Advanced Distribution Theory for SiZer

J. Hannig
Department of Statistics
Colorado State University
Fort Collins, CO 80523-1877

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

May 3, 2005

Abstract

SiZer is a powerful method for exploratory data analysis. In this paper, 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.

1 Introduction

SiZer has proved 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 for example Scott (1992), Wand and Jones (1995) and Fan and Gijbels (1996), for an introduction to this area. As noted in these monographs, 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 are given in Section 2) are considered here for their simplicity, ease of interpretation, and because they have been very widely studied.

Practical use of kernel methods, in both density estimation and regression, is profoundly affected by the choice of the window width (the tuning parameter which controls the amount of local averaging being used). When this is too small, the resulting estimated curve strongly feels 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, where one tries 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 what is "the best" method of doing this, which has appeared to slow actual use of these methods, for example through their implementation in software packages.

Scale space ideas (see Chaudhuri and Marron (2000) for broad discussion of these issues) have provided practical means of avoiding the problem of bandwidth selection. Scale space is a theoretical model for vision, that was 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, while 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, that 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 that is 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 about 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 paper, a much deeper distributional investigation is done, with the goal of improving the statistical performance of SiZer.

The SiZer method, as well as potential advantages from an improved distribution theory, are illustrated in Figure 1. The underlying regression function, shown as the thick black curve in Figure 1a, 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, $\sigma = 0.1$, (much higher than is typical in the wavelet literature) that was first used by Marron et al (1998) is used in the generation of the $n = 1024$ data points shown as green dots in Figure 1a.

The thin blue curves in Figure 1a show the scale space for this data set, i.e. 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 of these are 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 non-smooth features. However, for this 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 instead at understanding which of its features can be distinguished from the background noise, i.e. determining which aspects observable in the blue curves are important underlying structure, and which are spurious noise-driven artifacts.

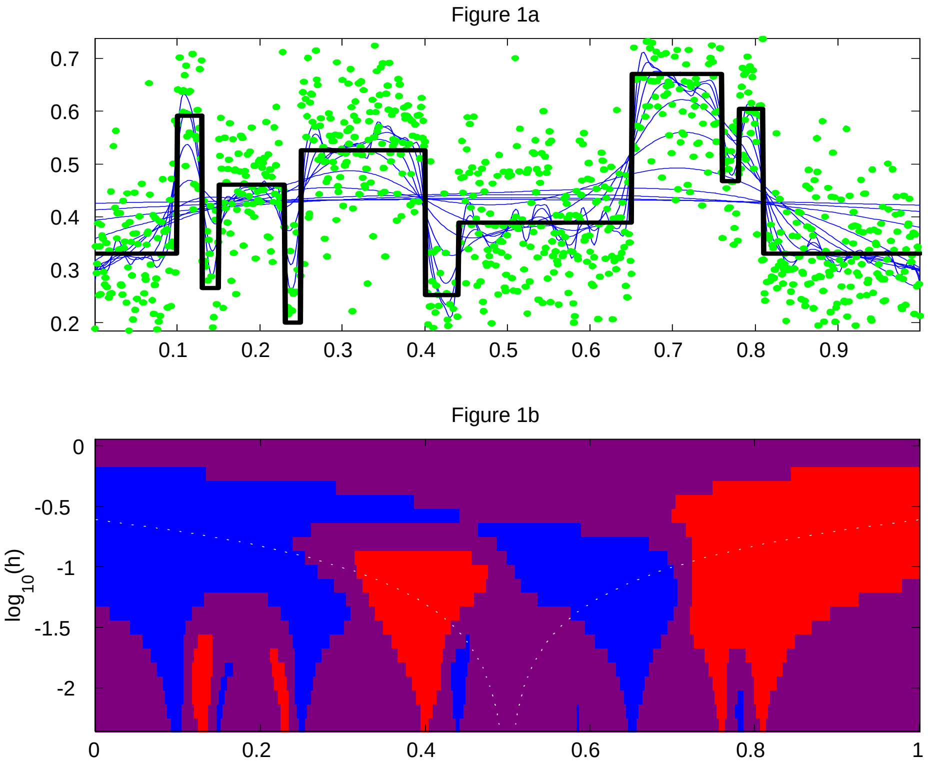
Figure 1a

Figure 1b

FIGURE 1: *Conventional SiZer analysis of the Donoho - Johnstone Blocks regression, with high noise. Shows good performance, plus a spurious feature. True regression, data and scale space shown in Figure 1a. SiZer analysis in Figure 1b.*

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 1a (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. In general, SiZer flags features of these various types using a color map, such as the one shown in Figure 1b.

The horizontal location in the SiZer map are the same as in the scale space plot above. 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, respectively). 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 1b, 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 paper.

Note that each jump in Figure 1a corresponds to a red or a blue (depending on the direction of the jump) region in the SiZer map in Figure 1b. 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 0.58. 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 into 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 paper, 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. Since the regression function is 0, the data are simply i.i.d. 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 0.05).

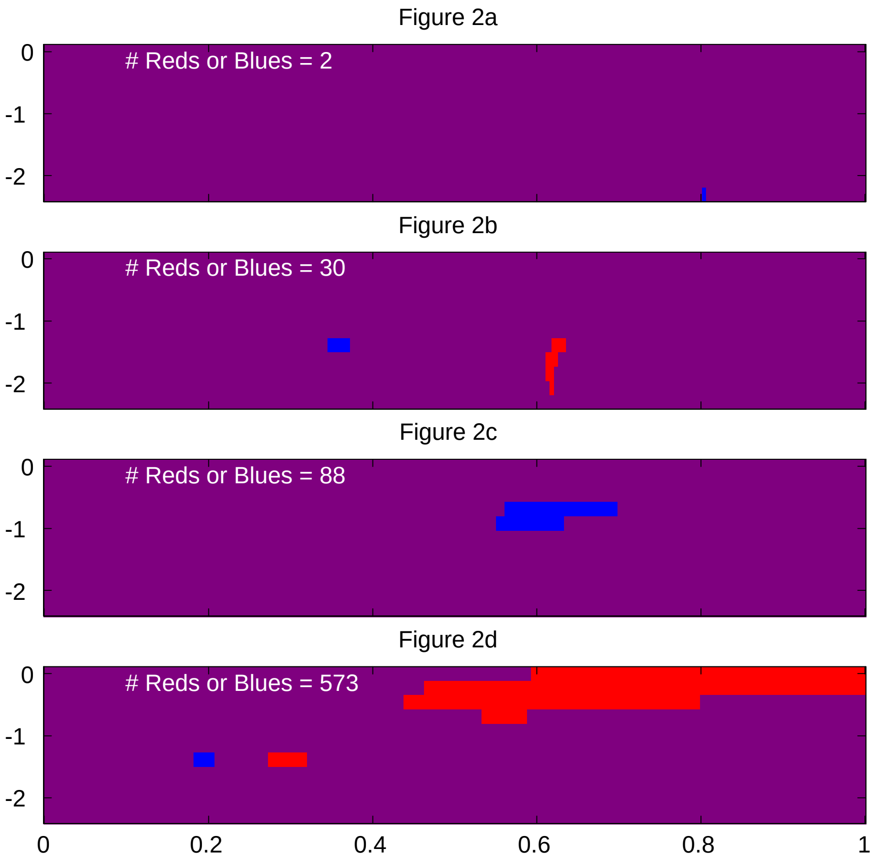

FIGURE 2: *Conventional SiZer maps, based on simulated null distributions, for* $n = 1600$ *equally spaced regression data points. Figures 2a, b, c and d are for 0.5, 0.75, 0.85 and 0.95, respectively, quantiles of distribution.*

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 1000 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 from the null distribution, it is desirable for the number of such pixels to be small. The first 405 of the 1000 ordered SiZer maps were completely purple (and are thus not shown to save space). Figure 2a shows the 500th of these (essentially the median of the population), where two pixels, at the finest scale, were flagged as significant. Figure 2b shows the 750th (the 3rd quartile), with substantially more significant pixels at medium fine scales. Figure 2c shows the 850th SiZer map, with quite a large blue region at medium coarse scales. Figure 2d 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 2a-c should be completely purple, and that Figure 2d might or might not have some color. The goal of this paper is to improve this performance, by using a better approximation of the underlying distribution theory.

A natural solution to this problem would be the use of simulation methods to compute the critical values needed for proper simultaneous adjustment. This idea was seriously considered in Section 3 of Chaudhuri and Marron (1999) and was implemented in early versions of the SiZer software. But there was a serious drawback: the simulation took hours, while 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 would no longer take hours, but it is still some minutes, enough to keep the method out of the class of *interactive* methods. The method proposed in this paper 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 1000 underlying data sets, 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 1000 maps are shown as Figures 3a, 3b, 3c and 3d respectively.

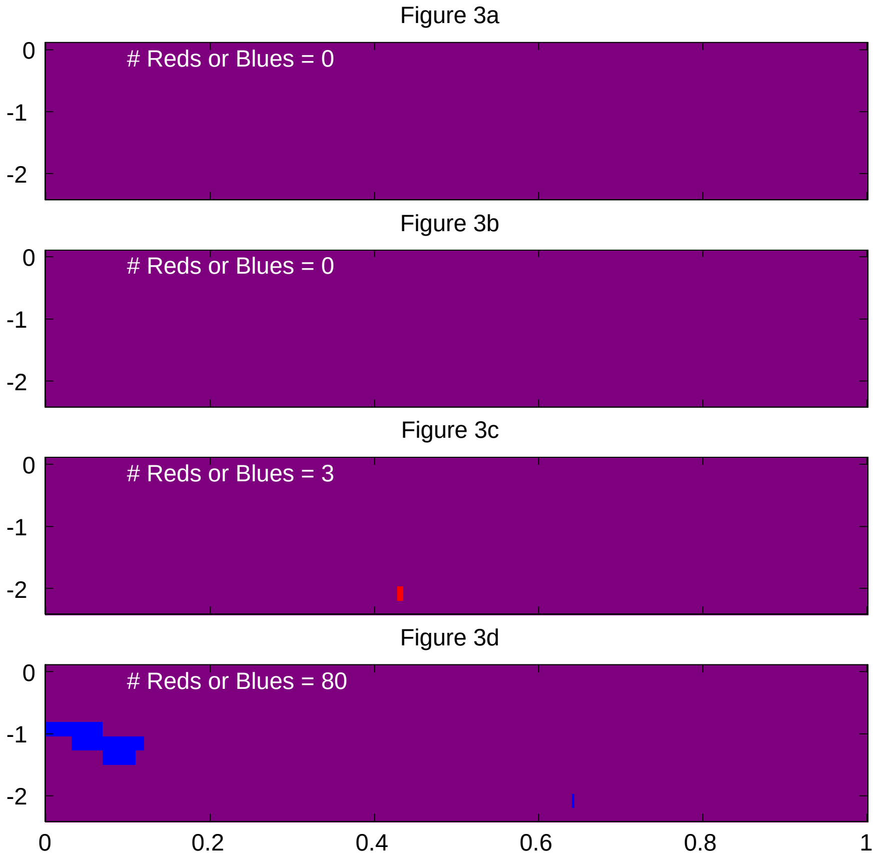

FIGURE 3: *SiZer maps for simulated null distributions, for* $n = 1600$ *equally spaced regression data points, based on the proposed row-wise procedure. Figures 3a, b, c and d are for the 0.5, 0.75, 0.85 and 0.95, respectively, quantiles of the distribution.*

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 3a and 3b (representing the first 3 quartiles), there were no spurious results. Even for the 850th ordered SiZer map, in Figure 3c, 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, that are 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 only adjusted row by row. Hence it is not surprising that some spurious structure manages to be flagged here, since each of the maps in Figure 3 includes 11 such rows (so just by chance, the test flags significant structure more often than 5% of the time).

To address this problem we also propose a global adjustment in Section 2. The corresponding globally adjusted version of Figure 3 has been plotted, but is not shown 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. Detailed analysis of the impact of these improvements is done in Section 3.

As one should expect, the improved size properties, further investigated 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 data sets in Section 3.3. The main lessons learned there are that while 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 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 view this as an acceptable price to pay for most exploratory data analyses. However, we anticipate that others will disagree, and furthermore recognize situations where statistical rigor is imperative, and thus our software allows a choice between row-wise and global implementations, together with an option to choose the level of significance α. Matlab software, based on these new ideas can be found at the web site:

http://www.stat.unc.edu/postscript/papers/marron/Matlab6Software/Smoothing/

In addition to this important row-wise vs. 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, some basics of kernel smoothing are first reviewed.

Convenient notation for density estimation is $X_1, ..., X_n$ 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 paper, 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 $(X_1, Y_1), ..., (X_n, Y_n)$. 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 $\varepsilon_1, ..., \varepsilon_n$ are independent and identically distributed. While the term "random design" means that $(X_1, Y_1), ..., (X_n, Y_n)$ are a random sample 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

$$\widehat{m}(x) = a_0, \text{ where } (a_0, a_1) = \underset{a_0, a_1}{\arg\min} \sum_{i=1}^{n} \{Y_i - [a_1 (X_i - x) + a_0]\}^2 K_h(x - X_i). \tag{2}$$

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 paper, for the same reasons as the kernel density estimator. As above, 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 in Fan and Marron (1994) is important, especially for larger data sets.

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 pre-set 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, i.e. at each pixel location in the SiZer map. Let $T_1, ..., T_g$, 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 i-th 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 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,

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

We will show in the appendix that the sequence $T_1, ..., T_g$ could be approximated by a stationary Gaussian process with mean zero, variance one and correlation

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

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

Similarly, the whole SiZer map is a matrix of pixels that were 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, since bandwidth is a multiplicative notion) and we can assume without loss of generality that the k-th row was calculated using the bandwidth hd^k, for some $h > 0$ and $0 < d < 1$.

Again the random field $T_{1,1}, ..., T_{g,r}$ could be approximated by a mean zero, variance one Gaussian random field with correlation

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

In what follows we will use $T_{i,j}$ to denote the test statistics of the SiZer map. In order 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 by using an appropriate Hungarian approximation technique (see Csörgő and Révész (1974/75) for the original paper 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 the equation (3). Chaudhuri and Marron (1999) used a "number of independent blocks" approach to give a first approximate solution. In this paper, much more precise approximations are developed. These come from

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

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

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

where Z is a standard Gaussian random variable.

Of course the main challenge is due to the fact that SiZer pixels are not independent. To that end consider a stationary, mean zero, variance one, Gaussian process $T_1, ..., T_g$, with a j step correlation denoted ρ_j. We will be interested in the distribution of $\max(T_1, ..., T_g)$. Berman (1964) has proved that if $\log(j)\rho_j \to 0$ the distribution function of $\max(T_1, ..., T_g)$ behaves asymptotically as the g-th power of the distribution function of a standard Gaussian random variable, i.e.,

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

Unfortunately this approximation is usually of little practical significance as the speed of convergence is very slow. In order to overcome this one needs to consider second order asymptotics. There are at least two alternate approximations in the literature based upon 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 $\max(\rho_1, ..., \rho_m) > 0$ then

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

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 will not reproduce it here. An interested reader can consult section 4.6 of Leadbetter, Lindgren and Rootzén (1983) for details.

The second approach, discussed in Hsing, Husler and Reiss (1996), is based on the observation that for dependent data it is often better to approximate $P[\max(T_1, ..., T_g) \leq x]$ by $\Phi(x)^{\theta g}$ where $\theta < 1$. Their main idea is to find θ

using asymptotic considerations. In order to get $\theta < 1$ they need the correlation ρ_j to increase to 1 with g for each fixed j.

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

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

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

$$\vartheta = P\left[V/2 + \sqrt{\delta_k} H_k \leq \delta_k \text{ for all } k \geq 1\right], \qquad (7)$$

where V is a standard exponential random variable and H_k is a mean zero Gaussian process independent of V that satisfies $EH_iH_j = \frac{\delta_i + \delta j - \delta_{|i-j|}}{2\sqrt{\delta_i \delta_j}}$. The authors then claim that under certain technical conditions on $\rho_{j,g}$ the distribution function $P[\max\left(\hat{T}_{1,g}, \ldots, \hat{T}_{g,g}\right) \leq x]$ 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 as neither of the methods clearly dominated the other two. 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, the implementation of the Rootzén method was seen to have 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 paper we only utilize the approach of Hsing et al (1996), which improves the size of SiZer dramatically, 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 $\rho_j = e^{-j^2\tilde{\Delta}^2/(4h^2)}\left[1 - \frac{j^2\tilde{\Delta}^2}{2h^2}\right]$. 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 appears in the appendix.

Theorem 1 *Consider a triangular array $\hat{T}_{i,g}$ of mean zero, 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

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

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) recommend that, in applications, Theorem 1 should be used to approximate $P\left[\max_{i=1,\ldots,g}\hat{T}_{i,g} \le x\right]$ 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

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

where the cluster index

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

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, 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 and variance 1 and j step correlation $\bar{\rho}_j = e^{-j^2\tilde{\Delta}^2/(4h^2)}\left(1 - j^2\tilde{\Delta}^2/h^2 + j^4\tilde{\Delta}^2/(12h^4)\right)$. 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 Figure 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 will need to study the asymptotic distribution of the maxima of the whole SiZer map, i.e.,

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

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 $T_{1,1}, ..., T_{g,r}$ are Gaussian, with mean 0 and variance 1 and correlation given by (5). In order to be able to make use of Theorem 1 we again set $\tilde{\Delta}/h = C/\sqrt{\log g}$. The following theorem is proved 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 in the appendix.

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

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

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

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

where

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

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 Hsing et al's recommendation and approximate the maximum of the SiZer map by

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

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 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 will not put the details here to save space.

It is worth pointing out that Theorem 2 could 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 available yet, as we would need second order extreme value theory for non-stationary 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 considerations. 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 in order 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.

The setting is again fixed design regression, for sample size $n = 1600$, based on an identically 0 regression function, with standard Gaussian noise. For each of 1000 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 1000 maxima is studied in Figure 4, where it is compared to 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 vs. 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, 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 if it represents a serious departure of the data distribution from the theoretical distribution. 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.

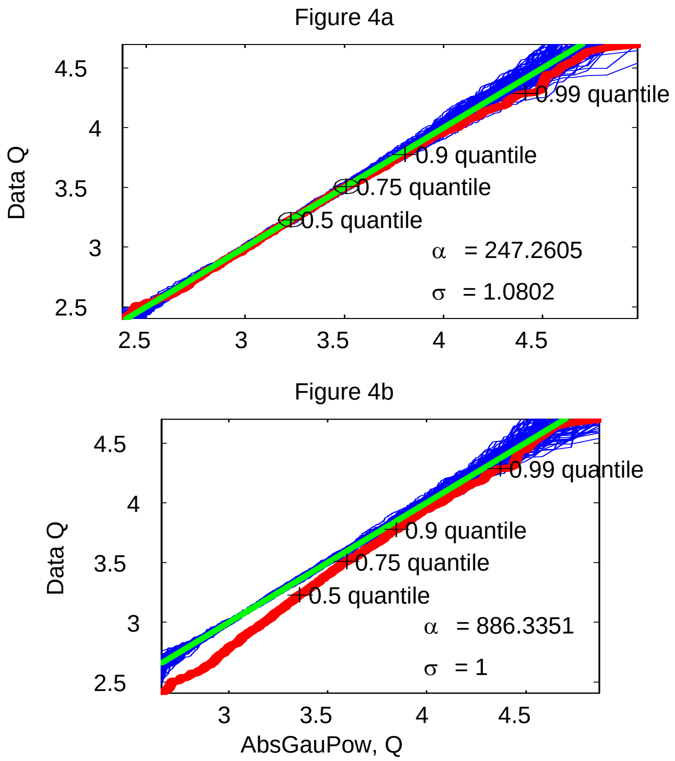

FIGURE 4 *QQ Plot showing that a power of Gaussian provides a good fit to the maxima of the 1000 simulated SiZer maps under the null hypothesis. This plot was generated using the same simulated data set as in figures 2 and 3. The parameters are obtained by quantile matching (4a) and by the theoretical considerations of Section 2.3 (4b). This shows that the global adjustment will be slightly conservative, due to the slow rate of convergence.*

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

The estimated value of $\sigma = 1.08$ is very good, because the underlying Gaussian distribution here 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 we decide to match. For example, if we set $\sigma = 1$ we get $\alpha = 552$. Moreover if we then decide to approximately match quantiles .8 and .95 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 et al (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 4b. We can see that while 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 4b is below the blue envelope for some of the quantiles suggest that in practice the global adjustment will be conservative. This conclusion is confirmed by the simulation results of Section 3.1 that shows that global adjustment is indeed slightly conservative.

Similar Q-Q plots have been constructed for other simulation settings (detailed in Section 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 vs. t distribution issues, which are discussed 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 α100% of the rows containing "false positives", i.e., pixels flagged as statistically significant when no signal is present in the data. The second, called "global" simultaneous inference, aims at having at most α100% 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 $\max(T_1, ..., T_g)$ is the same as the distribution of -$\min(T_1, ..., T_g)$. It follows that, if the data contains no signal, then the probability there is a spurious color on the gth row is

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

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 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 while the constant C_G is the same for all the rows. Again

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

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 the size properties are studied via a simulation study in Section 3.1. The amount of power that is sacrificed to get the size correct, is studied via 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 did an array of simulations against several variations of "the null hypothesis". We tried:

- Each of the settings of:

1. (KDE) kernel density estimation, for the Uniform(0,1) density,

2. (FDR-N) fixed design regression, for an equally spaced design, with standard Gaussian noise, but no signal,

3. (FDR-E) fixed design regression, for an equally spaced design, with standard Exponential noise,

4. (RDR-U) random design regression, where the X_i are chosen from the Uniform(0,1) density, and the Y_i are independent standard Gaussian.

5. (RDR-N) random design regression, where the X_i are chosen from the N(0,1) density, and the Y_i are independent standard Gaussian.

- For each of the above settings, the following sample sizes were tested:

1. $n = 100$,
2. $n = 400$,
3. $n = 1600$,
4. $n = 6400$.

For each of the 20 combinations above, 1000 pseudo data sets were drawn, and the various SiZer maps were calculated, and the numbers of red and blue pixels (ideally none, since there are no signals in any of these examples) was counted.

One way to summarize these numbers is row-wise in the SiZer maps. In particular, for each setting, each sample size, and each row, report the percentage of realizations of the data where there were some red or blue pixels in that row. Figure 5 shows these summaries. Notice that if no red or blue pixels are present in a particular row, the $\max_{i=1,\ldots,g} |T_i|$ for T_i's corresponding to this row was less then the preset value of the cut-off 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 seen 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 5a summarizes performance for the Kernel Density Estimation setting, Figure 5b does the same for the Fixed Design Regression Gaussian noise setting, Figure 5c is for the Uniform(0,1) Random Design Regression setting, Figure 5d is for the N(0,1) Random Design Regression setting and Figure 5e contains the Fixed Design Regression exponential noise setting. The coordinates (points on the horizontal axes) represent rows of the SiZer map, and thus are quantified via $\log_{10} h$ (only shown on 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 1000 replications) that showed 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 $\alpha = 0.05$ for the row-wise procedures such as the conventional SiZer, and our new Row-Wise SiZer. Hence, this level is shown by a horizontal black line.

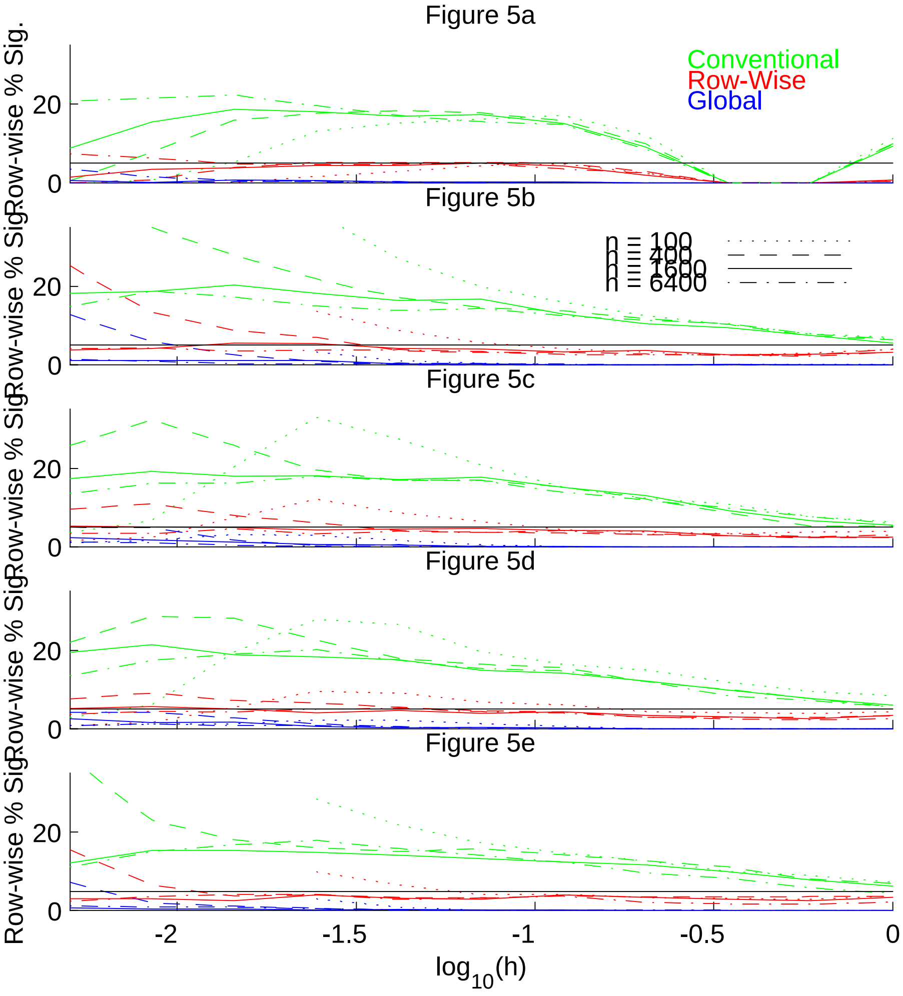

FIGURE 5 *Row-wise summaries of the percent of significant pixels for SiZer under the null hypothesis, allowing comparison of the different simultaneity adjustments and sample sizes. Clearly shows relationship between sample sizes. Figures 5a, b, c, d and e are for the settings of KDE, FDR-N, RDR-U, RDR-N and FDR-E respectively.*

Note that in almost every case the conventional SiZer flags significant structure far too often. This again verifies the main idea in this paper: it is well worth finding less crude approaches to this multiple comparison problem.

Similarly, in a large majority of the cases, the new Row-Wise SiZer is quite close to the desired $\alpha = 0.05$.

As expected, the global method are almost always quite far below the desired level, because they aim at a global size of $\alpha = 0.05$, which requires them to be deliberately conservative when studied in this row-wise sense.

A perhaps surprising feature in the KDE setting, studied in Figure 5a, is the 0 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, far too many percentages are 100%, since 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), page 31, where the "circular design" device is called the "wrap around" boundary condition). While this crude adjustment is reasonably effective at most scales, there are a few where it introduces artifacts such as the zeros shown in Figure 5a. Such boundary effects are not a serious issue for the regression settings, because the local linear smoother that is used in both performs an automatic first order boundary adjustment.

Another departure from the expected size occurs for the regression settings, shown in Figures 5b, 5c, 5d and 5e. 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. This idea is verified by the fact that it is generally worst for $n = 100$, better for $n = 400$, and the problem is nonexistent for $n = 1600$ and $n = 6400$. Exceptions include the FDRs in Figure 5b and 5e, 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 Figure 5c and 5d, 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 do any inference (thus most of the pixels are colored gray), and 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 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 1000) 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.

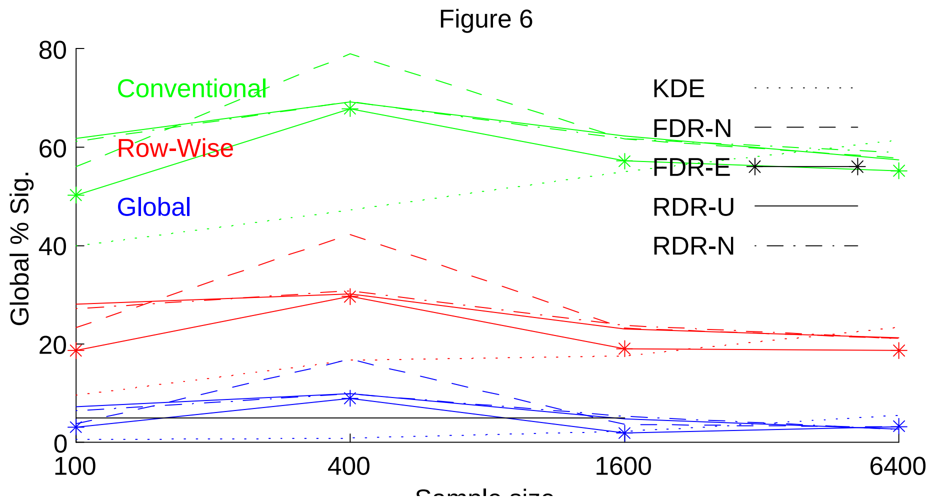

Figure 6: *The global size summaries showing the percent of significant pixels in the full SiZer maps, under the null hypothesis, grouped by settings.*

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 $\alpha = 0.05$, which not surprisingly shows that there is substantial difference between row-wise and global statistical inference. This is consistent with the global method appearing as generally too conservative in Figure 5.

Performance of the global SiZer approach, is quite dependent on the setting. For Kernel Density Estimation the method is generally conservative. This is caused by the boundary effect and adjustment discussed above, and 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 above (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 above, 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 ($n = 1600$, $n = 6400$), 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 there because the settings are in different panels. This time the panels show the sample sizes n, with $n = 100$, 400, 1600 and 6400 in Figures 7a, b, c and d respectively. 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

$\log_{10} h$.

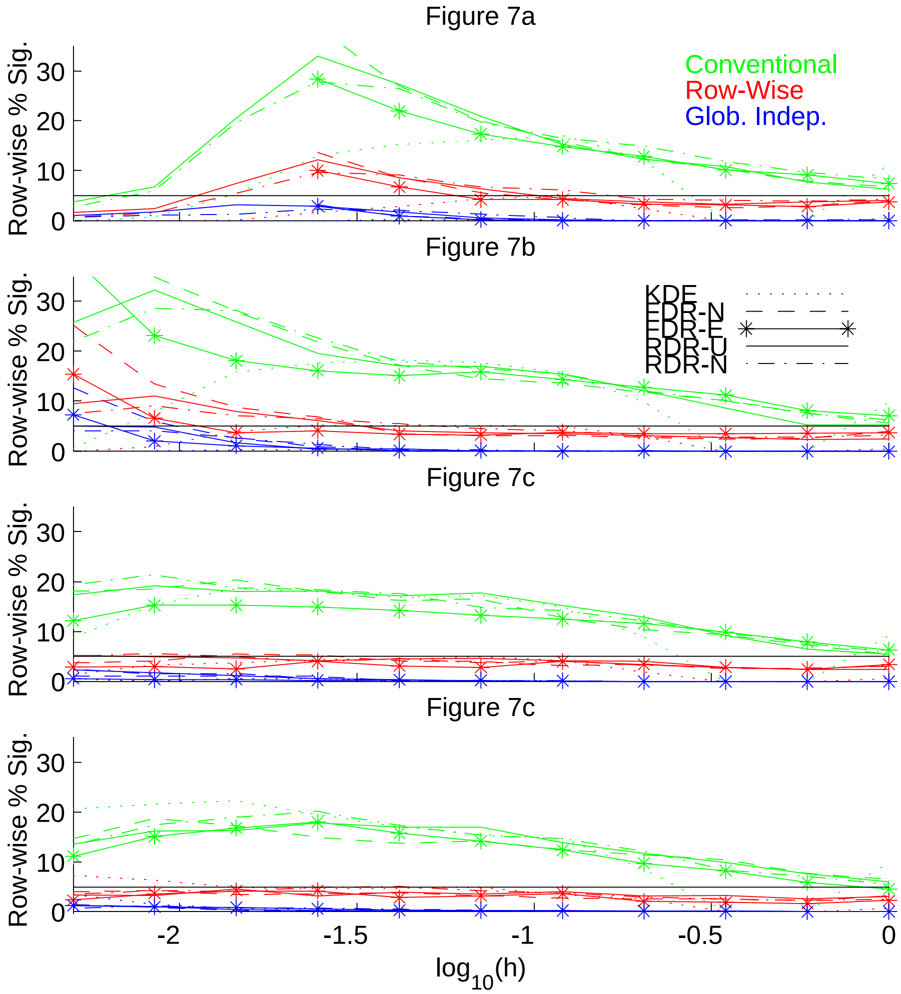
Figure 7a

Figure 7b

Figure 7c

Figure 7c

FIGURE 7 *Row-wise summaries of the percent of significant pixels for SiZer under the null hypothesis, allowing comparison of the different simultaneity adjustments and settings. This organization shows that settings are very similar. Figures 7a, b, c and d are for the sample sizes of* $n = 100$*,* 400*,* 1600 *and* 6400*, respectively.*

The main lesson of Figure 7 is that curves of the same color tend to be very close to each other, i.e. the settings are very similar. While there are important differences in the simultaneity type (expressed by colors), and the sample sizes (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 $n = 1600$ and $n = 6400$, the percentages virtually achieve their goal of $\alpha = 0.05$, uniformly over both rows and settings (except for density estimation at large scales).

A similar simulation study has been 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.

3.2 Power Simulations

The previous section showed that our global versions of SiZer were quite good at achieving the desired overall size for the statistical inference. In this and the next section, by analyzing some simulated and real data sets, it is seen that this could entail substantial cost in terms of power especially when using one of the global adjustments. This is to be expected as it is consistent with well established principles of hypothesis testing, in particular, the theory which 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 1a. Figure 8 allows direct comparison between the conventional SiZer shown in Figure 8a, the new Row-Wise SiZer shown in Figure 8b and the Global SiZer shown in Figure 8c.

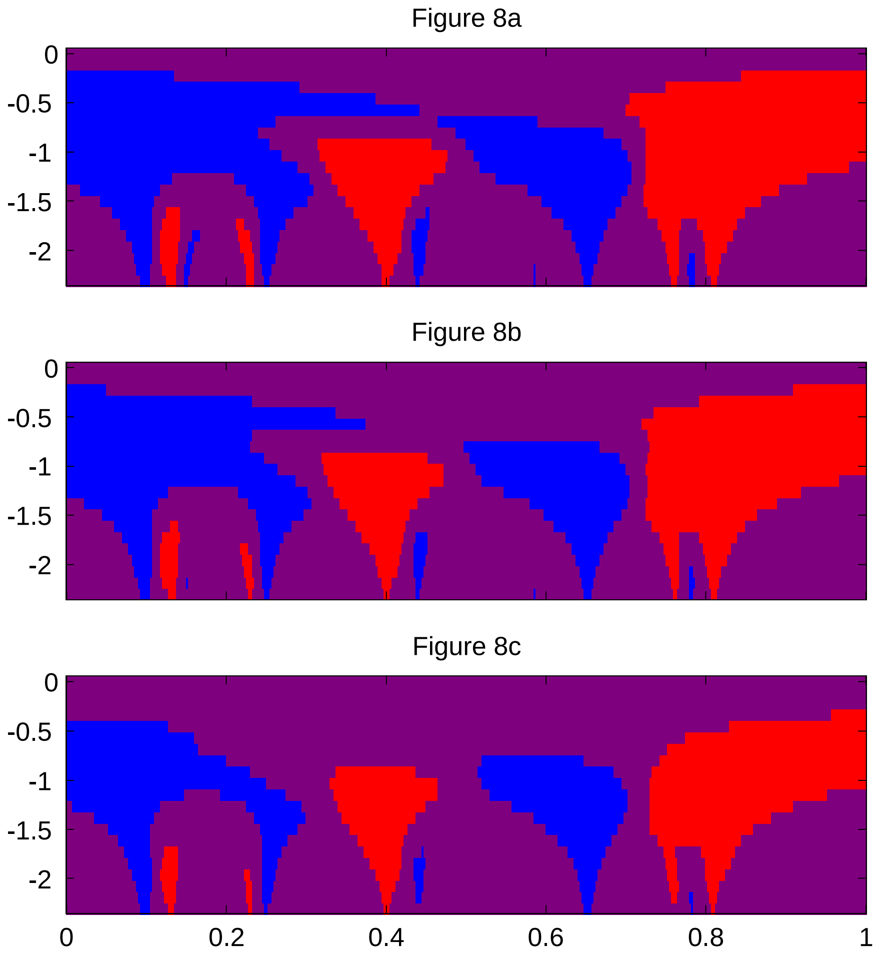

FIGURE 8: *Full range of SiZer analyses of the Donoho - Johnstone Blocks regression, with high noise. Figures 8a, b and c show conventional, row-wise and global SiZer versions, respectively.*

As shown in Figure 1b, the conventional SiZer flags all 11 jumps as statistically significant, but it also indicates a spurious jump near $x = 0.58$. As expected, the new Row-Wise SiZer (Figure 8b) flags fewer pixels as significant, but still finds all 11 jumps. The spurious jump near $x = 0.58$ is still present, but smaller. For the global method the spurious feature disappears, but also the jump near $x = 0.15$ disappears as well. This reflects the loss of power from insisting on global simultaneous inference.

If one were to use only the global analysis, the upward jump near $x = 0.78$, would be flagged as statistically significant by a very small blue region. Thinking 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) at all that is found by a global method, should be regarded as 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 window width 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, in particular the small blue region just left of $x = 0$, again reflecting some loss of power.

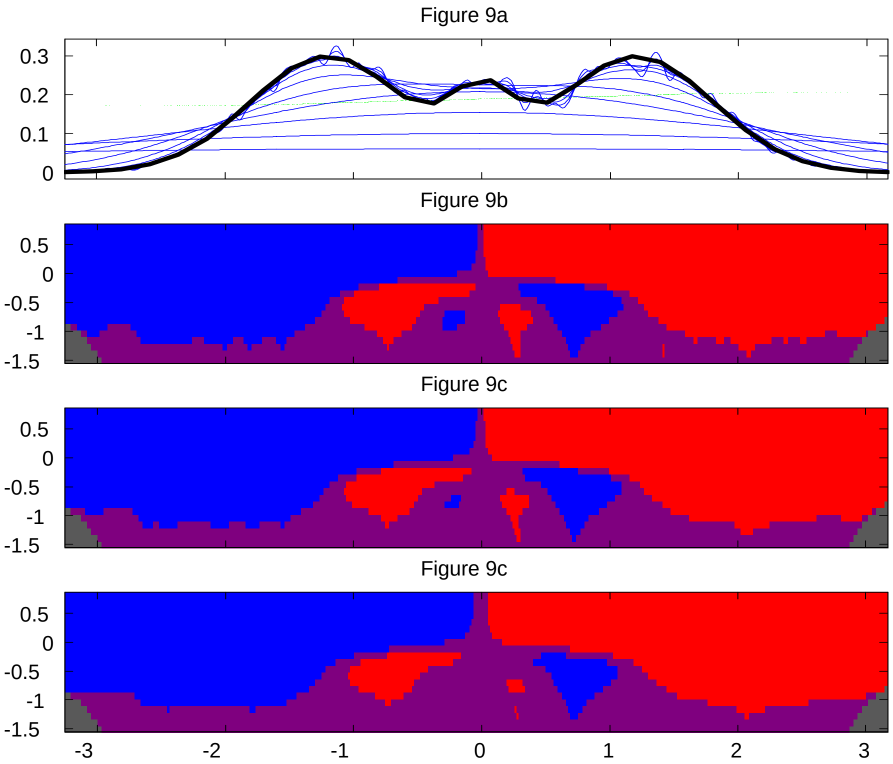
Figure 9a

Figure 9b

Figure 9c

Figure 9c

FIGURE 9: *Full range of SiZer analyses of the Trimodal mixture of Gaussians. Figures 9a, b, c and d show the scale space overlaid with the true density, conventional, row-wise and global SiZer versions, respectively.*

Similar plots have been constructed for all of the Marron-Wand Gaussian mixture densities, for the sample sizes $n = 100$, 1000, $10,000$. 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 a similar extent to that shown in

Figure 9). Sometimes this erosion was enough that significant features actually disappeared, as in Figure 9d, but most often they did not. Spurious features, such as the very small red region, near $x = 1.4$ in Figure 9b, were fairly rare, perhaps because at most locations, these densities are not close to flat (as at the null distributions studied in Section 3.1), but instead have substantial slope.

3.3 Real Data Examples

Another approach to studying the trade-off between size and power that is made by 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 Figures 9. The conventional SiZer analysis shows two significant modes, which has been independently confirmed by a parametric analysis as discussed in Schmitz and Marron (1992). The red region between modes is still present for the new Row-Wise SiZer shown in Figure 10c, and again, greater credence needs to be placed in this more precise version. Unfortunately this red region completely disappears in the global SiZer map. This loss of power is particularly unfortunate, since the bimodality is the important feature of this data set.

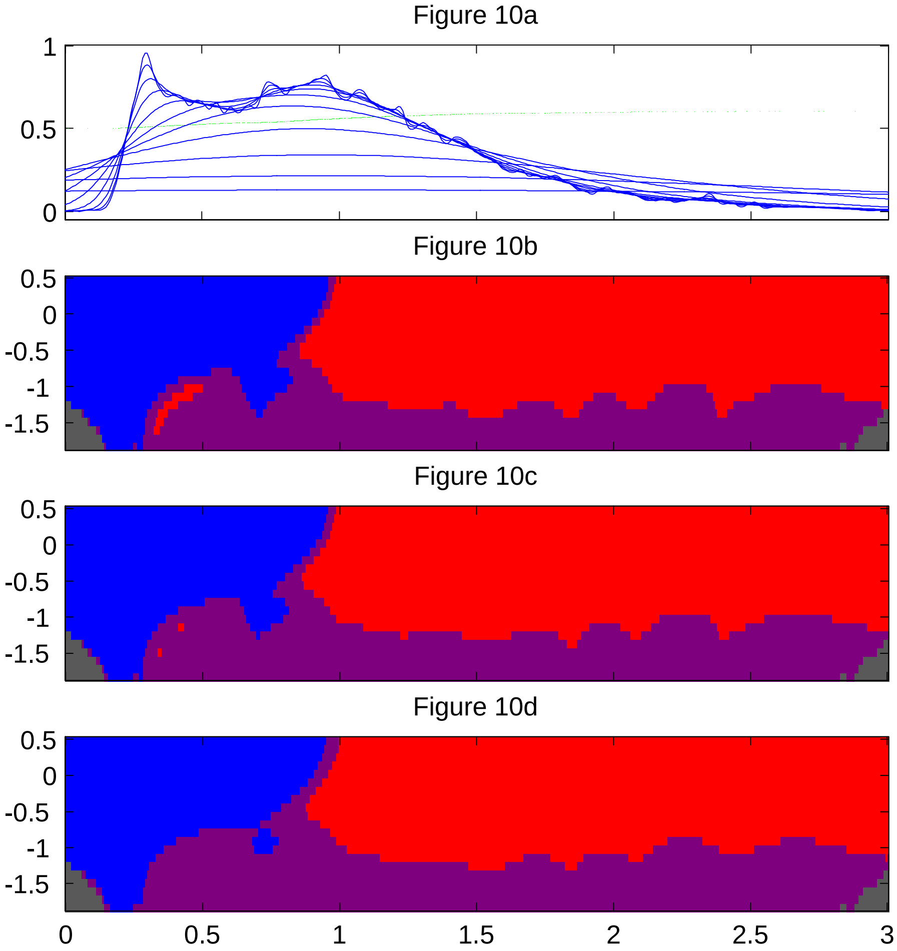

FIGURE 10: *Full range of SiZer analyses of the British Family Incomes data. Figures 10a, b, c and d show the scale space, conventional, row-wise and global SiZer versions, respectively.*

While 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 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.

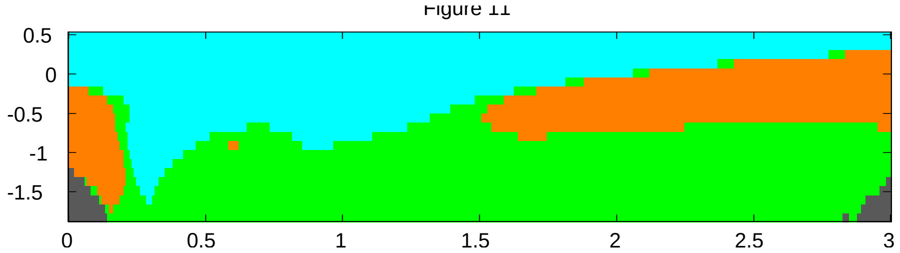

FIGURE 11: *Global curvature SiZer analysis of the British Family Incomes data. This finds the bimodality that is known to be an important feature of this data set.*

The bimodality of this data set is shown to be strongly significant, by the very small orange region near $x = 0.6$. While 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 regarded as 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 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, Massachusetts, 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, and the intensity of florescence 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" there are in this square root histogram. Here we treat this as a regression problem.

Figure 12 shows again the same three panels, comparing the different simultaneity methods. Figure 12b, 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 12c. 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.

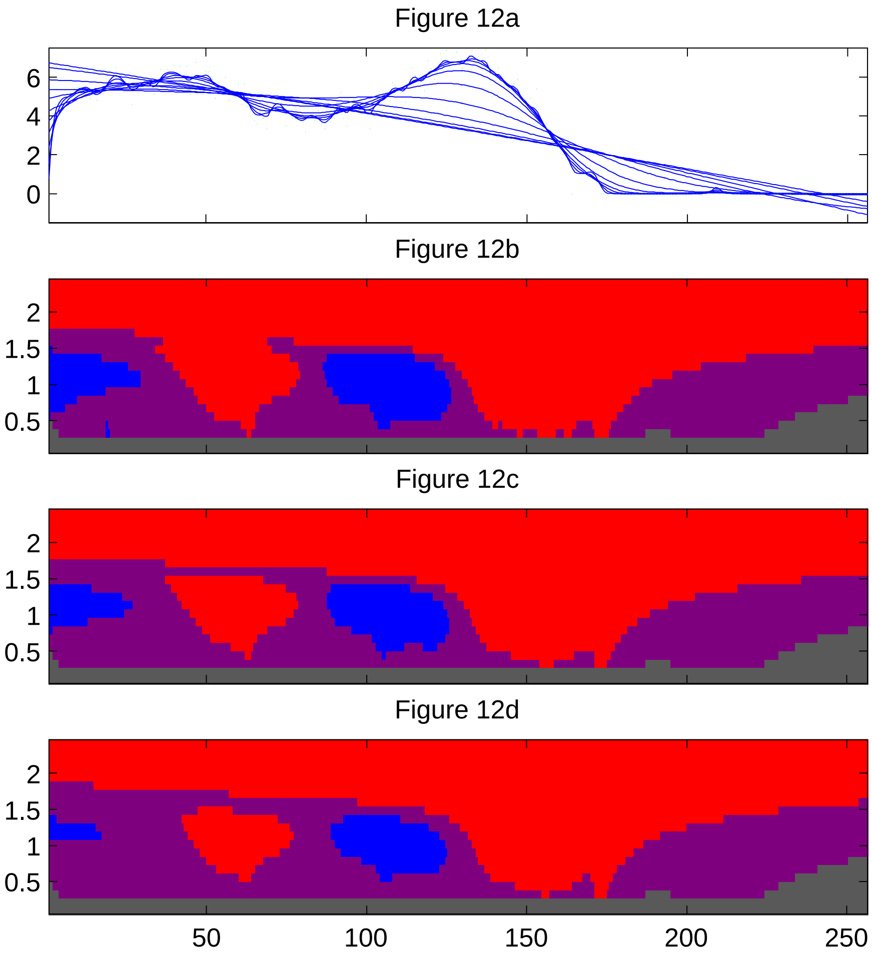

Figure 12a

Figure 12b

Figure 12c

Figure 12d

FIGURE 12: *Full range of SiZer analyses of a flow cytometry data set. Figures 11a, b, c and d show the scale space, conventional, row-wise and global SiZer versions, respectively.*

Based on this experience, and 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 needs to 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

While the methods developed in this paper are intended to enhance the applicability of the SiZer method, there are a number of remaining open problems, including:

1. Develop 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.

A Appendix

In the appendix we present proofs of the key results of this paper.

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

We will 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 setting have some very strong connections. For some interesting mathematics that makes these connections precise, see Nussbaum (1996), Brown and Low (1996), Brown at. al. (2002) and 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

$$\begin{aligned} a_1 &= -c^{-1}\left[\sum_{i=1}^{n} K_h(x-X_i)\right]\left[\sum_{i=1}^{n}(x-X_i)\,K_h(x-X_i)\,Y_i\right] \\ &+ c^{-1}\left[\sum_{i=1}^{n}(x-X_i)\,K_h(x-X_i)\right]\left[\sum_{i=1}^{n} K_h(x-X_i)\,Y_i\right], \qquad (15) \\ c &= \left[\sum_{i=1}^{n} K_h(x-X_i)\right]\left[\sum_{i=1}^{n}(x-X_i)^2\,K_h(x-X_i)\right] \\ &- \left[\sum_{i=1}^{n}(x-X_i)\,K_h(x-X_i)\right]^2 . \end{aligned}$$

is significantly different from 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 will usually assume $\Delta \to 0$.) If x is away from the boundary, 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 (15) disappears.

Denote $p = \tilde{\Delta}/\Delta$. The number 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

$$T_j \approx \sum_{q=1}^{n} W_{jp-q}^{h} Y_q. \tag{16}$$

The exact form of the W_{jp-q}^{h} is given in the first term of (15). For our purpose it suffices to realize that W_{jp-q}^{h} is proportional to $-(jp - q)\, K_{h/\Delta}(jp - q)$. Thus the weights W_q^h 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 independent, identically distributed random variables. If additionally the Y's have two finite moments, the linear approximation (16) greatly simplifies the distribution theory, because for h/Δ large enough the Cramèr-Wold device and Lindeberg-Feller Central Limit Theorem (see for example 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 $T_1, ..., T_g$ also depends on the correlation between them. This correlation is approximated by

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

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

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

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

which is (5).

In practice we do not know the standard deviation of the noise ε_i. This is needed to scale $T_{1,1}, ..., T_{g,r}$ to have variance 1. For this reason it must be estimated from the data introducing additional dependence as well as other issues. However, this is not a problem in theory, as consistent estimators of this standard deviation are available and therefore 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

Using 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

$$\sup_{j\geq l_g} |\rho_{j,g}| \log g \leq |\rho_{l_g,g}| \log g \to 0,$$

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

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

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

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

Thus

$$\begin{aligned}
\sum_{j=m}^{l_g} & \qquad g^{-(1-\rho_{j,g})/(1+\rho_{j,g})} \frac{(\log g)^{-\rho_{j,g}/(1+\rho_{j,g})}}{(1-\rho_{j,g})^{1/2}} \\
&= \sum_{j=m}^{l_g} g^{-(1-\rho_{j,g})/(1+\rho_{j,g})} \frac{(\log g)^{(1/2)(1-\rho_{j,g})/(1+\rho_{j,g})}}{((1-\rho_{j,g})\log g)^{1/2}} \\
&\le \max\left(l_g g^{-(1-\exp(-\varepsilon))/(1+\exp(-\varepsilon))} \frac{(\log g)^{(1/2)(1+2\exp(-3/2))/(1-2\exp(-3/2))}}{((1-\exp(-\varepsilon))\log g)^{1/2}}, \right. \\
&\qquad \left. \sum_{j=m}^{l_g} \exp\left[-\frac{3j^2C^2}{8}\left(1-\frac{\varepsilon}{2}\right)\right] \frac{\exp[(3j^2C^2)\log\log g/(4\log g)]}{((3j^2C^2)(1-\varepsilon/2)/4)^{1/2}} \right).
\end{aligned}$$

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, i.e., 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 $Z = H_1 = H_2 = ...$, where Z is a standard Gaussian random variable. Thus the problem in (7) transforms to

$$\vartheta = P\left[V/2 + k\sqrt{3\xi}Z \le 3\xi k^2 \text{ for all } k \ge 1\right], \tag{17}$$

where $\xi = C^2/4$. Since V is a non-negative random variable, the system of inequalities in (17) implies $Z < \sqrt{3\xi}$. Moreover under this condition $V/2 + k\sqrt{3\xi}Z - 3\xi k^2$ is decreasing as a function of k and therefore

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

The equation (9) follows immediately.

A.3 Proof of Theorem 2

Unlike the proof of Theorem 1, we cannot use an “off the shelf” theorem. Instead we will 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 zero, variance one, Gaussian random field with correlation given by (12). For simplicity denote

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

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

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

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

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

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

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

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

where

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

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 enough to prove

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

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

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

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

References

[1] Berman, S. (1964) Limit theorems for the maximum term in stationary sequences, *Annals of Mathematical Statistics*, 35, 502-516.

[2] Brown, L. D. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise, *Annals of Statistics*, 24, 2384-2398.

[3] Brown, L. D., Cai, T. T. and Low, M. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Annals of Statistics*, 30, 688-707.

[4] Chaudhuri, P. and Marron, J. S. (1999) SiZer for exploration of structure in curves, *Journal of the American Statistical Association*, 94, 807-823.

[5] Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation, *Annals of Statistics*, 28, 408-428.

[6] Chaudhuri, P. and Marron, J. S. (2002) Curvature vs. Slope Inference for Features in Nonparametric Curve Estimates, unpublished manuscript.

[7] Csörgő, M. and Révész, P. (1974/75) A new method to prove Strassen type laws of invariance principle I, II, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 31, 255–259, 261 − 269.

[8] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425-455.

[9] Durrett, R. (2005) *Probability: Theory and examples*, 3rd edition, Duxbury Press, Belmont, CA.

[10] Fan, J. and Gijbels, I. (1996) *Local polynomial modeling and its applications*, Chapman and Hall, London.

[11] Fan, J. and Marron, J. S. (1994) Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, 3, 35-56.

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

[13] Gong, W., Liu, Y., Misra, V. and Towsley, D. (2001) On the tails of web file size distributions, *Proceedings of 39-th Allerton Conference on Communication, Control, and Computing*. Oct. 2001. Internet available at: http://www-net.cs.umass.edu/networks/publications.html.

[14] Grama, I. and Nussbaum, M. (1998) Asymptotic equivalence for nonparametric generalized linear models, *Probability Theory and Related Fields*, 111, 167-214.

[15] Grama, I. and Nussbaum, M. (2002) Asymptotic equivalence for nonparametric regression, *Mathematical Methods of Statistics*, 1-36.

[16] Hsing, T., Husler, J. and Riess, R. D. (1996) The extremes of a triangular array of normal random variables, *Annals of Applied Probability*, 6, 671-686.

[17] Hannig, J. and Lee, T. C. M. (2005) Robust SiZer for Exploration of Regression

[18] Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer*, 1, 69-91.

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

[20] Li, W. V. and Shao Q.-M. (2002) A normal comparison inequality and its applications, *Probab. Theory Relat. Fields*, 122, 494-508.

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

[22] Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error, *Annals of Statistics*, 20, 712-736.

[23] 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, 1-9 (with discussion, and rejoinder 103-112).

[24] 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.

[25] 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.

[26] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b) Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics*, 11, 337-381.

[27] Nussbaum, M. (1996) Asymptotic equivalence of density estimation and Gaussian white noise, *Annals of Statistics*, 24. 2399-2430.

[28] Rootzen, H. (1983) The rate of convergence of extremes of stationary normal sequences, *Advances in Applied Probability*, 15, 54-80.

[29] Scott, D. W. (1992) *Multivariate density estimation, theory, practice and visualization*, John Wiley: New York.

[30] Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, *Econometric Theory*, 8, 476-488.

[31] Silverman, B. W. (1986) *Density estimation for statistics and data analysis*, Chapman and Hall: London.

[32] ter Haar Romeny, B. M. (2001) *Front-End Vision and Multiscale Image Analysis*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

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

[34] Wilhelm, J. R. (2002) *A simulation study on competing distributions for the maxima of stationary normal processes*, M. S. Thesis, Department of Statistics, Colorado State University.