Robust SiZer for Exploration of Regression Structures and Outlier Detection

Jan Hannig[*] & Thomas C. M. Lee

June 6, 2004; revised: December 7, 2004; February 21, 2005

Abstract

The SiZer methodology proposed by Chaudhuri & Marron (1999) is a valuable tool for conducting exploratory data analysis. In this article a robust version of SiZer is developed for the regression setting. This robust SiZer is capable of producing SiZer maps with different degrees of robustness. By inspecting such SiZer maps, either as a series of plots or in the form of a movie, the structures hidden in a data set can be more effectively revealed. It is also demonstrated that the robust SiZer can be used to help identifying outliers. Results from both real data and simulated examples will be provided.

KEY WORDS: local linear regression, M–Estimator, outlier identification, robust estimation, SiZer

[*]Corresponding author. Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A. Email: hannig@stat.colostate.edu

1 Introduction

Since its first introduction in Chaudhuri & Marron (1999, 2000), SiZer has proven to be a powerful methodology for conducting exploratory data analysis. Given a set of noisy data, its primary goal is to help the data analyst to distinguish between the structures that are "really there" and those that are due to sampling noise. This goal is achieved by the construction of a so–called *SiZer map*. In short, a SiZer map is a 2D image that summarizes the locations of all the statistically significant slopes where these slopes are estimated by smoothing the data with different bandwidths. The idea is that, say if at location x, all estimated slopes (with different bandwidths) to its left are significantly increasing while all estimated slopes to its right are significantly decreasing, then it is extremely likely that there is a "true bump" in the data peaked at x. For various Bayesian versions of SiZer, see Erästö & Holmström (2004) and Godtliebsen & Oigard (2005).

In this article some major modifications are made to the original SiZer of Chaudhuri & Marron (1999) for the regression setting. These include the replacement of the original local linear smoother with a robust M–type smoother and the use of a new robust estimate for the noise variance. In addition, a different definition for the so–called effective sample size is proposed. Since there is a *cutoff* parameter c (defined in Section 2.1) that one can choose for the M–type smoother, this enables the new SiZer to produce different SiZer maps with different levels of robustness. With these modifications, the new SiZer is able to produce improved SiZer maps that are better in terms of revealing the structures hidden in the data. In addition, the new SiZer can also be applied to help identifying outliers.

The new robust SiZer also has the following appealing feature. The data driven choice of bandwidth for the M-type nonparametric smoothers is not-well known or investigated in the literature. Moreover, the expensive computation needed for M-type estimation makes any standard cross-validation type techniques for bandwidth selection computationally difficult, while virtually nothing is known in the literature about the properties of such bandwidth selectors in the context of M-type nonparametric smoothing. Consequently, the multi-scale (or multi-bandwidth) approach of SiZer is particularly appealing here, as it eliminates the need for a choice of the "optimal" bandwidth.

To proceed we first present an example for which the SiZer maps produced by the original and the new SiZers are different. Displayed in the top panel of Figure 1 is a simulated noisy data set generated from the regression function shown in red. This regression function, modified from the "bumps" function of Donoho & Johnstone (1994), is an increasing linear trend superimposed with two sharp features located at $x = 0.3$ and $x = 0.7$. Also displayed in blue is a set of estimated regression functions computed with different bandwidths. The bottom panel displays a non–robust SiZer map (i.e., with cutoff $c = \infty$; see Section 2.1) obtained by applying the new SiZer to this data set. The horizontal axis of the map gives the x–coordinates of the data, while the vertical axis corresponds to the bandwidths used to compute the blue smoothed curves. These bandwidths are displayed on the log scale, with the smallest bandwidth at the bottom. The color of each pixel in the SiZer map indicates the result of a hypothesis test for testing the significance of the estimated slope computed with the bandwidth and at the location indexed by respectively the vertical and horizontal axes. Altogether there are four colors: blue and red indicate respectively the estimated slope is significantly increasing and decreasing, purple indicates the slope is not significant, while grey shows that there is not enough data for conducting reliable statistical inference.

This SiZer map correctly identifies the two bumps peaked at $x = 0.3$ and $x = 0.7$. However, due to the present of an outlier that was artificially introduced at $x = 0.75$, the map also suggests that there is a bump located at $x = 0.75$ when the data are examined at a relatively finer scale; i.e., when one uses relatively smaller bandwidths to smooth the data.

The robust version of the new SiZer is capable of eliminating various spurious features such as this "false bump". The top panel of Figure 2 displays the same noisy data set as in Figure 1, together with a family of robust M–type local fits. The cutoff parameter used in these M–type fits was $c = 1.345$. Displayed in the bottom panel is a robust SiZer map corresponds to these M–type local fits. One can see that the effect of the outlier was eliminated. In addition, as the top portion of this map is blue, it also correctly suggests that there is an increasing trend when the data are examined at a relatively coarser scale.

The corresponding SiZer map obtained from the original SiZer of Chaudhuri & Marron (1999) is given in Figure 3. This map was produced from codes provided by Professor Marron. Notice that it fails to detect the bump at $x = 0.7$, and misses the global increasing trend at the coarser scales. An explanation for this failure is given in Section 3.2

The rest of the article is organized as follows. The proposed SiZer is presented in Section 2. Section 3 discuss the issue of outlier identification. In Section 4 the proposed SiZer is applied to a difficult simulated and a real data sets. Conclusions are offered in Section 5. Technical and computational details are delayed to the appendix.

2 A Robust Version of SiZer

2.1 Background

We shall follow Chaudhuri & Marron (1999) and consider nonparametric smoothing using local linear regression. Suppose observed is a set of observations $\{X_i, Y_i\}_{i=1}^{n}$ satisfying

$$Y_i = m(X_i) + \epsilon_i,$$

where m is the regression function and the ϵ_i's are zero mean independent noise with common variance σ^2. The local linear regression estimate for $m(x)$ and $m'(x)$ at

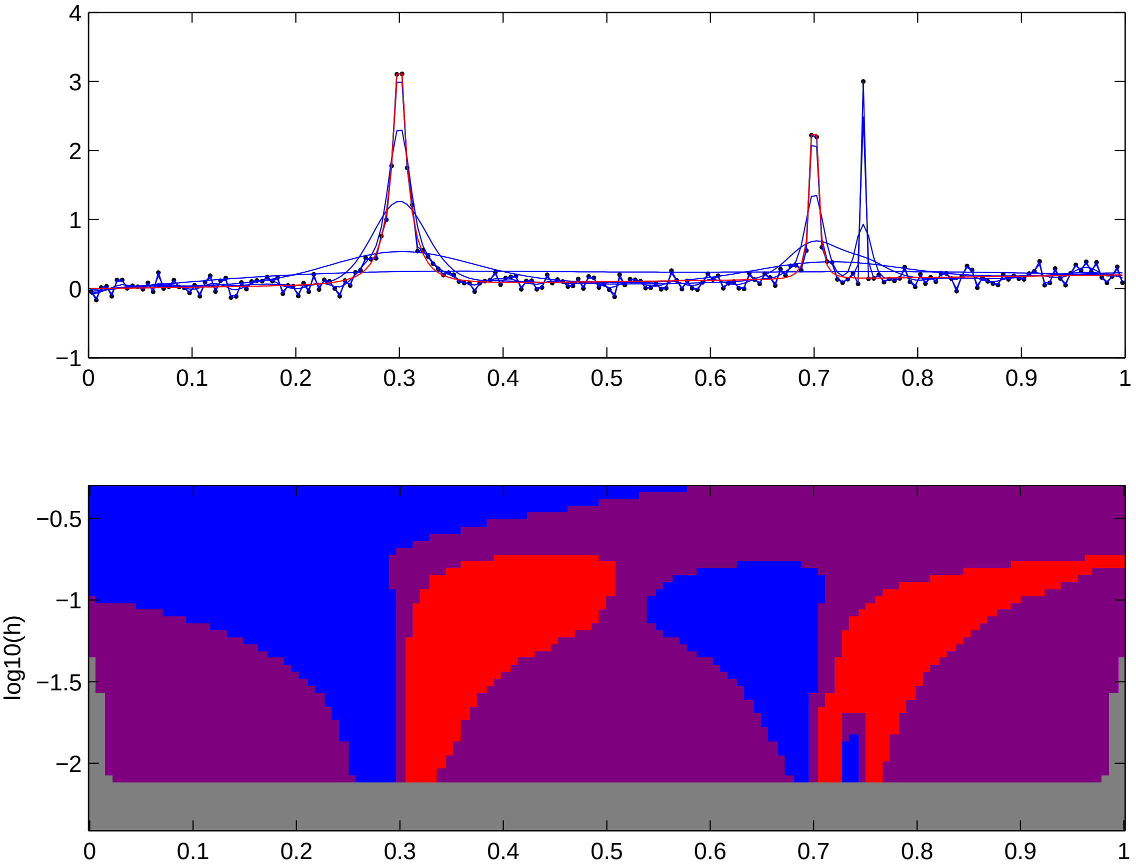

Figure 1: *Top: noisy data (black) generated by the red regression function* $m(x) = \frac{x}{5} + 4.2\left(1 + \left|\frac{x-0.3}{0.03}\right|\right)^{-4} + 5.1\left(1 + \left|\frac{x-0.7}{0.01}\right|\right)^{-4}$*, together with a family of local linear fits (blue). Bottom: corresponding non–robust SiZer map produced by the proposed SiZer with cutoff* $c = \infty$*.*

location x are given respectively by $\hat{m}_h(x) = \hat{a}_h$ and $\hat{m}'_h(x) = \hat{b}_h$, where

$$(\hat{a}_h, \hat{b}_h) = \arg\min_{a,b} \sum_{i=1}^{n} \left[Y_i - \{a + b(X_i - x)\}\right]^2 K_h(x - X_i). \tag{1}$$

In the above h is the bandwidth, K is the kernel function, and $K_h(x) = K(x/h)/h$. A Gaussian kernel is used throughout this article. Expressions for the asymptotic

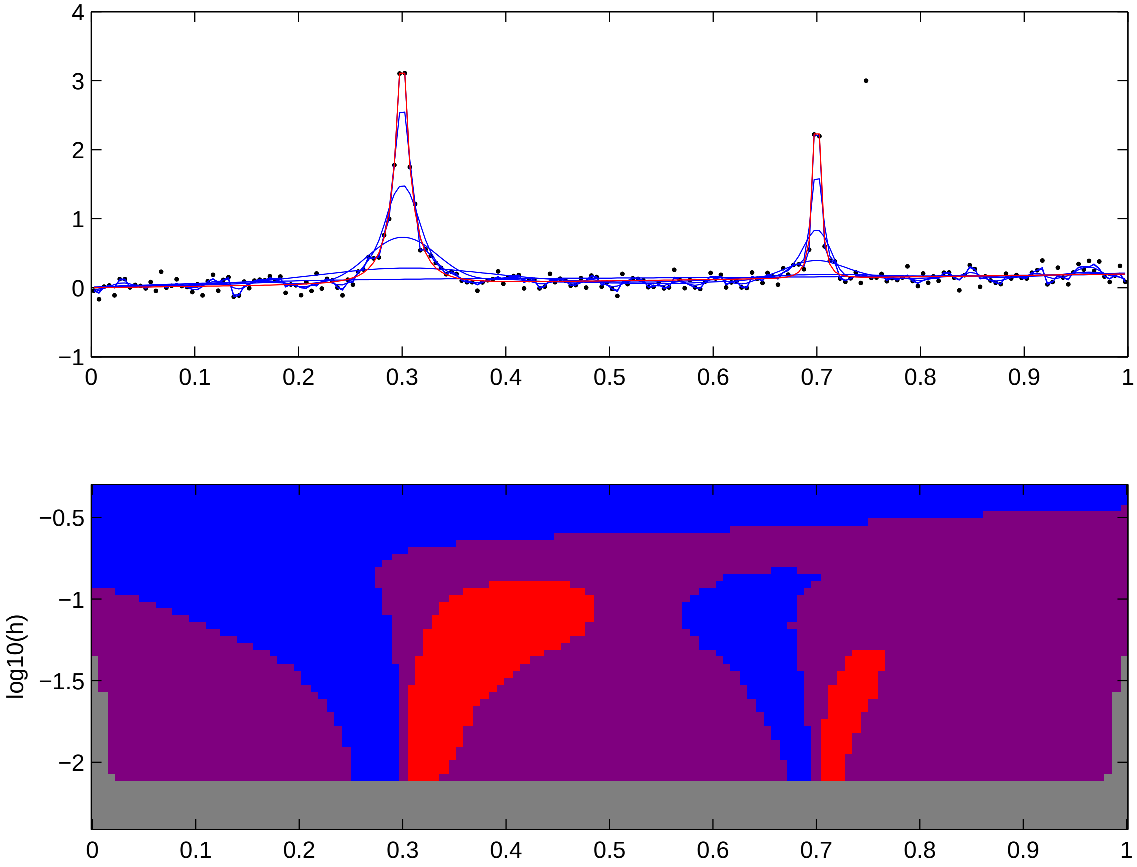

Figure 2: *Top: same noisy data (black) as in Figure 1, together with a family of robust local linear fits (blue). Bottom: SiZer map produced by the new robust SiZer with cutoff* $c = 1.345$.

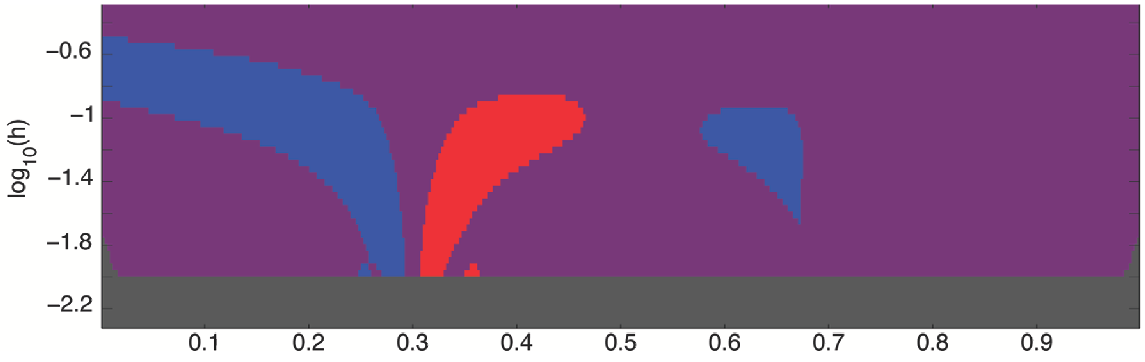

Figure 3: *SiZer map produced by the original SiZer of Chaudhuri & Marron (1999).*

variances for $\hat{m}_h(x)$ and $\hat{m}'_h(x)$ can be found, for examples, in Wand & Jones (1995) and Fan & Gijbels (1996). These expressions are required for the construction of a conventional SiZer map.

To construct a robust SiZer map, we need robust estimates for m and m', and we consider M–type local linear regression (e.g., Fan & Gijbels 1996, Section 5.5). Let $\hat{m}_{h,c}(x)$ and $\hat{m}'_{h,c}(x)$ be respectively the M–type robust estimates for $m(x)$ and $m'(x)$ with bandwidth h and cutoff c (see below). These estimates are defined as $\hat{m}_{h,c}(x) = \hat{a}_{h,c}$ and $\hat{m}'_{h,c}(x) = \hat{b}_{h,c}$, where now

$$(\hat{a}_{h,c}, \hat{b}_{h,c}) = \arg\min_{a,b} \sum_{i=1}^{n} \rho_c \left[\frac{Y_i - \{a + b(X_i - x)\}}{\hat{\sigma}} \right] K_h(x - X_i). \tag{2}$$

Here $\hat{\sigma}$ is a robust estimate of σ and $\rho_c(x)$ is the Huber loss function

$$\rho_c(x) = \begin{cases} x^2/2 & \text{if } |x| \le c, \\ \\ |x|c - c^2/2 & \text{if } |x| > c. \end{cases}$$

The cutoff $c > 0$ can be treated as a robustness parameter. Smaller values of c give more robust fits while larger values of c give less robust fits. In particular if $c \to \infty$ then $\hat{m}_{h,c} \to \hat{m}_h$ and $\hat{m}'_{h,c} \to \hat{m}'_h$. A typical choice for c is $c = 1.345$ (e.g., Huber 1981). For the proposed robust SiZer, c is treated in the same manner as h: a range of c values will be used. We will discuss the estimation of σ in Section 3.2. The estimates $\hat{m}_{h,c}$ and $\hat{m}'_{h,c}$ can be computed quickly using the method described in the appendix.

Besides the Huber loss function, the ideas presented above generalize straightforwardly to any other choice of loss function for M-estimation, but then the calculations presented in the appendix would need to be modified accordingly. We have chosen the Huber loss function for several reasons. First it is well-known and its properties are well studied. Secondly, it can be easily interpreted as an interpolation between L_1 and L_2 based inference. Moreover, we expect that, just as the case of choosing a kernel function in local linear smoothing, any reasonable choice of a loss function will lead to essentially the same results.

2.2 Asymptotic Variances for M–Type Estimates

To construct a robust SiZer map, we need to test if $\hat{m}'_{h,c}$ is significant for different combinations of h and c. Thus estimates for quantities like the variance of $\hat{m}'_{h,c}$ are required. This subsection provides convenient expressions for approximating these quantities. The following notation will be useful: $e_{i:p}$ is a p-dimensional column vector having 1 in the ith entry and zero elsewhere,

$$W = \mathrm{diag}\left\{K_h(X_i - x)\right\}, \quad \text{and} \quad X = \begin{pmatrix} 1 & \dots & 1 \\ X_1 - x & \dots & X_n - x \end{pmatrix}^T. \tag{3}$$

In the appendix the following approximation for the asymptotic variance of $\hat{m}_{h,c}$ is derived:

$$\mathrm{var}\{\hat{m}_{h,c}(x)\} \approx \sigma^2 e_{1:2}^T (X^T W X)^{-1} (X^T W^2 X)(X^T W X)^{-1} e_{1:2}\ r(c). \tag{4}$$

The corresponding expression for the asymptotic variance of $\hat{m}'_{h,c}(x)$ is essentially the same as (4) except that the two $e_{1:2}$'s are replaced by $e_{2:2}$. Note that these robust

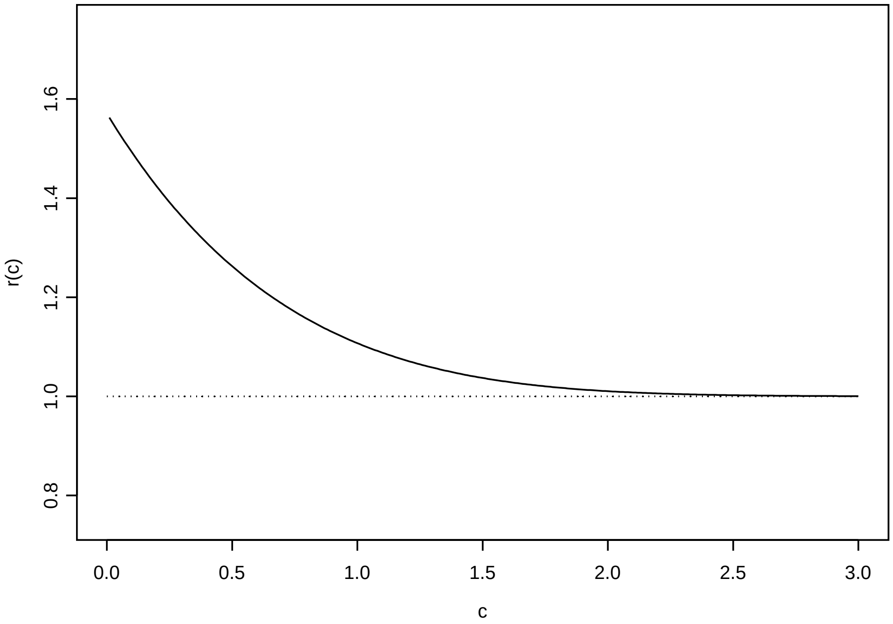

Figure 4: *Plot of* $r(c)$*. The horizontal dotted line is* $y = 1$*, the asymptote of* $r(c)$*.*

variance expressions (for $\hat{m}_{h,c}$ and $\hat{m}'_{h,c}$) only differ from the corresponding non–robust variance expressions (for $\hat{m}_h$ and $\hat{m}'_h$) by the quantity $r(c)$, which is derived to be

$$r(c) = \frac{c^2 - 2c\phi(c) - (c^2 - 1)\{2\Phi(c) - 1\}}{\{2\Phi(c) - 1\}^2}, \tag{5}$$

where $\phi(c)$ and $\Phi(c)$ are the density and the distribution function of the standard normal distribution respectively. A plot of $r(c)$ is given in Figure 4. Also notice that as $c \to \infty$, $r(c) \to 1$ and the robust variance expressions converge to the non–robust expressions.

Now the practical estimation of $\mathrm{var}\{\hat{m}_{h,c}(x)\}$ and $\mathrm{var}\{\hat{m}'_{h,c}(x)\}$ can be achieved by replacing σ^2 with a robust estimate $\hat{\sigma}^2$. We will discuss the choice of $\hat{\sigma}^2$ in Section 3.2.

2.3 Multiple Robust Slope Testing

For the construction of a SiZer map, every estimated slope $\hat{m}'_{h,c}(x)$ is classified into one of the following four groups: significantly increasing, significantly decreasing, not significant, and not enough data.

If an estimated slope is classified to the last group of not enough data, it means that the slope was estimated with too little data points and reliable hypothesis testing cannot be performed. This last group involves the concept of *effective sample size* (ESS). Our ESS definition is different from Chaudhuri & Marron (1999). Define $w_i(x)$ as the weight that the observation (X_i, Y_i) contributes to the non–robust local linear regression estimate $\hat{m}_h(x)$ for m at location x. That is, $\hat{m}_h(x) = \sum_{i=1}^{n} w_i(x) Y_i$ and $\sum w_i(x) = 1$. Exact expression for $w_i(x)$ is given for example by Equation (5.4) of Wand & Jones (1995). Then our ESS is defined as the number of elements in S, where S is the smallest subset of $[1, \ldots, n]$ such that $\sum_{i \in S} |w_i(x)| > 0.90$. Loosely, this ESS gives the smallest number of data points that constitutes 90% of the total weights. An estimated slope is classified to be not enough data if its ESS is less than or equal to 5. When comparing with the ESS definition of Chaudhuri & Marron (1999), we feel that ours is more natural, and agrees with the notion that ESS is the number of data points from which the estimate draws most of its information.

Now assume that the ESS of a $\hat{m}'_{h,c}(x)$ is large enough, and let $\hat{v}'_{h,c}(x)$ be an estimate of $\mathrm{var}\{\hat{m}'_{h,c}(x)\}$; i.e., expression (4) with σ^2 and $e_{1:2}$ replaced by $\hat{\sigma}^2$ and $e_{2:2}$ respectively. In the proposed robust SiZer the estimated slope $\hat{m}'_{h,c}(x)$ is declared to be significant if $|\hat{m}'_{h,c}(x)/\hat{v}'_{h,c}(x)| > C_R$, where C_R is the critical value. Since a large number of such statistical tests are to be conducted, one needs to perform multiple testing adjustment. We use the row–wise adjustment method proposed in Hannig & Marron (2004) to choose C_R. The method developed there is based on asymptotic consideration that are also valid in the present situation.

Let g be the number of pixels in a row in the SiZer map, Δ be the distance between neighboring locations at which the statistical tests are to be performed, and $\alpha = 0.05$ be the overall significance level of the tests. Hannig & Marron (2004) suggest the following value for C_R:

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

where

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

In Hannig & Marron (2004) the quantity $\theta(\Delta)$ is defined as the *clustering index* that measures the level of dependency between pixels.

To sum up, if the ESS of an estimated slope is less than or equal to 5, the corresponding pixel in the SiZer map will be colored grey. If the ESS is bigger than 5, then the corresponding pixel will be colored blue if the standardized slope $\hat{m}'_{h,c}(x)/\hat{v}'_{h,c}(x)$ is bigger than C_R, red if it is less than $-C_R$, and purple otherwise.

3 Outlier Identification

Barnett & Lewis (1978, page 4) define an outlier to be "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". They also state: "It is a matter of subjective judgment on the part of the observer whether or not he picks out some observation (or set of observations) for scrutiny". This agrees with the statement that the identification of outliers sometimes cannot be done by purely statistical techniques. Often, subjective decisions from experimental scientists are required. For alternative definitions of outliers, see Davies & Gather (1993) and references given therein.

The robust SiZer proposed above can be applied to help scientists to identify outliers. The general idea is as follows. First for all desired combinations of (h, c) we compute the *standardized residuals* (defined below). Then, for each pair of (h, c), apply a conventional outlier test to these standardized residuals to identify potential outliers. If any particular observation is classified as an outlier for most combinations of (h, c), then it is very likely that this observation is in fact an outlier. We illustrate this idea with the following example.

3.1 An Example

In the top panel of Figure 5 is a simulated noisy data set generated from the red regression function taken from Ruppert, Sheather & Wand (1995). As for Figure 1, the blue lines represent a family of estimated regression functions. In this data set, two outliers were artificially introduced, at $x = 0.25$ and $x = 0.75$. The bottom panel displays the corresponding SiZer map with a cutoff $c = \infty$; i.e., a non–robust map. In this map a new fifth color, black, is used to indicate the presence of probable outliers. For a given bandwidth h, if the result of an outlier test is significant when the test is applied to the observation (X_i, Y_i), then the pixel that is closest to (X_i, h) in the map will be colored black. The two long vertical black lines at $x = 0.25$ and $x = 0.75$ strongly suggest the presence of outliers at these two locations. To confirm this observation, four other SiZer maps were constructed using different cutoff values c. These SiZer maps are displayed in Figure 6. The same two vertical black lines remain in all these four maps. We have also computed other SiZer maps with other cutoffs. The results are summarized in a movie format, which can be downloaded from `http://www.stat.colostate.edu/∼tlee/robustsizer`.

There are other short black lines appearing in the lower part of these SiZer maps, which suggest that there is the possibility of other potential outliers. However, due to their short lengths and the fact that they do not appear in all the maps, these black lines are most likely caused by sampling noise rather than the presence of real outliers.

3.2 Variance Estimation and Outlier Testing

This subsection presents our method for estimating σ^2, and provides details of the outlier test used. We start by defining *standardized residuals*.

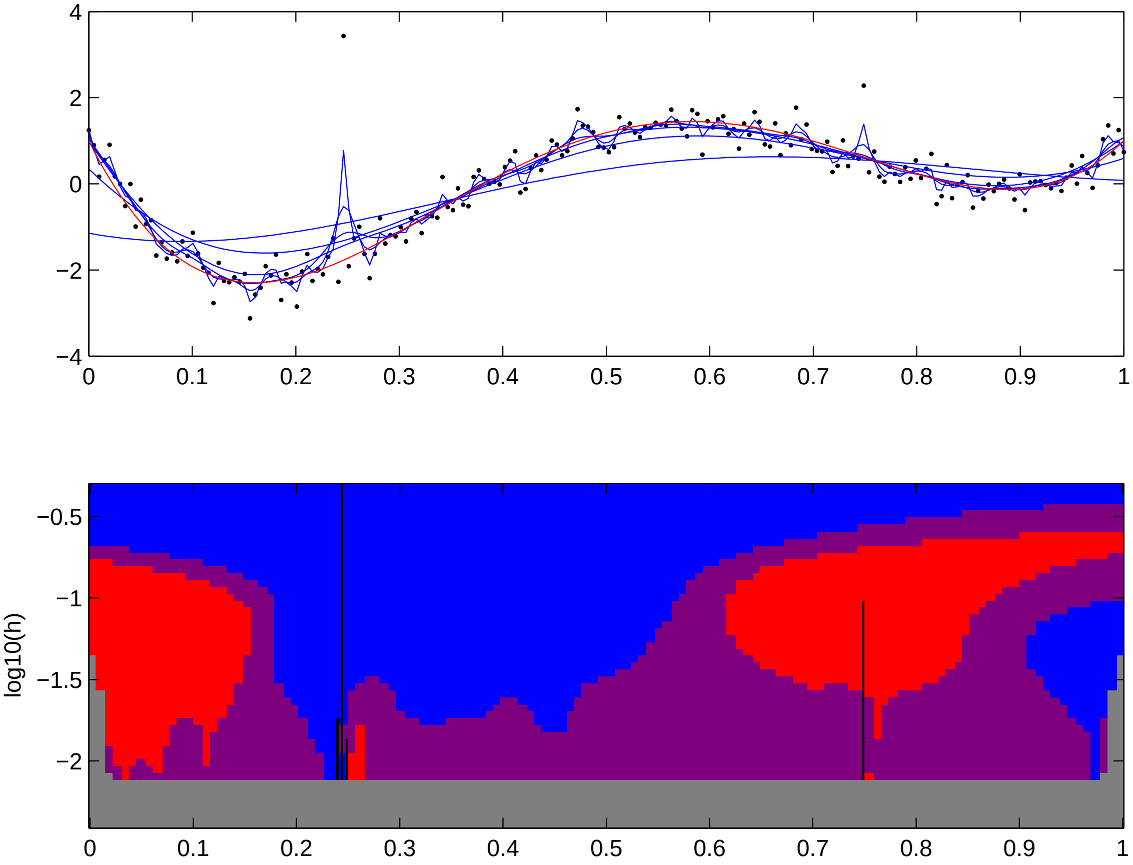

Figure 5: *Top: noisy data (black) generated by the red regression function, together with a family of local linear fits (blue). Bottom: corresponding non–robust SiZer map with cutoff* $c = \infty$.

In the appendix it is shown that the variance of the residuals $Y_i - \hat{m}_{h,c}(X_i)$ can be well approximated by

$$\mathrm{var}\{Y_i - \hat{m}_{h,c}(X_i)\} \approx \sigma^2 \left\{ 1 - 2w_i(X_i) + r(c) \sum_{j=1}^{n} w_j^2(X_i) \right\}, \qquad (6)$$

where the weights w_i were previously defined in Section 2.3. This motivates our

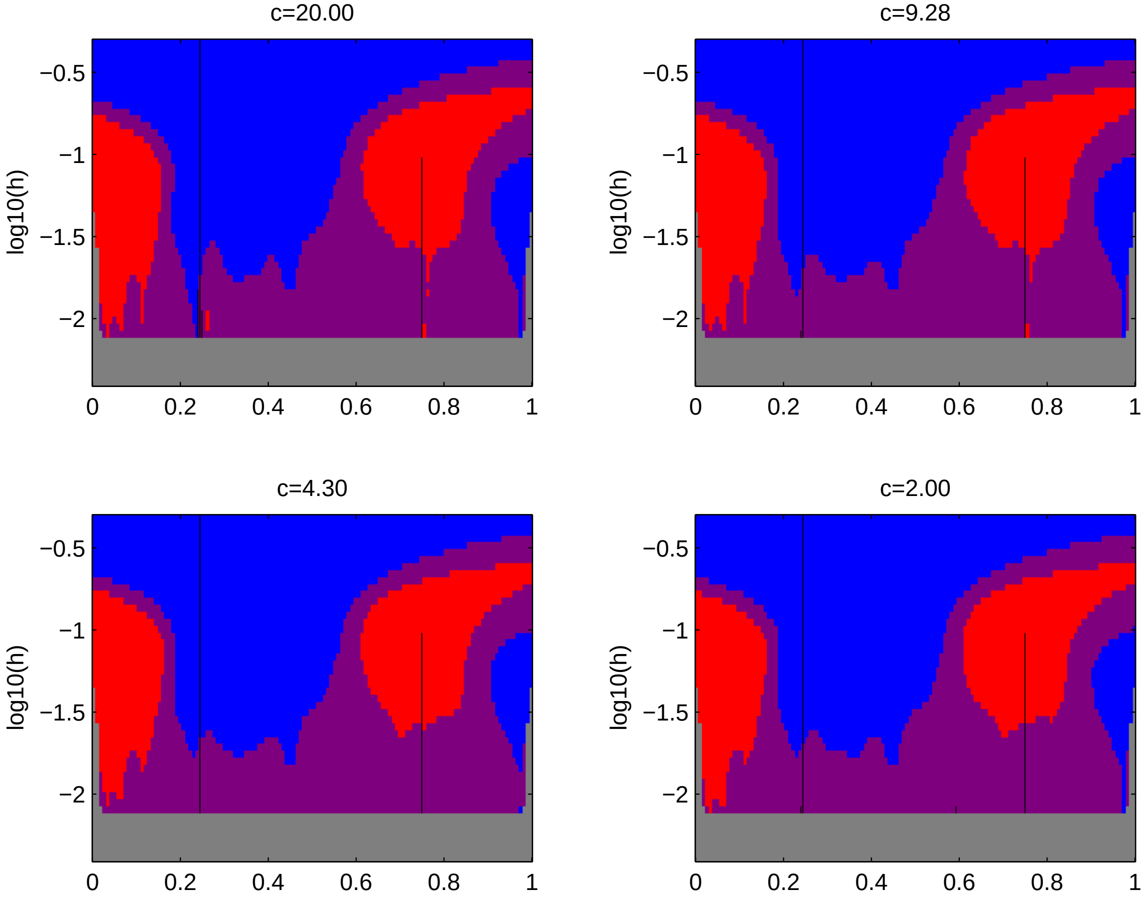

Figure 6: *Four robust SiZer maps obtained with different cutoff parameters. The actual cutoff parameters used are given at the top of each map.*

definition of *standardized residuals*:

$$\hat{\epsilon}_i = \frac{Y_i - \hat{m}_{h,c}(X_i)}{\left\{1 - 2w_i(X_i) + r(c)\sum_{j=1}^{n} w_j^2(X_i)\right\}^{1/2}}.$$

Throughout the whole article our robust estimate $\hat{\sigma}$ for σ is taken as the interquartile range of these standardized residuals divided by $2\Phi^{-1}(0.75)$. Of course both $\hat{\epsilon}_i$ and $\hat{\sigma}$ are functions of (h, c), but for simplicity we suppress this dependence in their notation.

For any given pair of (h, c), if the ESS of $\hat{m}_{h,c}(X_i)$ is greater than 5, then our robust SiZer flags (X_i, Y_i) as an outlier if $|\frac{\hat{\epsilon}_i}{\hat{\sigma}}| > t_{q,\nu}$, where $t_{q,\nu}$ is the q quantile of the t-distribution with ν degrees of freedom. Here q is set to $q = \frac{2\alpha}{n}$, where we choose $\alpha = 0.05$ and the divisor n is for the Bonferroni multiple testing adjustment. We define the degrees of freedom ν as the nearest integer to $n - \sum_{i=1}^{n} w_i(X_i)$. In the robust SiZer map the color black is used to indicate the presence of an outlier.

Following the ideas from Hannig & Marron (2004) we have also investigated other multiple outlier testing procedures that utilize the dependence structure of the residuals. In particular we have investigated an approximation to the distribution of the maximum of the residuals based on Rootzén (1983). We found that, due to the relatively high degree of independence amongst the residuals, this adjustment is almost identical to the Bonferroni adjustment. Figure 7 illustrates this finding. Displayed are the critical values for outlier testing obtained using both the Bonferroni (blue) and Rootzen's (red) multiple testing adjustments, plotted as a function of number of data points for a relatively small bandwidth. One can see that the two curves are almost on top of each other, suggesting that there is very little difference between the methods. Similar plots were also obtained for a wide variety of bandwidths. Therefore, we decided to use the relatively simpler Bonferonni multiple testing adjustment.

Now we are ready to provide an explanation of why the original SiZer of Chaudhuri & Marron (1999) failed to detect the bump located at $x = 0.7$ in the data set shown in Figure 1. In the original SiZer σ^2 is estimated locally, with a normalized "sum of squared residuals" type estimate. For x around 0.7, such an estimate of σ^2 was

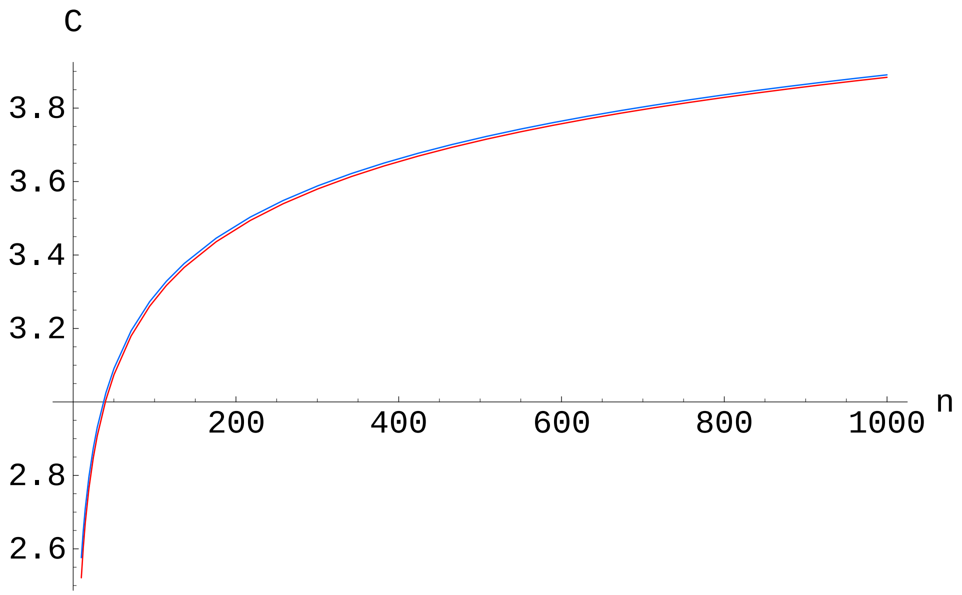

Figure 7: *Critical values for outlier testing obtained using both the Bonferroni (blue) and Rootzen's (red) multiple testing adjustments, plotted as a function of number of data points for a given bandwidth.*

badly inflated by the outlier at 0.75, which in turn deflated the test statistic. As a consequence, the hypothesis testing on the slopes was less likely to be significant, and hence missed the bump. On the other hand, for the proposed robust SiZer, σ^2 was estimated robustly and hence the effect of the outlier was minimized. Thus the bump at $x = 0.7$ was detected, even with $c = \infty$.

4 Further Examples

4.1 A Simulated Data Set with Multiple Outliers

It has been known that nonparametric curve estimates are most biased at bumps and valleys in the true regression function. Thus identifying outliers located in such regions is a challenging task. Another challenging task is the identification of multiple outliers that are clustered together. The following numerical experiment was performed to examine the effectiveness of the proposed robust SiZer under these two difficult situations. A simulated data set (of total 200 observations) was generated from a sine wave, where five outliers were added to a bump and another five outliers were introduced to a valley of the wave. This simulated data set is displayed in the top two panels of Figure 8. Two SiZer maps of this data set are also displayed in Figure 8, one with $c = \infty$ (i.e., non-robust) while the other with $c = 1.345$. When comparing these two SiZer maps, one can see that the robust map produces less spurious features, especially around $x = 0.63$ for small values of h, and it also better preserves the real features around $x = 0.4$ and $\log_{10} h = -1.5$. In addition, the robust SiZer correctly suggests the presence of the outliers.

4.2 Real Data: the Radar Glint Data Set

The proposed robust SiZer was also applied to the glint data set that was analyzed for example by Sardy, Tseng & Bruce (2001). This data set, displayed in Figure 9, are

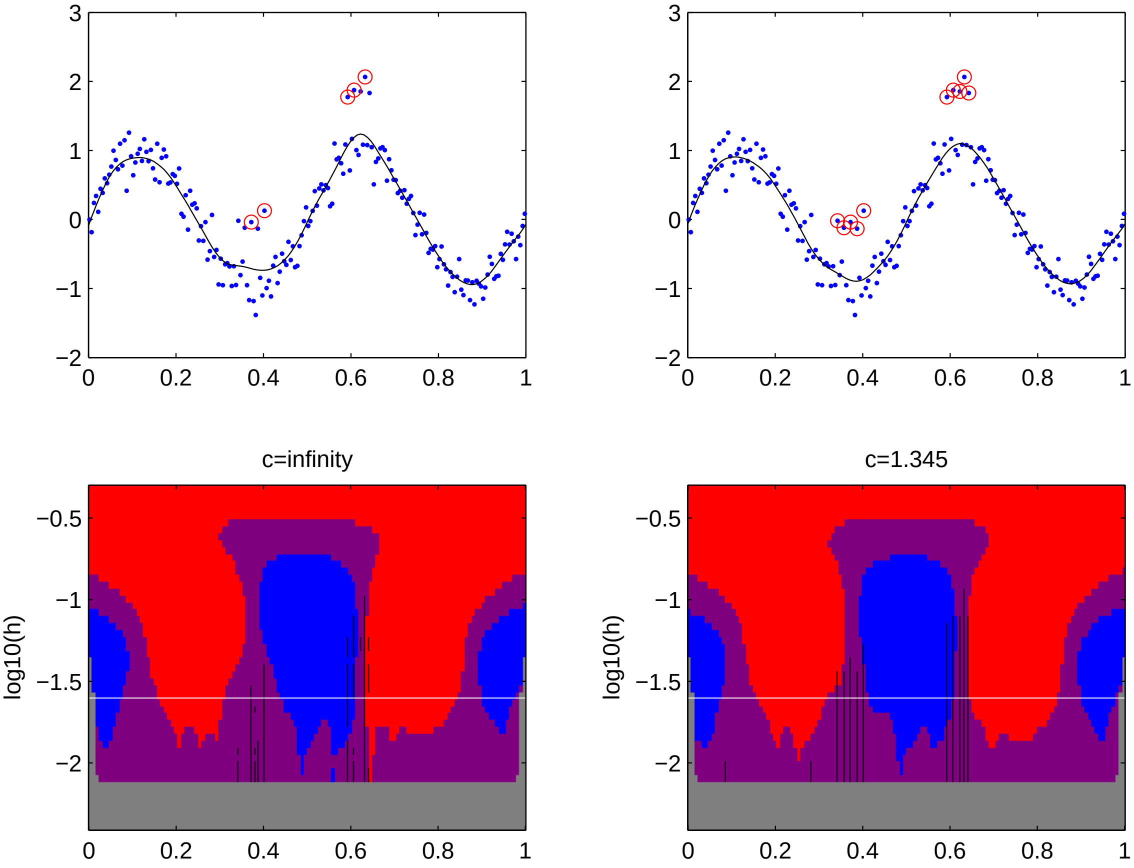

Figure 8: *Top two panels: a simulated data set with multiple outliers. The black line in the top-left panel is the curve estimate computed with* $c = \infty$ *and the bandwidth denoted by the while line in the SiZer map underneath. The black curve estimate in the top-right panel was computed with the same bandwidth but with* $c = 1.345$*. This bandwidth is chosen subjectively. Identified outliers are circled in red. The bottom panels display the corresponding SiZer maps computed with* $c = \infty$ *and* $c = 1.345$.

radar glint observations from a target captured at 512 angles. By visual inspection, one can see that there are some sharp features in the data, together with quite a few of potential outliers.

Five SiZer maps obtained with different cutoffs are also displayed in Figure 9. The outlier color, black, is not used in these maps. From these maps, one can conclude that there are two jumps in the data, located at around $x = 0.64$ and $x = 0.78$. There are also some fine structures present inside the range $(0.0, 0.5)$. These fine structures seem to be "real", as one needs to use a very small cutoff $c = 0.3$ to eliminate them.

For the purpose of outlier identification, Figure 10 displays a robust SiZer map obtained with $c = 1.345$ and uses the black outlier color. Also displayed are three curve estimates computed with three different bandwidths, with potential outliers highlighted for further inspection. We have chosen to display the estimates corresponding to different bandwidths separately rather than overlaying them in order to make the indication of potential outliers less cluttered. Similar plots with other cutoffs were also constructed. These results were again summarized in the form of movies, and can be downloaded in the same webpage previously listed in Section 3.1. Since there are many potential outliers, these movies provide a very useful visual summary for identifying them.

5 Conclusion

In this article a robust version of SiZer is proposed. One main feature of this robust SiZer is the use of M–type local smoothing. By varying the cutoff parameter of the M–type smoothing, various SiZer maps of various degrees of robustness can be produced. It is shown that with such a series of SiZer maps, structures hidden in a

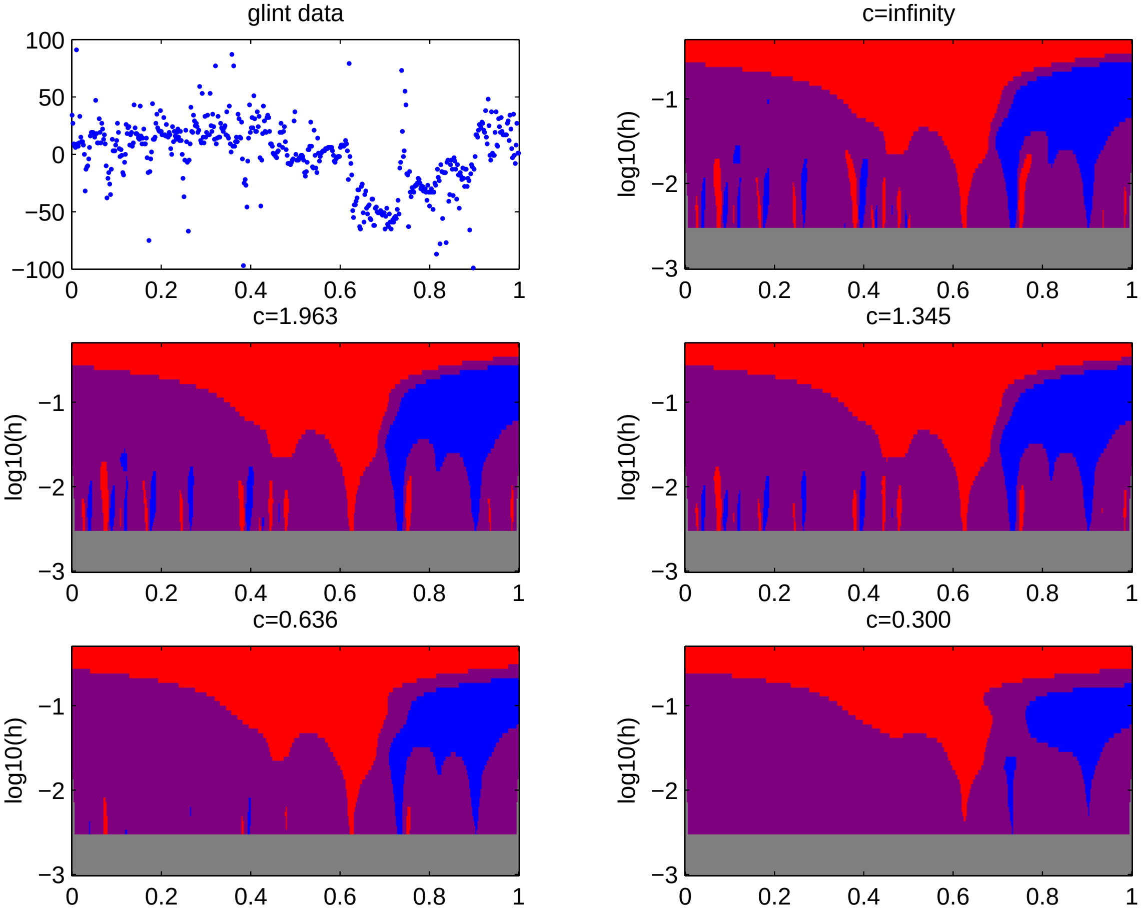

Figure 9: *Top-left panel: the glint data set. Remaining panels: five robust SiZer maps obtained with different cutoff parameters. The actual cutoff parameters used are given at the top of each map. The outlier color, black, is not used in these maps.*

data set can be more effectively revealed. It is also shown that the new robust SiZer can be applied to help identifying outliers.

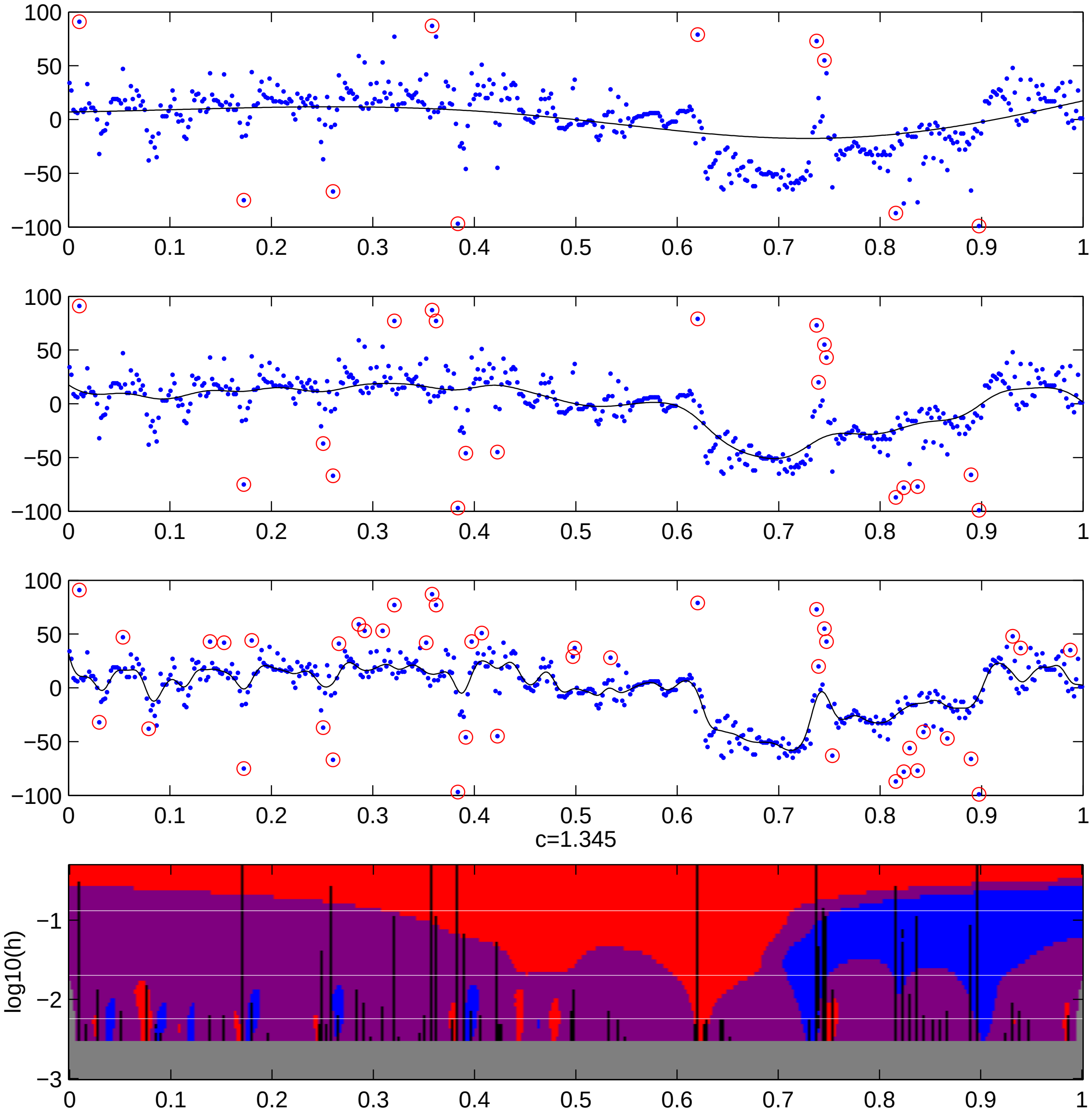

Figure 10: *First three panels: the glint data set (blue) with three robust curve estimates (black) computed with three different bandwidths. Identified outliers are circled in red. Bottom panel: corresponding robust SiZer map, in which the horizontal white lines indicate the bandwidths used to compute the robust curve estimates. The top, middle and bottom white lines correspond respectively to the curve estimates in the first, second and third panels.*

Acknowledgement

The authors would like to thank the referees for useful comments that lead to a much improved version of this article. The work of Hannig was supported in part by an IBM Faculty Award. The work of Lee was supported in part by U.S. National Science Foundation grants DMS-0203901.

6 Appendix

This appendix provides details behind our technical calculations and practical computations. In the subsequent derivations the robust estimates $\hat{m}_{h,c}$ and $\hat{m}'_{h,c}$ are denoted respectively as $\hat{m}^{(0)}_{h,c}$ and $\hat{m}^{(1)}_{h,c}$.

Derivation of (4): Our estimator of variance is based on the work of Welsh (1996). First we introduce some notation. The matrices N_p and T_p are both of size $(p+1)\times(p+1)$ with the (i,j)th element being $\int u_{i+j-2}K(u)\,du$ and $\int u_{i+j-2}K(u)^2\,du$ respectively. Since a Gaussian kernel is used these matrices can be calculated explicitly; e.g.,

$$N_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{and} \quad T_1 = \begin{pmatrix} \frac{1}{2\sqrt{\pi}} & 0 \\ 0 & \frac{1}{4\sqrt{\pi}} \end{pmatrix}.$$

Furthermore denote $\psi_c(x) = \rho'_c(x)$ and define $\chi_a(x) = 1 - 2[\Phi\{(x+d)/a\} - \Phi\{(x-d)/a\}]$. Here a is small positive number and d is chosen in such a way that $\int \chi_a(u)\,dF(u) = 0$, where $F(u)$ is the distribution function of the ϵ_i/σ. We use these functions to define

$$K = \begin{pmatrix} \sigma^{-2} \int \psi_c(u)^2 \, dF(u) & \sigma^{-1} \int \psi_c(u)\chi_a(u) \, dF(u) \\ \sigma^{-1} \int \psi_c(u)\chi_a(u) \, dF(u) & \int \chi_a(u)^2 \, dF(u) \end{pmatrix},$$

and

$$M = \begin{pmatrix} \sigma^{-2} \int \psi_c'(u) \, dF(u) & \sigma^{-1} \int u\psi_c'(u) \, dF(u) \\ \sigma^{-1} \int \chi_a(u)' \, dF(u) & \int u\chi_a'(u) \, dF(u) \end{pmatrix}.$$

Under some technical assumptions that are satisfied for our error distribution, Welsh (1996) shows that for a random design regression

$$\mathrm{var}\{\hat{m}_{h,c}^{(i)}(x)\} \approx n^{-1}h^{-2i-1} \, e_{1:2}^T M^{-1} K M^{-1} e_{1:2} \, g(x)^{-1} e_{(i+1):2}^T N_1^{-1} T_1 N_1^{-1} e_{(i+1):2}, \quad i = 1, 2, \tag{7}$$

where $g(x)$ is the density of the distribution of the design points and $e_{i:p}$ is a p-dimensional column vector of 0 with 1 on the ith position.

The proposed robust SiZer assumes that under the null hypothesis of no outliers the $F(u)$ is the standard normal distribution function. Thus we calculate that

$$K = \begin{pmatrix} \frac{c^2 - 2c\varphi(c) - (c^2-1)\{2\Phi(c)-1\}}{\sigma^2} & 0 \\ 0 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} \frac{2\Phi(c)-1}{\sigma^2} & 0 \\ 0 & \frac{4d\varphi\{d(1+a^2)^{-1/2}\}}{(1+a^2)^{3/2}} \end{pmatrix},$$

and by simplifying (7) we get, for $i = 1, 2$,

$$\mathrm{var}\{\hat{m}_{h,c}^{(i)}(x)\} \approx r(c)\sigma^2 \frac{(i!)^2 e_{(i+1):2}^T N_1^{-1} T_1 N_1^{-1} e_{(i+1):2}}{nh^{2i+1} g(x)}, \tag{8}$$

where the form of $r(c)$ is given in (5). It is worth pointing out that we can derive similar formulas using different choices of $F(u)$, ψ and ξ.

The formula (8) cannot be directly used as $g(x)$ is usually an unknown. To solve this problem consider the non-robust local polynomial regression; i.e., $c = \infty$. The variance of the non-robust estimator estimator is (e.g., formula (3.6) of Fan & Gijbels 1996)

$$\operatorname{var}\{\hat{m}^{(i)}_{h,\infty}(x)\} = \sigma^2 e^T_{(i+1):2}(X^TWX)^{-1}(X^TW^2X)(X^TWX)^{-1}e_{(i+1):2},$$

where X and W were defined in (3). Furthermore, Theorem 3.1 of Fan & Gijbels (1996) states that

$$\operatorname{var}\{\hat{m}^{(i)}_{h,\infty}(x)\} \approx \sigma^2 \frac{(i!)^2 e^T_{(i+1):2} N_1^{-1} T_1 N_1^{-1} e_{(i+1):2}}{nh^{2i+1} g(x)}. \tag{9}$$

Notice that $(X^TWX)^{-1}(X^TW^2X)(X^TWX)^{-1}$ depends only on the design points. Its asymptotic behavior is therefore not affected by the choice of c. Thus by comparing (8) and (9) we conclude (4).

Derivation of (6): Using the results, in particular Theorem 5.3, from Welsh (1989), we have $\hat{m}^{(0)}_{h,c}(x) - m(x) \approx b(x)$, where

$$b(x) = \frac{\sigma}{\int \psi_c'(u)\, dF(u)}\, e^T_{1:2}(X^TWX)^{-1}(X^TW)(\psi_c(\epsilon_1/\sigma), \ldots, \psi_c(\epsilon_n/\sigma))^T.$$

The variance of the regression residuals is

$$\operatorname{var}\{Y_i - \hat{m}^{(0)}_{h,c}(X_i)\} = \operatorname{var}(Y_i) - 2\operatorname{cov}\{Y_i, \hat{m}^{(0)}_{h,c}(X_i)\} + \operatorname{var}\{\hat{m}^{(0)}_{h,c}(X_i)\}.$$

The first term is σ^2 and the third term has been calculated before, so we calculate the second term

$$\begin{aligned}\operatorname{cov}\{Y_i, \hat{m}_{h,c}^{(0)}(X_i)\} &\approx \operatorname{cov}\{Y_i, b(X_i)\} \\ &= \sigma^2 e_{1:2}(X^TWX)^{-1}(X^TW)e_{i:n}\frac{\int u\psi_c(u)\,dF(u)}{\int \psi_c'(u)\,dF(u)} \qquad (10) \\ &= \sigma^2 e_{1:2}(X^TWX)^{-1}(X^TW)e_{i:n}.\end{aligned}$$

The last calculation follows from the fact that if $F(u)$ is the standard Gaussian distribution function then $\int u\psi_c(u)\,dF(u)/\int \psi_c'(u)\,dF(u) = 1$. Notice that the final result in (10) is the same as the covariance $\operatorname{cov}\{Y_i, \hat{m}_{h,\infty}^{(0)}(X_i)\}$ for a non-robust local linear estimator. Formula (6) now follows immediately by calculating $e_{1:2}(X^TWX)^{-1}(X^TW)e_{i:n} = w_i(X_i)$.

Computational Details: Here we provide details behind the practical implementation of the robust SiZer. First recall that the construction of a SiZer map for any given c requires the computation of the robust estimates $\hat{m}_{h,c}(x)$ and $\hat{m}_{h,c}'(x)$ in (2) for many different values of h. In our implementation the number of h we used was 50, equally–spaced in the log scale from $d/(2g)$ to $d/2$, where g is the number of pixels in a row in the SiZer map and $d = \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)$ is the range of the X_i's. We also used a fast iterative algorithm for computing $\hat{m}_{h,c}(x)$ and $\hat{m}_{h,c}'(x)$ for any given pair of h and c. This algorithm is similar to the one proposed in Lee & Oh (2004), and consists of the following steps:

1. Obtain an initial curve estimate $\hat{m}_{h,c}^{[0]}$ for m. This can be the solution to (1).

2. Set $\tilde{Y}_i^{[0]} = Y_i$ for $i = 1, \ldots, n$.

3. Iterate, until convergence, the following steps for $j = 0, 1, \ldots$:

   (a) Obtain a robust estimate $\hat{\sigma}^{[j+1]}$ of the noise standard deviation using the residuals $Y_i - \hat{m}_{h,c}^{[j]}(X_i)$, $i = 1, \ldots, n$. In our implementation we use 1.4826 times the median absolute deviation of these residuals.

   (b) For $i = 1, \ldots, n$, compute

   $$\tilde{Y}_i^{[j+1]} = \hat{m}_{h,c}^{[j]}(X_i) + \frac{\hat{\sigma}^{[j+1]}}{2}\psi_c\left(\frac{\tilde{Y}_i^{[j]} - \hat{m}_{h,c}^{[j]}(X_i)}{\hat{\sigma}^{[j+1]}}\right),$$

   where the function ψ_c is the derivative of ρ_c.

   (c) Calculate the $(j+1)$th iterative estimates $\hat{m}_{h,c}^{[j+1]}(x)$ and $\hat{m}_{h,c}^{\prime[j+1]}(x)$ as

   $$\left\{\hat{m}_{h,c}^{[j+1]}(x), \hat{m}_{h,c}^{\prime[j+1]}(x)\right\} = \arg\min_{a,b} \sum_{i=1}^{n} \left[\tilde{Y}_i - \{a + b(X_i - x)\}\right]^2 K_h(x - X_i). \tag{11}$$

4. Take the converged estimates $\hat{m}_{h,c}^{[\infty]}$ and $\hat{m}_{h,c}^{\prime[\infty]}$ as our final robust estimates for m and m' respectively.

Notice that this algorithm replaces the hard minimization problem in (2) with a series of quick least-squares type minimizations (11). Also, in practice this algorithm converges very quickly.

References

Barnett, V. & Lewis, T. (1978), *Outliers in Statistical Data*, John Wiley & Sons, Chichester.

Chaudhuri, P. & Marron, J. S. (1999), 'SiZer for exploration of structures in curves', *Journal of the American Statistical Association* **94**, 807–823.

Chaudhuri, P. & Marron, J. S. (2000), 'Scale space view of curve estimation', *The Annals of Statistics* **28**, 408–428.

Davies, L. & Gather, U. (1993), 'The identification of multiple outliers (with discussion)', *Journal of the American Statistical Association* **88**, 782–801.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.

Erästö, P. & Holmström, L. (2004), 'Bayesian multiscale smoothing for making inferences about features in scatter plots', *Journal of Computational and Graphical Statistics.* to appear.

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

Godtliebsen, F. & Oigard, T. A. (2005), 'A visual display device for significant features in complicated signals', *Computational Statistics and Data Analysis* **48**, 317–343.

Hannig, J. & Marron, J. S. (2004), ‘Advanced Distribution Theory for SiZer’. Unpublished manuscript.

Huber, P. J. (1981), *Robust Statistics*, John Wiley & Sons, New York.

Lee, T. C. M. & Oh, H.-S. (2004), ‘Fast computation of robust m-type penalized regression splines’. Unpublished manuscript.

Rootzén, H. (1983), ‘The rate of convergence of extremes of stationary normal sequences’, *Advances in Applied Probability* **15**(1), 54–80.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), ‘An effective bandwidth selector for local least squares regression’, *Journal of the American Statistical Association* **90**, 1257–1270.

Sardy, S., Tseng, P. & Bruce, A. (2001), ‘Robust wavelet denoising’, *IEEE Transactions on Signal Processing* **49**, 1146–1152.

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

Welsh, A. H. (1989), ‘On M-processes and M-estimation’, *The Annals of Statistics* **17**, 337–361.

Welsh, A. H. (1990), ‘Correction: “On M-processes and M-estimation”’, *The Annals of Statistics* **18**, 1500.

Welsh, A. H. (1996), ‘Robust estimation of smooth regression and spread functions and their derivatives’, *Statistica Sinica* **6**, 347–366.