps so as to recycle earlier data and improve efficiency.

It is possible to distribute design points more toward the center than the edges of confidence intervals, reflecting the relative likelihood that the true value of θ lies in different parts of the intervals. Moreover, particularly when reassessing an earlier decision, one need not place the design points at the same places as before. Changes such as these introduce only notational technicalities into the theoretical arguments in Sections 4 and 6 and have little effect on numerical properties.

3. Spatial problem. In the spatial case, f represents a response surface with a fault-type discontinuity in the $(x^{(1)}, x^{(2)})$-plane. The analogue of the representation at (2.1) in this case is

$$f(x) = g_1(x) + g_2(x) I\{x^{(2)} \le \psi(x^{(1)})\}, \tag{3.1}$$

where

$$\sup_{x,\omega} \max\{|D_\omega g_1(x)|, |D_\omega g_2(x)|\} < \infty, \qquad \inf_{x \in \mathcal{C}} |g_2(x)| > 0,$$

$x = (x^{(1)}, x^{(2)})$, the fault line is denoted by $\mathcal{C}$ and has equation $x^{(2)} = \psi(x^{(1)})$, and $D_\omega g(x)$ denotes the derivative of $g(x)$ in the direction of the unit vector ω. The model at (3.1) requires $\mathcal{C}$ to admit a single-valued Cartesian equation, although our methods are valid more generally.

We make no assumption about relative values of derivatives of g_1 and g_2 on either side of the fault line, and so the fault cannot necessarily be interpreted as the result of "slippage." We may observe f at arbitrary points x in the plane, subject to additive error. The x's are open to sequential choice, and the errors are independent and identically distributed. Using information obtained in this way we wish to estimate $\mathcal{C}$, or equivalently to estimate ψ.

As in the univariate case, it is instructive to consider the problem of approximating $\mathcal{C}$ when f may be observed deterministically, without stochastic error. This we do below, before developing the stochastic case by analogy.

If we are given a sequence of ν points along a given section of $\mathcal{C}$, approximately equally spaced, then $\mathcal{C}$ may be estimated with accuracy $O(\nu^{-k})$ by interpolation using a kth degree polynomial, provided its functional representation has at least k derivatives. We can of course improve on this rate if we have a parametric formula for $\mathcal{C}$, but otherwise the rate $O(\nu^{-k})$ is optimal, in a minimax sense, for approximating a k-times differentiable curve from ν approximately equally spaced points.

Of course, even in a deterministic setting we would be unlikely to be given points that are actually on the fault line. However, if we approximate the curve in a sequential manner then at any stage of the algorithm we shall have a good current approximation to both the location and the tangent to $\mathcal{C}$. To see how such an algorithm might proceed, suppose we wish to compute an approximation to $\mathcal{C}$ that is accurate to within $O(\delta)$, where δ will be taken to converge to 0. Assuming the curve has k bounded derivatives, we strike an arc, of radius $O(\delta^{1/k})$ and centered at

the current point, across the tangent approximation to the curve in the direction of travel. By placing $C_1|\log\delta|$ points sequentially across the arc, where $C_1 > 0$, and by measuring the response surface at those points and treating the approximation problem as one of estimating a changepoint in a univariate function defined on a line (on this occasion, the arc) and observed deterministically, without error, we may compute an approximation to the place at which the arc crosses the fault line, accurate to within $O(\delta^{C_2})$, for any given $C_2 \geq 2$, provided C_1 is sufficiently large. See Section 1 for discussion of the problem of sequentially estimating a changepoint on a line, using deterministic data.

This gives us a new current approximation to the fault line. By joining this point to the previous approximation and extrapolating in the direction of travel, we obtain a new approximation to the tangent. The error in the resulting approximation to slope is $O(\delta^{1/k})$, assuming $k \geq 2$. If the arc that we strike across the tangent subtends angle $\pi/2$ on either side of the point at which it intersects the tangent approximation, then it is sufficiently accurate for the next step of the algorithm.

Arguing in this way, in the context of direct observation (i.e., without random error) of a response surface, we can construct an algorithm that approximates a k-times differentiable fault line to within $O(\delta)$, uniformly along a bounded segment of its length, by using only $O(\delta^{-1/k}|\log\delta|)$ sampling operations. We may start the algorithm by constructing initial approximations to a point on the fault line, and to a tangent at that point, using transects placed across the curve. These initial steps cost only $O(|\log\delta|)$ sampling operations, and so do not affect the overall order of magnitude of expense.

The same approach may be employed when f is observed only with noise. The only significant change is that slightly more points need to be distributed across the arc when estimating the next point on the fault line and the gradient of the tangent at the next point. The increase is from $O(|\log\delta|)$ to at most $O(|\log\delta|^{1+\alpha})$, for $\alpha > 0$ arbitrarily small. (In fact, the factor $|\log\delta|^{\alpha}$ may be reduced to a power of $\log|\log\delta|$.) Therefore, for any $\alpha > 0$, we may approximate a k-times differentiable fault line to within $O(\delta)$ after only $\delta^{-1/k}|\log\delta|^{1+\alpha}$ sampling operations, when the response surface is observed with stochastic error. This result will be discussed in more detail in Section 4; see particularly Theorem 4.3. A numerical example will be given in Section 5.6.

4. Theoretical properties. It will be assumed that each test is conducted as described in Section 2.4, using $c_{\text{crit}} = m\xi(1-\xi)$ where $0 < \xi < \frac{1}{2}$. Furthermore, each test will be applied only against values $\theta = x_i$ that are sufficiently far from the endpoints of $\mathcal{I}$ that both $m_1(\theta)$ and $m_2(\theta)$, in the definition of $T(\theta)$ at (2.2), are averages of at least $N \equiv Cm$ data, for an arbitrarily small but fixed positive constant $C \leq \min(\xi, 1-\xi)$. Thus, an estimator $\hat{\theta}$ of θ within an interval will be defined by maximizing $T(\theta)$ over $\theta = x_i$ for $N \leq i \leq n - N$. For notational simplicity we shall treat N as though it were an integer.

Theorems 4.1 and 4.2 will address the one-dimensional problem, and Theorem 4.3 will illustrate application of Theorem 4.2 to the spatial problem.

Assume the sampled data are generated by the model described in Section 2.1, where in particular f satisfies (2.1), and that the errors are independent and identically distributed with zero mean and finite variance. Call these conditions (C_1). Divide the proposed sample size, n, into two parts, of respective sizes n_1 and n_2. The value of n_2 should be at least as large as δn for some $\delta \in (0, 1)$. Draw n_1 data in a single operation (that is, nonsequentially), and use them to construct a "preliminary" or "pilot" estimator $\tilde{\theta}$ of the changepoint θ, with the property that, for all $\alpha > 0$ and some $\beta > 0$,

$$P\bigl(|\tilde{\theta} - \theta_0| \le \alpha n^{-\beta}\bigr) \to 1. \tag{4.1}$$

Standard methods that guarantee (4.1) with $\beta < 1$ are discussed in papers cited in Section 1. The case $\beta \ge 1$ is not feasible unless an exceptionally fortuitous design sequence is selected.

Divide the second sample into ℓ subsamples of size m, where ℓ denotes the integer part of n_2/m. Use these to carry out the "sequential refinement with reassessment" algorithm described in Section 2.2, starting with $\mathcal{I}_1 = (\tilde{\theta} - n^{-\beta}, \tilde{\theta} + n^{-\beta})$ and producing the estimator $\hat{\theta}_{\mathrm{SRR}}$.

We claim it is possible to choose ℓ and m so that, for any given sequence $\rho = \rho(n) \downarrow 0$, and any model satisfying (C_1), $\hat{\theta}_{\mathrm{SRR}} = \theta_0 + O_p(e^{-\rho n})$ as $n \to \infty$. Indeed, $\hat{\theta}_{\mathrm{SRR}}$ will satisfy

$$P\bigl(|\hat{\theta}_{\mathrm{SRR}} - \theta_0| \le e^{-\rho n}\bigr) \to 1. \tag{4.2}$$

THEOREM 4.1. *Assume conditions* (C_1), *and given* $\rho = \rho(n) \downarrow 0$, *choose* $m = m(n)$ *and* $\lambda = \lambda(n)$ *to diverge to* ∞, *in such a manner that* $\lambda/m \to 0$ *and* $(m\rho)^{-1} \log(m/\lambda) \to \infty$. *Using these values, construct* $\hat{\theta}_{\mathrm{SRR}}$ *as suggested above. Then* (4.2) *holds.*

A refinement of the proof of Theorem 4.1 shows that, for appropriate choices of m and λ that are fixed and do not depend on n, there exists a fixed constant $\rho > 0$ with the property $\hat{\theta}_{\mathrm{SRR}} = \theta_0 + O_p(e^{-\rho n})$:

$$\lim_{C \to \infty} \liminf_{n \to \infty} P\bigl(|\hat{\theta}_{\mathrm{SRR}} - \theta_0| \le C e^{-\rho n}\bigr) = 1. \tag{4.3}$$

Choice of m, λ and ρ depends intimately on properties of the error distribution, however. Therefore, (4.3) is arguably not as significant as the result addressed in Theorem 4.1. Nevertheless, construction of a version of $\hat{\theta}_{\mathrm{SRR}}$ that gives (4.3) is straightforward if it is assumed that the errors are Gaussian.

Next we state analogous results which provide a rate of convergence for the probability at (4.2). This will prove helpful in addressing extensions of our methods to the spatial case. Construct $\hat{\theta}_{\mathrm{SRR}}$ as described earlier, by dividing the second potential sample size, n_2, into ℓ lots of size m each, with ℓ equal to the integer part of n_2/m.

THEOREM 4.2. *Assume conditions* (C_1) *and in addition that the error distribution has finite moment generating function in a neighborhood of the origin. Choose* $m = m(n)$ *and* $\lambda = \lambda(n)$ *such that*

$$n^{-1}m + m^{-1}\{\lambda + (\log n)^2\} + \lambda^{-1}\log n \to 0, \tag{4.4}$$

and for $C_1 > 0$ *put* $\rho = \rho(n) \equiv C_1 m^{-1}\log(m/\lambda)$, *which converges to* 0. *Then*

$$P\big(|\hat{\theta}_{\mathrm{SRR}} - \theta_0| \le e^{-\rho n}\big) = 1 - O(n^{-C}) \tag{4.5}$$

for all $C > 0$.

For example, suppose we take $\lambda(n) \asymp (\log n)^{1+\alpha}$ and $m(n) \asymp (\log n)^{2+\beta}$, where $0 < \alpha < 1 + \beta$, $\beta > 0$, and the notation $a(n) \asymp b(n)$ means that $a(n)/b(n)$ is bounded away from zero and infinity as $n \to \infty$. Then (4.4) holds. This choice shows that we may ensure (4.5) with $\rho = (\log n)^{-\gamma}$ for any $\gamma > 2$. In particular, the extra conservatism of procedures that have polynomially small chances of error involve a deterioration in the convergence rate by only a logarithmic factor applied to ρ.

Section 3 introduced sequential methods for approximating a smooth fault line in a regression surface, assuming the surface could be observed without error. It was argued that the algorithm, and its accuracy and cost of sampling, are almost identical in the case of stochastic error. Theorem 4.3 below verifies this claim.

Indeed, suppose we may observe the response surface with error: $Y = f(x) + \varepsilon$, where f satisfies (3.1), the function ψ defining the fault line $\mathcal{C}$ has k bounded derivatives, and the errors ε are independent and identically distributed with zero mean and finite moment generating function in the neighborhood of the origin. Call these conditions (C_2). Assume too that we have constructed an initial estimate of a point on the fault line and of the slope at that point, which are accurate to within $C_1\delta^{C_2}$ and $C_1\delta^{C_3}$, respectively, for any $C_1 > 0$ and some $C_2 > 1$ and $C_3 > 0$, with probability $1 - O(\delta^C)$ for all $C > 0$, where $\delta \to 0$. (In view of Theorem 4.2, this order of accuracy may be achieved at the expense of only $|\log\delta|^{1+\alpha}$ sampling operations, for any $\alpha > 0$, by sampling along a transect of the fault line.) Strike an arc of radius $\delta^{1/k}$ across the tangent in the direction of travel, with its center at the previously computed approximation to a point on $\mathcal{C}$, and subtending angle $\pi/2$ radians on either side of the point at which it intersects the tangent estimate. By distributing $|\log\delta|^{1+\alpha}$ points sequentially within the arc, where $\alpha > 0$ is fixed but otherwise arbitrary, and by using either of the methods suggested in Section 2, we may locate the point at which the arc crosses the fault line to within $O(\delta^C)$, and with probability $1 - O(\delta^C)$, for all $C > 0$. (This result follows from Theorem 4.2.)

Repeating this sequence of steps and noting that only polynomially many steps are required, whereas the error of approximation is of the stated size with probability $1 - O(\delta^C)$ for all $C > 0$, we see that with the latter probability, after only $\delta^{-1/k}|\log\delta|^{1+\alpha}$ sampling operations, we have computed $\delta^{1/k}$ points

that are each within $O(\delta^C)$ of the true fault line, for all $C > 0$, and are equally spaced except for errors that equal $O(\delta^C)$ for all $C > 0$. Interpolating among these points, and exploiting the fact that f has k bounded derivatives, we obtain an approximation to the fault line that is accurate to $O(\delta)$. We have proved the following result.

THEOREM 4.3. *If conditions* (C_2) *hold and* $\alpha > 0$ *then we may develop a sequential approximation to the fault line that, with probability* $1 - O(\delta^C)$ *for all* $C > 0$, *is accurate to* $O(\delta)$ *uniformly along any given, bounded segment of the line and employs no more than* $\delta^{-1/k} |\log \delta|^{1+\alpha}$ *sampling operations.*

Indeed, the factor $|\log \delta|^{1+\alpha}$, for any $\alpha > 0$, may be reduced to

$$|\log \delta| (\log |\log \delta|)^{\beta},$$

for some $\beta > 0$, by refining the same argument. These sampling rates compare favorably with those in more conventional problems, where a function with k bounded derivatives can be estimated, with accuracy no better than $O(n^{-k/(k+1)})$, from n random (e.g., Poisson-distributed) design points in the plane. See, for example, Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995). Solving the equation $n^{-k/(k+1)} = \delta$ for n, we see that such nonsequential sampling procedures require at least $O(\delta^{-1-(1/k)})$ sampling operations in order to achieve $O(\delta)$ accuracy; sequential sampling has reduced this to $O(\delta^{-1/k})$, times a logarithmic factor, for an approximation of $O(\delta)$. [A logarithmic factor must be appended to the sample size $O(\delta^{-1-(1/k)})$ in order that the rate $O(\delta)$ be achievable uniformly along a given segment of the fault line. Otherwise the rate is available only in a pointwise sense.]

5. Numerical studies.

5.1. *Simulation set-up.* We shall treat the problem of sequentially estimating θ when $f(x) = f(x|\theta) \equiv I(x > \theta)$. Suppose the true value of θ is $\theta_0 = \frac{1}{2}$ and consider drawing data $Y = f(x) + \varepsilon$, where the errors ε are independent and $x \in \mathcal{I} = [0, 1]$. We present below simulation studies for errors having the Normal distribution with mean 0 and standard deviation $\sigma = 0.7$.

We shall compare a nonsequential estimation method, using the likelihood ratio test described in Section 2.4, with our SRR method. Both techniques will be applied to a common (but varying) number of sampled data, n. Of course, the nonsequential method uses n observations at once when applying the likelihood ratio test; the SRR method employs the test using only a fraction of n each time. The nonsequential method involves distributing n equally spaced points x_i within $\mathcal{I}$ and estimating θ as the value of x_i at which $T(\theta)$, defined at (2.2), achieves its maximum value. To ensure good performance of both approaches we

took θ only as close to the ends of $\mathcal{I}$ as was possible without reducing the number of data on which $T(\theta)$ was based.

The SRR method requires us to specify λ, ℓ, the proportion of data used to construct the pilot estimator and also the critical point $c_{\text{crit}} = m\xi(1 - \xi)$. For each chosen combination of parameters we performed $N = 1000$ independent simulations. When implementing the SRR method we used $\frac{1}{2}n$ points to produce the pilot estimator. The latter was computed using the conventional nonsequential likelihood ratio approach discussed in the previous paragraph. The other $\frac{1}{2}n$ points were employed to improve the estimator, using our sequential method with ℓ steps based on m points each, so that $n = 2\ell m$.

5.2. *Comparison of sequential and nonsequential methods.* We shall report results that compare the nonsequential and SRR methods for the following values of parameters: $\lambda = 15$, $\xi = 0.1$ and $\ell = 10$. To ensure adequate quantities of data were used when computing the log-likelihood ratio, we did not permit i/m to get closer than 0.1 to the endpoints 0 and 1 of $\mathcal{I}$. Figure 2(a) plots the ratio of the standard errors for the sequential and nonsequential estimates obtained from the 1000 independent simulations against the value of m in the range 50 to 250, in steps of 5. (The value of n in each case was $2\ell m$.) Specifically, for each estimator type (i.e., sequential or nonsequential) and each value of m, we computed the standard error from the 1000 independent simulations. Then, for each given value of m, to construct the ratio we divided the standard deviation for the sequential method by its counterpart for the nonsequential approach. It is clear from the figure that for $m \geq 75$ the SRR method performs substantially better than the nonsequential one.

Indeed, the improved performance is available much more generally than this. The increase in standard deviation of the SRR method at $m = 70$ is the result of a single aberrant dataset out of the $N = 1000$ that we simulated in that setting. It can be removed by slightly increasing λ, ξ and m. We have not done so, however, since the uncharacteristic decline in performance demonstrated by the "blip" in Figure 2(a) serves a didactic purpose, showing that properties of the SRR method depend to some extent on choice of the tuning parameters.

The fluctuations that lead to the blip are indeed caused by very rare events, as panel (b) of Figure 2 shows. There we plot values of the ratios of robust scale estimators. Here each scale estimate is defined as the median of absolute differences between estimates of θ and the true value of θ. The value of the ratio is depicted by the unbroken line in the figure. The sequential method is seen to give improved performance by a factor of about 2.6 for $m = 30$, rising to 10^6 for $m = 100$ and to 10^{10} for $m \geq 200$.

It is readily seen from these results that the SRR method improves strikingly on even the best possible deterministic result, based on distributing n evenly spaced points in $\mathcal{I}$ and observing f without noise. Even taking an extremely conservative view, the error of the best deterministic approximation can be no less than n^{-1} times that of the absolutely best possible nonsequential estimator when noise

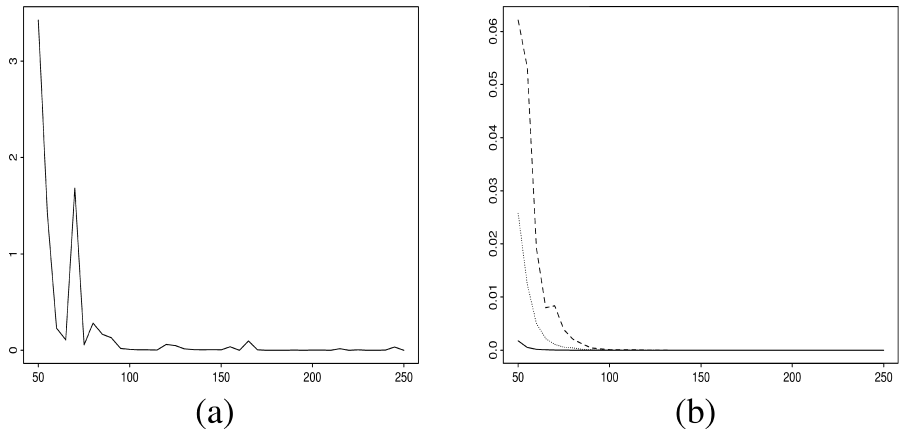

FIG. 2. (a) *Ratio of standard errors for sequential and nonsequential estimates*; (b) *median absolute deviation ratios* (*unbroken line*), *and their counterparts for* 90% *quantiles* (*dotted line*) *and* 99% *quantiles* (*dashed line*), *for sequential and nonsequential methods. In each panel the vertical axis shows the value of the ratio*, *and the horizontal axis shows* m. *Each sample size was* $n = 20m$.

is present. However, as we have just seen, the SRR estimator is far more accurate than this.

Some idea of the effects of stochastic variability can be gained by looking at ratios of high-level quantiles of absolute values of the differences between estimates and the true value of θ. Figure 2(b) shows plots of these ratios for 90% quantiles (dotted line) and 99% quantiles (dashed line). In particular, the ratio of the 99% quantile is below 0.063 for all $m \geq 50$. In that sense, the error of the SRR estimator is more than 15 times less than that of its nonsequential counterpart, for 99% of samples whenever $m \geq 50$.

5.3. *Further analysis of SRR method.* Implementation of the SRR method relies on choice of several parameters. Below we report on a comparison of results obtained when some parameters are varied while others are kept fixed.

Changes in ξ of course influence the level of the likelihood ratio test. Choosing ξ too large results in too many refinement steps being rejected, which worsens overall performance of $\hat{\theta}_{\mathrm{SRR}}$. To explore this property, two series of simulations were undertaken, one using $\xi = 0.1, 0.12, 0.14, \ldots, 0.3$, where i/m was not permitted to be closer than 0.1 to the ends of $\mathcal{I}$, and the other taking $\xi = 0.02, 0.04, \ldots, 0.3$, where i/m was kept at least 0.02 from the ends. (For simplicity we shall not mention any further the latter requirement, which had only a very minor impact on performance.) Values of m ranged from 30 to 250. We assessed performance using both standard deviations and median absolute deviations. In most cases it was found that the sequential method gave better results for values of ξ near the lower end of its range.

Our results also showed that the relationship between ℓ and ξ, for fixed m, had surprisingly little impact on performance. For example, taking $m = 50$ and

varying ℓ from 5 to 50 we observed that the smallest robust scale estimates, and the smallest quantiles of absolute differences, were obtained for ξ in the range 0.1–0.16, without showing any obvious trends. However, it was seen that when the standard deviation criterion was applied, rather than mean absolute deviation, slightly higher values of ξ were needed to achieve optimal performance.

Choice of λ for the sequential method was explored for $m = 50$, 100 and 200 and $\xi = 0.1$. Optimal performance using either the standard deviation criterion, or that based on *maximum* absolute difference, was obtained for $\lambda = 13$, 19 and 29, corresponding, respectively, to the values chosen for m. However, when employing *mean* absolute difference the optimal values of λ were substantially smaller, at $\lambda = 7$, 13 and 13, respectively. These properties result from the fact that standard deviation is affected by a very small number of large deviations. It was found too that, while the optimal value of λ increased with m, the optimal value of λ/m (proportional to the widths of the intervals $\mathcal{I}_k$) decreased with increasing m. That is, it was advantageous to decrease interval lengths with increasing m.

5.4. *Influence of the pilot estimator.* The reassessment part of our sequential method ensures that the method successfully overcomes inaccuracies in intermediate estimation steps when estimating θ. In particular, the SRR estimator is surprisingly robust against poor choice of the pilot. To illustrate this property we took $\ell = 10$, $m = 50$, $\lambda = 15$ and $\xi = 0.1$, resulting in $n = 2\ell m = 1000$. But we calculated the pilot estimator using only 50 points, one-twentieth of the full dataset; the pilot was thus very highly variable. Nevertheless, the sequential method produced particularly reliable final results. For the setting just described, Figure 3 shows 10,000 plots of estimates as functions of k, the stage of the reassessment procedure.

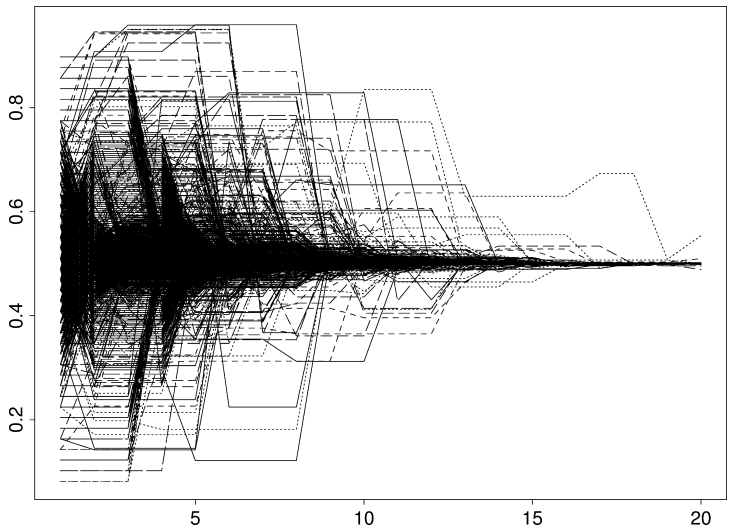

FIG. 3. *Ten thousand individual estimates as functions of the stage of the SRR method. Sample size was* $n = 1000$.

5.5. *Variants of the reassessment method.* We simulated two variants of our SRR method. One involved the modification that if the likelihood ratio test did not produce a significant result at a given step, it was reapplied on a substantially enlarged interval, rather than simply using the interval associated with the preceding step. This gave results very similar to standard SRR. The other variant involved keeping interval length constant at that where the nonsignificant value of the likelihood ratio statistic was encountered when working through the reassessment steps. This gave worse results than conventional SRR.

5.6. *Simulation of the spatial problem.* We implemented the method suggested in Section 3, using a smooth quadratic fault line $\mathcal{C}$ and data generated by the model $Y_i = f(x_i) + \varepsilon_i$. The function f was as defined at (3.1), with $\psi(x) = 0.8x^2 + 0.1$, $g_1 \equiv 0$ and $g_2 \equiv 1$. The function ψ is illustrated in either panel of Figure 4. For simplicity we used the same error distribution as in Section 5.2 and also the same tuning parameters: $\lambda = 15$, $\xi = 0.1$ and $\ell = 10$.

The initial estimate was chosen by applying the SRR method to the one-dimensional changepoint problem on the left-hand vertical edge of the unit square $\mathcal{S} = [0, 1]^2$. Then a semicircle was drawn, centered at the initial estimate and with its axis horizontal. The next estimate was found by applying the SRR method to the one-dimensional problem on the semicircle. From the first two estimates of points on $\mathcal{C}$ one may obtain an approximation to the tangent. Each subsequent estimate was computed by striking an arc (of radius 0.02 and subtending angle $2\pi/3$) across the most recent tangent estimate and solving the one-dimensional changepoint problem on the arc, using the SRR method. In this way the algorithm worked its way along $\mathcal{C}$ from the bottom left to the top right of $\mathcal{S}$, stopping as soon as the estimate exited the square.

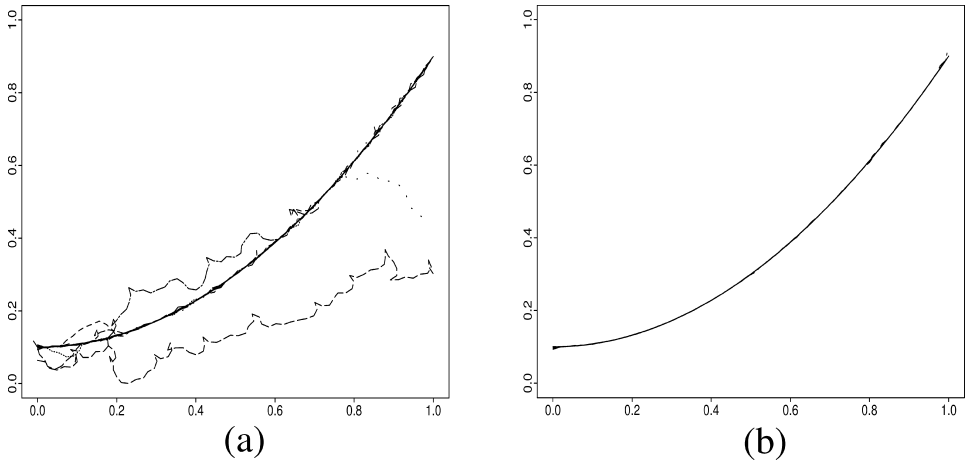

FIG. 4. *Plots of the fault line* $\mathcal{C}$ *and of* 100 *sequential estimates for* (a) $m = 35$ *and* (b) $m = 50$.

Panels (a) and (b) of Figure 4 each show the results of 100 simulations for $m = 35$ and $m = 50$, respectively. The latter values were chosen since they lie on either side of the smallest value (approximately $m = 40$) for which the algorithm loses contact with $\mathcal{C}$, within $\mathcal{S}$, less than 1% of the time. In particular, when $m = 35$ the algorithm strays well away from $\mathcal{C}$ on two occasions out of 100, and on a few other occasions it meanders some distance from $\mathcal{C}$ but manages to return. However, for $m = 50$ it hardly departs from $\mathcal{C}$ for any part of any of the 100 estimates.

6. Proofs.

6.1. *Preliminaries for proofs of Theorems* 4.1 *and* 4.2. Suppose we are conducting the test on an interval $\mathcal{J}$ of bounded length $\eta = \eta(n)$. In the following discussion we regard $\mathcal{J}$ and η as nonstochastic, although in practice they would involve stochastic effects. There, the probabilities considered below would be interpreted conditional on the past. The bounds obtained would nevertheless be the same deterministic bounds, available with probability 1 in the probability space generated by past events.

Assume initially that θ_0, denoting the true value of θ, is an element of $[x_N, x_{m-N})$, and let ${\theta_0}'$ be the design point (x_{i_0}, say) such that $x_{i_0} \le \theta_0 < x_{i_0+1}$. Let $\theta_1 = x_{i_1}$ denote any design point for which $N \le i_1 \le m - N$. It may be proved that $T(\theta_0) = T({\theta_0}') = T(\theta_1) + T_1 + T_2$, where $T_1 = (S_1 - S_2)\{2\Delta - \nu(S_1 + S_2)\}$, S_1 and S_2 equal the averages of Y_i over $N \le i \le i_0$ and $i_0 + 1 \le i \le m - N$, respectively, Δ equals the sum of Y_i over $N \le i \le i_1$ minus the sum over $N \le i \le i_0$, $\nu = i_1 - i_0$, $\delta_1 = \nu/(i_0 - N + 1)$, $\delta_2 = \nu/(m - N - i_0)$ and

$$\begin{aligned} T_2 = {} & -2\Delta\{S_1\delta_1(1+\delta_1)^{-1} + S_2\delta_2(1-\delta_2)^{-1}\} \\ & + \Delta^2\{(i_0 - N + 1)^{-1}(1+\delta_1)^{-1} + (m - N - i_0)^{-1}(1-\delta_2)^{-1}\} \\ & + (i_0 - N + 1)\delta_1^2(1+\delta_1)^{-1}S_1^2 + (m - N - i_0)\delta_2^2(1-\delta_2)^{-1}S_2^2. \end{aligned}$$

Define $D_1 = \sum_{N \le i \le i_0} \varepsilon_i$ and

$$D_2(i_1) = \begin{cases} \displaystyle\sum_{i=i_0+1}^{i_1} \varepsilon_i, & \text{if } i_0 < i_1, \\ \displaystyle\sum_{i=i_1+1}^{i_0} \varepsilon_i, & \text{if } i_0 > i_1, \\ 0, & \text{if } i_0 = i_1. \end{cases}$$

If $m = m(n) \to \infty$ then, since the ε_i's have zero mean and finite variance, we have for all $\zeta > 0$,

$$P(|D_1| > m\zeta) + P\left\{\sup_{N \le i_1 \le m-N} |D_2(i_1)| > m\zeta\right\} \to 0. \tag{6.1}$$

We may deduce from this property and the definition of T_2 that $T_2 = T_3 + T_4$, where T_3 is nonstochastic and equals $O(\nu^2/m)$ uniformly in $N \le i_1 \le m - N$, T_4 is stochastic and vanishes if $\nu = 0$, and for all $\zeta > 0$,

$$P\left\{\sup_{N \le i_1 \le m-N} |T_4(i_1)/\nu| > \zeta\right\} \to 0. \tag{6.2}$$

[If $V = V(i_1)$ is a random variable that vanishes when $i_1 = i_0$, we interpret V/ν as zero if $i_1 = i_0$, i.e., if $\nu = 0$.]

Similarly, $T_1 = -|\nu|\gamma^2 + T_5 + T_6$, where T_5 is nonstochastic and equals $O(|\nu|\eta)$, and T_6 is stochastic and satisfies $T_6(i_0) = 0$ and, for $k = 6$,

$$P\left\{\sup_{N \le i_1 \le m-N} |T_k(i_1)/\nu| > \zeta\right\} \le o(1) + P\left\{\sup_{N \le i_1 \le m-N} |D_2(i_1)/\nu| > C_1\zeta\right\}, \tag{6.3}$$

the constant $C_1 > 0$ not depending on m, n or ζ.

We may deduce from (6.2) and (6.3) that

$$T(\theta_0) = T(\theta_1) - |i_1 - i_0|\gamma^2 + T_7 + T_8,$$

where T_7 is nonstochastic and equals $O\{(i_1 - i_0)^2 m^{-1} + |i_1 - i_0|\eta\}$, and T_8 is stochastic and satisfies $T_8(i_0) = 0$ and (6.3). It follows from these properties that if $\eta(n) \to 0$ then

$$\lim_{\zeta \to \infty} \limsup_{n \to \infty} \sup_{x_N \le \theta_0 < x_{m-N}} P_{\theta_0}(|\hat{\theta} - \theta_0| > \eta\zeta m^{-1}) = 0, \tag{6.4}$$

where P_{θ_0} denotes probability measure under the model $f(\cdot|\theta_0)$ for f.

More simply, it may be proved that if $c_{\text{crit}} = mc$, where $c = \xi(1 - \xi)$ and $\xi \in (\xi', \frac{1}{2})$, and provided $\eta(n) \to 0$, then for any $\xi_1 \in (\xi, \frac{1}{2})$ and $\xi_2 \in (\xi', \xi)$,

$$\inf_{x_{\xi_1 m} \le \theta_0 < x_{m-\xi_1 m}} P_{\theta_0}(\sup T > c_{\text{crit}}) \to 1, \tag{6.5}$$

$$\sup_{\theta_0 \le x_{\xi_2 m} \text{ or } \theta_0 \ge x_{m-\xi_2 m}} P_{\theta_0}(\sup T > c_{\text{crit}}) \to 0, \tag{6.6}$$

where $\sup T$ denotes the supremum of $T(x_i)$ over $N \le i \le m - N$. [The role of ξ' is to ensure that each series in the definition of $T(\theta)$ is based on at least Cm data, for some $C > 0$. Note too that, if θ_0 is fixed at a number which divides $\mathcal{I}$ in the proportion $p : (1 - p)$, then the ratio of $\sup_\theta T(\theta)$ to $mp(1 - p)\gamma^2$ converges to 1 in probability.]

6.2. *Proof of Theorem* 4.1. The sequential refinement with reassessment method involves a sequence of ℓ tests, the jth of which we may take to give a result R_j which equals 1 if the corresponding version of $\sup T$ exceeds c_{crit} and equals 0 otherwise. Thus, the sequence of tests produces a vector $R = (R_1, \ldots, R_\ell)$ of 0's and 1's. Results (6.4)–(6.6) imply the following property, which we call (P_1). Conditional on $R_j = 1$ and $R_{j+1} = 0$, and for $k \ge 2$, the probability that

"$R_{j+2} = \cdots = R_{j+k+1} = 0$ and $R_{j+k+2} = 1$" is bounded above by π_1^k, where $\pi_1 > 0$ does not depend on j and $\pi_1 = \pi_1(n) \to 0$.

To derive (P_1), note that, in view of the "reassessment" aspect of the SRR method, a sequence $R_{j+2} = \cdots = R_{j+k+1} = 0$ may be interpreted as a sequence of k pairs of independent tests, in identical settings and in a reassessment cycle of the algorithm, where the two test results are conflicting. The test pairs give results $(R_{r_i}, R_{j+k+2-i}) = (1, 0)$, for $1 \le i \le k$, where $r_1 < \cdots < r_k = j$. If for the ith pair of tests in the reassessment cycle, giving result $(R_{r_i}, R_{j+k+2-i})$, the value of θ_0 is within the central proportion $1 - 2\xi_2$ of the interval, then, for the $(i+1)$st pair of tests, the probability that θ_0 is within the central proportion $1 - 2\xi_1$ is close to 1, and therefore the probability that $(R_{r_{i+2}}, R_{j+k+2-(i+2)}) = (1, 1)$ is close to 1. Hence, the probability that $(R_{r_{i+2}}, R_{j+k+2-(i+2)}) = (1, 0)$ is close to 0. On the other hand, if for the pair $(R_{r_i}, R_{j+k+2-i})$ the value of θ_0 is not within the central proportion $1-2\xi_2$ of the interval, then the probability that $(R_{r_{i+1}}, R_{j+k+2-(i+1)}) = (0, 0)$ is close to 1, and so the probability that $(R_{r_{i+1}}, R_{j+k+2-(i+1)}) = (1, 0)$ is close to 0. Property (P_1) follows from these results.

Property (P_1) implies that runs of 0's in the vector R are relatively short. In particular, the probability that the length of an arbitrary run of 0's exceeds 3 converges to 0 as $n \to \infty$. Call this property (P_2).

Results (6.4)–(6.6) imply that if θ_0 is in the central proportion $1 - 2\xi_2$ of the interval on the occasion of the jth test then, with probability close to 1, both (a) θ_0 is in the central proportion λm^{-1} of the interval on the occasion of the $(j+1)$st test, and (b) $R_{j+1} = 1$. If (a) holds then, with probability close to 1, $R_{j+2} = 1$. Moreover, (6.4)–(6.6) imply that if θ_0 is not in the central proportion $1-2\xi_2$ of the interval on the occasion of the jth test then the probability that $R_j = 0$ is close to 1. It follows that sequences of 1's in the vector R are relatively long, with the probability of not only the length of an arbitrary sequence exceeding C, but also the number of tests in the sequence for which θ_0 is in the central proportion λm^{-1} of the interval exceeding C, converging to 1 for any $C > 0$. Call this property (P_3).

Together, properties (P_2) and (P_3) imply that, for some $\delta > 0$, the probability that, among the intervals remaining at the end of the algorithm for the SRR method, there are at least $\delta\ell$ for which θ_0 is in the central proportion λm^{-1} of the interval, converges to 1 as $n \to \infty$. (The intervals that remain at the end of the algorithm are those that correspond to tests that gave the result $R_j = 1$ and which were not overridden in a reassessment cycle of the algorithm.)

Theorem 4.1 follows from the latter result and the fact that the intervals that remain at the end of the algorithm are nested. Indeed, this property implies that, with probability tending to 1 as $n \to \infty$, θ_0 is contained in an interval centered on $\hat{\theta}_{\mathrm{SRR}}$ and of width no more than $2t$, where $t = (\lambda/m)^{\delta\ell}$. (Here, δ is as in the previous paragraph.) Since ℓ is no smaller than a constant multiple of n/m, then, for some $C > 0$, t is not of larger order than $s \equiv \exp\{-C(n/m)\log(m/\lambda)\}$. The definitions of m and λ in the theorem imply that $s = o(e^{-\rho n})$.

6.3. *Proof of Theorem* 4.2. If the errors are independent and Gaussian, and if $\zeta \geq 1$, then the left-hand sides of (6.1) and (6.2) are both equal to $O\{m^{-1/2}\exp(-C_1 m)\}$ for some $C_1 > 0$ not depending on ζ. Still in the Gaussian case, if we put $\zeta = (C_2 \log n)^{1/2}$ on the left-hand side of (6.3) then the right-hand side may be taken as $O(n^{-C_2 C_3})$, where $C_3 > 0$ does not depend on C_2.

This leads to the following analogue of (6.4), valid for Gaussian errors: if ζ_n is any sequence of positive constants diverging to infinity, and if $m^{-1}\log n \to 0$, then

$$\sup_{x_N \leq \theta_0 < x_{m-N}} P_{\theta_0}\{|\hat{\theta} - \theta_0| > \eta \zeta_n m^{-1}(\log n)^{1/2}\} = O(n^{-C}) \tag{6.7}$$

for all $C > 0$. Likewise, provided $m^{-1}\log n \to 0$, the following versions of (6.5) and (6.6) hold in the Gaussian case:

$$\inf_{x_{\xi_1 m} \leq \theta_0 < x_{m-\xi_1 m}} P_{\theta_0}(\sup T > c_{\text{crit}}) = 1 + O(n^{-C}), \tag{6.8}$$

$$\sup_{\theta_0 \leq x_{\xi_2 m} \text{ or } \theta_0 \geq x_{m-\xi_2 m}} P_{\theta_0}(\sup T > c_{\text{crit}}) = O(n^{-C}) \tag{6.9}$$

for all $C > 0$.

To obtain analogous results for non-Gaussian errors we employ Gaussian approximations to processes of partial sums. In particular, defining $U_i = \sum_{j \leq i} \varepsilon_j$, and writing σ^2 for the variance of the errors ε_i, there exists a standard Brownian motion W such that

$$P\left\{\max_{1 \leq i \leq n} |U_i - \sigma W(i)| > c_1 \log n + x\right\} \leq c_2 \exp(-c_3 x) \tag{6.10}$$

for all $x > 0$, where c_1, c_2, c_3 depend only on the error distribution. See, for example, Shorack and Wellner [(1986), page 66ff.]; we have used the fact that the distribution of the errors has a moment generating function in a neighborhood of the origin.

Since the intercept term in the quantity "$c_1 \log n + x$" on the left-hand side of (6.10) is proportional to $\log n$, and since $\exp(-c_3\zeta_n) = O(n^{-C})$ for all $C > 0$ if $\zeta_n / \log n \to \infty$, then in view of (6.10) the additional complication of non-Gaussian errors may by incorporated by considering deviations that are of larger order than $\log n$ rather than just $(\log n)^{1/2}$. Arguing thus we may show that, provided $m^{-1}(\log n)^2 \to 0$, (6.8) and (6.9) hold without change in the non-Gaussian case and (6.7) continues to hold provided we remove the exponent from $(\log n)^{1/2}$ on the left-hand side.

In consequence, if $\zeta_n \to \infty$ and we take $\lambda = \zeta_n \log n$ in the proof of Theorem 4.1, and choose m to diverge to infinity sufficiently fast for $m^{-1}\{\lambda + (\log n)^2\} \to 0$, then all the probability approximations stated in that proof are accurate to order n^{-C} for all $C > 0$. In particular, probabilities that were close to 0 or 1 are now within $O(n^{-C})$ of those respective quantities, for all $C > 0$. As a result, with probability $1 - O(n^{-C})$ for all $C > 0$, $\hat{\theta}_{\text{SRR}}$ is within $C_2(\lambda/m)^{C_1 n/m}$ of θ_0 for some $C_1, C_2 > 0$. This establishes (4.5) in the case of $\bar{\theta} = \hat{\theta}_{\text{SRR}}$.

Acknowledgments. The authors are grateful to D. M. Titterington for stimulating discussions. The very helpful comments of two referees are gratefully acknowledged.

REFERENCES

CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D. (1994), eds. *Change-Point Problems.* IMS, Hayward, CA.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*, rev. ed. Wiley, New York.

GHOSH, M., MUKHOPADHYAY, N. and SEN, P. K. (1997). *Sequential Estimation.* Wiley, New York.

GIJBELS, I., HALL, P. and KNEÏP, A. (1999). On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.* **51** 231–251.

HALL, P. and RAIMONDO, M. (1997). Approximating a line thrown at random onto a grid. *Ann. Appl. Probab.* **7** 648–665.

HALL, P. and RAIMONDO, M. (1998). On global performance of approximations to smooth curves using gridded data. *Ann. Statist.* **26** 2206–2217.

HALL, P. and RAU, C. (2000). Tracking a smooth fault line in a response surface. *Ann. Statist.* **28** 713–733.

KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statist.* **82**. Springer, Berlin.

LOADER, C. L. (1996). Change-point estimation using nonparametric regression. *Ann. Statist.* **24** 1667–1678.

MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524.

MÜLLER, H.-G. and SONG, K.-S. (1997). Two-stage change-point estimators in smooth regression models. *Statist. Probab. Lett.* **34** 323–335.

PRONZATO, L., WYNN, H. P. and ZHIGLJAVSKY, A. A. (2000). *Dynamical Search. Applications of Dynamical Systems in Search and Optimization.* Chapman and Hall, London.

QIU, P. (1998). Discontinuous regression surfaces fitting. *Ann. Statist.* **26** 2218–2245.

QIU, P. and YANDELL, B. (1997). Jump detection in regression surfaces. *J. Comput. Graph. Statist.* **6** 332–354.

RAIMONDO, M. (1996). Modèles en rupture, situations non ergodique et utilisation de méthode d'ondelette. Ph.D. dissertation, Univ. Paris VII.

RUDEMO, M. and STRYHN, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Statist.* **21** 41–55.

RUPPERT, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis* (B. K. Ghosh and P. K. Sen, eds.) 503–529. Dekker, New York.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

TITTERINGTON, D. M. (1985a). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53** 141–170.

TITTERINGTON, D. M. (1985b). General structure of regularization procedures in image reconstruction. *Astronom. and Astrophys.* **144** 381–387.

WANG, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika* **82** 385–397.

ZHIGLJAVSKY, A. A. (1991). *Theory of Global Random Search.* Kluwer, Dordrecht.

CENTRE FOR MATHEMATICS
AND ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 0200
AUSTRALIA
E-MAIL: peter.hall@anu.edu.au

INSTITUT FÜR MATHEMATISCHE
STATISTIK UND VERSICHERUNGSLEHRE
UNIVERSITÄT BERN
SIDLERSTRASSE 5
CH-3012 BERN
SWITZERLAND
E-MAIL: ilya.molchanov@stat.unibe.ch