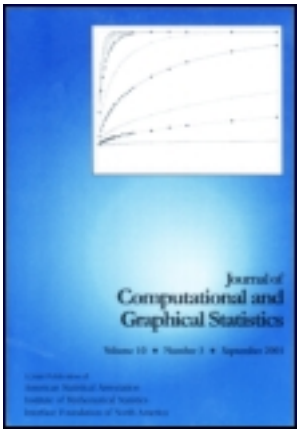

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

Multiscale Exploratory Analysis of Regression Quantiles Using Quantile SiZer

Cheolwoo Park, Thomas C. M. Lee and Jan Hannig

Cheolwoo Park is Assistant Professor, Department of Statistics, University of Georgia, Athens, GA 30602-1952 . Thomas C. M. Lee is Professor, Department of Statistics, University of California at Davis, Davis, CA 95616 . Jan Hannig is Associate Professor, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260 .

Available online: 01 Jan 2012

To cite this article: Cheolwoo Park, Thomas C. M. Lee and Jan Hannig (2010): Multiscale Exploratory Analysis of Regression Quantiles Using Quantile SiZer, Journal of Computational and Graphical Statistics, 19:3, 497-513

To link to this article: http://dx.doi.org/10.1198/jcgs.2010.09120

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Multiscale Exploratory Analysis of Regression Quantiles Using Quantile SiZer

Cheolwoo PARK, Thomas C. M. LEE, and Jan HANNIG

The SiZer methodology proposed by Chaudhuri and Marron (1999) is a valuable tool for conducting exploratory data analysis. Since its inception different versions of SiZer have been proposed in the literature. Most of these SiZer variants are targeting the mean structure of the data, and are incapable of providing any information about the quantile composition of the data. To fill this need, this article proposes a quantile version of SiZer for the regression setting. By inspecting the SiZer maps produced by this new SiZer, real quantile structures hidden in a dataset can be more effectively revealed, while at the same time spurious features can be filtered out. The utility of this quantile SiZer is illustrated via applications to both real data and simulated examples. This article has supplementary material online.

Key Words: Effective sample size; Multiple slope testing; Nonparametric quantile regression; Robust variance estimation; Running regression quantile; SiZer.

1. INTRODUCTION

Given a sequence of realizations of the random variables (X, Y), quantile regression aims to estimate the conditional αth quantile of the response Y given the covariate X. By changing the value of $0 < \alpha < 1$, quantile regression can be applied to explore the different behaviors of the conditional distribution $f_{Y|X}$ of Y given X at its center, lower, and upper tails. Therefore quantile regression is capable of providing more information about $f_{Y|X}$ than the usual regression which only targets the conditional expectation $E(Y|X)$. Many useful nonparametric methods for conducting quantile regression analysis have been proposed in the literature. For example, Lejeune and Sarda (1988) proposed the use of moving parabolic fitting, Bhattacharya and Gangopadhyay (1990), Truong (1989), and Yu (1999) investigated kernel-type methods, and Chaudhuri and Loh (2002) considered

Cheolwoo Park is Assistant Professor, Department of Statistics, University of Georgia, Athens, GA 30602-1952 (E-mail: *cpark@uga.edu*). Thomas C. M. Lee is Professor, Department of Statistics, University of California at Davis, Davis, CA 95616 (E-mail: *tcmlee@ucdavis.edu*). Jan Hannig is Associate Professor, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3260 (E-mail: *hannig@email.unc.edu*).

Journal of Computational and Graphical Statistics, Volume 19, Number 3, Pages 497–513
DOI: 10.1198/jcgs.2010.09120

the use of regression trees. In addition, local polynomial regression has been studied by Chaudhuri (1991a, 1991b) and Su and Ullah (2008), while spline methods were considered by Koenker and Bassett (1978), Hendricks and Koenker (1992), Koenker, Ng, and Portnoy (1994), Takeuchi et al. (2006), and Yuan (2006). Typically these methods produce a nonparametric estimate of the quantile of the conditional distribution of Y given X in the form of a smooth curve. Such a nonparametric quantile curve can then be plotted for visual inspection, say for checking for the existence of any local maximum. However, although this curve can suggest the existence or even the locations of possible local maxima, it does not carry the necessary information to help judge whether the local maxima are statistically significant features, or are just spurious structures caused by sampling errors.

In the nonparametric regression context, the SiZer methodology of Chaudhuri and Marron (1999, 2000) is a powerful multiscale tool for handling this issue. 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, through the construction of a so-called *SiZer map*. In short, a SiZer map is a two-dimensional (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 a more detailed description and a thorough survey of the SiZer methodology, see Section 2.

To date all SiZer related methods proposed in the literature target the "mean structure" of the data. The main contribution of this article is the proposal of a new SiZer that targets the "quantile structure" of the data. This new SiZer is termed *quantile SiZer*, and as compared to the original SiZer of Chaudhuri and Marron (1999), major modifications made by it include the use of a local linear quantile smoother and a robust estimator for the noise variance function. With these modifications the proposed quantile SiZer is capable of producing new *quantile SiZer maps* that target different quantiles of the distribution. By inspecting an array of such SiZer maps, real structures hidden in both the mean and variance of the data can be more easily detected, while at the same time spurious features can be filtered out in an effective manner. For example, by observing the quantile SiZer maps for both the 0.1th and 0.9th quantiles, information about the "spread" of the data can be identified.

The proposed quantile SiZer also eliminates the need for the choice of a "best" smoothing parameter which is required in most of the aforementioned nonparametric conditional quantile estimation methods. Unlike nonparametric conditional mean estimation, fast and reliable automatic methods for smoothing parameter selection for nonparametric quantile regression are still largely missing. Hence, a single reliable nonparametric estimate of a conditional quantile curve cannot be easily obtained. The SiZer's philosophy of using multiple bandwidths rather than one, therefore, is particularly appealing.

The rest of the article is organized as follows. It starts with Section 2 in which a brief review of the conventional SiZer of Chaudhuri and Marron (1999) is provided. Then the proposed quantile SiZer is presented in Section 3. In Section 4 the utility of the proposed

quantile SiZer is illustrated with several simulated and real datasets. Concluding remarks are offered in Section 5, while technical and computational details are provided as Supplemental Materials.

2. BACKGROUND

2.1 CONVENTIONAL SIZER

As mentioned before, the conventional SiZer is a multiscale methodology for helping the data analyst to detect real structures hidden in the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Suppose for the moment these data satisfy the usual constant noise variance model:

$$Y_i = m(X_i) + e_i, \qquad e_i \sim \text{iid } N(0, \sigma^2).$$

The SiZer methodology begins with nonparametrically estimating the conditional mean $m(X)$ and slope $m'(X)$ of the data at different scales (or resolutions). Chaudhuri and Marron (1999) achieved this by applying local linear regression to the data with different bandwidths. Then statistical hypothesis tests are applied to all the estimated slopes (with different values of X and bandwidths), to test if they are statistically significant. Results of such tests are summarized in a SiZer map, providing a visual device for user inspection.

To understand how the test results are encoded in a SiZer map, consider Figure 1. The bottom panel is a simulated noisy dataset generated from the regression function displayed as the thick solid line. Also displayed as thin solid lines are a set of estimated regression functions obtained by local linear regression with different bandwidths. The top panel displays the conventional SiZer map that corresponds to this dataset. 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 thin smoothed curves. These bandwidths are displayed on the log scale, with the smallest bandwidth at the bottom. The gray-level 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 gray-levels: darkest and lightest indicate respectively the estimated slope is significantly increasing and decreasing, second lightest indicates the slope is not significant, while second darkest shows that there are not enough data for conducting reliable statistical inference.

The top portion of the SiZer map in Figure 1 is completely darkest, indicating that the underlying regression function has an overall increasing trend. In the middle part of the SiZer map, the gray-level switches from lightest to darkest to lightest and to darkest again. This suggests that the underlying function, when observed at a medium resolution level, first decreases, then increases, then decreases, and finally increases again. Also, those locations in the SiZer map at which the gray-level changes correspond to the inflection points of the regression function. Once the user is familiar with the SiZer map gray-level labeling scheme, a lot of useful information about the *mean structure* of the regression function can be extracted. As mentioned earlier, the proposed quantile SiZer aims to produce SiZer maps that are capable of revealing important information about the *quantile structure*.

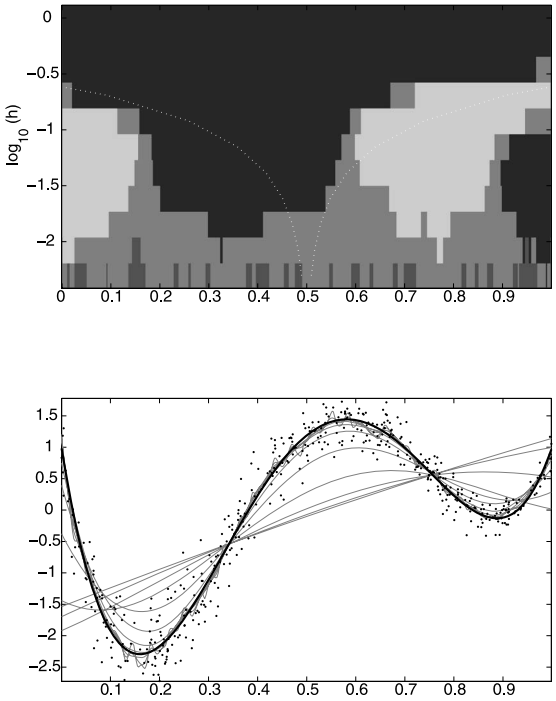

Figure 1. Top: Conventional SiZer map of Chaudhuri and Marron (1999) that corresponds to the dataset displayed in the bottom panel. Bottom: Noisy data generated by the regression function $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ (thick line), together with a family of local linear fits (thin lines).

2.2 OTHER SIZERS

It has been a decade since the above conventional SiZer was made. During this period different versions and improvements of SiZer have been developed. On the theoretical side an important contribution is the work of Hannig and Marron (2006), in which a new inference method for reducing spurious pixels in the SiZer map was developed. This new inference method has been successfully adopted by other SiZer tools, such as the *robust SiZer* of Hannig and Lee (2006). In the regression setting, this robust version can be used for identifying outliers, and it is also capable of producing SiZer maps with different degrees of robustness.

Park, Marron, and Rondonotti (2004) considered dependent observations. Since then other SiZer tools that target time series data have been proposed, including those of Rondonotti, Marron, and Park (2007), Olsen, Sorbye, and Godtliebsen (2008), and Park, Hannig, and Kang (2009). For multivariate problems, Godtliebsen, Marron, and Chaudhuri (2002) and Ganguli and Wand (2004) considered, respectively, bivariate density estimation and bivariate smoothing, while Ganguli and Wand (2007) and Gonzalez-Manteiga, Martinez-Miranda, and Raya-Miranda (2008) studied generalized additive models. In addition, various Bayesian versions of SiZer have also been proposed, by, for example, Erästö and Holmström (2005), Godtliebsen and Oigard (2005), Oigard, Rue, and Godtliebsen

(2006), and Erästö and Holmström (2007); the last two articles also considered time series data.

There are also other SiZer tools that were developed for other more specialized tasks. For example, hazard rate estimation was studied by Marron and de Una-Alvarez (2004), change point and jump detection was considered by Kim and Marron (2006) and Olsen, Chaudhuri, and Godtliebsen (2008), and curve comparison was investigated by Park and Kang (2008). Moreover, the multiscale visualization idea behind SiZer has gained popularity in network traffic modeling; for example, see the work of Rolls, Michailidis, and Hernández-Campos (2005) and Park et al. (2005). Lastly, other contributions to the SiZer family include the smoothing spline SiZer of Marron and Zhang (2005) and the wavelet SiZer of Park et al. (2007).

The aforementioned work represents a large library of SiZers and related tools developed for multiscale analysis of different estimation problems. However, these tools mostly focus on the "mean" structures of the data. Therefore the proposal of the quantile SiZer of the current article fills an important gap in the SiZer methodology.

3. METHODOLOGY

3.1 NONPARAMETRIC QUANTILE REGRESSION

First we follow Chaudhuri and 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

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \tag{3.1}$$

where m is the regression function and the ϵ_i's are zero-mean independent noise with common variance 1. It is known as the location-scale model in quantile regression (e.g., Koenker 2005). Its limitations and the possibility of generalizing it to other more general models are discussed in Section 5.

Let h be a bandwidth, K be a kernel function, and write $K_h(x) = K(x/h)/h$. Throughout this article a Gaussian kernel is used. The local linear regression estimates for $m(x)$ and $m'(x)$ at location x are given respectively by $\hat{m}_h(x) = \hat{a}_h$ and $\hat{m}'_h(x) = \hat{b}_h$, where

$$(\hat{a}_h, \hat{b}_h) = \arg\min_{a,b} \sum_{i=1}^{n} \big[Y_i - \{a + b(X_i - x)\}\big]^2 K_h(x - X_i). \tag{3.2}$$

Notice that both $\hat{a}_h$ and $\hat{b}_h$ are functions of x, but for simplicity, this dependence is suppressed in their notation. Expressions for the asymptotic variances for $\hat{m}_h(x)$ and $\hat{m}'_h(x)$ can be found, for example, in the books by Wand and Jones (1995) and Fan and Gijbels (1996). These expressions are required for the construction of a conventional SiZer map.

To construct a quantile SiZer map, we need a corresponding nonparametric running regression quantile estimator (e.g., Koenker 2005). A natural estimator can be obtained by modifying the L_2 loss function in (3.2) as follows. Denote our estimates for the conditional

αth quantile of Y given X and its derivative as $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$, respectively. These estimates are defined as $\hat{\beta}_{h,\alpha}(x) = \hat{a}_{h,\alpha}$ and $\hat{\beta}'_{h,\alpha}(x) = \hat{b}_{h,\alpha}$, where now

$$(\hat{a}_{h,\alpha}, \hat{b}_{h,\alpha}) = \arg\min_{a,b} \sum_{i=1}^{n} \rho_\alpha\big[Y_i - \{a + b(X_i - x)\}\big] K_h(x - X_i) \tag{3.3}$$

with $\rho_\alpha(x)$ as the so-called check loss function

$$\rho_\alpha(x) = \begin{cases} x\alpha, & \text{if } x \geq 0 \\ x(\alpha - 1), & \text{if } x < 0. \end{cases}$$

Similarly to $\hat{a}_h$ and $\hat{b}_h$, the dependence of $\hat{a}_{h,\alpha}$ and $\hat{b}_{h,\alpha}$ on x is suppressed in the notation.

The estimator defined in (3.3) has been studied, for example, by Chaudhuri (1991a, 1991b) and Yu and Jones (1998). In all our numerical work, the estimates $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$ are computed using the fast method proposed by Oh, Lee, and Nychka (2010). For completeness, this method is outlined in the Supplemental Materials.

3.2 Asymptotic Variances for Running Regression Quantile Estimates

A quantile SiZer map is a graphical device summarizing if $\hat{\beta}'_{h,\alpha}(x)$ is statistically significant for different values of h and x. Thus to construct such a map, estimates for quantities like the variance of $\hat{\beta}'_{h,\alpha}(x)$ are required. This subsection provides convenient expressions for approximating these quantities. The following notation will be useful: $\mathbf{e}_{i:p}$ is a p-dimensional column vector having 1 in the ith entry and zero elsewhere,

$$\mathbf{W} = \operatorname{diag}\{K_h(X_i - x)\} \qquad \text{and} \qquad \mathbf{X} = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 - x & \cdots & X_n - x \end{pmatrix}^T. \tag{3.4}$$

In the Supplemental Materials the following approximations for the asymptotic expectation and variance of $\hat{\beta}_{h,\alpha}(x)$ are derived:

$$\begin{aligned} E\{\hat{\beta}_{h,\alpha}(x)\} &\approx m(x) + \sigma(x)\Phi^{-1}(\alpha), \\ \operatorname{var}\{\hat{\beta}_{h,\alpha}(x)\} &\approx \sigma^2(x) e_{1:2}^T (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} e_{1:2} r(\alpha), \end{aligned} \tag{3.5}$$

where the quantity $r(\alpha)$ is defined in (3.7) below. Similarly the corresponding expressions for $\hat{\beta}'_{h,\alpha}(x)$ are

$$\begin{aligned} E\{\hat{\beta}'_{h,\alpha}(x)\} &\approx m'(x) + \sigma'(x)\Phi^{-1}(\alpha), \\ \operatorname{var}\{\hat{\beta}'_{h,\alpha}(x)\} &\approx \sigma^2(x) e_{2:2}^T (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} e_{2:2} r(\alpha). \end{aligned} \tag{3.6}$$

Note that these variance expressions (for $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$) only differ from the corresponding usual local linear variance expressions (for $\hat{m}_h(x)$ and $\hat{m}'_h(x)$) by the quantity $r(\alpha)$, which is derived to be

$$r(\alpha) = \alpha(1 - \alpha)\phi(\Phi^{-1}(\alpha))^{-2}, \tag{3.7}$$

where $\phi(x)$ and $\Phi(x)$ are the density and the distribution function of the standard normal distribution, respectively. A plot of $r(\alpha)$ is given in Figure 2. Notice that the minimum

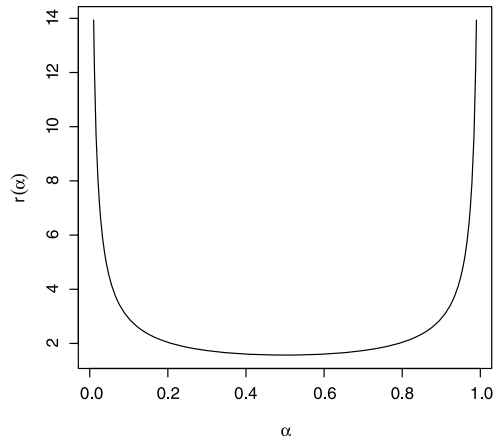

Figure 2. Plot of $r(\alpha)$ for $0.01 \le \alpha \le 0.99$.

value of $r(\alpha)$ is 1.571 (occurs when $\alpha = 0.5$), which agrees with the general fact that standard errors for "quantile-based estimation" are larger than those for "mean-based estimation." Therefore, when comparing to the conventional "mean" SiZer, the quantile SiZer will detect less significant features if the same level of significance is used.

Since the only unknown in (3.5) and (3.6) is $\sigma^2(x)$, the practical estimation of $\text{var}\{\hat{\beta}_{h,\alpha}(x)\}$ and $\text{var}\{\hat{\beta}'_{h,\alpha}(x)\}$ can be achieved by replacing $\sigma^2(x)$ with an appropriate robust estimate $\hat{\sigma}^2(x)$. Such an estimate can be obtained in several ways. In this article, we first calculate the scaled differenced series

$$e_i = \frac{\sqrt{\pi}}{2}(Y_i - Y_{i-1}), \qquad i = 2, \ldots, n,$$

and then apply local linear smoothing to the absolute values of these scaled differences, using the same bandwidth as we calculate the quantile estimate (3.3). We have investigated other options such as applying local median smoothing or using standardized residuals, but local linear smoothing on these scaled differences gave the most satisfactory results.

3.3 EFFECTIVE SAMPLE SIZE AND MULTIPLE SLOPE TESTING

For the construction of a SiZer map, every estimated slope $\hat{\beta}'_{h,\alpha}(x)$ 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 few data points and reliable hypothesis testing conclusion cannot be obtained.

This last group involves the concept of *effective sample size* (ESS) defined by Chaudhuri and Marron (1999). There are different definitions of ESS and we define ESS as in the article by Hannig and Lee (2006). Denote $w_i(x)$ as the weight that the observation (X_i, Y_i) contributes to the local linear regression estimate $\hat{m}_h(x)$ (3.2) for m at location x. That is, $\hat{m}_h(x) = \sum_{i=1}^n w_i(x) Y_i$ and $\sum w_i(x) = 1$. An exact expression for $w_i(x)$ can be found, for example, in equation (5.4) of Wand and 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. In this article an estimated slope is classified to be not enough data if its ESS is less than or equal to 5.

Now assume that the ESS of a $\hat{\beta}'_{h,\alpha}(x)$ is large enough, and let $\hat{v}'_{h,\alpha}(x)$ be an estimate of $\text{var}\{\hat{\beta}'_{h,\alpha}(x)\}$, that is, expression (3.6) with $\sigma^2(x)$ replaced by $\hat{\sigma}^2(x)$. In the proposed quantile SiZer the estimated slope $\hat{\beta}'_{h,\alpha}(x)$ is declared to be significant if $|\hat{\beta}'_{h,\alpha}(x)/\hat{v}'_{h,\alpha}(x)| > C_R$, where C_R is the critical value. Since a large number of such statistical tests are to be conducted, multiple testing adjustment is required. We use the following row-wise adjustment method proposed by Hannig and Marron (2006) to choose C_R. The method developed by them is based on asymptotic considerations 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 two successive neighboring locations at which the statistical tests are to be performed, and $\alpha = 0.05$ be the overall significance level of the tests. Hannig and Marron (2006) suggested 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 the article by Hannig and Marron (2006) the quantity $\theta(\Delta)$ is defined as the *clustering index* that measures the level of dependency between pixels. This adjusts the SiZer map to have a correct family-wise error rate in each row of each picture. One could also adjust family-wise error rate of a whole picture or even a sequence of plots. However, our experience suggests that doing so would lead to a significant loss of power.

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 have the second darkest gray-level. If the ESS is bigger than 5, then the corresponding pixel gray-level will be darkest if the standardized slope $\hat{\beta}'_{h,\alpha}(x)/\hat{v}'_{h,\alpha}(x)$ is bigger than C_R, lightest if it is less than $-C_R$, and second lightest otherwise.

4. PRACTICAL PERFORMANCE

In this section the utility of the quantile SiZer methodology is illustrated with simulated data, as well as three real datasets.

4.1 SIMULATED DATA

We tested our quantile SiZer with two mean functions and three noise variance functions. The two mean functions are

$$m_1(x) = 0 \qquad \text{and} \qquad m_2(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4,$$

all with domain [0, 1]. The first mean function is a zero constant function and it will be useful for testing if the quantile SiZer is capable of picking up quantile structures that are due to nonconstant noise variance. The second function, which is the same as the true function used in Figure 1, contains a good mix of different degrees of curvatures. It was first used by Ruppert, Sheather, and Wand (1995) for evaluating bandwidth selection methods in the context of local linear regression, and by Hannig and Lee (2006) for testing SiZer map construction. The three variance functions are

$$\sigma_1(x) = 1, \qquad \sigma_2(x) = 2.5 - 2x, \qquad \text{and} \qquad \sigma_3(x) = -4x^2 + 4x + 0.5.$$

The first one is constant throughout the whole domain, the second decreases linearly, while the last one is quadratic.

From each of the six combinations of mean function and variance function, an artificial dataset of size $n = 512$ was first simulated, where the design points x were generated from unif [0, 1]. Then the quantile SiZer was applied to construct SiZer maps that correspond to $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9. For space consideration, only SiZer maps for $\alpha = 0.9$ are displayed in this article; these SiZer maps are displayed in Figures 3 and 4, while the remaining SiZer maps can be found on the website *http://aaron.stat.uga.edu/~cpark/Sizer/QS/*. From these plots, one could see that the quantile SiZer is capable of capturing the important real features of both the mean and variance functions, and simultaneously suppressing spurious features that are due to sampling errors. For example, in the left column of Figure 3 the data are purely noise and the quantile SiZer does not detect any spurious feature. In the middle and right columns of Figure 3, the general trends of the variance functions $\sigma_2(x)$ (decreasing) and $\sigma_3(x)$ (increasing and then decreasing) are detected at the coarse levels (i.e., large bandwidths), with no false positive shown. Note that such features can be captured by neither the original SiZer nor the quantile SiZer with $\alpha = 0.5$, as there is no trend in the mean structure. Lastly, in Figure 4 real features resulting from the superpositions of test function $m_2(x)$ with the variance functions $\sigma_1(x)$, $\sigma_2(x)$, and $\sigma_3(x)$

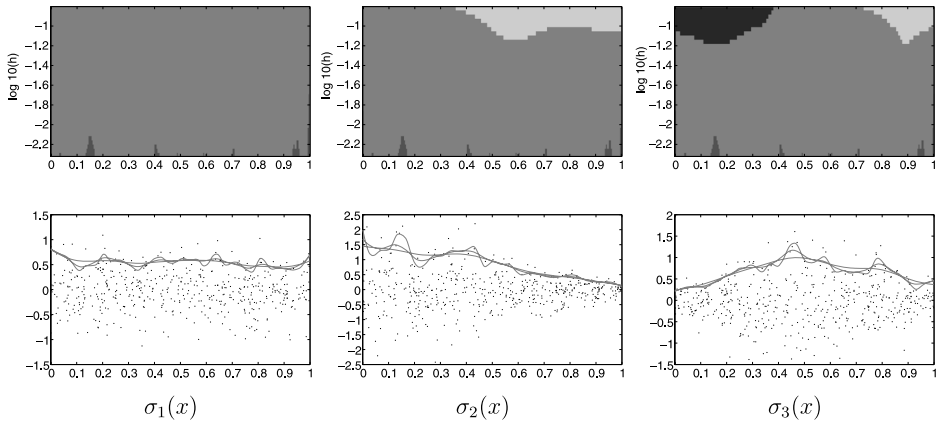

Figure 3. Quantile SiZer maps with $\alpha = 0.9$ (top row) of artificial data (dots in bottom row) generated from mean function $m_1(x)$ and variance functions $\sigma_1(x)$, $\sigma_2(x)$, and $\sigma_3(x)$. In each panel of the bottom row, the three lines are running regression quantile estimates (3.3) computed with three different bandwidths.

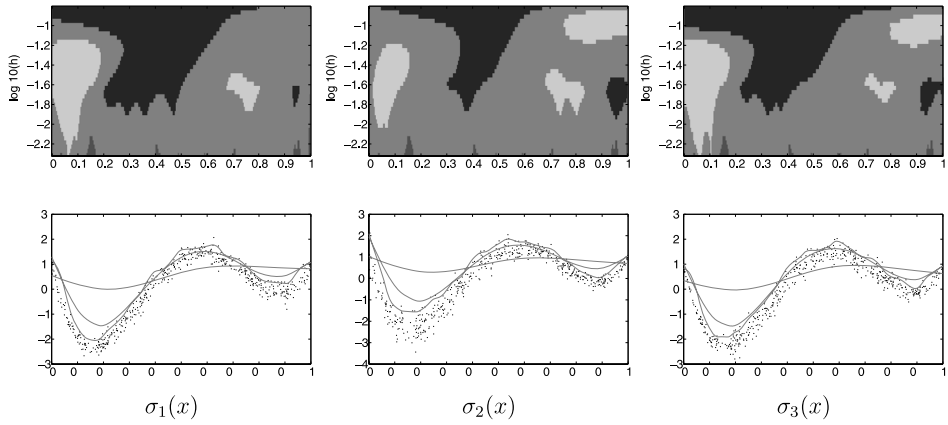

Figure 4. Similar to Figure 3 but for mean function $m_2(x)$.

are correctly identified. For example, for $x < 0.5$, there are more features revealed in the left quantile SiZer map than the middle quantile SiZer map, while for $x > 0.5$, the reverse is true. It is because the variance function $\sigma_1(x)$ of the left dataset is constant, while the variance function $\sigma_2(x)$ of the middle dataset is decreasing.

We have repeated this experiment a number of times with other simulated datasets. Similar SiZer maps were obtained so the results presented in Figures 3 and 4 are representative.

4.2 REAL DATASETS

We have also constructed different quantile SiZer map for three real datasets. The first one is the "cars" dataset analyzed, for example, by Hawkins (1994) and Ng (1996). This dataset has 392 observations on the response, fuel consumption (measured in miles/gallon), and the covariate, power output (measured in HP). This dataset is plotted in Figure 5, together with a fitted nonparametric running median curve. The fitted curve seems to suggest that there is a strong increasing trend toward the right end of the data.

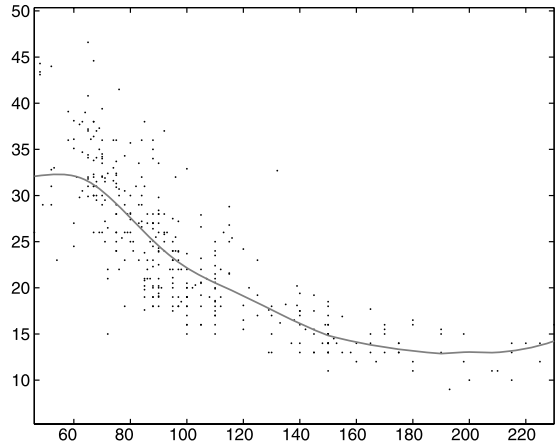

Figure 5. The "Cars" dataset with running median estimate.

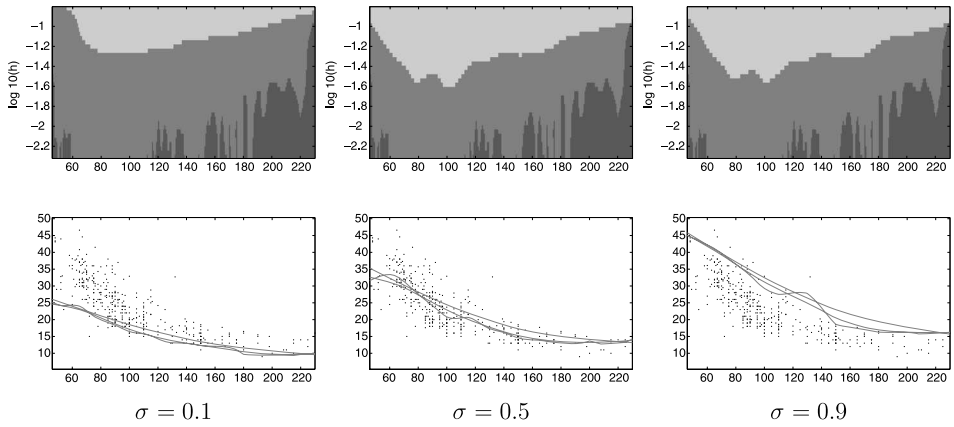

Figure 6. Quantile SiZer maps for the "Cars" dataset.

However, as argued by both Hawkins (1994) and Ng (1996), the relationship between the response and covariate should be monotonically decreasing. This implies that this increasing trend is a spurious feature. The quantile SiZer was applied to this dataset, and the SiZer maps for $\alpha = 0.1, 0.5$, and 0.9 are displayed in Figure 6. All of these SiZer maps suggested an overall decreasing trend in the data.

The second dataset that we consider is the "U.S. Girls" dataset studied, for example, by Yu, Lu, and Stander (2003). This dataset contains the weights (kg) and ages for a sample of 4011 US girls. The data are plotted in the bottom panel of Figure 7, together with a few nonparametric running regression quantile estimates obtained with different bandwidths but the same $\alpha = 0.97$. Notice that, due to a potential outlier around age 17 and weight 150 kg, some running regression quantiles exhibit a decreasing trend for age greater than 17. Such a decreasing trend seems to be a spurious feature, as in reality there is no reason to expect that a girl's weight would start to decrease after age 17. We have constructed a quantile SiZer map for $\alpha = 0.97$; see the top panel of Figure 7. This SiZer map does not support the existence of this decreasing trend.

In the third real data example we demonstrate that the quantile SiZer is capable of detecting real quantile features while the ordinary SiZer fails to do so. Here we are interested in detecting, if any, significant changes in the volatility in the Standard and Poors 500 Index (S&P 500) from January 4, 1989, to October 19, 2001, at daily frequency. The log returns of this data series are displayed in the top-left panel of Figure 8. This same data series was also analyzed by Davis, Lee, and Rodriguez-Yam (2008), where three change points located at time indices 197, 726, and 2229 were found. However, visual inspection seems to suggest that the first detected change point at $t = 197$ is a false positive. This observation is also supported by Andreou and Ghysels (2002) and hence here we take $t = 726$ and $t = 2229$ as the only statistically significant change points.

The ordinary SiZer was applied to this dataset and no significant structure was found; see the top-right panel of Figure 8. A possible explanation for this is that the mean is effectively zero across time and hence no features are detected by the ordinary SiZer. Displayed in the bottom row of Figure 8 are two quantile SiZer maps that correspond to $\alpha = 0.1$

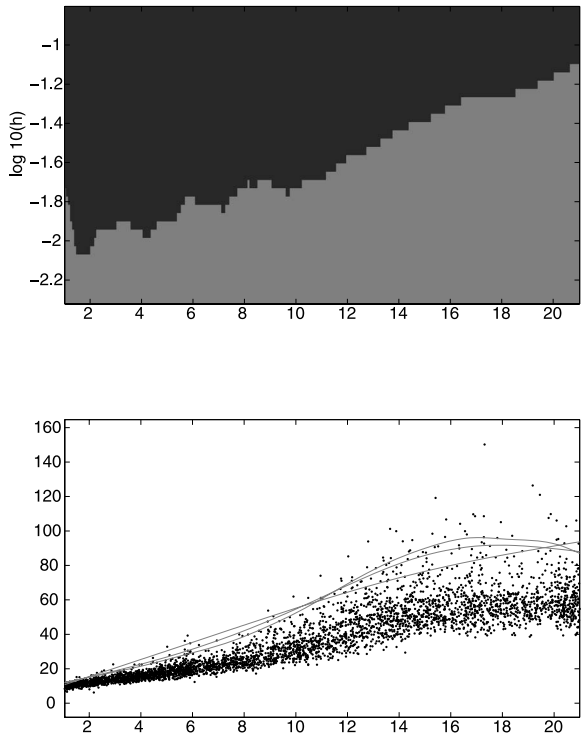

Figure 7. Bottom panel: The "U.S. Girls" dataset with running regression quantile estimates with $\alpha = 0.97$. The horizontal axis is age (year) and the vertical axis is weight (kg). Top panel: Corresponding quantile SiZer map.

and 0.9. These two SiZer maps strongly suggest there are changes in the volatility near $t = 726$ and $t = 2229$.

5. CONCLUDING REMARKS

In this article a quantile version of SiZer is proposed. This new SiZer is capable of producing SiZer maps aiming at exploring different quantile structures. Through applications to simulated examples and real datasets, it is shown that with these SiZer maps, quantile structures hidden in a dataset can be effectively revealed without paying the price of showing spurious features. It has also been shown that the quantile SiZer can detect quantile features that the ordinary SiZer fails to.

Throughout the whole article the tests in the proposed quantile SiZer are derived under the null distribution of constant mean and variance in the location-scale model (3.1). In principle one could develop other quantile SiZers for more general models, such as when the distribution of the noise ϵ depends on X_i and/or follows a non-Gaussian distribution. These generalizations, however, bring in new computational challenges. For example, the fast algorithm described in the Supplemental Materials cannot be used to calculate $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$, and their variances would probably need to be estimated with computationally intensive procedures such as the bootstrap. In order to investigate the performance of the

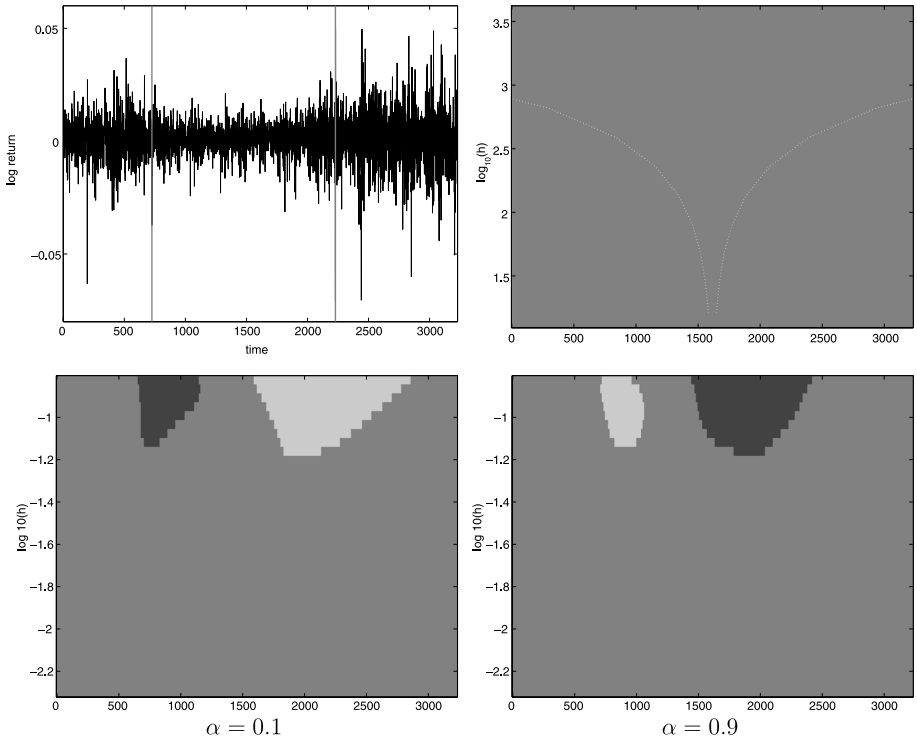

Figure 8. Top-left: The log returns of the S&P 500 series data. The two vertical lines indicate the change points at $t = 726$ and $t = 2229$ detected by Davis, Lee, and Rodriguez-Yam (2008). Top-right panel: Ordinary SiZer map. Bottom row: Quantile SiZer maps for $\alpha = 0.1$ and 0.9.

proposed quantile SiZer when the location-scale assumption is violated, the following experiment was conducted. Artificial datasets were generated from the settings in Section 4.1, but with the following model:

$$Y_i = m_2(X_i) + \sigma_j(X_i)E_{X_i}, \qquad j = 1, 2, 3, \tag{5.1}$$

where E_{X_i} follows a Laplace distribution with mean X_i, and E_{X_i} and E_{X_j} are independent if $i \neq j$. The variance of E_{X_i} was chosen in such a way that the resulting signal-to-noise ratios are comparable to those in Section 4.1. The generated datasets, together with the corresponding $\alpha = 0.9$ quantile SiZer maps, are displayed in Figure 9. Although the noise assumption is wrong, the proposed quantile SiZer did produce SiZer maps that reflect well the true structures of the data.

Another interesting point suggested by a referee is an alternative way of defining a quantile SiZer map. Instead of fixing α and having the bandwidth vary in the vertical axis, one could produce a SiZer map with α changing in the vertical axis for a fixed bandwidth. In this way information from different values for α (for the same fixed bandwidth) can be grasped at the same time. Figure 10 provides such an example. The data were generated from $m_1(x)$ and $\sigma_3(x)$. This SiZer map correctly suggests that there is no significant feature near the mean of $m_1(x)$, while the quadratic trend of $\sigma_3(x)$ is apparent.

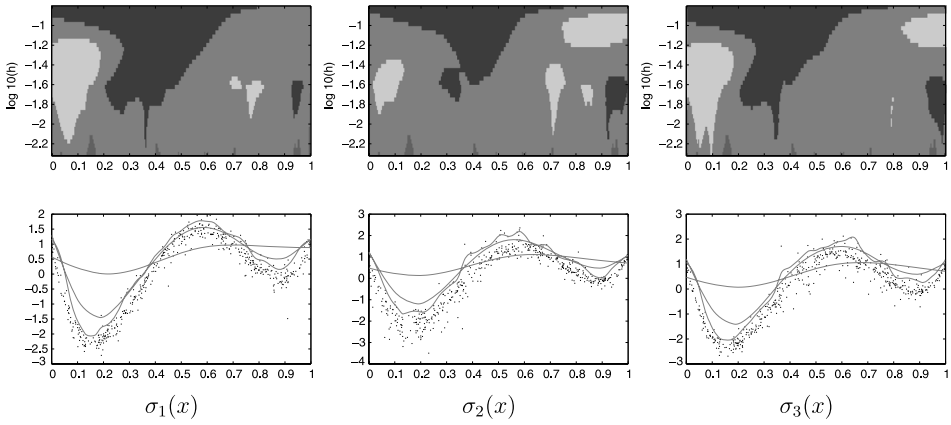

Figure 9. Similar to Figure 4 but for the data generated from the Laplace distribution.

SUPPLEMENTAL MATERIALS

Appendix: We provide fast computation of $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$. We also derive asymptotic expectations and variances for $\hat{\beta}_{h,\alpha}(x)$ and $\hat{\beta}'_{h,\alpha}(x)$. (appendix.pdf)

Matlab programs and datasets: We provide Matlab programs and datasets for quantile SiZer. (qsizerpg.zip)

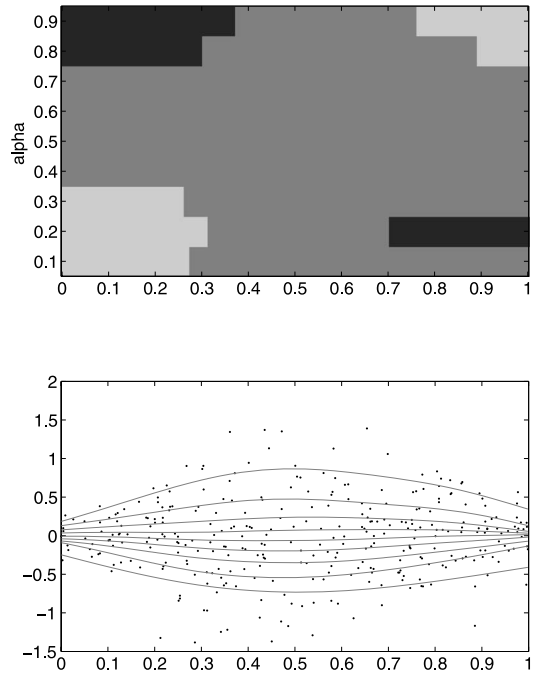

Figure 10. An alternative quantile SiZer map with $h = 0.13$.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Keming Yu for sending us the real datasets used in Section 4.2, and to the reviewers and the associate editor for many useful comments that led to this much improved version of the article. The work of Hannig was supported in part by the National Science Foundation under grants 0504737 and 0707037. The work of Lee was supported in part by the Hong Kong Research Grants Council under CERG 401507, a Chinese University of Hong Kong Direct Grant, and the National Science Foundation under grant 0707037. The work of Park was supported in part by National Security Agency grant No. H982300810056. This study was also supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

[Received July 2009. Revised April 2010.]

REFERENCES

Andreou, E., and Ghysels, E. (2002), "Detecting Multiple Breaks in Financial Market Volatility Dynamics," *Journal of Applied Econometrics*, 17, 579–600. [507]

Bhattacharya, P. K., and Gangopadhyay, A. K. (1990), "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile," *The Annals of Statistics*, 18, 1400–1415. [497]

Chaudhuri, P. (1991a), "Global Nonparametric Estimation of Conditional Quantile Functions and Their Derivatives," *Journal of Multivariate Analysis*, 39, 246–269. [498,502]

——— (1991b), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *The Annals of Statistics*, 19, 760–777. [498,502]

Chaudhuri, P., and Loh, W. Y. (2002), "Nonparametric Estimation of Conditional Quantiles Using Quantile Regression Trees," *Bernoulli*, 8, 561–576. [497]

Chaudhuri, P., and Marron, J. S. (1999), "SiZer for Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823. [497-501,503]

——— (2000), "Scale Space View of Curve Estimation," *The Annals of Statistics*, 28, 408–428. [498]

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2008), "Break Detection for a Class of Nonlinear Time Series Models," *Journal of Time Series Analysis*, 29, 834–867. [507,509]

Erästö, P., and Holmström, L. (2005), "Bayesian Multiscale Smoothing for Making Inferences About Features in Scatter Plots," *Journal of Computational and Graphical Statistics*, 14, 569–589. [500]

——— (2007), "Bayesian Analysis of Features in a Scatter Plot With Dependent Observations and Errors in Predictors," *Journal of Statistical Computation and Simulation*, 77, 421–434. [501]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall. [501]

Ganguli, B., and Wand, M. P. (2004), "Feature Significance in Geostatistics," *Journal of Computational and Graphical Statistics*, 13, 954–973. [500]

——— (2007), "Feature Significance in Generalized Additive Models," *Statistics and Computing*, 17, 179–192. [500]

Godtliebsen, F., and Oigard, T. A. (2005), "A Visual Display Device for Significant Features in Complicated Signals," *Computational Statistics and Data Analysis*, 48, 317–343. [500]

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2002), "Significance in Scale Space for Bivariate Density Estimation," *Journal of Computational and Graphical Statistics*, 11, 1–21. [500]

Gonzalez-Manteiga, W., Martinez-Miranda, M. D., and Raya-Miranda, R. (2008), "SiZer Map for Inference With Additive Models," *Statistics and Computing*, 18, 297–312. [500]

Hannig, J., and Lee, T. C. M. (2006), "Robust SiZer for Exploration of Regression Structures and Outlier Detection," *Journal of Computational and Graphical Statistics*, 15, 101–117. [500,503,505]

Hannig, J., and Marron, J. S. (2006), "Advanced Distribution Theory for SiZer," *Journal of the American Statistical Association*, 101, 484–499. [500,504]

Hawkins, D. M. (1994), "Fitting Monotonic Polynomials to Data," *Computational Statistics*, 9, 233–247. [506, 507]

Hendricks, W., and Koenker, R. (1992), "Hierarchical Spline Model for Conditional Quantiles and the Demand for Electricity," *Journal of the American Statistical Association*, 87, 58–68. [498]

Kim, C. S., and Marron, J. S. (2006), "SiZer for Jump Detection," *Journal of Nonparametric Statistics*, 18, 13–20. [501]

Koenker, R. (2005), *Quantile Regression*, New York: Cambridge University Press. [501]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [498]

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680. [498]

Lejeune, M. G., and Sarda, P. (1988), "Quantile Regression: A Nonparametric Approach," *Computational Statistics and Data Analysis*, 6, 229–239. [497]

Marron, J. S., and de Una-Alvarez, J. (2004), "SiZer for Length Biased, Censored Density and Hazard Estimation," *Journal of Statistical Planning and Inference*, 121, 149–161. [501]

Marron, J. S., and Zhang, J. T. (2005), "SiZer for Smoothing Splines," *Computational Statistics*, 20, 481–502. [501]

Ng, P. T. (1996), "An Algorithm for Quantile Smoothing Splines," *Computational Statistics and Data Analysis*, 22, 99–118. [506,507]

Oh, H.-S., Lee, T. C. M., and Nychka, D. (2010), "Fast Nonparametric Quantile Regression With Arbitrary Smoothing Methods," unpublished manuscript, Seoul National University. [502]

Oigard, T. A., Rue, H., and Godtliebsen, F. (2006), "Bayesian Multiscale Analysis for Time Series Data," *Computational Statistics and Data Analysis*, 51, 1719–1730. [500,501]

Olsen, L. R., Chaudhuri, P., and Godtliebsen, F. (2008), "Multiscale Spectral Analysis for Detecting Short and Long Range Change Points in Time Series," *Computational Statistics and Data Analysis*, 52, 3310–3330. [501]

Olsen, L. R., Sorbye, S. H., and Godtliebsen, F. (2008), "A Scale-Space Approach for Detecting Non-Stationarities in Time Series," *Scandinavian Journal of Statistics*, 35, 119–138. [500]

Park, C., and Kang, K.-H. (2008), "SiZer Analysis for the Comparison of Regression Curves," *Computational Statistics and Data Analysis*, 52, 3954–3970. [501]

Park, C., Godtliebsen, F., Taqqu, M., Stoev, S., and Marron, J. S. (2007), "Visualization and Inference Based on Wavelet Coefficients, SiZer and SiNos," *Computational Statistics and Data Analysis*, 51, 5994–6012. [501]

Park, C., Hannig, J., and Kang, K.-H. (2009), "Improved SiZer for Time Series," *Statistica Sinica*, 19, 1511–1530. [500]

Park, C., Hernández-Campos, F., Marron, J. S., and Smith, F. D. (2005), "Long-Range Dependence in a Changing Internet Traffic Mix," *Computer Networks*, 48, 401–422. [501]

Park, C., Marron, J. S., and Rondonotti, V. (2004), "Dependent SiZer: Goodness-of-Fit Tests for Time Series Models," *Journal of Applied Statistics*, 31, 999–1017. [500]

Rolls, D. A., Michailidis, G., and Hernández-Campos, F. (2005), "Queueing Analysis of Network Traffic: Methodology and Visualization Tools," *Computer Networks*, 48, 447–473. [501]

Rondonotti, V., Marron, J. S., and Park, C. (2007), "SiZer for Time Series: A New Approach to the Analysis of Trends," *Electronic Journal of Statistics*, 1, 268–289. [500]

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270. [505]

Su, L., and Ullah, A. (2008), "Nonparametric Prewhitening Estimators for Conditional Quantiles," *Statistica Sinica*, 18, 1131–1152. [498]

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006), "Nonparametric Quantile Estimation," *The Journal of Machine Learning Research*, 7, 1231–1264. [498]

Truong, Y. K. N. (1989), "Asymptotic Properties of Kernel Estimators Based on Local Medians," *The Annals of Statistics*, 17, 606–617. [497]

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall. [501,503]

Yu, K. (1999), "Smoothing Regression Quantile by Combining k-nn With Local Linear Fitting," *Statistica Sinica*, 9, 759–771. [497]

Yu, K., and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237. [502]

Yu, K., Lu, Z., and Stander, J. (2003), "Quantile Regression: Applications and Current Research Areas," *Journal of the Royal Statistical Society, Ser. D*, 52, 331–350. [507]

Yuan, M. (2006), "GACV for Quantile Smoothing Splines," *Computational Statistics and Data Analysis*, 50, 813–829. [498]