An Exploratory Data Analysis in Scale-Space for Interval-Valued Data

Cheolwoo Park

Department of Statistics, University of Georgia, Athens, GA 30602, USA

Yongho Jeon

Department of Applied Statistics, Yonsei University, Seoul, 120-749, Korea

Kee-Hoon Kang

Department of Statistics, Hankuk University of Foreign Studies, Yongin, 449-791, Korea

Abstract

We propose an exploratory data analysis approach when data are observed as intervals in a nonparametric regression setting. The interval-valued data contain richer information than single-valued data in the sense that they provide both center and range information of the underlying structure. Conventionally, these two attributes have been studied separately as traditional tools can be readily used for single-valued data analysis. We propose a unified data analysis tool that attempts to capture the relationship between response and covariate by simultaneously accounting for variability present in the data. It utilizes a kernel smoothing approach, which is conducted in scale-space so that it considers a wide range of smoothing parameters rather than selecting an optimal value. It also visually summarizes the significance of trends in the data as a color map across multiple locations and scales. We demonstrate its effectiveness as an exploratory data analysis tool for interval-valued data using simulated and real examples.

Keywords: Exploratory data analysis, Interval-valued data, Nonparametric regression, Scale-Space, Visualization.

1 Introduction

With the rapid advancement of computing technology and storage capacity, both the size of data and the complexity of their structure have significantly increased. These enormous data are sometimes converted into new types of data such as intervals, histogram, and trees for an effective summary (Billard and Diday, 2003; Wang and Marron, 2007; Noirhomme-Fraiture and Brito, 2011). An analysis of these types of data using traditional statistical approaches often encounters unsatisfactory outcomes, and thus it is imperative to develop appropriate statistical methodologies for these data sets.

Interval-valued data are observed with lower and upper bounds, representing uncertainty or variability. Interval-valued data often arise in sampling or aggregation in large data sets. This aggregation reduces large, complex data sets to a size that is more manageable for practitioners by keeping only the extracted information. Interval-valued data offer richer and more complex information than single-valued data because they contain both trend and variation. Examples of interval-valued data include blood pressure values reported in medical records of patients, the maximum and minimum stock prices in a day, and selling prices of cars or houses.

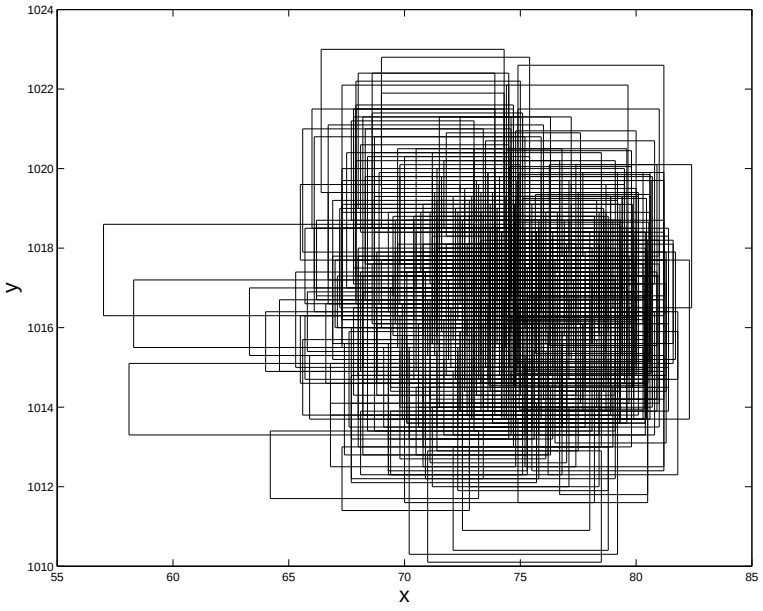


Figure 1: Scatterplot of the Hawaii climate data in 2013. The x- and y-axes represent temperature in Fahrenheit and sea level pressure in millibars as interval-valued data generated from daily means collected at 29 climate/weather stations.

To provide the motivation of the proposed work, we introduce a real example analyzed in Section 4.2. The example concerns the relationship between the daily mean sea level pressure and the daily mean temperature in Hawaii. The relationship between these two variables is of interest to climatologists, e.g. see Bayr and Dommenget (2013) and references therein. Our goal is to conduct a quick exploratory data analysis as an initial step without considering any other factors. The original data (publicly available from the National Climatic Data Center at `http://www.ncdc.noaa.gov/`) were collected in the single-valued form in 2013 from 29 stations, yielding 117,709 observations. From these single-valued data we calculate the 10th and 90th percentiles of the 29 stations each day for both temperature and pressure, producing a total of 365 interval-valued observations. Figure 1 depicts the aggregated interval-valued data. It is hard to see whether there exists any trend between the two variables from the scatterplot. Also, it is not clear whether the trend, if any, is statistically significant, or is just created by the artifacts of sampling noise or variation in the intervals.

Interval-valued data analysis (Billard and Diday, 2003) has gained considerable attention in regression (de Carvalho et al., 2004; Lima Neto et al., 2004, 2005; Lima Neto and de Carvalho F. A. T., 2010; Blanco-Fernández et al., 2011; Yang et al., 2011; Blanco-Fernández et al., 2013; Jeon et al., 2015), multivariate analysis (Lauro and Palumbo, 2000; Palumbo and Lauro, 2003; Douzal-Chouakria et al., 2011; Le-Rademacher and Billard, 2012), and time series contexts (Maia et al., 2008; Arroyo et al., 2011). In parametric regression problems, Billard and Diday (2000) propose a center method that fits an ordinary regression model to centers of intervals and Lima Neto and de Carvalho (2008) fit two separate regression models for centers and ranges of the intervals. Bivariate models have also been considered. Lima Neto et al. (2009, 2011) develop bivariate generalized linear models for symbolic data and Silva et al. (2011) propose a copula-based regression model. Ahn et al. (2012) apply a resampling scheme to account for the variation in the interval-valued data and conduct statistical inference in a parametric regression setting. Most of the existing statistical methods for interval-valued data aim to predict future observations, but little work has been done for exploratory data analysis. Hence, it is necessary to develop a proper tool to explore interval-valued data, which would assist prediction and statistical inference in later steps.

We take a kernel-based nonparametric approach for exploring trends in interval-valued data. One could apply a nonparametric smoothing technique to lower and upper bounds, or to centers and ranges separately, but it does not fully utilize the information available in the data, and it could be difficult to interpret separate analyses. Therefore, it would be desirable to develop a unified data analysis tool and extract meaningful trends in the interval-valued

data by taking the internal variation into account.

In nonparametric kernel smoothing problems, the selection of smoothing parameters has been a critical issue, which could be more challenging for interval-valued data because the range of the intervals as well as the noise error are extra sources of variation. In order to circumvent this difficulty we take a scale-space approach (Lindeberg, 1994) and develop a *SiZer* (SIgnificant ZERo crossing of the derivatives) tool (Chaudhuri and Marron, 1999) for interval-valued data. SiZer investigates significant features in the data at multiple smoothing levels instead of choosing a single, optimal one, and visually summarizes its statistical inference results as a color map, called *SiZer map* for easy and quick interpretation. Therefore, it allows data analysts to discover all the information in the data that is available at each smoothing level.

Since the seminal work of Chaudhuri and Marron (1999), SiZer tools have been extended to various statistical methods and applied to a broad set of real applications. For example, Hannig and Lee (2006) study median regression function and detect outliers in the data, and Park et al. (2010) generalize it to the quantile function. SiZer tools have been developed for jump points detection (Kim and Marron, 2006), survival analysis (Marron and de Uña Álvarez, 2004), generalized linear models (Li and Marron, 2005; Ganguli and Wand, 2007; Park and Huh, 2013), smoothing spline (Marron and Zhang, 2005), additive models (González-Manteiga et al., 2008), and comparison of multiple curves (Park and Kang, 2008; Park et al., 2015). For time series data, Park et al. (2004), Rondonotti et al. (2007), and Park et al. (2007, 2009a,b) apply a scale-space approach while accounting for serial correlation in the data. Also, Bayesian multiscale smoothing techniques are used in Erästö and Holmström (2005), Godtliebsen and Øigård (2005), Øigård et al. (2006), Erästö and Holmström (2007), and Sorbye et al. (2009). The SiZer idea has been extended to two dimensional imaging data as well (Godtliebsen et al., 2002, 2004; Duong et al., 2008; Vaughan et al., 2012; Holmström and Pasanen, 2012).

The objective of the proposed work is to develop a SiZer tool, which conducts exploratory data analysis using nonparametric kernel smoothing at multiple scales and offers statistical inference for finding meaningful structure in interval-valued data. In order to make statistical inference that accounts for the variation in interval-valued data we propose three different ways of constructing a confidence interval at each location and at each scale by combining bootstrap and Monte Carlo resampling schemes. These resampling approaches enable one to fully make use of the variability in interval-valued data and obtain the sampling distribution of relevant smoothing estimators.

The remainder of the paper is organized as follows. Section 2 reviews the conventional SiZer for single-valued data and introduces interval-valued data analysis using centers and ranges separately. Section 3 proposes a new SiZer tool for interval-valued data. Section 4 presents simulation results and real data analysis with the proposed tool. The paper concludes with discussion in Section 5.

2 Conventional SiZer

Suppose that n pairs of single-valued data $\{(X_i, Y_i), i = 1, ..., n\}$ are independently observed. A nonparametric regression setting is given as

$$Y_i = f(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, ..., n, \tag{2.1}$$

where f is an unknown regression function, $\sigma^2(x) = Var(Y_i|X_i = x)$ is a conditional variance function, and ϵ_i independently follows $N(0, 1)$.

The conventional SiZer (Chaudhuri and Marron, 1999) is an effective multiscale tool for discovering any important features hidden in the single-valued data. Its target changes from an underlying true function (e.g. f in (2.1)) to a smoothed function (e.g. f_h below) depending on a scale parameter (e.g. bandwidth), and analysis is done across multiple locations and scales. It utilizes the local linear smoothing method (Fan and Gijbels, 1996) to estimate the unknown function f in the model (2.1) and its derivative f' at a given location x for a given scale (bandwidth) h. More specifically, one can obtain $(\hat{\beta}_0, \hat{\beta}_1)$ at (x, h) by minimizing

$$\sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1(x - X_i))]^2 K_h(x - X_i) \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$ and K is the standard normal density function. Then, since $\hat{\beta}_0 \approx f_h(x) = \int f(u)K_h(x - u)du$ and $\hat{\beta}_1 \approx f'_h(x) = \int f'(u)K_h(x - u)du$, $(\hat{\beta}_0, \hat{\beta}_1)$ is an estimate of $(f_h(x), f'_h(x))$.

SiZer conducts statistical inference on the slopes to find meaningful trends and reports its significance testing results in a SiZer map. Since a scale-space approach assumes that the truth exists at each scale (Lindeberg, 1994), SiZer studies the estimated slopes with different locations and bandwidths and conducts the statistical hypothesis tests $H_0 : f'_h(x) = 0$ at each (x, h). The corresponding $100(1 - \alpha)\%$ confidence band is given as

$$\hat{f}'_h(x) \pm q_h \widehat{SD}(\hat{f}'_h(x)) \tag{2.3}$$

where the estimate of the standard deviation (SD) is given by the conditional weighted sample variances (Fan and Gijbels, 1996). The modified Gaussian quantile with multiple testing adjustment is given as (Hannig and Marron, 2006),

$$q_h = \Phi^{-1}\left(\left(1-\frac{\alpha}{2}\right)^{1/(\theta g)}\right) \tag{2.4}$$

where Φ is the cumulative distribution function of the standard normal, g is the number of pixels in each row of a SiZer map, and the cluster index θ is given as

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

which measures the equivalent number of independent observations. Here, $\tilde{\Delta}$ denotes the distance between the pixels of the SiZer map. See Hannig and Marron (2006) for more details. The nominal level α=0.05 is used in our numerical examples.

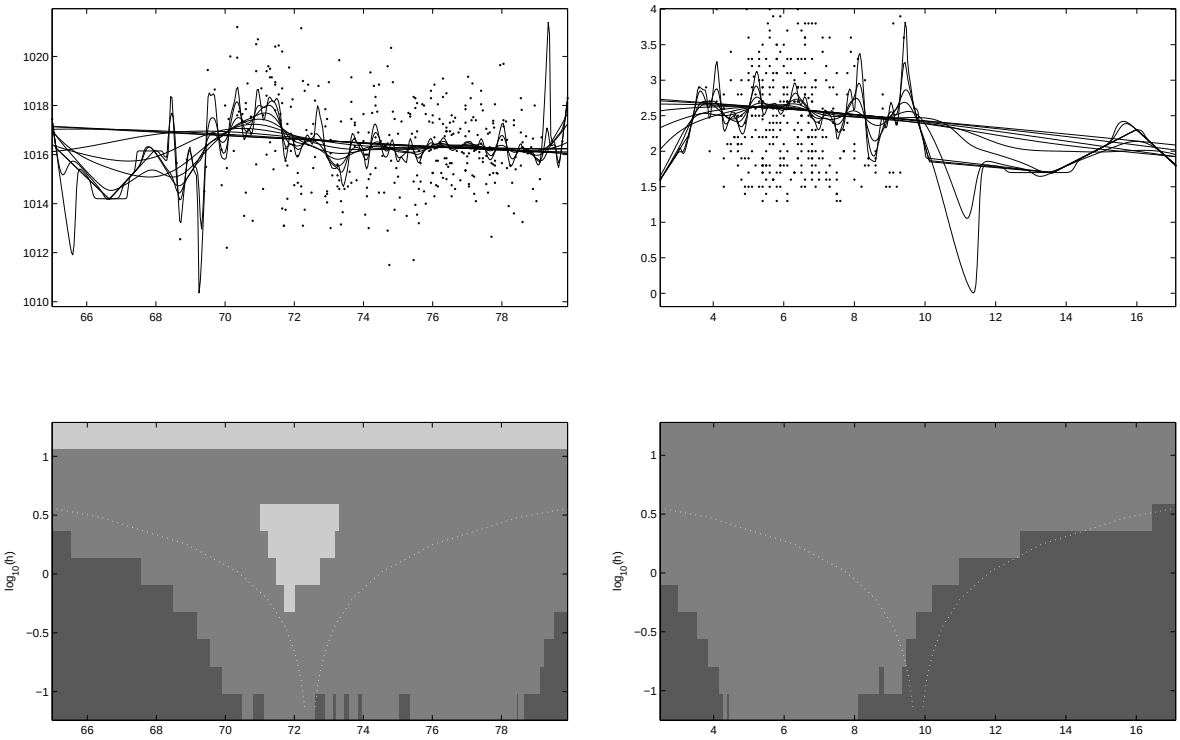
SiZer uses colors to present the statistical test results in a SiZer map. The pixel at (x, h) is colored black if the confidence band in (2.3) is above zero, implying that the smoothed function $f_h(x)$ is increasing at the corresponding point x and scale h. It is colored white if the confidence band is below zero, implying that the smoothed function $f_h(x)$ is decreasing at (x, h). If the confidence band contains zero, the slope at (x, h) is not significantly above or below zero and the pixel is colored intermediate gray. SiZer utilizes the effective sample size (ESS)

$$ESS(x,h) = \frac{\sum_{i=1}^{n} K_h(x - X_i)}{K_h(0)}, \tag{2.5}$$

to determine whether the number of observations is sufficient to make any statistical decision at (x, h). If $ESS(x, h) < 5$, the pixel is colored darker gray and shows no testing result.

The conventional SiZer for single-valued data can also be used for interval-valued data. For example, it can be applied to centers and ranges separately. In what follows, we apply the conventional SiZer to the Hawaii climate data introduced in Section 1.

Figure 2(a) depicts SiZer plots using the mid-points of the interval-valued data to investigate the center relationship between the daily sea level pressure and the daily temperature for 2013. In the top panel, the horizontal axis represents the temperature and the vertical axis the sea level pressure. The dots display the mid-points of the intervals and the thin curves the family of smooths, which are the local linear smooths $\hat{f}_h(x)$ with different h values (i.e. at different scales). These curves show an overall decreasing trend and some oscillating trends that vary depending on the scale, but it is not obvious which features are



(a) Center (b) Range

Figure 2: We apply the conventional SiZer to the centers (a) and the ranges (b) of the Hawaii climate data, sea level pressure vs. temperature, depicted in Figure 1. The dotted white curves in the SiZer maps display the window widths for each scale, $\pm 2h$.

really present. The bottom panel displays the SiZer map summarizing statistical inference on the slopes of the smoothed curves in the top panel. The horizontal axis represents the same horizontal locations as in the family of smooths in the top panel. The vertical axis represents the level of smoothing on a log scale, $\log_{10} h$. Note that each row of the SiZer map corresponds to one of the thin curves in the top panel. In the map, the pixels at the coarsest scale (the largest h) are colored white, which confirms that the overall decreasing trend is statistically significant at the large scales. Another white feature appears in the middle at medium scales, which suggests that the decreasing trend in those regions is also statistically significant at the corresponding scales. The other regions are colored either intermediate or darker gray, which shows no significance or no decision, respectively. Figure 2(b) depicts SiZer plots using the ranges of the interval-valued data to investigate the range relationship between the sea level pressure and the temperature. Note that the horizontal axes in these plots are different from those in Figure 2(a) because the ranges of sea level pressure and temperature are the response and the covariate, respectively. The family of smooths in the top panel shows an overall decreasing trend and some oscillating curves. However, none of these features turn out to be statistically significant in the SiZer map in the bottom panel because only intermediate or darker gray colors are present in the map. Therefore, we can conclude that there is no particular relationship between the ranges of sea level pressure and temperature.

This analysis demonstrates that the conventional SiZer is still useful for analyzing interval-valued data and finding the relationship between response and covariate using their centers and ranges. However, it would be desirable to develop a unified scale-space tool that takes both centers and ranges into account simultaneously.

3 SiZer for Interval-Valued Data

Let $\{(X_i, Y_i), i = 1, ..., n\}$ be interval-valued observations with $X_i = [X_{Li}, X_{Ui}]$ $(X_{Li} \leq X_{Ui})$ and $Y_i = [Y_{Li}, Y_{Ui}]$ $(Y_{Li} \leq Y_{Ui})$. Assume that the points within the intervals are uniformly distributed. In this section, we propose three SiZer inference approaches for interval-valued data, namely MC, BMC-q, BMC-SD SiZers. In particular, each approach reflects the internal variation in the interval-valued data into the confidence band in (2.3) differently. Ahn et al. (2012) apply a resampling scheme to interval-valued data to fully utilize the variability within intervals and make inference on regression coefficients in a parametric setting. However, their approach is limited in the sense that it cannot be used for single-valued data; this

is because their inference relies on lengths of intervals. We design SiZer inference that could be applied to both single-valued and interval-valued data using Monte Carlo and/or Bootstrap resampling approaches. We note that it is a common practice to assume a uniform distribution for the internal distribution (Billard and Diday, 2007), but a non-uniform (e.g. normal) distribution can be applied to the proposed SiZers and we suggest it as our future work.

3.1 MC SiZer

This method uses a Monte Carlo resampling approach to estimate $\hat{f}'_h(x)$ and $\widehat{SD}(\hat{f}'_h(x))$ in (2.3) for interval-valued data. More specifically, for $m = 1, \ldots, B_1$,

(i) generate a single-valued random regression sample $(X_{i,m}, Y_{i,m})$ by assuming the uniform distribution within the intervals $X_i = [X_{Li}, X_{Ui}]$ and $Y_i = [Y_{Li}, Y_{Ui}]$ for each $i = 1, \ldots, n$.

(ii) Apply the local linear smoothing technique in (2.2) and obtain $\hat{f}'_{h,m}(x)$ and $\widehat{SD}(\hat{f}'_{h,m}(x))$.

Then, the final estimates are given as

$$\hat{f}'_h(x) = \frac{1}{B_1}\sum_{m=1}^{B_1} \hat{f}'_{h,m}(x), \quad \widehat{SD}(\hat{f}'_h(x)) = \frac{1}{B_1}\sum_{m=1}^{B_1} \widehat{SD}(\hat{f}'_{h,m}(x)). \tag{3.1}$$

Since single-valued points are randomly generated from the observed intervals, it accounts for the internal variation within them. We use the same quantile q_h in (2.4) in the MC SiZer inference.

3.2 BMC-q SiZer

The quantile q_h in (2.4) is theoretically driven using the Gaussian approximation when data are single-valued (Hannig and Marron, 2006). It is not guaranteed that the familywise error rate would be controlled using this quantile for interval-valued data. **Hence, we consider a tractable approach directly from the given data instead of Gaussian approximation.** The BMC-q SiZer empirically approximates the quantile q_h using bootstrap sampling as well as estimates $\hat{f}'_h(x)$ and $\widehat{SD}(\hat{f}'_h(x))$ using Monte Carlo sampling as in (3.1). **The bootstrap method has proven to be a powerful tool in many applications, and is accepted as an alternative to asymptotic approaches. There is a rich set of**

literature on bootstrap; for example, Mammen (1992)) and Horowitz (2001) summarize its theoretical properties, including consistency, and illustrate cases when bootstrap works well and when it does not. Our algorithm is given as follows.

(i) Sample the interval-valued data (X_i^*, Y_i^*) from the original data (X_i, Y_i), $i = 1, \ldots, n$ with replacement.

(ii) For $m = 1, \ldots, B_1$, generate a single-valued random regression sample $(X_{i,m}^*, Y_{i,m}^*)$ by assuming the uniform distribution within the intervals $X_i^* = [X_{Li}^*, X_{Ui}^*]$ and $Y_i^* = [Y_{Li}^*, Y_{Ui}^*]$ for each $i = 1, \ldots, n$. Apply the local linear smoothing technique in (2.2) and obtain $\hat{f}_{h,m}'^*(x)$ and

$$\hat{f}_h'^*(x) = \frac{1}{B_1} \sum_{m=1}^{B_1} \hat{f}_{h,m}'^*(x).$$

(iii) Calculate the Z^*-statistic:

$$Z^*(x, h) = \frac{\hat{f}_h'^*(x) - \hat{f}_h'(x)}{\widehat{SD}(\hat{f}_h'(x))}$$

where $\hat{f}_h'(x)$ and $\widehat{SD}(\hat{f}_h'(x))$ are given in (3.1).

(iv) Repeat (i)-(iii) B_2 times.

(v) Using the B_2 repetitions, obtain the empirical distribution of $\max_x |Z^*(x, h)|$ and the quantile q_h.

The BMC-q SiZer also uses $\hat{f}_h'(x)$ and $\widehat{SD}(\hat{f}_h'(x))$ in (3.1) for its confidence bands. A similar idea for estimating q_h based on bootstrap sampling can be found in Chaudhuri and Marron (1999).

3.3 BMC-SD SiZer

The BMC-SD SiZer uses the bootstrap sampling method to estimate $\widehat{SD}(\hat{f}_h'(x))$. The algorithm is given as follows.

(i) Sample the interval-valued data (X_i^*, Y_i^*) from the original data (X_i, Y_i), $i = 1, \ldots, n$ with replacement.

(ii) For $m = 1, \ldots, B_1$, generate a single-valued random regression sample $(X^*_{i,m}, Y^*_{i,m})$ by assuming the uniform distribution within the intervals $X^*_i = [X^*_{Li}, X^*_{Ui}]$ and $Y^*_i = [Y^*_{Li}, Y^*_{Ui}]$ for each $i = 1, \ldots, n$. Apply the local linear smoothing technique in (2.2) and obtain $\hat{f}'^*_{h,m}(x)$ and

$$\hat{f}'^*_h(x) = \frac{1}{B_1} \sum_{m=1}^{B_1} \hat{f}'^*_{h,m}(x).$$

(iii) Repeat (i) and (ii) B_2 times.

(iv) Using the B_2 repetitions, obtain the standard deviation of $\hat{f}'_h(x)$:

$$\widehat{SD}(\hat{f}'_h(x)) = \text{standard deviation } (\hat{f}'^*_h(x)).$$

The BMC-SD SiZer also uses $\hat{f}'_h(x)$ in (3.1) and the theoretically driven quantile q_h given in (2.4) for its confidence bands. One can develop a SiZer that approximates both SD and q_h, but we choose not to implement it due to heavy computation load.

Remarks. In our numerical study, we use $B_1 = B_2 = 100$. Also, the bandwidths used in our numerical examples are 11 equally spaced values on a logarithmic scale of the range of x. For three proposed SiZers, the effective sample size (ESS) is estimated by the Monte Carlo sampling approach. For $m = 1, \ldots, B_1$,

(i) generate a single-valued random regression sample $(X_{i,m}, Y_{i,m})$ by assuming the uniform distribution within the intervals (X_i, Y_i) for each $i = 1, \ldots, n$.

(ii) Calculate

$$ESS_m(x, h) = \frac{\sum_{i=1}^n K_h(x - X_{i,m})}{K_h(0)}.$$

Then, the final $ESS(x, h)$ is given as

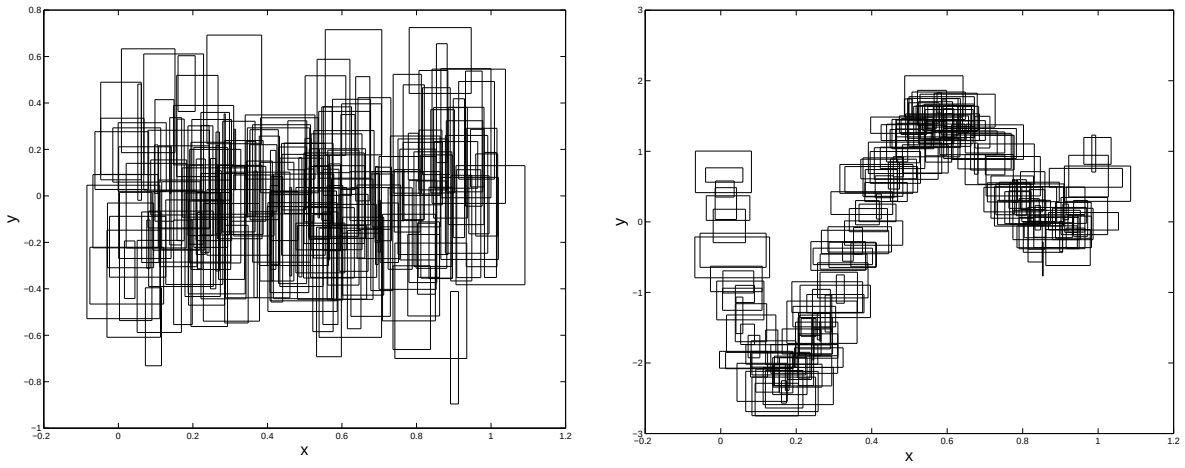
$$ESS(x, h) = \frac{1}{B_1} \sum_{m=1}^{B_1} ESS_m(x, h).$$

If $ESS(x, h) < 5$, then the pixel is colored darker gray and no statistical decision is made at the corresponding pixel.

4 Numerical Examples

In Section 4.1, we examine the performance of three proposed SiZer tools, MC, BMC-q, BMC-SD, under various simulation settings. In Section 4.2, we analyze the real example introduced in Section 1.

4.1 Simulation



(a) No trend (b) Nonlinear trend

Figure 3: The simulated examples with uniform ranges and constant variance function.

We randomly generate the centers of X_i, $X_{i,C}$, from $U(0,1)$, and the centers of Y_i, $Y_{i,C}$, from

$$Y_{i,C} = f(X_{i,C}) + \sigma(X_{i,C})\epsilon_{i,C},$$

where $\epsilon_{i,C}$ independently follows the standard normal distribution. The regression function f is chosen from the following two functions:

$$\text{(C1) } f(x) = 0 \text{ and (C2) } f(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4.$$

The variance function σ^2 is also chosen from the following two functions:

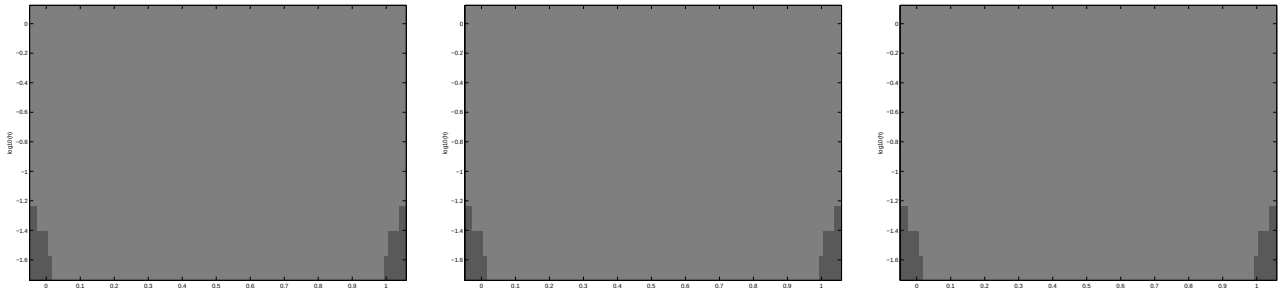
$$\text{(V1) } \sigma^2(x) = 0.2^2 \text{ and (V2) } \sigma^2(x) = -x^2 + x + 0.1.$$

We consider two scenarios for ranges: an independent case and a dependent case between the ranges of X and Y:

(R1) $X_{i,R} \sim U(0, 0.1)$ and $Y_{i,R} \sim U(0.1, 0.3)$ and they are independent of each other.

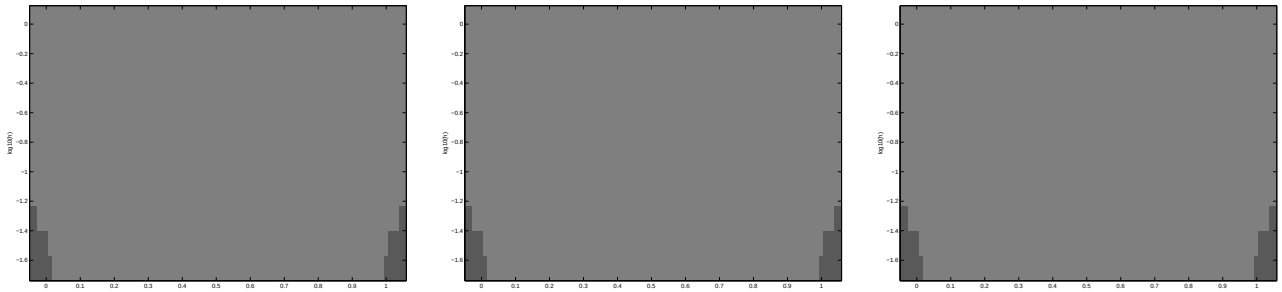
(R2) $X_{i,R} \sim U(0, 0.1)$ and $Y_{i,R} = -X_{i,R}^2 + X_{i,R} + 0.1 + \epsilon_{i,R}$

where $\epsilon_{i,R}$ independently follows $N(0, 0.01^2)$. Figure 3 displays the simulated data with (C1)-(V1)-(R1) (no trend, constant variance function, independent ranges) and (C2)-(V1)-(R1) (nonlinear trend, constant variance function, independent ranges). Each example has the sample sizes $n = 100$, and we repeat each simulation combination 100 times and report the SiZer map with the majority voting scheme (e.g. if the pixel is flagged as increasing 90 times and not significant 10 times, then it is colored black).



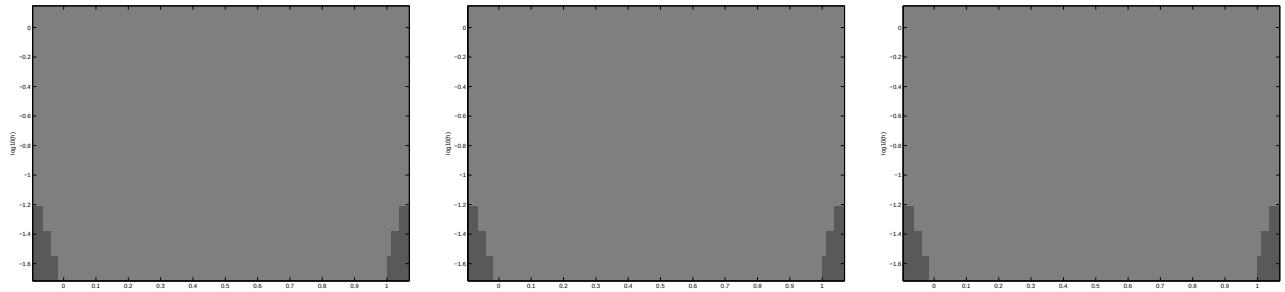
(a) MC (b) BMC-q (c) BMC-SD

Figure 4: Three proposed SiZer maps for the case of no trend (C1), constant variance function (V1), and independent ranges (R1).



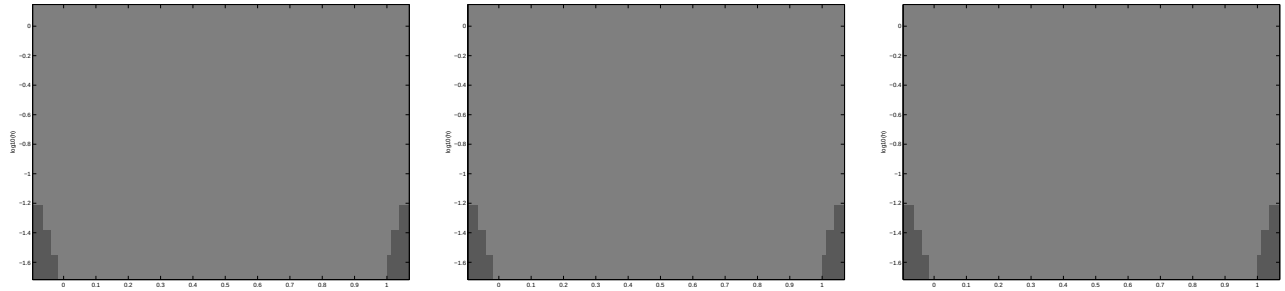
(a) MC (b) BMC-q (c) BMC-SD

Figure 5: Three proposed SiZer maps for the case of no trend (C1), quadratic variance function (V2), and independent ranges (R1).



(a) MC (b) BMC-q (c) BMC-SD

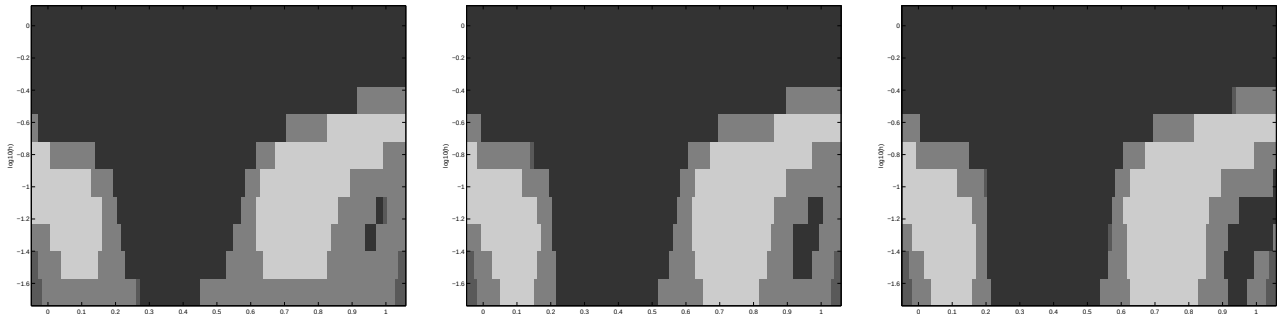
Figure 6: Three proposed SiZer maps for the case of no trend (C1), constant variance function (V1), and dependent ranges (R2).



(a) MC (b) BMC-q (c) BMC-SD

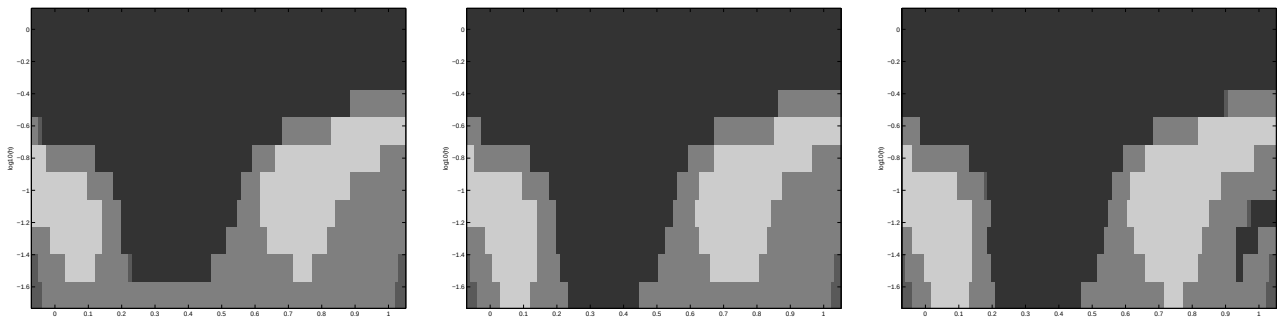
Figure 7: Three proposed SiZer maps for the case of no trend (C1), quadratic variance function (V2), and dependent ranges (R2).

In Figure 4, three proposed SiZer maps are depicted for the case of no trend (C1), constant variance function (V1), and independent ranges (R1). Almost all pixels are colored intermediate gray in the three maps, which provides strong evidence of no significant features across all locations and scales. This overall insignificance agrees with the expected result, as the regression function was designed not to display any trend. The darker gray in both bottom corners indicates that there are not sufficient intervals on those boundaries at small scales to make SiZer inference. Figures 5–7 show similar SiZer maps with different variance functions and ranges. From these plots we conclude that all three SiZer tools are robust to different types of variance functions and range relationships.



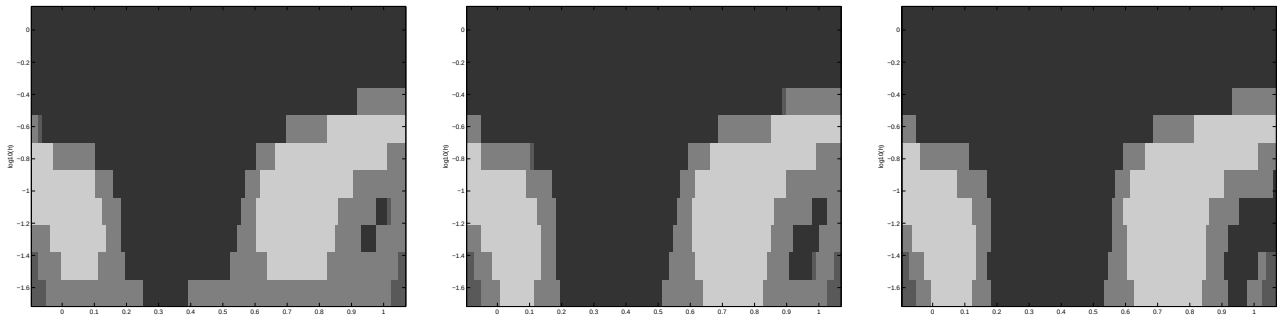
(a) MC (b) BMC-q (c) BMC-SD

Figure 8: Three proposed SiZer maps for the case of nonlinear trend (C2), constant variance function (V1), and independent ranges (R1).



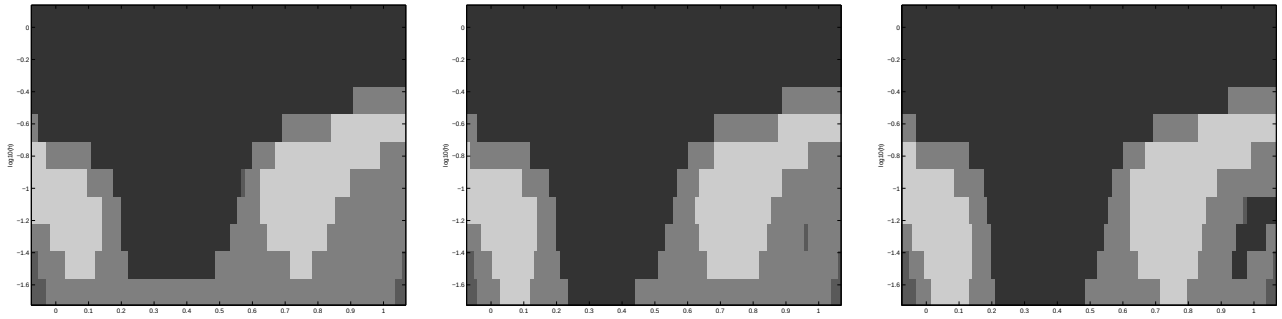
(a) MC (b) BMC-q (c) BMC-SD

Figure 9: Three proposed SiZer maps for the case of nonlinear trend (C2), quadratic variance function (V2), and independent ranges (R1).



(a) MC (b) BMC-q (c) BMC-SD

Figure 10: Three proposed SiZer maps for the case of nonlinear trend (C2), constant variance function (V1), and dependent ranges (R2).



(a) MC (b) BMC-q (c) BMC-SD

Figure 11: Three proposed SiZer maps for the case of nonlinear trend (C2), quadratic variance function (V2), and dependent ranges (R2).

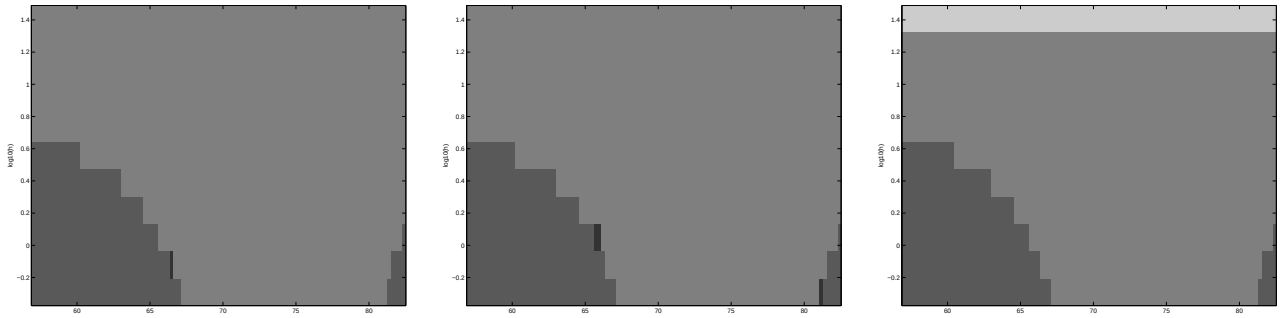
Figures 8-11 display three proposed SiZer maps when the center relationship is nonlinear. From the SiZer maps in Figure 8, it is evident that the overall increasing trend (black) is found at large scales and the oscillating trend (decreasing-increasing-decreasing-increasing) is found at smaller scales in all three maps. We note that the size of the significant feature depends on the degree of increasing/decreasing trend in the data (refer to Figure 3(b)). Although three SiZer maps suggest the similar conclusion, it can be seen that the MC SiZer shows the least significant features, and the BMC-SD SiZer shows the most. For the case of the quadratic variance function and independent ranges, the difference among three SiZer maps can be clearly noted because only the BMC-SD SiZer detects the small increasing trend around $x = 0.95$ while the other two SiZers fail to flag this feature as significant in their maps. It suggests that the standard deviation estimated from the bootstrap sampling is more robust to different types of variance functions when small features appear in the data than from the original estimate in (2.4). For the case of dependent ranges (Figures 10 and 11), we find similar results as in the independent case; i.e. more significant features are detected in the BMC-SD SiZer map than in the MC, and only the BMC-SD SiZer finds the small increasing trend around $x = 0.95$.

The simulation study demonstrates that the three proposed SiZer tools do not falsely find the features and can detect real features across multiple scales in various simulation settings. Among the three, it is clear that the BMC-SD SiZer has the most power.

4.2 Real Data Analysis

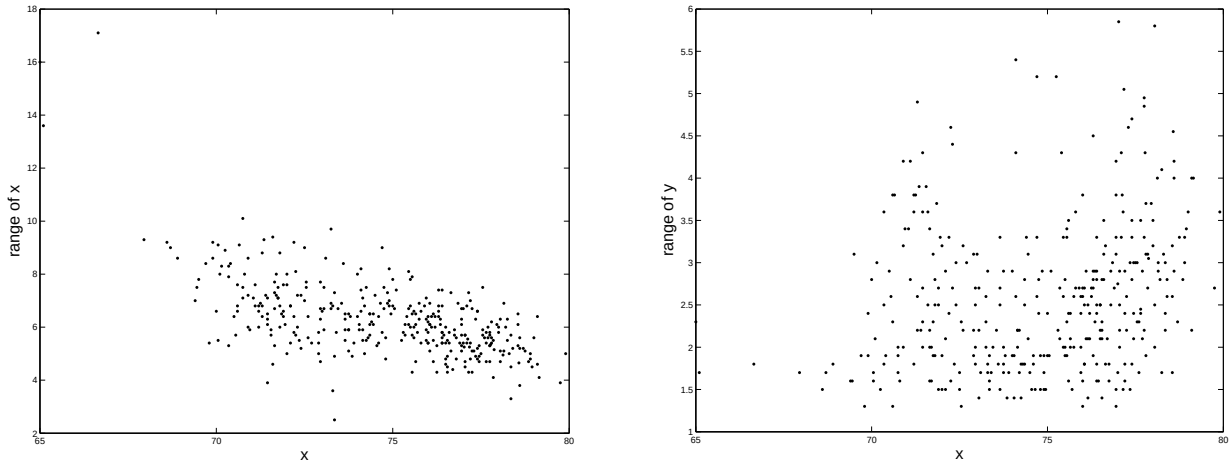
We analyze two real interval-valued examples in this section. As in Section 4.1, we compare three proposed SiZers, MC, BMC-q, and BMC-SD.

The first example regards the Hawaii climate data depicted in Figure 1. Note that the conventional SiZer analysis using the centers and ranges of the interval-valued is shown in Figure 2. Figure 12 displays three proposed SiZer maps for the data. Both the MC and BMC-q SiZers mostly show intermediate or darker gray colors, which indicates an absence of statistically significant features in the data. However, the BMC-SD SiZer colors the top row of the map white, indicating an overall decreasing trend at the coarsest scale, which agrees with the conventional SiZer using centers in Figure 2(a). However, none of the three SiZer maps show the feature found in the middle of Figure 2(a). This is also supported by Figure 13 because the range values of temperature and sea level pressure are large around temperature=72. Therefore, we conclude that the feature found in the middle might not



(a) MC (b) BMC-q (c) BMC-SD

Figure 12: Three proposed SiZer maps for the Hawaii climate data.



(a) Range of Temperature vs Temperature (b) Range of Sea Level Pressure vs Temperature

Figure 13: Ranges versus Temperature in the Hawaii climate data.

be real because of the internal variation. Regarding the overall decreasing trend found in the conventional SiZer map using centers, the conclusion is split between the three SiZers. One can argue that the decreasing trend may be spurious because the ranges of temperature shows an apparent decreasing trend as temperature increases in Figure 13(a), and this change of variation should be accounted for instead of interpreting the trend in centers only. On the other hand, BMC-SD SiZer, which attempts to accurately calculate the standard deviation in (2.3), flags the overall trend significant. Hence, this feature could be on the borderline of the significance and should be carefully examined in a future study.

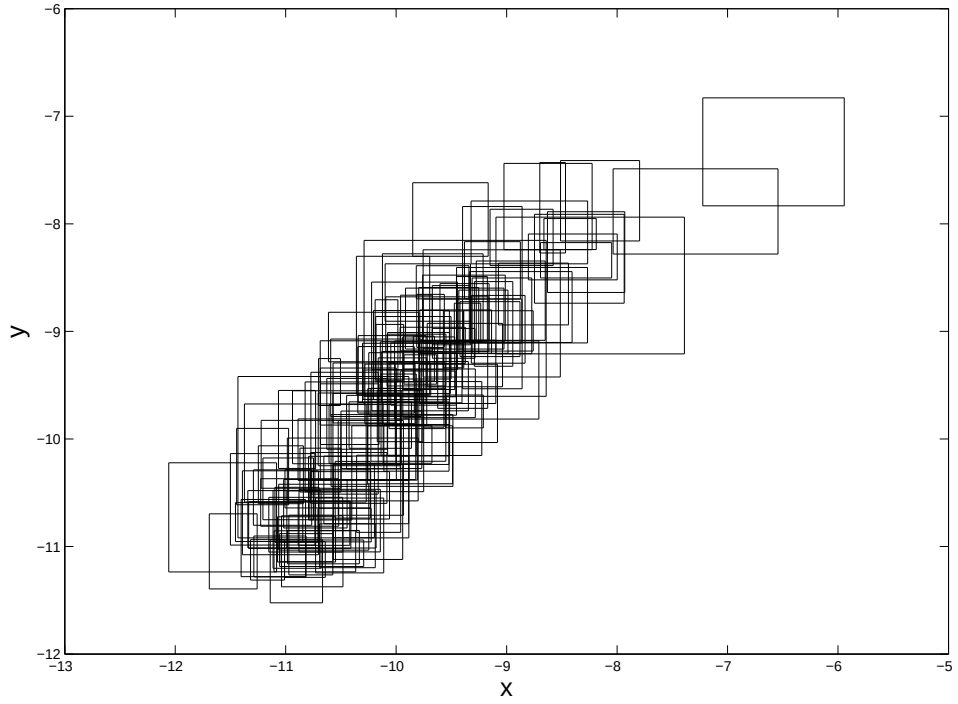
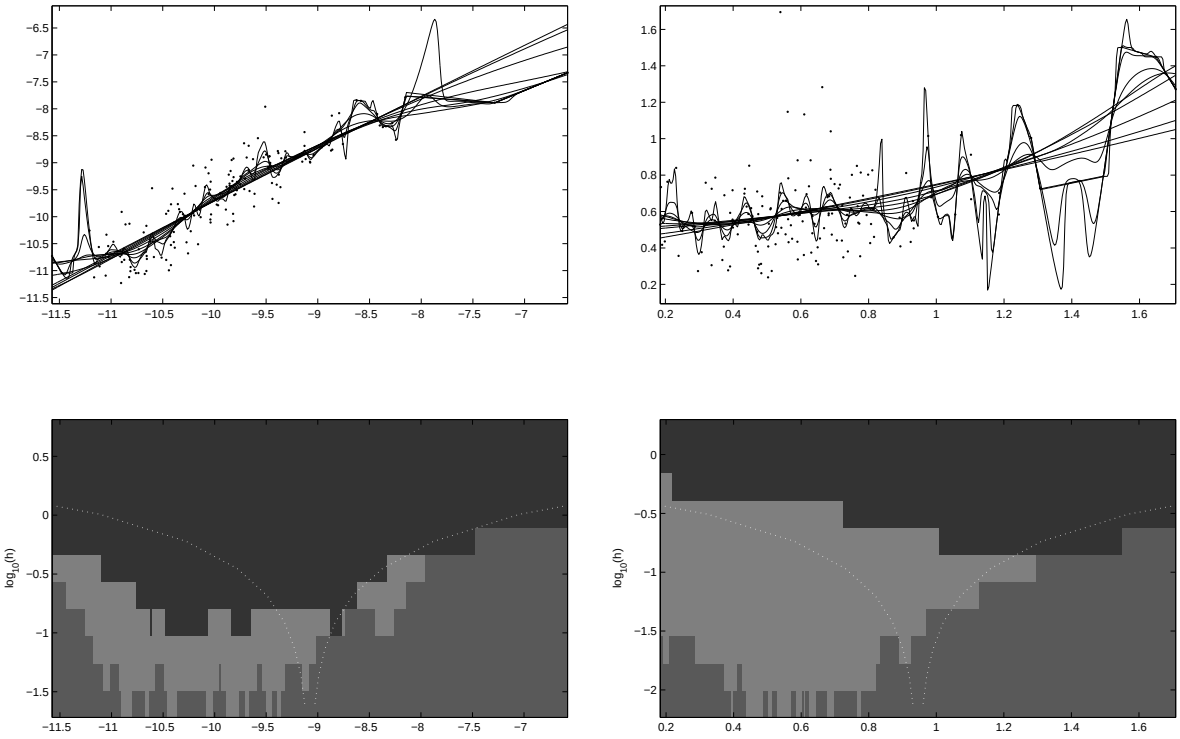


Figure 14: Display of stock indices interval-valued data.

In the second example, we study the relationship between two stock indices Dow Jones Industrials (DJ30) and Spanish IBEX (IBEX35) using the data analyzed in Cipollini et al. (2013). In particular, we investigate the relationship between the realized variances of the two stock indices, which is one of volatility measures in their paper. The original data are observed in the single-valued form from January 1996 to February 2009, which yields 3,411 daily realized variance observations. From these single-valued data we create 158 interval-valued observations by calculating the first and third quartiles of each month for both stock indices. Figure 14 depicts the aggregated interval-valued data on a logarithm scale. There

is a clear monotonically increasing trend between two realized variances. It is of interest if this trend is statistically significant or created by the artifacts of sampling noise or variation in the intervals.

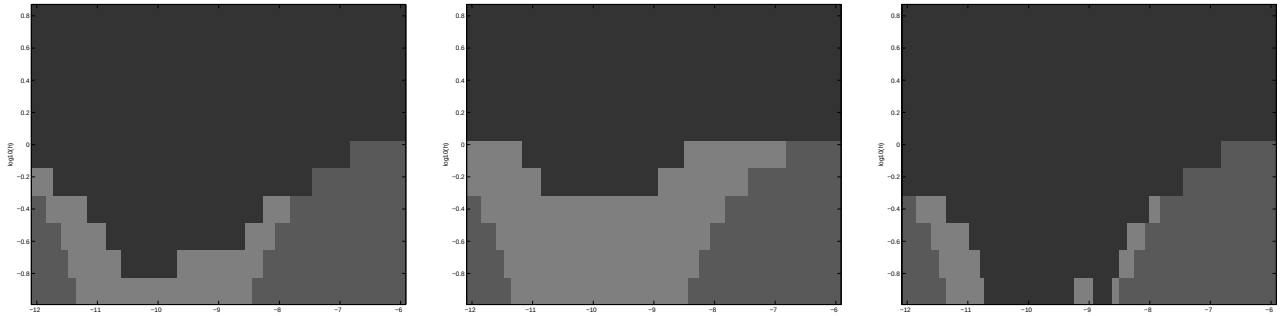


(a) Center (b) Range

Figure 15: We apply the conventional SiZer to the centers and the ranges of the stock indices data.

Figure 15(a) depicts SiZer plots using the mid-points of the interval-valued data to investigate the center relationship between two stock indices. In the top panel, the horizontal axis represents the logarithm of the realized variance for Dow Jones Industrial and the vertical axis for Spanish IBEX. The family of smooths in the top panel shows an overall increasing trend for all scales. In the SiZer map in the bottom panel, the pixels at large and medium scales are colored black, which confirms that the strong increasing trend across locations is statistically significant at those scales. The pattern in the SiZer map suggests a nonlinear trend because significant features appear from large to middle scales at the beginning, increase to small scales in the middle region, and then decrease in the later half. The pixels in the later half at small and medium scales are colored darker gray, which indicates insuf-

ficient data points for statistical decision. Figure 15(b) depicts SiZer plots using the ranges of the interval-valued data to investigate the range relationship between two stock indices. The family of smooths in the top panel shows an overall increasing trend, and the SiZer map shows its statistical significance at large scales. It can be seen that this trend is not as strong as that in the center relationship because it is not significant at medium scales. The oscillating trends at smaller scales in the top panel correspond to either intermediate or darker gray, which suggests that they are not significant or there are insufficient data points for statistical inference.



(a) MC (b) BMC-q (c) BMC-SD

Figure 16: Three proposed SiZer maps for the two stock indices data.

Figure 16 shows SiZer maps for three proposed tools. All three SiZer maps also show that there exists a strong increasing trend when accounting for variability within the intervals. The significant features in the BMC-q SiZer map are rather flat, which might imply a linear trend, whereas the MC and BMC-SD SiZer maps display similar patterns as that of the conventional SiZer for the centers. The BMC-SD SiZer shows the most significant features at the finer scales, which is consistent with the results from the simulation study in Section 4.1. We suggest a rigorous test of (non)linearity for interval-valued data as our future work. In conclusion, the SiZer analysis is able to detect the overall increasing relationship between Dow Jones Industrials and Spanish IBEX stock indices.

5 Discussion

We have proposed three exploratory data analysis tools for interval-valued data in scale-space. They all repeatedly utilize Monte Carlo sampling to account for the internal varia-

tion, and aggregate the estimates to produce SiZer maps that summarize statistical inference across multiple locations and scales. MC SiZer uses the known formulae for the quantile and standard deviation of the derivative estimate in the confidence band (2.3), which are originally derived for the conventional SiZer. These calculations might be inaccurate because they are derived based on the Gaussian approximation, and thus can be improved via bootstrap sampling. BMC-q SiZer obtains the quantile that controls the familywise error rate from the empirical distribution of the data, and BMC-SD SiZer directly calculates the standard deviation of $\hat{f}'_h(x)$ from the bootstrap samples. Our simulation study in Section 4 shows that BMC-SD SiZer is most powerful among the three SiZers, and MC SiZer is least powerful. We suggest a more thorough theoretical and empirical study of these three SiZers and a development of other SiZers for interval-valued data as our future work.

Another interesting future direction is to develop a scale-space tool based on the nonparametric approach proposed by Jeon et al. (2015). This approach directly works with interval-valued data rather than generateing single-valued data. It utilizes mixture of Gaussian densities to obtain the conditional distribution of the response given a predictor. Once this distribution is obtained, they estimate a regression function (trend) via expectation. Because the estimation depends on a smoothing parameter, a scale-space idea is applicable to this approach. If a scale-space idea were implemented, it could potentially find meaningful local features hidden in the data, as demonstrated in their paper. However, its implementation might encounter two serious obstacles. First, it is not straightforward to estimate the derivative of a regression function given the conditional distribution. It is even more challenging to theoretically derive the quantile or the standard deviation of the derivative. Second, a computational method such as bootstrapping can be used to construct a confidence interval, but the computational burden could become impractical because the approach is already computationally intensive for the estimation of a trend itself. This heavy computation might work against the purpose of exploratory data analysis. Nevertheless, a development of SiZer using the Jeon et al.'s approach would be an interesting and non-trivial research question, and a good addition to scale-space tools.

Acknowledgments

The second author was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2012R1A1A1012043). The third author was supported by Basic Science Research Pro-

gram through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2011-0012288).

References

Ahn, J., Peng, M., Park, C., and Jeon, Y. (2012). A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining*, 5:336–348.

Arroyo, J., Espínola, R., and Maté, C. (2011). Different approaches to forecast interval time series: A comparison in finance. *Computational Economics*, 37:169–191.

Bayr, T. and Dommenget, D. (2013). The tropospheric land-sea warming contrast as the driver of tropical sea level pressure changes. *Journal of Climate*, 26:1387–1402.

Billard, L. and Diday, E. (2000). Regression analysis for interval-valued data. In Kiers, H. A. L., Rassoon, J.-P., Groenen, P. J. F., and Schader, M., editors, *Data Analysis, Classification, and Related Methods*, pages 369–374. Springer-Verlag, Berlin.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the America Statistical Association*, 98:470–487.

Billard, L. and Diday, E. (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, New York.

Blanco-Fernández, A., Colubi, A., and González-Rodríguez, G. (2013). Linear regression analysis for interval-valued data based on set arithmetic: A review. In Borgelt, C., Gil, M. A., ao M.C. Sousa, J., and Verleysen, M., editors, *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, volume 285 of *Studies in Fuzziness and Soft Computing*, pages 19–31. Springer-Verlag, Berlin.

Blanco-Fernández, A., Corral, N., and González-Rodríguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics and Data Analysis*, 55:2568–2578.

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

Cipollini, F., Engle, R. F., and Gallo, G. M. (2013). Semiparametric vector MEM. *Journal of Applied Econometrics*, 28:1067–1086.

de Carvalho, F. A. T., Lima Neto, E. . A., and Tenorio, C. P. (2004). A new method to fit a linear regression model for interval-valued data. In Biundo, S., Fruhwirth, T., and Palm, G., editors, *Lecture Notes in Computer Science, KI2004 Advances in Artifical Intelligence*, pages 295–306. Springer-Verlag, Brelin.

Douzal-Chouakria, A., Billard, L., and Diday, E. (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining*, 4:229–246.

Duong, T., Cowling, A., Koch, I., and Wand, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics and Data Analysis*, 52:4225–4242.

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.

Erästö, P. and Holmström, L. (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.

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

Ganguli, B. and Wand, M. P. (2007). Feature significance in generalized additive models. *Statistics and Computing*, 17:179–192.

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.

Godtliebsen, F., Marron, J. S., and Chaudhuri, P. (2004). Statistical significance of features in digital images. *Image and Vision Computing*, 22:1093–1104.

Godtliebsen, F. and Øigård, T. A. (2005). A visual display device for significant features in complicated signals. *Computational Statistics and Data Analysis*, 48:317–343.

González-Manteiga, W., Martínez-Miranda, M., and Raya-Miranda, R. (2008). SiZer map for inference with additive models. *Statistics and Computing*, 18:297–312.

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.

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, 101:484–499.

Holmström, L. and Pasanen, L. (2012). Bayesian scale space analysis of differences in images. *Technometrics*, 54:16–29.

Horowitz, J. L. (2001). The bootstrap. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, pages 3159–3228. Elsevier, Amsterdam.

Jeon, Y., Ahn, J., and Park, C. (2015). A nonparametric kernel approach to interval-valued data analysis. *To appear in Technometrics*.

Kim, C. S. and Marron, J. S. (2006). SiZer for jump detection. *Journal of Nonparametric Statistics*, 18:13–20.

Lauro, C. N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational Statistics*, 15:73–87.

Le-Rademacher, J. and Billard, L. (2012). Symbolic-covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*, 21:413–432.

Li, R. and Marron, J. S. (2005). Local likelihood SiZer map. *Sankhya*, 67:476–498.

Lima Neto, E. A., Cordeiro, G. M., and de Carvalho, F. A. T. (2011). Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation*, 81:1727–1744.

Lima Neto, E. A., Cordeiro, G. M., de Carvalho, F. A. T., Anjos, U. U., and Costa, A. G. (2009). Bivariate generalized linear model for interval-valued variables. In *Proceedings 2009 IEEE International Joint Conference on Neural Networks*, volume 1, pages 2226–2229, Atlanta, USA.

Lima Neto, E. A. and de Carvalho, F. A. T. (2008). Center and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis*, 52:1500–1515.

Lima Neto, E. A., de Carvalho, F. A. T., and Tenorio, C. P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In Webb, G. and Yu, X., editors, *Lecture Notes in Computer Science, AI 2004 Advances in Artificial Intelligence*, pages 526–537. Springer-Verlag, Brelin.

Lima Neto, E. A. and de Carvalho F. A. T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis*, 54:333–347.

Lima Neto, E. A., de Carvalho F. A. T., and Freire, E. S. (2005). A new method to fit a linear regression model for interval-valued data. In Furbach, U., editor, *Lecture Notes in Computer Science, KI: Advances in Artificial Inteligence*, pages 92–106. Springer-Verlag, Brelin.

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

Maia, A. L. S., de Carvalho, F. A. T., and Ludermir, T. B. (2008). Forecasting models for interval-valued time series. *Neurocomputing*, 71:3344–3352.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. Springer, New York.

Marron, J. S. and de Uña Álvarez, J. (2004). SiZer for length biased, censored density and hazard estimation. *Journal of Statistical Planning and Inference*, 121:149–161.

Marron, J. S. and Zhang, J. T. (2005). SiZer for smoothing splines. *Computational Statistics*, 20:481–502.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, 4:157–170.

Øigård, T. A., Rue, H., and Godtliebsen, F. (2006). Bayesian multiscale analysis for time series data. *Computational Statistics and Data Analysis*, 51:1719–1730.

Palumbo, F. and Lauro, C. (2003). A PCA for interval-valued data based on midpoints and radii. In Yanai, H., Okada, A., Shigemasu, K., Kano, Y., and Meulman, J., editors, *New Developments in Psychometrics*, pages 641–648. Springer, Tokyo.

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.

Park, C., Hannig, J., and Kang, K. H. (2009a). Improved SiZer for time series. *Statistica Sinica*, 19:1511–1530.

Park, C., Hannig, J., and Kang, K. H. (2015). Nonparametric comparison of multiple regression curves in scale-space. *To appear in Journal of Computational and Graphical Statistics*.

Park, C. and Huh, J. (2013). Statistical inference and visualization in scale-space using local likelihood. *Computational Statistics and Data Analysis*, 57:336–348.

Park, C. and Kang, K. H. (2008). SiZer analysis for the comparison of regression curves. *Computational Statistics and Data Analysis*, 52:3954–3970.

Park, C., Lee, T., and Hannig, J. (2010). Multiscale exploratory analysis of regression quantiles using quantile SiZer. *Journal of Computational and Graphical Statistics*, 19:497–513.

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.

Park, C., Vaughan, A., Hannig, J., and Kang, K. H. (2009b). SiZer for the comparison of time series. *Journal of Statistical Planning and Inference*, 139:3974–3988.

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.

Silva, A. O., Lima Neto, E. A., and Anjos, U. U. (2011). A regression model to interval-valued variables based on copula approach. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland.

Sorbye, S. H., Hindberg, K., Olsen, L., and Rue, H. (2009). Bayesian multiscale feature detection of log-spectral densities. *Computational Statistics and Data Analysis*, 53:3746–3754.

Vaughan, A., Jun, M., and Park, C. (2012). Statistical inference and visualization in scale-space for spatially dependent images. *Journal of the Korean Statistical Society*, 41:115–135.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35:1849–1873.

Yang, C.-Y., Jeng, J.-T., Chuang, C.-C., and Tao, C. (2011). Constructing the linear regression models for the symbolic interval-values data using PSO algorithm. In *System Science and Engineering (ICSSE), 2011 International Conference on*, pages 177–181. IEEE.