Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Advanced Distribution Theory for SiZer

J Hannig & J. S Marron

To cite this article: J Hannig & J. S Marron (2006) Advanced Distribution Theory for SiZer, Journal of the American Statistical Association, 101:474, 484-499, DOI: 10.1198/016214505000001294

To link to this article: https://doi.org/10.1198/016214505000001294

Advanced Distribution Theory for SiZer

J. HANNIG and J. S. MARRON

SiZer is a powerful method for exploratory data analysis. In this article approximations to the distributions underlying the simultaneous statistical inference are investigated, and large improvements are made in the approximation using extreme value theory. This results in improved size, and also in an improved global inference version of SiZer. The main points are illustrated with real data and simulated examples.

KEY WORDS: Extreme value theory; Kernel smoothing; Multiple testing adjustment; SiZer.

1. INTRODUCTION

SiZer has proven to be a valuable technique for exploratory data analysis by smoothing methods. These methods include histograms and smoother approaches to understanding the structure of one-dimensional distributions (called the "density estimation setting" here) and scatterplot smoothers (called the "regression setting" here). (See, e.g., Scott 1992; Wand and Jones 1995; Fan and Gijbels 1996 for an introduction to this area.) As noted by earlier authors, many smoothing (i.e., estimation) schemes have been proposed. (See Marron 1996 for an overview of the many criteria that have been used to compare different smoothing methods.) Kernel based-methods (definitions of which are given in Sec. 2) are considered here for their simplicity and ease of interpretation, and because they have been very widely studied.

The practical use of kernel methods in both density estimation and regression is profoundly affected by the choice of the window width (the tuning parameter that controls the amount of local averaging used). When this is too small, the resulting estimated curve is strongly affected by sampling variation and is wiggly, reflecting spurious artifacts of the sampling process. For too large a window width, the curve estimate smooths away important underlying features. There is a large literature on data-based selection of the window width, in which attempts are made to estimate it from the data (see Jones, Marron, and Sheather 1996a,b); however, the problem is very challenging. There are limits on how well this selection can be done in practice, and there has never been a consensus on "the best" method of doing this, which has appeared to hinder the actual use of these methods through, for example, their implementation in software packages.

Scale-space ideas (see Chaudhuri and Marron 2000 for a broad discussion of these issues) have provided practical means of avoiding the problem of bandwidth selection. Scale space is a theoretical model for vision constructed in the computer vision community. The model is simply a family of Gaussian window smooths, indexed by the window width. It is a model for vision in the sense that large values of the window width correspond to standing back and viewing a scene macroscopically, whereas small values correspond to a zoomed-in view. (See Lindeberg 1994 and ter Haar Romeny 2001 for access to the scale-space literature.) A fundamental concept of scale space, which is the heart of SiZer, is that instead of trying to choose a single "best scale" (i.e., best window width), one should use all of them (i.e., study the full family of smooths). This is clear in a vision modeling context, because different levels of resolution of an image (i.e., smooths with different window widths) contain different types of useful information.

SiZer is a combination of the scale-space idea of simultaneously considering a family of smooths, with the statistical inference needed for exploratory data analysis, in the presence of noise. In particular, SiZer addresses the question of "which features observed in a smooth are really there?," meaning representing important underlying structure, not artifacts of the sampling noise.

For reasonable statistical inference using SiZer, care needs to be taken regarding the multiple comparison issue. In particular, the visual display of SiZer can be viewed as a summary of a large number (hundreds) of hypothesis test results. Current implementations of SiZer address this issue using the fairly crude "independent blocks" idea, developed in section 3 of Chaudhuri and Marron (1999). In this article we conduct a much deeper distributional investigation, with the goal of improving the statistical performance of SiZer.

The SiZer method and potential advantages from an improved distribution theory are illustrated in Figure 1. The underlying regression function, shown as the thick black curve in Figure 1(a), is the Blocks example from Donoho and Johnstone (1994), which appears to be rather challenging to estimate by smoothing methods because of the 11 sharp jumps. To make the problem even more challenging, a high level of Gaussian noise, $\sigma = .1$ (much higher than is typical in the wavelet literature), first used by Marron, Adak, Johnstone, Neumann, and Patil (1998), is used in generating of the $n = 1{,}024$ data points shown as green dots in Figure 1(a).

The thin blue curves in Figure 1(a) show the scale space for this dataset, that is, the family of smooths for a wide range of different window widths. Some of these are seriously oversmoothed, showing strong rounding of the corners; some are undersmoothed, showing spurious wiggles. None is very good at attaining the goal of recovering the thick black curve. Wavelets (see, e.g., Donoho and Johnstone 1994) are a compelling approach to the problem of recovering curves such as this with nonsmooth features. But for these data even wavelets give poor signal recovery, because the noise level is so high.

SiZer has a somewhat different goal. Instead of trying to recover the underlying black curve as well as possible, it aims

J. Hannig is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (E-mail: *jan.hannig@colostate.edu*). J. S. Marron is Amos Hawley Professor of Statistics and Operations Research, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *marron@email.unc.edu*). Jan Hannig's research is supported in part by National Science Foundation under grant DMS-05-04737. J. S. Marron's research is supported in part by National Science Foundation under grant DMS-03-08331.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000001294

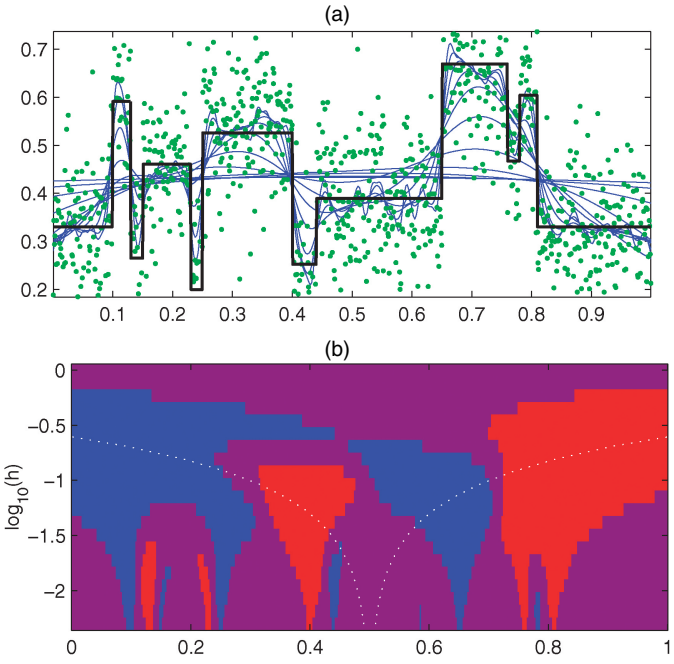

Figure 1. Conventional SiZer Analysis of the Donoho–Johnstone Blocks Regression, With High Noise, Showing Good Performance, Plus a Spurious Feature. (a) True regression, data and scale space. (b) SiZer analysis.

instead at understanding which of its features can be distinguished from the background noise, that is, determining which aspects observable in the blue curves are important underlying structure and which are spurious noise-driven artifacts.

SiZer focuses on finding regions of statistically significant slope in the blue curves. Slope works well in the example of Figure 1, because the interesting features there are the 11 jumps (elsewhere the regression is flat). With the high noise level used in Figure 1(a) (making signal recovery challenging, even by the best wavelet methods), determining which jumps are statistically significant turns out to be attainable by SiZer. In other cases of data analysis using smoothing methods, bumps are of interest. Bumps are also determined by slope, because the curve slopes up on one side and down on the other side. In general, SiZer flags features of these various types using a color map, such as the one shown in Figure 1(b).

The horizontal location in the SiZer map are the same as in the scale-space plot in Figure 1(a). The vertical locations correspond to the window widths of the family of blue curves, shown on the log scale. Each pixel shows a color that essentially gives the result of a hypothesis test for the slope of the blue curve, at the point indexed by the horizontal location, and at the scale (window width) corresponding to that row. When the slope is significantly positive (negative) the pixel is colored blue (red, resp.). When the slope is not significant (as happens in regions where sampling noise is dominant), the color purple is used. There is a fourth SiZer color that does not appear in Figure 1(b), which is gray, used to show pixel locations where the data are too sparse for reasonable statistical inference. (For the exact rule on labeling pixels gray, see Chaudhuri and Marron 1999.) The rule on labeling pixels gray is not changed by the theory developed in this article.

Note that each jump in Figure 1(a) corresponds to a red or a blue (depending on the direction of the jump) region in the SiZer map in Figure 1(b). Thus SiZer has correctly found all 11 of the jumps in the thick black curve, so for the specific goal of finding important features, it substantially outperforms wavelet methods. (See Marron et al. 1998 for discussion of some wavelet analysis results.)

A very careful look at the SiZer map shows a small, unexpected feature: a tiny blue region at the finest scales (the bottom of the map) near .58. This is suggesting the slope is statistically significant, when, in fact, the underlying target curve is flat. Such features have been observed in a number of other cases as well. This has not presented a serious obstacle to data analysis by SiZer, because analysts have learned to not put too much credence on such very small features when they are flagged by SiZer. But it is still very desirable to eliminate these, to give a more precise analysis. This goal is attained in the present article, by developing an improved distribution theory.

First, we take a deeper look at the extent of the problem of small spurious features appearing in the SiZer map, by studying some simulations. Figures 2 and 3 show some SiZer maps for simulated data from the null distribution in the case of equally spaced design regression. Because the regression function is 0, the data are simply iid standard Gaussian random variables. In this situation the SiZer map should ideally be completely purple, except for perhaps $\alpha 100\%$ of the cases in the size-α case (here α is always taken to be .05).

The SiZer maps shown in Figure 2 illustrate the population of SiZer maps for this underlying distribution. They were drawn from a simulated sample of 1,000 such SiZer maps. The population was ordered in terms of number of pixels that flag significant structure by being red or blue. Because these were drawn

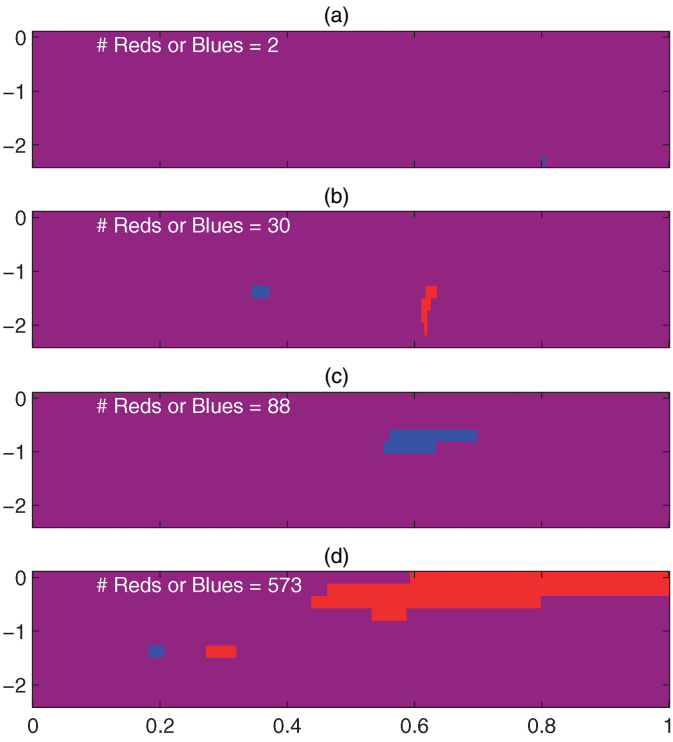

Figure 2. Conventional SiZer Maps, Based on Simulated Null Distributions, for $n = 1{,}600$ Equally Spaced Regression Data Points. (a) .5 quantile of the distribution; (b) .75 quantile of the distribution; (c) .85 quantile of the distribution; (d) .95 quantile of the distribution.

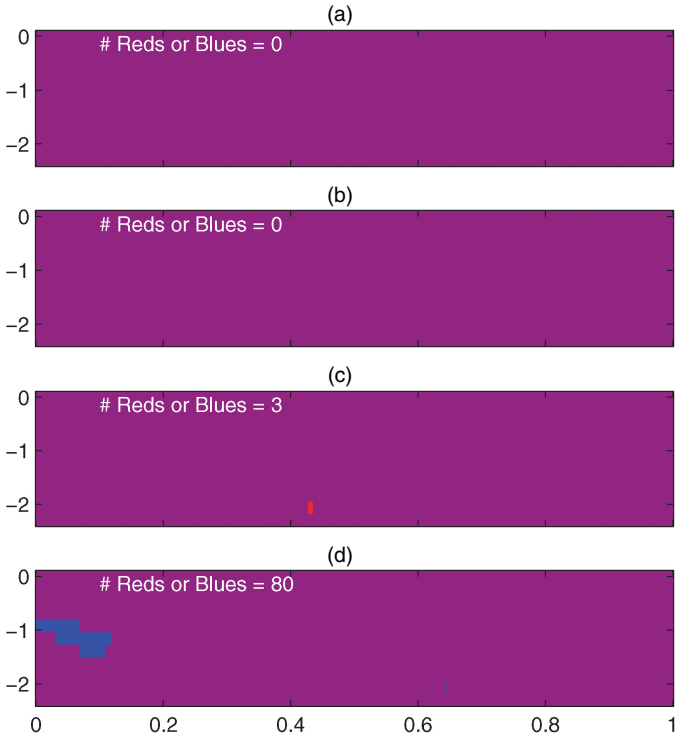

Figure 3. SiZer Maps for Simulated Null Distributions, for n = 1,600 Equally Spaced Regression Data Points, Based on the Proposed Row-Wise Procedure. (a) .5 quantile of the distribution; (b) .75 quantile of the distribution; (c) .85 quantile of the distribution; (d) .95 quantile of the distribution.

from the null distribution, it is desirable that the number of such pixels is small. The first 405 of the 1,000 ordered SiZer maps were completely purple (and are thus not shown to save space). Figure 2(a) shows the 500th of these (essentially the median of the population), where two pixels, at the finest scale, were flagged as significant. Figure 2(b) shows the 750th (the third quartile), with substantially more significant pixels at medium-fine scales. Figure 2(c) shows the 850th SiZer map, with quite a large blue region at medium coarse scales. Figure 2(d) is the 950th member of the ordered population, showing an even larger red region at the coarsest scales, plus the suggestion of a small mode at medium scales. There appears to be a relationship between the number of spurious pixels and the scales at which they appear, which is not surprising because at coarse scales, adjacent pixels are strongly correlated.

This suggests a serious need for improvement in the size characteristics of the conventional SiZer. The ideal here is that Figures 2(a)–2(c) should be completely purple and that Figure 2(d) might or might not have some color. The goal of this article is to improve this performance by using a better approximation of the underlying distribution theory.

A natural solution to this problem would be to use simulation methods to compute the critical values needed for proper simultaneous adjustment. This idea was seriously considered by Chaudhuri and Marron (1999, sec. 3) and was implemented in early versions of the SiZer software. But there was a serious drawback: The simulation took hours, whereas the crude approximation came up in only a few seconds. In applications of SiZer, the interactive capabilities of the crude approximation were preferred so uniformly that the simulation version was simply phased out as the software was adapted over the years. Of course computers are faster now, so the simulation no longer takes hours, but it still does take some minutes, enough to keep the method out of the class of *interactive* methods. The method proposed in this article has the advantage of achieving very good distributional properties in a really interactive way.

The results of the main proposed solution (later referred to as row-wise adjustment) are shown in Figure 3. The format is the same as Figure 2, based on the same 1,000 underlying datasets, but this time an improved version of the SiZer map is used. Again the maps were ordered, and the 500th, 750th, 850th, and 950th of the 1,000 maps are shown as Figures 3(a), 3(b), 3(c), and 3(d).

The SiZer maps in Figure 3 flag far less spurious structure than was found for the corresponding population quantiles in Figure 2. In particular, in Figures 3(a) and 3(b) (representing the first three quartiles) there were no spurious results. Even for the 850th ordered SiZer map in Figure 3(c), the spurious structure is quite small. Hence the improved SiZer map studied in Figure 3 clearly has better size properties than the original SiZer shown in Figure 2; however, these results are still not completely satisfactory.

This size problem is driven by a number of factors studied in Section 2, the most important of which is that the simultaneous inference is only row-wise in nature. This means that the SiZer inference in Figure 3 is adjusted only row by row. Hence it is not surprising that some spurious structure manages to be flagged here, because each of the maps in Figure 3 includes 11 such rows (so, just by chance, the test flags significant structure more than 5% of the time).

To address this problem, we also propose a global adjustment in Section 2. We plotted the corresponding globally adjusted version of Figure 3 but do not include it here (to save space), because each of the panels is completely purple, indicating that the size problem has been solved.

The distribution theory that drives the improvements in the statistical performance of SiZer shown in Figures 2 and 3 is developed in Section 2, with the main recommendations summarized in Section 2.5. A detailed analysis of the impact of these improvements is provided in Section 3.

As expected, the improved size properties, investigated further in Section 3.1, come at a some cost in terms of power. Power issues are studied for simulated data in Section 3.2 and for real datasets in Section 3.3. The main lessons learned there are that although the loss of power appears to be minimal for the row-wise adjustment, it is very significant for the global adjustment.

In our personal opinion, the substantial loss of power by the global method makes the row-wise improved SiZer more useful for data analysis than the global versions. The reason for this is that away from the null distribution (i.e., when the underlying target curve actually has some interesting structure), the spurious features of the type illustrated in Figures 2 and 3 tend to come up far less frequently than suggested by the size analysis. We consider this an acceptable price to pay for most exploratory data analyses. However, we anticipate that others will disagree, and recognize situations where statistical rigor is imperative, and thus our software

allows a choice between row-wise and global implementations, together with an option of choosing the level of significance α. Matlab software based on these new ideas can be found at *http://www.stat.unc.edu/postscript/papers/marron/Matlab6Software/Smoothing/*.

In addition to this important row-wise versus global issue, there are also a number of other points, such as the impact of smoothing boundary effects, that are also discussed in Section 2.

2. IMPROVED DISTRIBUTIONS

To aid in the development of the distributional properties of SiZer, we begin by reviewing some basics of kernel smoothing. Convenient notation for density estimation is $X_1, \ldots, X_n$ for a random sample from a probability density $f(x)$. The kernel density estimate of f is

$$\widehat{f_h}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{1}$$

where K_h is a "kernel function," indexed by a "window-width" h. The estimator $\widehat{f_h}(x)$ is simply interpreted as "putting probability mass $1/n$ in a region near each data point," where the window width controls the critical amount of spread of this mass. The window width h is important enough to appear as a subscript in $\widehat{f_h}$. In all examples in this article, K_h is taken to be the Gaussian density function, with standard deviation h, because of its very natural scale-space interpretations. It is also important to point out that the scale-space ideas naturally lead to making inference about the smoothed density $\int f(t) K_h(x - t)\, dt$ rather than about the density $f(x)$. (See Chaudhuri and Marron 1999, 2000 for discussion on these subjects.)

Our notation for regression data is $(X_1, Y_1), \ldots, (X_n, Y_n)$. Such data arise in several ways and admit several mathematical models. The term "equally spaced design regression" is used to mean that the X_i are deterministic and equally spaced (in order) and that $Y_i = m(X_i) + \varepsilon_i$, where m is the regression function, and where $\varepsilon_1, \ldots, \varepsilon_n$ are iid. The term "random design" means that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are random samples from a bivariate distribution, with $E(Y_i | X_i) = m(X_i)$, so that again m is the regression function. For random design regression, it can also be useful to think of "residuals," defined as $\varepsilon_i = Y_i - m(X_i)$. For both settings, a common estimator is the local linear smoother, defined at each location x as

$$\begin{aligned} \widehat{m}(x) &= a_0, \quad \text{where} \\ (a_0, a_1) &= \arg\min_{a_0, a_1} \sum_{i=1}^{n} \big\{ Y_i - [a_1(X_i - x) + a_0] \big\}^2 \\ &\qquad \times K_h(x - X_i). \end{aligned} \tag{2}$$

This estimator is simply interpreted as providing a local linear fit, in a window centered at x determined by K_h, which is then "moved along" over the range of x values. Again there are many competing estimators, but the local linear smoother is the focus of this article, for the same reasons as the kernel density estimator. As before, the kernel window function K_h is the Gaussian density function, with standard deviation h.

Because SiZer requires evaluation of a number of smooths (indexed by the window width h), the fast-binned implementation discussed by Fan and Marron (1994) is important, especially for larger datasets.

2.1 SiZer Distribution Theory

Like other hypothesis tests, part of the performance of SiZer is driven by the distribution of SiZer under the null hypothesis of "no signal." It is desired to set the size of the test (i.e., the probability of "false positives") to be a small preset value α. There are two natural approaches to addressing the multiple comparison problems. The first, called "row-wise" simultaneous inference, seeks to have at most $\alpha 100\%$ of the rows containing "false positives." The second, called "global" simultaneous inference, aims at having at most $\alpha 100\%$ of the SiZer maps containing false positives.

To analyze the "row-wise" problem, fix a particular row of the SiZer map. The row contains colored pixel values, which report the results of a family of hypothesis tests. The distribution theory for each row is that of a sequence of test statistics (modeled as random variables) at each grid point in the domain of the smoother, that is, at each pixel location in the SiZer map. Let $T_1, \ldots, T_g$, where g is the number of grid points, denote these test statistics. The pixels are equidistant, and we can assume without loss of generality that the ith pixel is in the location $i\tilde{\Delta}$ for some $\tilde{\Delta} > 0$. It is worth pointing out that the locations (and number) of grid points of the SiZer map can differ significantly from the location and number of design points $(X_1, \ldots, X_n)$.

At the ith pixel in this given row of the SiZer, the color blue (significantly increasing) is used when $T_i > C$, and the color red is used when $T_i < -C$. The overall size of the row-wise simultaneous SiZer inference will be α when C is chosen such that, under the null distribution of the target curve being constant,

$$P\big[\{T_i > C \text{ or } T_i < -C\} \text{ for some } i\big] = \alpha. \tag{3}$$

In the Appendix we show that the sequence $T_1, \ldots, T_g$ can be approximated by a stationary Gaussian process with mean 0, variance 1, and correlation

$$\operatorname{corr}(T_i, T_{i+j}) \approx e^{-j^2 \tilde{\Delta}^2/(4h^2)} \left[1 - \frac{j^2 \tilde{\Delta}^2}{2h^2} \right], \tag{4}$$

where h is the bandwidth associated with the SiZer map row. In this approximation, boundary issues that introduce nonstationarity are ignored. In what follows we refer to the test statistics of a fixed SiZer row as T_i.

Similarly, the whole SiZer map is a matrix of pixels generated based on a matrix of test statistics,

$$\begin{pmatrix} T_{1,1} & \cdots & T_{g,1} \\ \vdots & \ddots & \vdots \\ T_{1,r} & \cdots & T_{g,r} \end{pmatrix}.$$

Each row of the matrix corresponds to a particular bandwidth, and each column corresponds to a particular location. SiZer bandwidths are chosen on a logarithmic scale (sensible, because bandwidth is a multiplicative notion), and we can assume without loss of generality that the kth row was calculated using the bandwidth hd^k for some $h > 0$ and $0 < d < 1$.

Again the random field $T_{1,1}, \ldots, T_{g,r}$ can be approximated by a mean-0, variance-1 Gaussian random field with correlation

$$\begin{aligned}&\operatorname{corr}(T_{i,k}, T_{i+j,l})\\ &\quad\approx e^{-j^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))}\\ &\qquad\times\left[1-\frac{j^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2}.\end{aligned}\tag{5}$$

In what follows we use $T_{i,j}$ to denote the test statistics of the SiZer map. To use the theorems derived in the next sections, we assume that the approximations in (4) and (5) are sufficient in the interchange of limits. Conditions on such approximations could be obtained using an appropriate Hungarian approximation technique (see Csörgő and Révész 1974/75 for the original work on this subject).

2.2 Row-Wise Extreme Value Theory for SiZer

The row-wise simultaneous inference used in SiZer depends on finding approximate solutions, in C, to (3). Chaudhuri and Marron (1999) used a "number of independent blocks" approach to give a first approximate solution. In this article, much more precise approximations are developed. These come from

$$\begin{aligned}&P\big[\{T_i > C \text{ or } T_i < -C\} \text{ for some } i\big]\\ &\quad= P[|T_i| > C \text{ for some } i]\\ &\quad= 1 - P[|T_i| < C \text{ for all } i]\\ &\quad= 1 - P\Big[\max_i |T_i| < C\Big].\end{aligned}$$

If $T_1, \ldots, T_g$ were independent, then the distribution needed is simply a power of the distribution of the absolute value of a Gaussian random variable, because

$$P[|T_i| < C \text{ for all } i] = \prod_{i=1}^{g} P[|T_i| < C] = P[|Z| < C]^g,$$

where Z is a standard Gaussian random variable.

Of course, the main challenge is due to the fact that SiZer pixels are not independent. Toward that end, consider a stationary, mean-0, variance-1 Gaussian process $T_1, \ldots, T_g$, with a j step correlation denoted by ρ_j. We are interested in the distribution of $\max(T_1, \ldots, T_g)$. Berman (1964) has proven that if $\log(j)\rho_j \to 0$, then the distribution function of $\max(T_1, \ldots, T_g)$ behaves asymptotically as the gth power of the distribution function of a standard Gaussian random variable, that is,

$$\big|P[\max(T_1, \ldots, T_g) \le z] - \Phi(z)^g\big| \to 0 \quad \text{as } g \to \infty. \tag{6}$$

Unfortunately, this approximation is usually of little practical significance, because the speed of convergence is very slow. To overcome this, one needs to consider second-order asymptotics. There are at least two alternate approximations in the literature based on more detailed asymptotics, with the aim of improving the small-sample properties of (6). The first approach, discovered by Rootzén (1983), shows that if the time series is m dependent, if $g(1 - \Phi(x_g)) \to \kappa$, and if $\max(\rho_1, \ldots, \rho_m) > 0$, then

$$P[\max(T_1, \ldots, T_g) \le x_g] - \Phi(x_g)^g \sim e^{-\kappa} R_g \quad \text{as } g \to \infty,$$

where R_g is positive quantity depending only on the ρ_j's, g, and κ. The formula for R_g is very complicated, so we do not reproduce it here. (An interested reader can consult sec. 4.6 of Leadbetter, Lindgren, and Rootzén 1983 for details.)

The second approach, discussed by Hsing, Husler, and Reiss (1996), is based on the observation that for dependent data, it is often better to approximate $P[\max(T_1, \ldots, T_g) \le x]$ by $\Phi(x)^{\theta g}$, where $\theta < 1$. Their main idea is to find θ using asymptotic considerations. To get $\theta < 1$, the correlation ρ_j must increase to 1 with g for each fixed j.

To achieve this, Hsing et al. (1996) embedded the series in a triangular array $\hat{T}_{j,g}$, where rows are indexed by g. For each fixed g, the random variables $\hat{T}_{j,g}$, $j = 1, 2, \ldots$, compose a mean-0, variance-1 stationary Gaussian series with the j step correlations $\rho_{j,g}$ satisfying

$$\log(g)(1 - \rho_{j,g}) \to \delta_j \quad \text{as } g \to \infty, \text{ for all } j,$$

where $\delta_j \in (0, \infty]$. They define

$$\vartheta = P\big[V/2 + \sqrt{\delta_k} H_k \le \delta_k \text{ for all } k \ge 1\big], \tag{7}$$

where V is a standard exponential random variable and H_k is a mean-0 Gaussian process independent of V that satisfies $EH_iH_j = (\delta_i + \delta j - \delta_{|i-j|})/(2\sqrt{\delta_i\delta_j})$. Hsing et al. then claimed that under certain technical conditions on $\rho_{j,g}$, the distribution function $P[\max(\hat{T}_{1,g}, \ldots, \hat{T}_{g,g}) \le x]$ could be approximated by $\Phi(x)^{\vartheta g}$. The parameter ϑ has been called the "cluster index."

Wilhelm (2002) performed an extensive simulation study comparing the three possible approaches for a wide class of stationary Gaussian processes. The simulation study proved inconclusive, because no method clearly dominated the others. In fact, none of the approaches seemed to give reliable answers in the case of highly dependent stationary series. In our simulation study, described in Section 3.1, implementation of the Rootzén method had even worse performance than the conventional SiZer approach, based on the independent block calculation, whose size characteristics are illustrated in Figure 2. Thus in the remainder of this article we use only the approach of Hsing et al. (1996), which dramatically improves the size of SiZer, as seen in Figure 3.

In the particular case of SiZer, as noted in Section 2.1, it is reasonable to assume that under the null hypothesis, $T_1, \ldots, T_g$ are Gaussian, with mean 0 and variance 1 and j step correlation $\rho_j = e^{-j^2\tilde{\Delta}^2/(4h^2)}[1 - j^2\tilde{\Delta}^2/(2h^2)]$. A natural way to embed our SiZer row into a triangular array compatible with Hsing et al. (1996) is to assume that $\tilde{\Delta}/h = C/\sqrt{\log g}$. This choice leads us to the following theorem, the proof of which is given in the Appendix.

Theorem 1. Consider a triangular array $\hat{T}_{i,g}$ of mean-0, variance-1 Gaussian random variables. For each fixed g the random series $\hat{T}_{1,g}, \ldots, \hat{T}_{g,g}$ is stationary with j step correlation

$$\rho_{j,g} = e^{-j^2C^2/(4\log g)}\left[1 - \frac{j^2C^2}{2\log g}\right],$$

where $C > 0$. Then

$$\lim_{g\to\infty} P\Big[\max_{i=1,\ldots,g} \hat{T}_{i,g} \le u(x)\Big] = e^{-\vartheta e^{-x}}, \tag{8}$$

where

$$\vartheta = 2\Phi\left(\frac{\sqrt{3}C}{2}\right) - 1 \tag{9}$$

and

$$u(x) = \sqrt{2\log g} + \frac{x}{\sqrt{2\log g}} - \frac{\log\log g + \log 4\pi}{\sqrt{8\log g}}.$$

Hsing et al. (1996) recommended that in applications, Theorem 1 should be used to approximate $P[\max_{i=1,\ldots,g} \hat{T}_{i,g} \le x]$ by $\Phi(x)^{\vartheta g}$ rather than by the limiting Gumbel distribution. Their reasoning is based on the fact that the Gaussian power distribution converges to the Gumbel distribution of Theorem 1 and the empirical fact that the Gaussian power distribution often fits better then the limiting Gumbel distribution. Following their recommendation, we conclude that in the case of SiZer,

$$P\Big[\max_{i=1,\ldots,g} T_i \le x\Big] \approx \Phi(x)^{\theta g}, \tag{10}$$

where the cluster index is

$$\theta = 2\Phi\left(\sqrt{3\log g}\frac{\tilde{\Delta}}{2h}\right) - 1. \tag{11}$$

Recall that $\tilde{\Delta}$ is the distance between the pixels of the SiZer map, g is the number of pixels on each row, h is the bandwidth used for the fixed row studied, and Φ is the standard normal distribution function.

Chaudhuri and Marron (2002) have shown that in a number of real data situations, an interesting structure can be found in the data using a curvature-based version of SiZer. In some cases this discovered structure is not flagged as statistically significant by the slope version. Hence we derive an analogous formula that can be used for this curvature version. Using a similar approximation as for the slope version of SiZer, we conclude that under the null hypothesis, the curvature SiZer version test statistics $\bar{T}_1, \ldots, \bar{T}_g$ are approximately Gaussian, with mean 0, variance 1, and j step correlation $\bar{\rho}_j = e^{-j^2\tilde{\Delta}^2/(4h^2)}(1 - j^2\tilde{\Delta}^2/h^2 + j^4\tilde{\Delta}^2/(12h^4))$. This leads to the cluster index of

$$\bar{\theta} = 2\Phi\left(\sqrt{5\log g}\frac{\tilde{\Delta}}{2h}\right) - 1.$$

Detailed discussion, with examples, are of some interest. However, they are not included here (except for Fig. 11), because the general ideas are the same as for the slope version of SiZer, so it does not seem to be worth the space.

2.3 Global Extreme Value Theory for SiZer

We need to study the asymptotic distribution of the maxima of the whole SiZer map, that is,

$$\max_{i=1,\ldots,g}\max_{j=1,\ldots,r} T_{i,j}.$$

The main result of this section shows that the maximum of the SiZer map behaves asymptotically as if the rows were independent.

In the particular case of SiZer, as noted in Section 2.1, it is reasonable to assume that under the null hypothesis, $T_{1,1}, \ldots, T_{g,r}$ are Gaussian, with mean 0, variance 1, and correlation given by (5). To be able to make use of Theorem 1, we again set $\tilde{\Delta}/h = C/\sqrt{\log g}$. The following theorem is proven by comparing the maximum of a SiZer map with the maximum of a similar map where the rows are assumed to be independent. The comparison is done using a powerful generalization of Slepian's lemma due to Li and Shao (2002). The proof of the theorem is also given in the Appendix.

Theorem 2. Consider a triangular array of matrices $\hat{T}_{i,j,g}$ of mean-0, variance-1 Gaussian random variables. For each fixed g, the random variables have correlation

$$\begin{aligned}&\operatorname{corr}(\hat{T}_{i,k,g}, \hat{T}_{i+j,l,g})\\ &\quad = e^{-j^2C^2/(2\log(g)(d^{2k}+d^{2l}))}\\ &\qquad \times\left[1 - \frac{j^2C^2}{\log(g)(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2},\end{aligned} \tag{12}$$

where $C > 0$ and $0 < d < 1$. Then

$$\lim_{g\to\infty} P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r} \hat{T}_{i,j,g} \le u(x)\Big] = e^{-(\vartheta_1+\cdots+\vartheta_r)e^{-x}},$$

where

$$\vartheta_k = 2\Phi\left(\frac{\sqrt{3}C}{2d^k}\right) - 1, \qquad k = 1, \ldots, r,$$

and

$$u(x) = \sqrt{2\log g} + \frac{x}{\sqrt{2\log g}} - \frac{\log\log g + \log 4\pi}{\sqrt{8\log g}}.$$

We again follow the recommendation of Hsing et al. and approximate the maximum of the SiZer map by

$$P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r} T_{i,j} < x\Big] \approx \Phi(x)^{(\theta_1+\cdots+\theta_r)g}, \tag{13}$$

where

$$\theta_k = 2\Phi\left(\sqrt{3\log g}\frac{\tilde{\Delta}}{2hd^k}\right) - 1. \tag{14}$$

Here $\tilde{\Delta}$ is the distance between the pixels of the SiZer map, g is the number of pixels in each row, r is the number of rows, hd^k is the bandwidth used to calculate the kth row, and Φ is the standard normal distribution function.

An analogous expression obtained by replacing $\sqrt{3}$ by $\sqrt{5}$ in (14) could be derived for the curvature version of SiZer. However, we omit the details here to save space.

It is worth pointing out that Theorem 2 can be thought of as a first-order approximation. In fact, the SiZer rows are correlated, and it would be beneficial to study the second-order asymptotic properties of the maximum of the SiZer map. However, the probability theory necessary for this is not yet available, because we would need second-order extreme value theory for nonstationary Gaussian random fields.

2.4 Empirical Verification of the Gaussian Power Distribution

The approximation of the distribution of the row-wise and global maximum by a power of a standard Gaussian distribution in (10) and (13), respectively, is based on asymptotic consider-

ations. The asymptotic is considered as the number of pixels g approaches infinity. This section investigates the properties of this for the most typical value $g = 400$. A similar study could be done for the row-wise maximums, but we omit it to save space.

Here we use the graphical device of the quantile–quantile (Q–Q) plot to study how well the Gaussian power distribution fits the simulated data that was studied in Figures 2 and 3. (See Fisher 1983 for an overview of Q–Q plots and a number of related graphical devices.)

Again the setting is fixed design regression for sample size $n = 1{,}600$, based on an identically 0 regression function, with standard Gaussian noise. For each of 1,000 realizations, we compute the maximum over all of the pixels in the SiZer map of the test statistics used to do inference (i.e., decide on the SiZer color). The distribution of these 1,000 maxima is illustrated in Figure 4, where it is compared with the theoretical Gaussian power distribution.

The Q–Q plot is a plot of the data quantiles (just the ordered data values) on the vertical axis versus the corresponding theoretical quantiles from the Gaussian power distribution on the horizontal axis. Connecting the dots give the red curve. If the theoretical distribution were correct and there was no sampling variation, then the red curve would lie exactly on the 45-degree line, shown in green. Sampling variation leads to some departure from the green line. An important question is whether the amount of variation is explainable by the sampling process, or whether it represents a serious departure of the data distribution from the theoretical distribution.

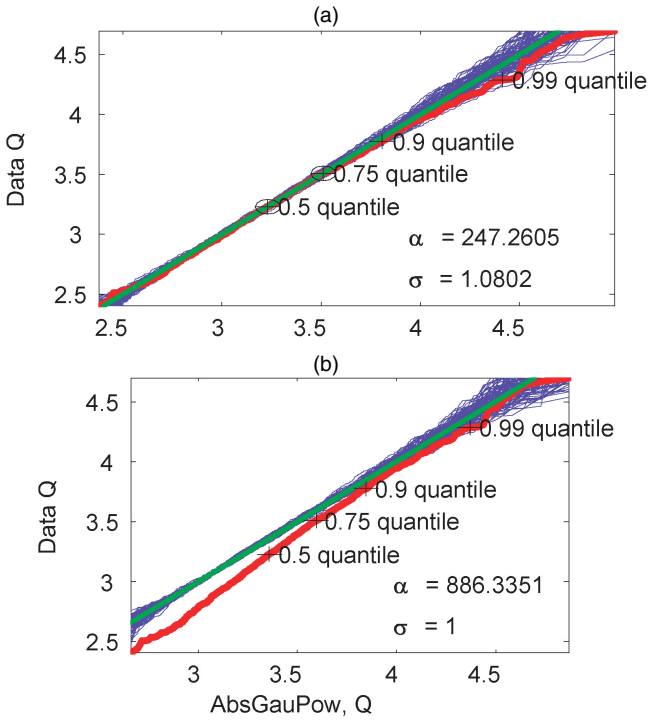

Figure 4. Q–Q Plot Showing That a Power of Gaussian Provides a Good Fit to the Maxima of the 1,000 Simulated SiZer Maps Under the Null Hypothesis. This plot was generated using the same simulated dataset as in Figures 2 and 3. The parameters are obtained by quantile matching (a) and by the theoretical considerations of Section 2.3 (b). This shows that the global adjustment will be slightly conservative, due to the slow rate of convergence.

This issue is addressed by the family of blue curves, which are 100 simulated Q–Q plots, from data having the theoretical distribution. If the red curve lies nearly completely inside the blue envelope, then we can conclude that the theoretical fit is good.

The theoretical distribution considered in Figure 4 is a member of a parametric family. In particular, the Gaussian power distribution is parameterized by a scale parameter σ (the standard deviation of the underlying Gaussian distribution), and a shape parameter α (the power of the Gaussian cdf, i.e., the number of independent Gaussians to maximize). These parameters are estimated in Figure 4(a) by quantile matching. In particular, they are solutions of the equations that make the Gaussian power distribution correct at the .5 and .75 quantiles (these were chosen to give good visual impression).

The estimated value of $\sigma = 1.08$ is very good, because here the underlying Gaussian distribution has standard deviation 1. The estimated value of $\alpha = 247.3$ appears to be unstable, being greatly affected by small changes in the value of σ and the quantiles that we decide to match. For example, if we set $\sigma = 1$, then we get $\alpha = 552$. Moreover, if we then decide to approximately match quantiles .8 and .95, then we get $\alpha = 689$. In all of these cases, the Q–Q plot shows a reasonable fit. This phenomenon is related to the "distributional fragility" ideas of Gong, Liu, Misra, and Towsley (2001).

The value of α based on the asymptotic theory of Section 2.3 and calculated from (13) is $\alpha = g\theta = 886.3$. The fit of this distribution is shown in Figure 4(b). We can see that although the fit is very good in the tail of the distribution, it is not very good in the body of the distribution. This is caused by the fact that even though we can approximate the distribution of the maximum as if the rows were independent asymptotically, this approximation is slow to converge. The fact that the red curve in Figure 4(b) is below the blue envelope for some of the quantiles suggests that in practice the global adjustment will be conservative. This conclusion is confirmed by the simulation results of Section 3.1 demonstrating that global adjustment is indeed slightly conservative.

Similar Q–Q plots have been constructed for other simulation settings (detailed in Sec. 3.1). The results were generally similar (i.e., the Gaussian power distribution gave a good fit) for the density estimation settings and for the larger sample sizes. For the smaller sample sizes, in the regression settings, there were no values of the parameters σ and α that left the red curve within the blue envelope. The values that gave the best visual fit resulted in estimates of σ that were far from 1 and unreasonable values of α. This occasional poor performance seems to be due to Gaussian versus t distribution issues, which we discuss further in Section 3.1.

2.5 Proposed Improvements

As mentioned at the beginning of Section 2, there are two natural goals when considering the size of SiZer. The first, called "row-wise" simultaneous inference, seeks to have at most $\alpha 100\%$ of the rows containing "false positives," that is, pixels flagged as statistically significant when no signal is present in the data. The second, called "global" simultaneous inference, aims at having at most $\alpha 100\%$ of the SiZer maps containing false positives.

The row-wise adjustment follows directly from the mathematical considerations of Section 2.2. Define

$$C_R = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{1/(\theta g)}\right),$$

where θ was defined in (11). Then color the ith pixel in the jth row blue if the corresponding $T_i > C_R$ and red if $T_i < -C_R$. Notice that under the null hypothesis, the distribution of $\max(T_1, \ldots, T_g)$ is the same as the distribution of $-\min(T_1, \ldots, T_g)$. It follows that if the data contain no signal, then the probability that there is a spurious color on the gth row is

$$\begin{aligned}
&P[T_i < -C_R \text{ or } T_i > C_R \text{ for some } i = 1, \ldots, g] \\
&\quad \le P[\min(T_1, \ldots, T_g) < -C_R] + P[\max(T_1, \ldots, T_g) > C_R] \\
&\quad = 2\big(1 - P[\max(T_1, \ldots, T_g) < C_R]\big) \\
&\quad \approx 2(1 - \Phi(C_R)^{\theta g}) \\
&\quad = \alpha.
\end{aligned}$$

Thus no more than about $\alpha 100\%$ of the rows will have spurious colors, as desired.

Global adjustment is based on Section 2.3. Define

$$C_G = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{1/((\theta_1+\cdots+\theta_r)g)}\right),$$

and recall that the θ_k's were defined by (14). Then color the ith pixel, in the jth row, blue if the corresponding $T_{i,j} > C_R$ and red if $T_{i,j} < -C_R$. It is worth pointing out that the constants C_R are different for each row, whereas the constant C_G is the same for all the rows. Again,

$$\begin{aligned}
&P\big[T_{i,j} < -C_G \text{ or } T_{i,j} > C_G \\
&\qquad \text{for some } i = 1, \ldots, g,\ j = 1, \ldots, r\big] \\
&\quad \le P[\min(T_{1,1}, \ldots, T_{g,r}) < -C_G] \\
&\qquad + P[\max(T_{1,1}, \ldots, T_{g,r}) > C_G] \\
&\quad \approx 2\big(1 - \Phi(C_G)^{(\theta_1+\cdots+\theta_r)g}\big) = \alpha.
\end{aligned}$$

Thus no more than about $\alpha 100\%$ of the SiZer maps will have spurious colors, as desired.

3. ANALYSIS OF IMPROVEMENTS

In this section we investigate the properties of these improvements of SiZer. First, we study the size properties through a simulation study in Section 3.1. We study the amount of power sacrificed to get the size correct, through simulation in Section 3.2 and through some real data examples in Section 3.3.

3.1 Size Simulations

To compare the size performance of the conventional SiZer with our new row-wise and global versions of SiZer, we performed an array of simulations against several variations of "the null hypothesis":

1. We tried the following settings:
   a. KDE: kernel density estimation for the Uniform(0, 1) density
   b. FDR–N: fixed design regression for an equally spaced design with standard Gaussian noise but no signal
   c. FDR–E: fixed design regression for an equally spaced design with standard exponential noise
   d. RDR–U: random design regression, where the X_i's are chosen from the Uniform(0, 1) density and the Y_i's are independent standard Gaussian.
   e. RDR–N: random design regression, where the X_i's are chosen from the N(0, 1) density and the Y_i's are independent standard Gaussian.
2. For each of the foregoing settings, we tested the following sample sizes:
   a. $n = 100$
   b. $n = 400$
   c. $n = 1{,}600$
   d. $n = 6{,}400$.

For each of the resulting 20 combinations, 1,000 pseudo-datasets were drawn, the various SiZer maps were calculated, and the red and blue pixels (ideally none, because there are no signals in any of these examples) were counted.

One way of summarizing these numbers is row-wise in the SiZer maps: for each setting, each sample size, and each row report the percentage of realizations of the data in which there were some red or blue pixels in that row. Figure 5 shows these summaries. Notice that if no red or blue pixels were present in a particular row, then the $\max_{i=1,\ldots,g} |T_i|$ for T_i's corresponding to this row was less then the preset value of the cutoff C. In particular, if we set C using the row-wise approximation of Section 2 and the simulated proportion of red and blue pixels is equal to the nominal value $\alpha = .05$, then we have some evidence that our approximation is working. As shown in Figure 5, this is often the case.

Instead of showing long tables of numbers, the main ideas are made more accessible by displaying the results with a parallel coordinate plot (see Inselberg 1985). Figure 5(a) summarizes performance for the KDE setting, Figure 5(b) does the same for the FDR–N setting, Figure 5(c) is for the RDR–U setting, Figure 5(d) is for the RDR–N setting, and Figure 5(e) contains the FDR–E setting. The coordinates (points on the horizontal axes) represent rows of the SiZer map and thus are quantified by $\log_{10} h$ (shown only in the bottom panel, to avoid overplotting with the figure titles), just as on the vertical axes of the SiZer maps. The vertical axes are the percentage of rows (across the 1,000 replications) showing some significant structure (i.e., red or blue pixels). Each curve represents one setting (indicated by color as shown) and one sample size (indicated by line type as shown). The curves are piecewise linear, with nodes at each row of the map (i.e., each window width h). The heights at the nodes contain the useful information, and the connecting line segments simply make it easier to understand the relationships.

Ideally, all of these values should be close to $\alpha = .05$ for the row-wise procedures, such as the conventional SiZer and our new Row-Wise SiZer. Hence this level is represented by a horizontal black line.

Note that in almost every case the conventional SiZer flags significant structure far too often. This again verifies the main idea in this article: It is well worth finding less crude approaches

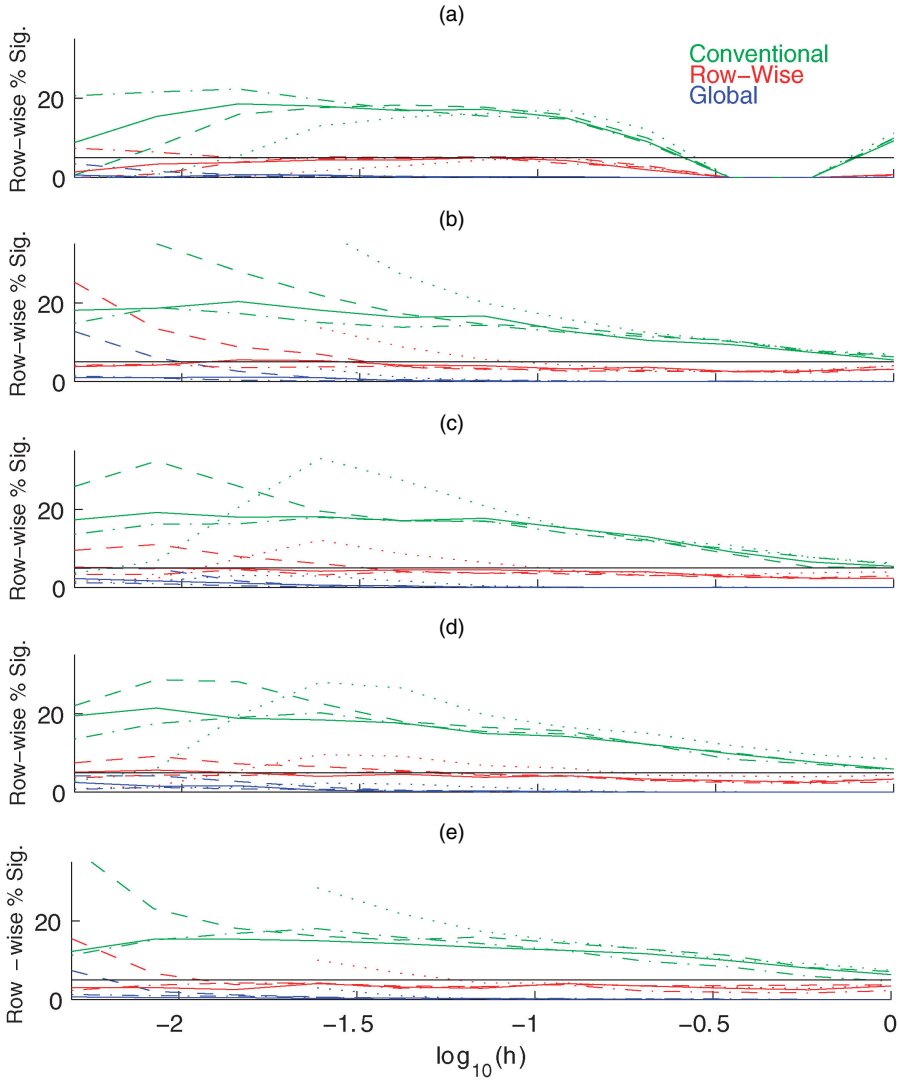

Figure 5. Row-Wise Summaries of the Percentage of Significant Pixels for SiZer Under the Null Hypothesis, Allowing Comparison of the Different Simultaneity Adjustments and Sample Sizes. (a) KDE; (b) FDR–N; (c) RDR–U; (d) RDR–N; (e) FDR–E. The plots clearly show the relationships between sample sizes. (· · · · · · · · · · ·, n = 100; — — — —, n = 400; ————————, n = 1,600; — · — · — · —, n = 6,400.)

to this multiple comparison problem. Similarly, in a large majority of the cases, the new Row-Wise SiZer is quite close to the desired $\alpha = .05$.

As expected, the global methods are almost always quite far below the desired level, because they aim at a global size of $\alpha = .05$, which requires that they be deliberately conservative when studied in this row-wise sense.

A perhaps surprising feature in the KDE setting, studied in Figure 5(a), is the zero values everywhere for the second- and third-coarsest scales. This is due to the crude type of boundary adjustment used. Boundary adjustment is essential for estimating the Uniform(0, 1) density with kernel estimates, because these methods tend to "round off the corners" at both edges. If the summaries of Figure 5 are computed with no adjustment, then far too many percentages are 100%, because every realization of most rows has some significant pixels flagged at the edges. To avoid this boundary problem, the simple "circular design" device was used. Here the data are treated as periodic, and shifted copies of the data are added at each end. (See Silverman 1986, p. 31, where the "circular design" device is called the "wrap around" boundary condition.) Although this crude adjustment is reasonably effective at most scales, there are a few where it introduces artifacts, such as the 0's shown in Figure 5(a). Such boundary effects are not a serious issue for the regression settings, because the local linear smoother used in both performs an automatic first-order boundary adjustment.

Another departure from the expected size occurs for the regression settings, shown in Figures 5(b), 5(c), 5(d), and 5(e). These are substantial increases in the percentage of realizations flagged as significant at finer scales. At these scales there can be few points in the kernel window. Therefore, the fact that SiZer uses a local estimator of variance implies that the underlying null distributions are better approximated by a t distribution than by the Gaussian distribution. This idea is verified by the fact that it is generally the worst for $n = 100$ and better for $n = 400$, and the problem is nonexistent for $n = 1{,}600$ and $n = 6{,}400$. Exceptions include the FDRs in Figures 5(b) and 5(e), where the dotted curves for $n = 100$ disappear for fine scales (because there are never enough data points in the kernel windows, i.e., the SiZer color is always gray), and

the RDRs in Figures 5(c) and 5(d), where the dotted curves for $n = 100$ actually go down for finer scales, because there are typically just a few locations where the data are rich enough to allow any inference (thus most of the pixels are gray). In those remaining locations, the SiZer color is often completely purple.

A simple approach to this problem is to replace the Gaussian distribution with the t distribution. This was attempted, but the results were too conservative to be useful. The reason for this seems to be the complicated interaction of the t distribution with the correlation structure.

The comparison in Figure 5 is for the row-wise size of the statistical inference. But also of keen interest is the global size for the multiple comparison problem over the entire map, not just within individual rows. Global size for the same simulation settings is studied in Figure 6.

Figure 6 is a parallel coordinate display of the percent of realizations (out of 1,000) for which there were some significant pixels in the SiZer map. Again, color is used to indicate SiZer type, with the same color scheme. The coordinates now are taken to be the sample size n, different from SiZer map row as in Figure 5, to highlight the perhaps surprising impact of n on the results. Line type is now used to show the setting.

In this sense, the size problems of the conventional SiZer map are even worse than in the row-wise sense indicated in Figure 5 (note the larger vertical axis). The new Row-Wise SiZer is also always far above the nominal level of $\alpha = .05$, which, not surprisingly, shows that there is substantial difference between row-wise and global statistical inference. This is consistent with the global method appearing generally too conservative in Figure 5.

Performance of the global SiZer approach, is quite dependent on the setting. For KDE, the method is generally conservative. This is caused by the boundary effect and adjustment discussed earlier, as well as by data sparseness issues at the finest scales. In particular, the 0's at the second- and third-coarsest scales are present for all SiZers. This means that at those scales, the boundary adjustment used effectively wipes out any trend possibly present in the data. For regression, the percentages are often too large. For $n = 400$, the percentage of maps flagged as significant increases substantially, because of the t effect described earlier (most of which occurs at the finest scales where there are relatively few points in each kernel window, so the number of degrees of freedom can be as low as 4). As noted earlier, many of the curves are lower for $n = 100$, because of data sparsity effects. As expected, the t effect is no longer present for large sample sizes ($n = 1{,}600$, $n = 6{,}400$), and the global SiZer has excellent size performance for all five regression settings.

Figure 7 is a reorganization of the parallel coordinates plot in Figure 5, which highlights an important lesson about how the settings compare that is obscured in that figure because the settings are in different panels. In Figure 7 the panels show the sample sizes n, with $n = 100$, 400, 1,600, and 6,400. As in Figures 5 and 6, color represents SiZer type, using the same scheme. The line type is consistent with Figure 6, representing the setting. Again the coordinates represent rows of the SiZer map and are indexed by $\log_{10} h$.

The main lesson of Figure 7 is that curves of the same color tend to be very close to each other; that is, the settings are very similar. Although there are important differences in the simultaneity type (expressed by colors) and sample size (different panels), the settings are similar. This validates the approach of using the common mathematical structure, as developed in Section 2.1.

Another useful feature of the view shown in Figure 7 is that it provides another way of seeing that the row-wise method is best in this sense, and that the best results are for the larger sample sizes. In particular, it is very clear that for high sample sizes of $n = 1{,}600$ and $n = 6{,}400$, the percentages virtually achieve their goal of $\alpha = .05$, uniformly over both rows and settings (except for density estimation at large scales).

A similar simulation study was carried out to investigate the size properties of the curvature version of SiZer. The results were similar to those summarized in Figures 5–7 for the slope version of SiZer and are not explicitly reported, to save space. The main differences between the results were that both the boundary effect in the kernel density estimation and the t effect for the small sample sizes of regression were even more severe in the curvature version than in the slope version.

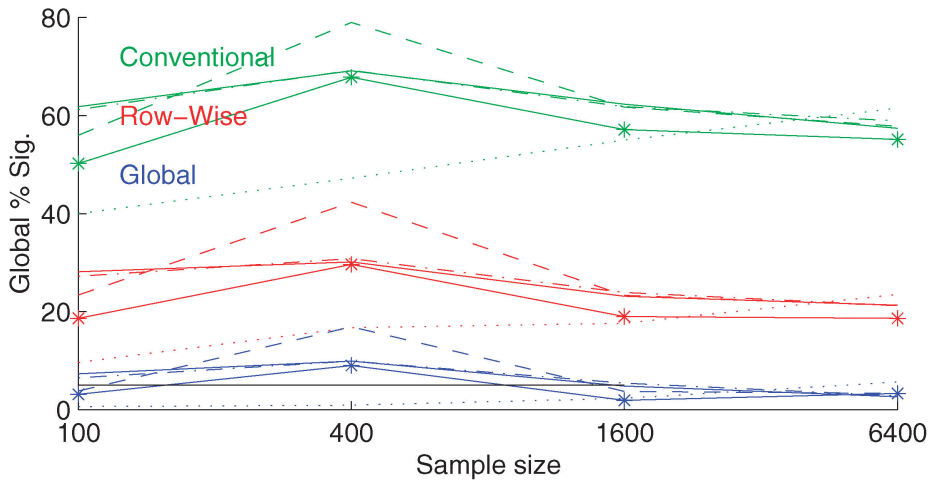

Figure 6. Global Size Summaries Showing the Percentage of Significant Pixels in the Full SiZer Maps, Under the Null Hypothesis, Grouped by Settings (KDE · · · · · · · · · · ; FDR–N — — — —; FDR–E ✳————✳; RDR–U ————; RDR–N · — · — · —).

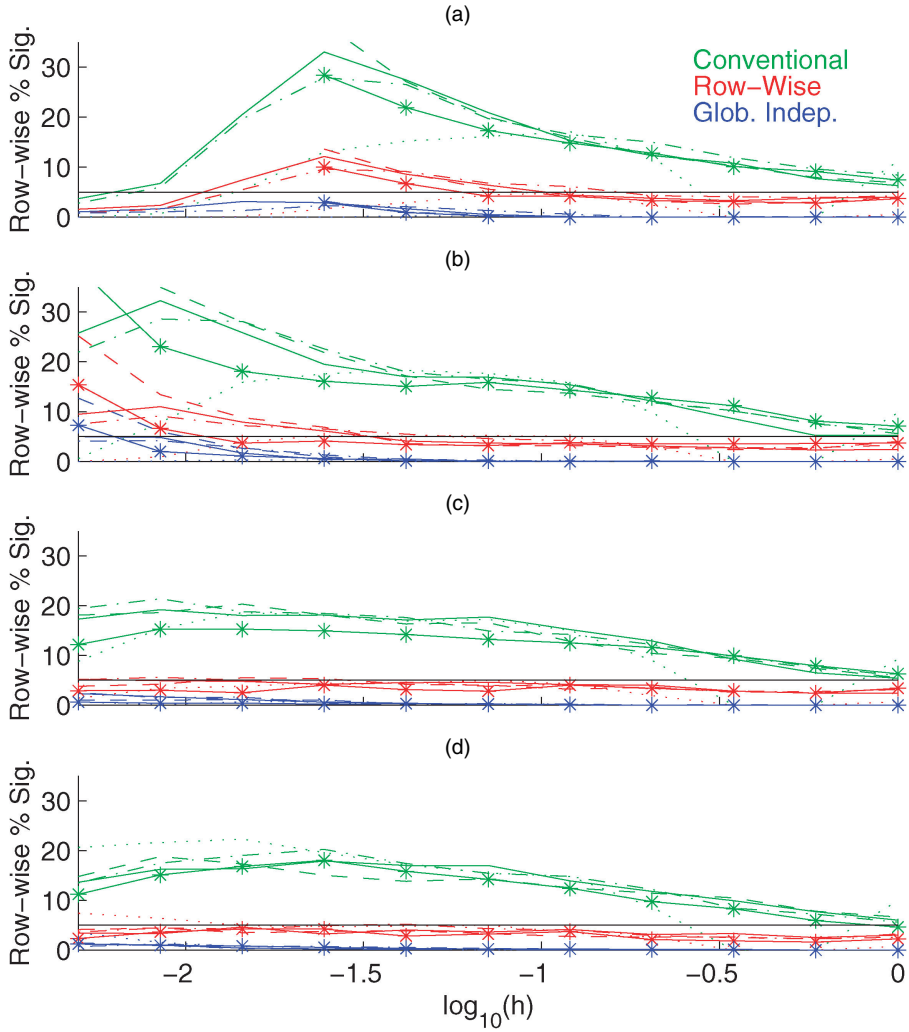

Figure 7. Row-Wise Summaries of the Percentage of Significant Pixels for SiZer Under the Null Hypothesis, Allowing Comparison of the Different Simultaneity Adjustments and Settings (KDE · · · · · · · · · · ; FDR–N — — — —; FDR–E ✳————✳; RDR–U ————; RDR–N · — · — · —). (a) n = 100; (b) n = 400; (c) n = 1,600; (d) n = 6,400. This organization shows that settings are very similar.

3.2 Power Simulations

The previous section demonstrates that our global versions of SiZer are quite good at achieving the desired overall size for the statistical inference. In this and the next section, by analyzing some simulated and real datasets, we show that this could entail substantial cost in terms of power, especially when using one of the global adjustments. This is to be expected, because it is consistent with well-established principles of hypothesis testing, particularly the theory that establishes the trade-off between size and power. The original SiZer had an inflated type I error, which resulted in more power (smaller type II error).

The first example is the same as shown in Figure 1, the Donoho–Johnson blocks regression function, with high noise, as shown in Figure 1(a). Figure 8 allows direct comparison between the conventional SiZer, the new Row-Wise SiZer and the global SiZer.

As shown in Figure 1(b), the conventional SiZer flags all 11 jumps as statistically significant, but it also indicates a spurious jump near $x = .58$. As expected, the new Row-Wise SiZer [Fig. 8(b)] flags fewer pixels as significant, but still finds all 11 jumps. The spurious jump near $x = .58$ is still present, but smaller. For the global method, the spurious feature disappears, as does the jump near $x = .15$. This reflects the loss of power from insisting on global simultaneous inference.

If one were to use only the global analysis, then the upward jump near $x = .78$, would be flagged as statistically significant by a very small blue region. From the viewpoint of conventional SiZer, it might be tempting to ignore this. However, an important lesson is that any significant pixel (regardless of how small it is) found by a global method should be considered important underlying structure.

Figure 9 shows a simulated density estimation example. In addition to the same three panels as in Figure 8, Figure 9 contains an additional panel showing the family of density estimators for wide range of different windows widthh and the underlying true density shown as a thick black curve. The underlying density is the trimodal Gaussian mixture density from Marron and Wand (1992), and the sample size is $n = 10{,}000$. Both the conventional and new Row-Wise SiZer show three statistically significant modes. However, the conventional SiZer also flags a spurious fine-scale feature near $x = 1.4$, which correctly disappears for the new row-wise version. The global SiZer shows some loss of significant structure, particularly the

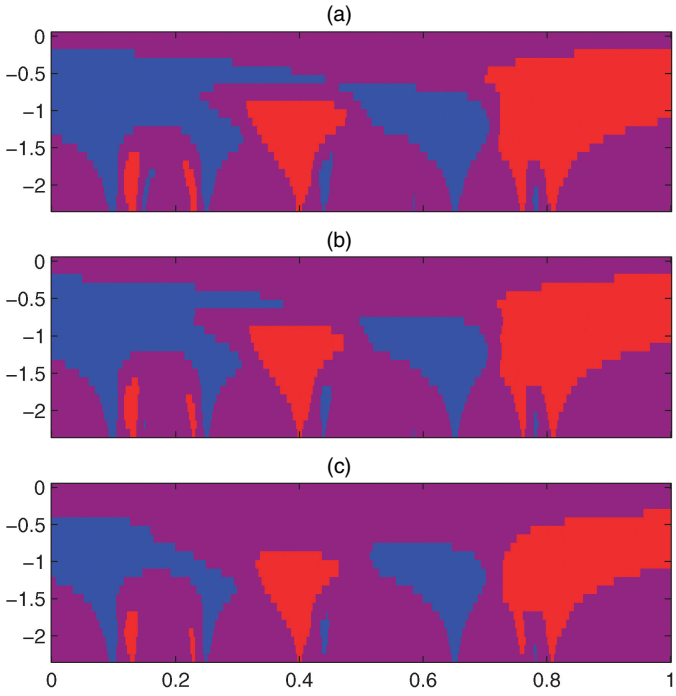

Figure 8. Full Range of SiZer Analyses of the Donoho–Johnstone Blocks Regression, With High Noise. (a) Conventional SiZer; (b) row-wise and global SiZer; (c) global SiZer.

small blue region just left of $x = 0$, again reflecting some loss of power.

Similar plots have been constructed for all of the Marron–Wand Gaussian mixture densities for sample sizes $n = 100$, 1,000, 10,000. Overall, the different versions of SiZer tended to flag very similar structure as being statistically significant. There was generally substantial erosion of the red and blue regions for the methods with better size properties (to an extent similar to that shown in Fig. 9). Sometimes this erosion was such that significant features actually disappeared, as in Figure 9(d), but most often they did not. Spurious features, such as the very small red region near $x = 1.4$ in Figure 9(b), were fairly rare, perhaps because at most locations, these densities are not close to flat (as at the null distributions studied in Sec. 3.1), but instead have substantial slope.

3.3 Real Data Examples

Another approach to studying the trade-off between size and power involved in these different versions of SiZer is through the analysis of real data. Figure 10 shows the density estimation example of the 1975 British Family Incomes data that was carefully analyzed by Schmitz and Marron (1992), again using similar four panels as in Figure 9. The conventional SiZer analysis shows two significant modes, which has been independently confirmed by a parametric analysis as discussed by Schmitz and Marron (1992). The red region between modes is still present for the new Row-Wise SiZer shown in Figure 10(c), and again, greater credence needs to be placed on this more precise version. Unfortunately, this red region completely disappears in the global SiZer map. This loss of power is particularly unfortunate because bimodality is the important feature of this dataset.

Although the global slope version was unable to find the important bimodal characteristics of the British Family Incomes data in Figure 10, it is interesting to note that the global

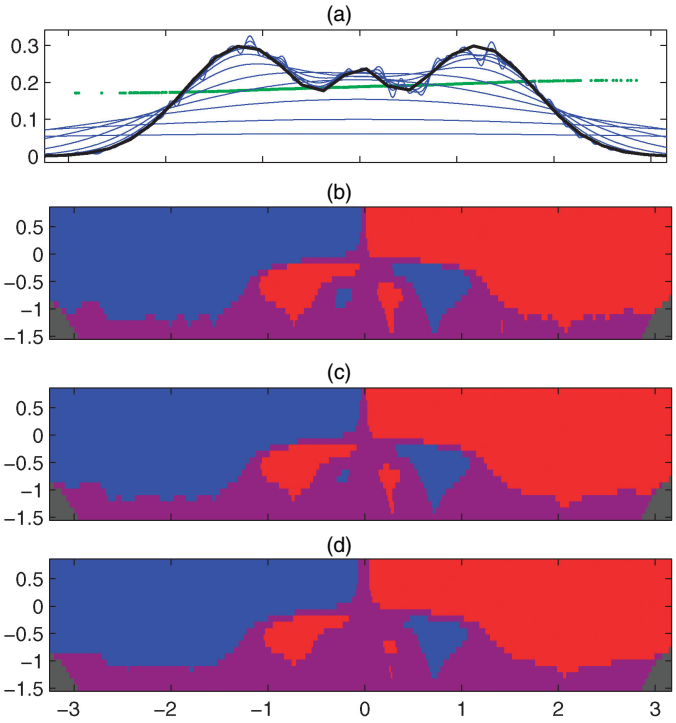

Figure 9. Full Range of SiZer Analyses of the Trimodal Mixture of Gaussians. The plots show the scale space overlaid with the (a) true density and the (b) conventional, (c) row-wise, and (d) global SiZer versions.

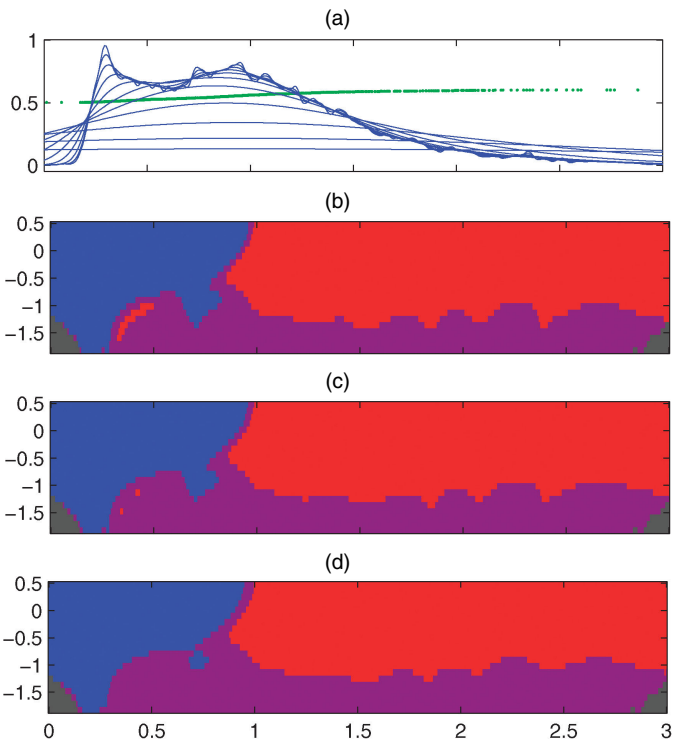

Figure 10. Full Range of SiZer Analyses of the British Family Incomes Data. (a) Scale space; (b) conventional SiZer; (c) row-wise SiZer; (d) global SiZer.

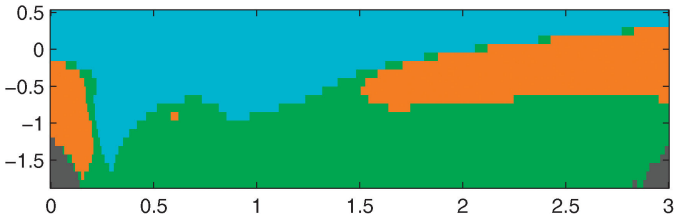

Figure 11. Global Curvature SiZer Analysis of the British Family Incomes Data. This figure shows the bimodality known to be an important feature of this dataset.

curvature version of SiZer does flag this feature of the data as statistically significant, as shown in Figure 11. The conventional curvature version of SiZer was proposed by Chaudhuri and Marron (2002). Here we improve the simultaneity using ideas from Section 2.

To clearly distinguish it from the slope version of SiZer, the curvature version uses a different color scheme. Pixels with significant concavity (second derivative strongly negative) are indicated by cyan (light blue). Those with significant convexity are colored orange. Locations in scale space where there is no significant curvature are colored green. Again, gray is used in regions where the data are too sparse.

The bimodality of this dataset is shown to be strongly significant by the very small orange region near $x = .6$. Although the region is very small, again it is important to keep in mind that when using global version of SiZer, any significance at all should be considered strong evidence.

Figure 12 shows an example from flow cytometry, where the presence and percentage of florescence-marked antibodies on cells are measured. The medical goal is the determination of

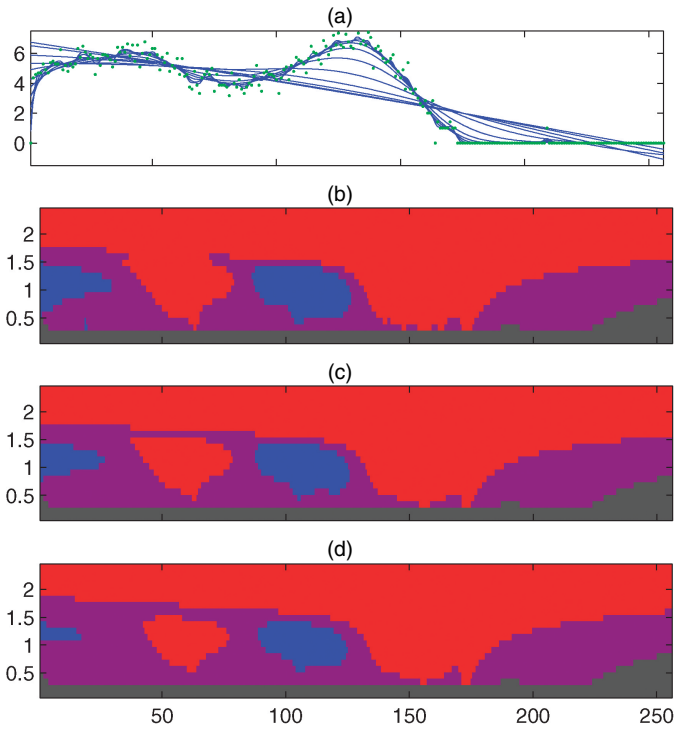

Figure 12. Full Range of SiZer Analyses of a Flow Cytometry Dataset. (a) Scale space; (b) conventional SiZer; (c) row-wise SiZer; (d) global SiZer.

quantities such as the percentage of lymphocytes among cells. The data come from the laboratory of Drs. S. Mentzer and J. Rawn, Brigham, and Women's Hospital, Boston, and we are grateful to M. P. Wand for putting us in contact with them. In a single experiment, many cells are run through a laser, the intensity of the fluorescence of each cell is measured, and the data are stored as 256 bin counts, where bins are called "channels." These bin counts are traditionally viewed on the square-root scale. An important question is how many "bumps" are in this square root histogram. Here we treat this as a regression problem.

Figure 12 shows the same three panels, comparing the different simultaneity methods. Figure 12(b) (conventional SiZer) shows two clear modes and a small fine-scale feature near $x = 20$. This small feature is already seen to be spurious by the new Row-Wise SiZer map in Figure 12(c). This time the effect of the global version is representative of many of the examples we have seen: the significant red and blue regions are somewhat eroded, but indicate essentially the same lessons.

Based on this experience, and on a number of other examples studied during this research, we recommend that the default version of SiZer be the new row-wise approach. This choice is made to give reasonable power, but it must be kept in mind that the statistical inference is not completely valid in the classical sense, which is often acceptable in exploratory data analysis situations. When statistical rigor is essential (e.g., before making a large investment of research effort in understanding "phenomena found"), it is recommended that the global version be used.

4. FUTURE WORK

Although the methods developed in this article are intended to enhance the applicability of the SiZer method, a number of open problems remain, including the following:

1. Development of the probability theory needed to improve the global approximation of Section 2.3, by an approximation that takes the full random field distribution of the SiZer inference into account
2. More careful boundary adjustment, as discussed in Section 3.1
3. Improved incorporation of the t distribution for regression settings, with careful accounting of the correlation structure, as discussed in Section 3.1.

APPENDIX: PROOFS

Here we present proofs of the key results of this article.

A.1 Derivation of (4) and (5)

We derive the correlation in the case of equally spaced regression, but our formulas also apply to other settings, including random design regression and density estimation, because these settings have some very strong connections. (For some interesting mathematics that makes these connections precise, see Nussbaum 1996; Brown and Low 1996; Brown, Cai, and Low 2002; Grama and Nussbaum 1998, 2002.) This equivalence between settings is also demonstrated empirically in Figure 7.

When dealing with regression data, SiZer uses the local linear smoother defined by (2). To color the pixels, SiZer checks whether

the estimate of the first derivative,

$$\begin{aligned} a_1 = {} & -c^{-1}\left[\sum_{i=1}^{n} K_h(x-X_i)\right]\left[\sum_{i=1}^{n}(x-X_i)K_h(x-X_i)Y_i\right] \\ & + c^{-1}\left[\sum_{i=1}^{n}(x-X_i)K_h(x-X_i)\right] \\ & \times\left[\sum_{i=1}^{n} K_h(x-X_i)Y_i\right], \qquad \text{(A.1)} \\ c = {} & \left[\sum_{i=1}^{n} K_h(x-X_i)\right]\left[\sum_{i=1}^{n}(x-X_i)^2 K_h(x-X_i)\right] \\ & - \left[\sum_{i=1}^{n}(x-X_i)K_h(x-X_i)\right]^2, \end{aligned}$$

is significantly different than 0. In the particular case of fixed design regression, the design points X_i satisfy $X_i = i\Delta$, where $\Delta > 0$. (In the asymptotic calculations, we usually assume that $\Delta \to 0$.) If x is away from the boundary, then it follows from symmetry of the kernel that

$$\sum_{i=1}^{n}(x-X_i)K_h(x-X_i) \approx 0.$$

This means that the second term in (A.1) disappears.

Denote $p = \tilde{\Delta}/\Delta$, where p is "the number of data points per SiZer column." For simplicity of notation, we can assume that p is a positive integer. This is supported by the fact that SiZer colors the pixel gray if the data are too sparse.

Thus T_j is proportional to the estimate of the first derivative a_1 calculated for $x = j\tilde{\Delta} = jp\Delta$. In particular,

$$T_j \approx \sum_{q=1}^{n} W^h_{jp-q} Y_q. \qquad \text{(A.2)}$$

The exact form of the W^h_{jp-q} is given in the first term of (A.1). For our purposes, it suffices to realize that W^h_{jp-q} is proportional to $-(jp-q)K_{h/\Delta}(jp-q)$. Thus the weights W^h_q are proportional to the derivative of the Gaussian kernel with standard deviation h/Δ.

If the null hypothesis of no signal is true, then the Y_i are iid random variables. If in addition the Y's have two finite moments, then the linear approximation (A.2) greatly simplifies the distribution theory, because for h/Δ large enough, the Cramèr–Wold device and Lindeberg–Feller central limit theorem (see, e.g., Durrett 2005) give an approximate Gaussian distribution, with mean 0 (under the SiZer null hypothesis) and variance 1, by appropriate scaling.

The full joint distribution of $T_1, \ldots, T_g$ also depends on the correlation between them. This correlation is approximated by

$$\begin{aligned} \operatorname{corr}(T_i, T_{i+j}) &= \frac{\sum_q W^h_{q-jp} W^h_q}{\sum_q (W^h_q)^2} \\ &\approx \frac{\int (x-jp)K_{h/\Delta}(x-jp)\,x\,K_{h/\Delta}(x)\,dx}{\int x^2 K_{h/\Delta}(x-jp)^2\,dx} \\ &= e^{-(jp\Delta)^2/(4h^2)}\left[1 - \frac{(jp\Delta)^2}{2h^2}\right], \end{aligned}$$

where the second step follows by replacing the sums by integral approximations. Equation (4) now follows by observing that $p\Delta = \tilde{\Delta}$.

Similarly, if we consider correlation between pixels at different SiZer rows, then we get

$$\begin{aligned} &\operatorname{corr}(T_{i,k}, T_{i+j,l}) \\ &= \frac{\sum_q W^{hd^k}_{q-jp} W^{hd^l}_q}{[\sum_q (W^{hd^k}_q)^2 \sum_q (W^{hd^l}_q)^2]^{1/2}} \\ &\approx \frac{\int (x-jp)K_{hd^k/\Delta}(x-jp)\,x\,K_{hd^l/\Delta}(x)\,dx}{[(\int x^2 K_{hd^k/\Delta}(x-jp)^2\,dx)(\int x^2 K_{hd^l/\Delta}(x-jp)^2\,dx)]^{1/2}} \\ &= e^{-j^2\tilde{\Delta}^2/(2h^2(d^{2k}+d^{2l}))}\left[1 - \frac{j^2\tilde{\Delta}^2}{h^2(d^{2l}+d^{2k})}\right]\left(\frac{2d^{k+l}}{d^{2k}+d^{2l}}\right)^{3/2}, \end{aligned}$$

which is (5).

In practice, we do not know the standard deviation of the noise ε_i. This is needed to scale $T_{1,1}, \ldots, T_{g,r}$ to have variance 1. For this reason, it must be estimated from the data introducing additional dependence as well as other issues. But this is not a problem in theory, because consistent estimators of this standard deviation are available, and thus the calculations presented in this section will still be valid asymptotically. This is confirmed by our simulation reported in Section 3.1, where the estimation of the standard deviation from the data seems to create problems only for small sample sizes at fine scales.

A.2 Proof of Theorem 1

Following Hsing et al. (1996), we see that (8) follows as long as we can verify the conditions of Hsing et al.'s theorem 2.2.

To that effect, first notice that

$$\lim_{g\to\infty} \log(g)(1-\rho_{j,g}) = \frac{3j^2C^2}{4},$$

which verifies the first condition of theorem 2.2. Verification of the remaining conditions is fairly routine. Our calculations are quite similar to the calculations performed in section 3 of Hsing et al. (1996) for a different stationary process. Set

$$l_g = (\log g)^{1/2}\log(\log g).$$

If $\log\log g > \sqrt{6}/C$, then

$$\sup_{j\ge l_g} |\rho_{j,g}| \log g \le |\rho_{l_g,g}| \log g \to 0,$$

and the second condition of theorem 2.2 follows. To verify the last condition, fix a small $\varepsilon > 0$ and notice that if $j^2C^2/(4\log g) > \varepsilon$, then

$$-2e^{-3/2} \le \rho_{j,g} \le e^{-\varepsilon}.$$

On the other hand, if $j^2C^2/(4\log g) \le \varepsilon$, then

$$\frac{3j^2C^2}{4\log g}\left(1 - \frac{\varepsilon}{2}\right) \le 1 - \rho_{j,g} \le \frac{3j^2C^2}{4\log g}.$$

Thus

$$\begin{aligned} &\sum_{j=m}^{l_g} g^{-(1-\rho_{j,g})/(1+\rho_{j,g})} \frac{(\log g)^{-\rho_{j,g}/(1+\rho_{j,g})}}{(1-\rho_{j,g})^{1/2}} \\ &\quad = \sum_{j=m}^{l_g} g^{-(1-\rho_{j,g})/(1+\rho_{j,g})} \frac{(\log g)^{(1/2)(1-\rho_{j,g})/(1+\rho_{j,g})}}{((1-\rho_{j,g})\log g)^{1/2}} \\ &\quad \le \max\Bigg(l_g\, g^{-(1-\exp(-\varepsilon))/(1+\exp(-\varepsilon))} \\ &\qquad \times \frac{(\log g)^{(1/2)(1+2\exp(-3/2))/(1-2\exp(-3/2))}}{((1-\exp(-\varepsilon))\log g)^{1/2}}, \end{aligned}$$

$$\sum_{j=m}^{l_g} \exp\left[-\frac{3j^2C^2}{8}\left(1-\frac{\varepsilon}{2}\right)\right] \times \frac{\exp[(3j^2C^2)\log\log g/(4\log g)]}{((3j^2C^2)(1-\varepsilon/2)/4)^{1/2}}\Bigg),$$

and the final condition of theorem 2.2 is readily verified.

To finish the proof of the theorem, we need to determine the value of ϑ; that is, we need to calculate the probability in (7). This could be a rather difficult task in general; however, in this case we are helped by the fact that $EH_iH_j = \frac{\delta_i+\delta j-\delta_{|i-j|}}{2\sqrt{\delta_i\delta_j}} = \frac{i^2+j^2-|i-j|^2}{2ij} = 1$ and therefore $Z = H_1 = H_2 = \cdots$, where Z is a standard Gaussian random variable. Thus the problem in (7) transforms to

$$\vartheta = P\big[V/2 + k\sqrt{3\xi}Z \le 3\xi k^2 \text{ for all } k \ge 1\big], \tag{A.3}$$

where $\xi = C^2/4$. Because V is a nonnegative random variable, the system of inequalities in (A.3) implies that $Z < \sqrt{3\xi}$. Moreover, under this condition, $V/2 + k\sqrt{3\xi}Z - 3\xi k^2$ is decreasing as a function of k, and therefore,

$$\begin{aligned}\vartheta &= P\big[V/2 + \sqrt{3\xi}Z \le 3\xi\big] \\ &= E\big(P\big[V \le 2(3\xi - \sqrt{3\xi}Z)|Z\big]\big) \\ &= \int_{-\infty}^{\sqrt{3\xi}} \big(1 - e^{-2(3\xi-\sqrt{3\xi}z)}\big)\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \\ &= 2\Phi(\sqrt{3\xi}) - 1.\end{aligned}$$

Equation (9) follows immediately.

A.3 Proof of Theorem 2.

Unlike in the proof of Theorem 1, here we cannot use an "off the shelf" theorem. Instead, we compare the maximum of the random field with the maximum of the random field, with the rows assumed to be independent using an improved version of Slepian's lemma due to Li and Shao (2002).

Recall that $\hat{T}_{1,1,g}, \ldots, \hat{T}_{g,r,g}$ is a mean 0, variance 1 Gaussian random field with correlation given by (12). For simplicity, denote

$$\hat{r}_{(i,k),(j,l),g} = \operatorname{corr}(\hat{T}_{i,k,g}, \hat{T}_{j,l,g}).$$

Define $\tilde{T}_{1,1,g}, \ldots, \tilde{T}_{g,r,g}$ as a mean 0, variance 1 Gaussian random field with correlation

$$\operatorname{corr}(\tilde{T}_{i,k,g}, \tilde{T}_{i+j,l,g}) = \delta_{k,l}e^{-j^2C^2/(4d^{2k}\log g)}\left[1 - \frac{j^2C^2}{2d^{2k}\log g}\right],$$

where $\delta_{k,l}$ is the Kronecker delta, that is, $\delta_{k,l} = 1$ if $k = l$ and 0 otherwise. Again denote

$$\tilde{r}_{(i,k),(j,l),g} = \operatorname{corr}(\tilde{T}_{i,k,g}, \tilde{T}_{j,l,g}).$$

We define the $\tilde{T}_{i,j,g}$ in such a way that $\hat{r}_{(i,k),(j,k),g} = \tilde{r}_{(i,k),(j,k),g}$.

Because the number of rows r is fixed, independence of the rows and Theorem 1 immediately imply that

$$\lim_{g\to\infty} P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r}\tilde{T}_{i,r,g} \le u(x)\Big] = e^{-(\vartheta_1+\cdots+\vartheta_r)e^{-x}},$$

where

$$\vartheta_k = 2\Phi\left(\frac{\sqrt{3}C}{2d^k}\right) - 1, \qquad k = 1, \ldots, r,$$

and

$$u(x) = \sqrt{2\log g} + \frac{x}{\sqrt{2\log g}} - \frac{\log\log g + \log 4\pi}{\sqrt{8\log g}}.$$

Therefore, to finish the proof of Theorem 2, it is sufficient to prove that

$$\lim_{g\to\infty}\Big|P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r}\hat{T}_{i,j,g} \le u(x)\Big] - P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r}\tilde{T}_{i,j,g} \le u(x)\Big]\Big| = 0.$$

Notice that there is $0 < D < 1$ such that $|\hat{r}_{(i,k),(j,l),g}| \le D < 1$ for all g, i, and j and $k \ne l$. This and theorem 2.1 of Li and Shao (2002) imply that

$$\begin{aligned}&\Big|P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r}\hat{T}_{i,j,g} \le u(x)\Big] - P\Big[\max_{i=1,\ldots,g}\max_{j=1,\ldots,r}\tilde{T}_{i,j,g} \le u(x)\Big]\Big| \\ &\le \frac{1}{8}\sum_{i,j,k,l}\big|\hat{r}_{(i,k),(j,k),g} - \tilde{r}_{(i,k),(j,k),g}\big|e^{-u(x)^2/(1+|\hat{r}_{(i,k),(j,k)}|)} \\ &\le K_1r^2ge^{-u(x)^2/(1+D)}\sum_{j=1}^{\infty}e^{-K_2j^2/\log g}\left(1 + \frac{K_3j^2}{\log g}\right) \\ &\le K_4r^2g^{1-2/(1+D)}(\log g)^{3/2} \\ &\to 0\end{aligned}$$

as $g \to \infty$. Here $K_1, \ldots, K_4$ are suitable positive constants. This concludes the proof.

[Received January 2004. Revised June 2005.]

REFERENCES

Berman, S. (1964), "Limit Theorems for the Maximum Term in Stationary Sequences," *The Annals of Mathematical Statistics*, 35, 502–516.

Brown, L. D., Cai, T. T., and Low, M. (2002), "Asymptotic Equivalence Theory for Nonparametric Regression With Random Design," *The Annals of Statistics*, 30, 688–707.

Brown, L. D., and Low, M. (1996), "Asymptotic Equivalence of Nonparametric Regression and White Noise," *The Annals of Statistics*, 24, 2384–2398.

Chaudhuri, P., and Marron, J. S. (1999), "SiZer for Exploration of Structure in Curves," *Journal of the American Statistical Association*, 94, 807–823.

——— (2000), "Scale Space View of Curve Estimation," *The Annals of Statistics*, 28, 408–428.

——— (2002), "Curvature versus Slope Inference for Features in Nonparametric Curve Estimates," unpublished manuscript.

Csörgő, M., and Révész, P. (1974/75), "A New Method to Prove Strassen Type Laws of Invariance Principle I, II," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31, 255–259, 261–269.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Durrett, R. (2005), *Probability*: *Theory and Examples* (3rd ed.), Belmont, CA: Duxbury Press.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall.

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Fisher, N. I. (1983), "Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography," *International Statistical Review*, 51, 25–58.

Gong, W., Liu, Y., Misra, V., and Towsley, D. (2001), "On the Tails of Web File Size Distributions," in *Proceedings of 39th Allerton Conference on Communication, Control, and Computing*, October 2001, available at: *http://www-net.cs.umass.edu/networks/publications.html*.

Grama, I., and Nussbaum, M. (1998), "Asymptotic Equivalence for Nonparametric Generalized Linear Models," *Probability Theory and Related Fields*, 111, 167–214.

——— (2002), "Asymptotic Equivalence for Nonparametric Regression," *Mathematical Methods of Statistics*, 1–36.

Hannig, J., and Lee, T. C. M. (2006), "Robust SiZer for Exploration of Regression, Structures and Outlier Detection," *Journal of Computational and Graphical Statistics*, to appear.

Hsing, T., Husler, J., and Riess, R. D. (1996), "The Extremes of a Triangular Array of Normal Random Variables," *The Annals of Applied Probability*, 6, 671–686.

Inselberg, A. (1985), "The Plane With Parallel Coordinates," *The Visual Computer*, 1, 69–91.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996a), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401–407.

——— (1996b), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," *Computational Statistics*, 11, 337–381.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Berlin: Springer-Verlag.

Li, W. V., and Shao, Q.-M. (2002), "A Normal Comparison Inequality and Its Applications," *Probability Theory and Related Fields*, 122, 494–508.

Lindeberg, T. (1994), *Scale-Space Theory in Computer Vision*, Dordrecht: Kluwer.

Marron, J. S. (1996), "A Personal View of Smoothing and Statistics," in *Statistical Theory and Computational Aspects of Smoothing*, eds. W. Härdle and M. Schimek, Berlin: Physica-Verlag, pp. 1–9.

Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M., and Patil, P. (1998), "Exact Risk Analysis of Wavelet Regression," *Journal of Computational and Graphical Statistics*, 7, 278–309.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

Nussbaum, M. (1996), "Asymptotic Equivalence of Density Estimation and Gaussian White Noise," *The Annals of Statistics*, 24, 2399–2430.

Rootzen, H. (1983), "The Rate of Convergence of Extremes of Stationary Normal Sequences," *Advances in Applied Probability*, 15, 54–80.

Schmitz, H. P., and Marron, J. S. (1992), "Simultaneous Estimation of Several Size Distributions of Income," *Econometric Theory*, 8, 476–488.

Scott, D. W. (1992), *Multivariate Density Estimation, Theory, Practice and Visualization*, New York: Wiley.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

ter Haar Romeny, B. M. (2001), *Front-End Vision and Multiscale Image Analysis*, Dordrecht: Kluwer Academic Publishers.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

Wilhelm, J. R. (2002), "A Simulation Study on Competing Distributions for the Maxima of Stationary Normal Processes," unpublished master's thesis, Colorado State University.