Nonparametric Comparison of Multiple Regression Curves in Scale-Space

Cheolwoo Park[*] Jan Hannig[†] Kee-Hoon Kang[‡]

Abstract

This paper concerns testing the equality of multiple curves in a nonparametric regression context. The proposed test forms an ANOVA type test statistic based on kernel smoothing and examines the ratio of between and within group variations. The empirical distribution of the test statistic is derived using a permutation test. Unlike traditional kernel smoothing approaches, the test is conducted in scale-space so that it does not require the selection of an optimal smoothing level, but instead considers a wide range of scales. The proposed method also visualizes its testing results as a color map and graphically summarizes the statistical differences between curves across multiple locations and scales. A numerical study using simulated and real examples is conducted to demonstrate the finite sample performance of the proposed method.

Keywords: Comparison of multiple curves, Kernel smoothing, Scale-Space, Visualization.

1 Introduction

The comparison of several populations often arises in function estimation such as densities, regression curves and survival functions. Among them, the problem of testing the equality of nonparametric regression curves has been widely studied. Suppose that we have k different samples and $n = \sum_{i=1}^{k} n_i$ independent observations (X_{ij}, Y_{ij}) from the nonparametric regression models:

$$Y_{ij} = f_i(X_{ij}) + \sigma_i(X_{ij})\varepsilon_{ij}, \quad j = 1, \dots, n_i, \ i = 1, \dots, k, \tag{1.1}$$

where X_{ij}'s are covariates, ε_{ij}'s are independently distributed random errors with mean 0 and variance 1; $f_i(X_i) = E(Y_i|X_i)$ is the unknown regression function and $\sigma_i^2(X_i) = Var(Y_i|X_i)$ is

[*] Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A. Email: cpark@uga.edu

[†] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, U.S.A. Email: hannig@email.unc.edu

[‡] Corresponding author. Department of Statistics, Hankuk University of Foreign Studies, Yongin, 449-791, Korea. Email: khkang@hufs.ac.kr

the conditional variance function of the ith sample ($i = 1, \ldots, k$). The general interests of the problems are "are the functions in the model (1.1) really different?", if then, "can we identify the locations where the differences are?" Therefore, the equality of k regression curves at a location x is hypothesized as

$$H_0 : f_1(x) = f_2(x) = \cdots = f_k(x) = f(x) \quad \text{vs. } H_1 : f_i(x) \neq f_j(x) \;\; \text{for some } x \text{ and } \quad i \neq j. \qquad (1.2)$$

For a motivation of the proposed work, we introduce two real examples analyzed in Section 4. The first example concerns monthly household expenditures on several commodities in Dutch guilders. The data were collected from April 1984 to September 1987, and the average was taken over the 42 months for each household. The dataset is divided into three groups by the number of members in the household: two, three or four members. We analyze the data using the model (1.1) with $k = 3$. We are interested in comparing the relationship between the expenditure on food and the total monthly expenditure for different family sizes. The second example is a random sample from the working population in Belgium for the year 1994. The data set consists of information on 893 males and 579 females. Our interest is to compare the relationship between wage (gross hourly wage rate in euro) on a log scale and years of experience across five different education levels, i.e. $k = 5$ in the model (1.1).

Much work has been done on this testing problem in nonparametric contexts. Hall and Hart (1990) proposed a bootstrap test for the comparison of two regression curves. Härdle and Marron (1990) took a semiparametric approach based on kernel smoothing under shape invariance. Delgado (1993); Kulasekera (1995); Kulasekera and Wang (1997); Neumeyer and Dette (2003) utilized empirical process approaches to investigate the overall inequality. Bowman and Young (1996) used kernel-based reference bands for detecting a difference in two regression curves. Pardo-Fernández et al. (2007) proposed two types of test statistics that are based on the estimation of the distribution of the residuals in two populations. We note that most of these approaches involve selecting an appropriate smoothing level, which has been a hurdle to the application of smoothers (see e.g. Chaudhuri and Marron (1999)). However, it might be difficult to for a single bandwidth to accurately estimate the multiple curves especially when each function has a different degree of smoothness. In addition, the optimal bandwidth for curve estimation could be different from the one for curve comparison, which might result in missing important local differences among the curves. Therefore, a unitary bandwidth value is incapable of accurately estimating the coefficients with different degrees of smoothness and adequately discovering the scale-dependent variations in the regression relationship.

In this paper, we take a kernel-based nonparametric approach for the comparison of multiple curves. The main key that distinguishes the proposed approach from existing methods is to investigate the differences of two or more regression curves at multiple locations and resolutions using a so-called *scale-space* approach (Lindeberg, 1994). Then, the testing results are summarized as a visual map, called *SiZer map*, to allow data analysts an easy interpretation. Chaudhuri and Marron (1999) proposed SiZer (SIgnificant ZERo crossing of the derivatives) as a scale-space-based exploratory data analysis tool for finding meaningful features in a single curve. The core of the SiZer approach is to simultaneously study a curve at different smoothing scales instead of trying to find the true underlying curve. At each scale, SiZer addresses the question of which features of the smoothed curve at that particular scale represent a statistically significant structure. This approach circumvents the difficulty of determining the optimal smoothing level and allows one to extract all the information that is available at each individual level of scale.

SiZer tools have been extensively developed and applied to various fields. Hannig and Marron (2006) improved SiZer inference to reduce type I errors. Hannig and Lee (2006) proposed a robust version of SiZer that examines the median regression function and later Park et al. (2010) extended it to the quantile function. SiZer has been also applied to time series data (Park et al., 2004; Rondonotti et al., 2007; Park et al., 2007, 2009a). In addition, 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) and additive models (González-Manteiga et al., 2008). Additionally, various Bayesian versions of SiZer have also been proposed as an approach to Bayesian multiscale smoothing (Erästö and Holmström, 2005; Godtliebsen and Oigard, 2005; Oigard et al., 2006; Erästö and Holmström, 2007; Sørbye et al., 2009). The scale-space has been also extended to two dimensions; Ganguli and Wand (2007) considered a additive model for generalized linear models, and Godtliebsen et al. (2002) and Duong et al. (2008) studied a density estimation. Godtliebsen et al. (2004) analyzed image data under the independent errors and Vaughan et al. (2012) extended it to the spatially dependent case. Note that these SiZer tools aim to discover the important features in a single curve. Recently, Park and Kang (2008) and Park et al. (2009b) proposed SiZer tools for comparing multiple curves with independent and dependent errors, respectively. The basic idea is to convert the problem into the comparison of two curves using residuals. One shortcoming of this approach is that the local and scale information of the differences among the curves is lost because the sets of residual curves are compared each other rather than the original curves.

The objective of this paper is to develop a SiZer tool that is capable of directly comparing multiple curves. The proposed SiZer is based on an ANOVA type test statistic and simultaneously investigates the differences of the smoothed curves for a wide range of scales. The testing results are summarized in a SiZer map to visualize the statistically significant differences among these curves. This approach enables one to compare several curves directly and get the information on their local differences at different scales, which reflects the original formulation of SiZer.

The remainder of the paper is organized as follows. Section 2 proposes SiZer for the comparison of multiple curves. Section 3 conducts a simulation study for the proposed tool. In Section 4, two real examples are analyzed using the proposed SiZer. Appendix provides derivation of the test statistic introduced in Section 2.

2 Proposed SiZer

In this section, we propose a SiZer tool for simultaneously comparing k regression curves. In the proposed SiZer, which utilizes kernel smoothing techniques, the main focus lies in investigating the smoothed curves indexed by the bandwidth h instead of the true underlying curve. Therefore, we consider the following null hypothesis in scale-space.

$$H_0 : f_{1,h}(x) = f_{2,h}(x) = \cdots = f_{k,h}(x) = f_h(x) \tag{2.1}$$

where

$$f_{i,h}(x) = \int f_i(u) K_h(x-u)du.$$

Here, h controls the smoothing level, $K_h(\cdot) = K(\cdot/h)/h$ and K is a symmetric density function.

Under the model (1.1), let

$$\hat{f}_{i,h}(x) = \sum_{j=1}^{n_i} K_{ij}(x) Y_{ij}$$

be a local polynomial estimator (Fan and Gijbels, 1996) of the regression function f_i in (1.1) based on the ith sample. Among local polynomial estimators local constant and local linear are a popular choice. For the local constant estimator, also known as Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964), the weights are given as

$$K_{ij}(x) = \frac{K_h(x - X_{ij})}{\sum_{j=1}^{n_i} K_h(x - X_{ij})}.$$

The weights of the local linear estimator are given as

$$K_{ij}(x) = \frac{S_{i,2}(x)K_h(x - X_{ij}) - S_{i,1}(x)(x - X_{ij})K_h(x - X_{ij})}{S_{i,0}(x)S_{i,2}(x) - (S_{i,1}(x))^2}$$

where $S_{i,l}(x) = \sum_{j=1}^{n_i}(x - X_{ij})^l K_h(x - X_{ij})$. Also, let

$$\hat{f}_h(x) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \tilde{K}_{ij}(x)Y_{ij}$$

be the pooled estimator of the common regression function $f(x)$ under the null hypothesis (1.2). For the local constant estimator $\tilde{K}_{ij}(x) = K_{ij}(x)$ and

$$\tilde{K}_{ij}(x) = \frac{S_2(x)K_h(x - X_{ij}) - S_1(x)(x - X_{ij})K_h(x - X_{ij})}{S_0(x)S_2(x) - (S_2(x))^2}$$

where $S_l(x) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x - X_{ij})^l K_h(x - X_{ij})$ for the local linear estimator. The local constant estimator is easy to calculate, but it is known that it has a boundary issue (Fan and Gijbels, 1996). In our numerical analysis in Sections 3 and 4, we use the local linear estimator.

In comparing the k smoothed curves, we apply a similar idea as in ANOVA in which the means of multiple normal populations are compared using the ratio of between and within group variations. This motivates us to consider the following test statistic for the null hypothesis (2.1) under the model (1.1):

$$F_h(x) = \frac{\sum_{i=1}^{k}(\hat{f}_{i,h}(x) - \hat{f}_h(x))^2/(c_1 \cdot df_1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \hat{f}_{i,h}(X_{ij}))^2 K_h(x - X_{ij})/(c_2 \cdot df_2)}, \tag{2.2}$$

where (c_1, df_1) and (c_2, df_2) are scale factors and degrees of freedom. In the test statistic $F_h(x)$, it can be seen that the numerator measures the variation between k groups and the denominator measures the variation within groups.

The numerator and denominator of the test statistic (2.2) may be represented by the quadratic form with suitable weight matrices. Therefore, we obtain the degrees of freedom d_1 and d_2, and scale factors c_1 and c_2 using the Satterthwaite approximation (Satterthwaite, 1946), i.e., by solving the equation related to the first two moments conditions. We provide its details for both local constant and local linear estimators in the Appendix.

Using the $F_h(x)$ statistic in (2.2), we simultaneously test (2.1) at multiple locations, say $x_1, \ldots, x_g$ for a given h. In order to find a critical value at each scale, we approximate $P(\max_{l=1,\ldots,g} F_h(x_l) < q_h)$ using a permutation idea. We obtain the empirical distribution of the maximum of the pointwise F's to address the multiple comparisons adjustment. In what follows we illustrate how to empirically determine a critical value q_h. For a given h,

(i) pool the k datasets and permute them;

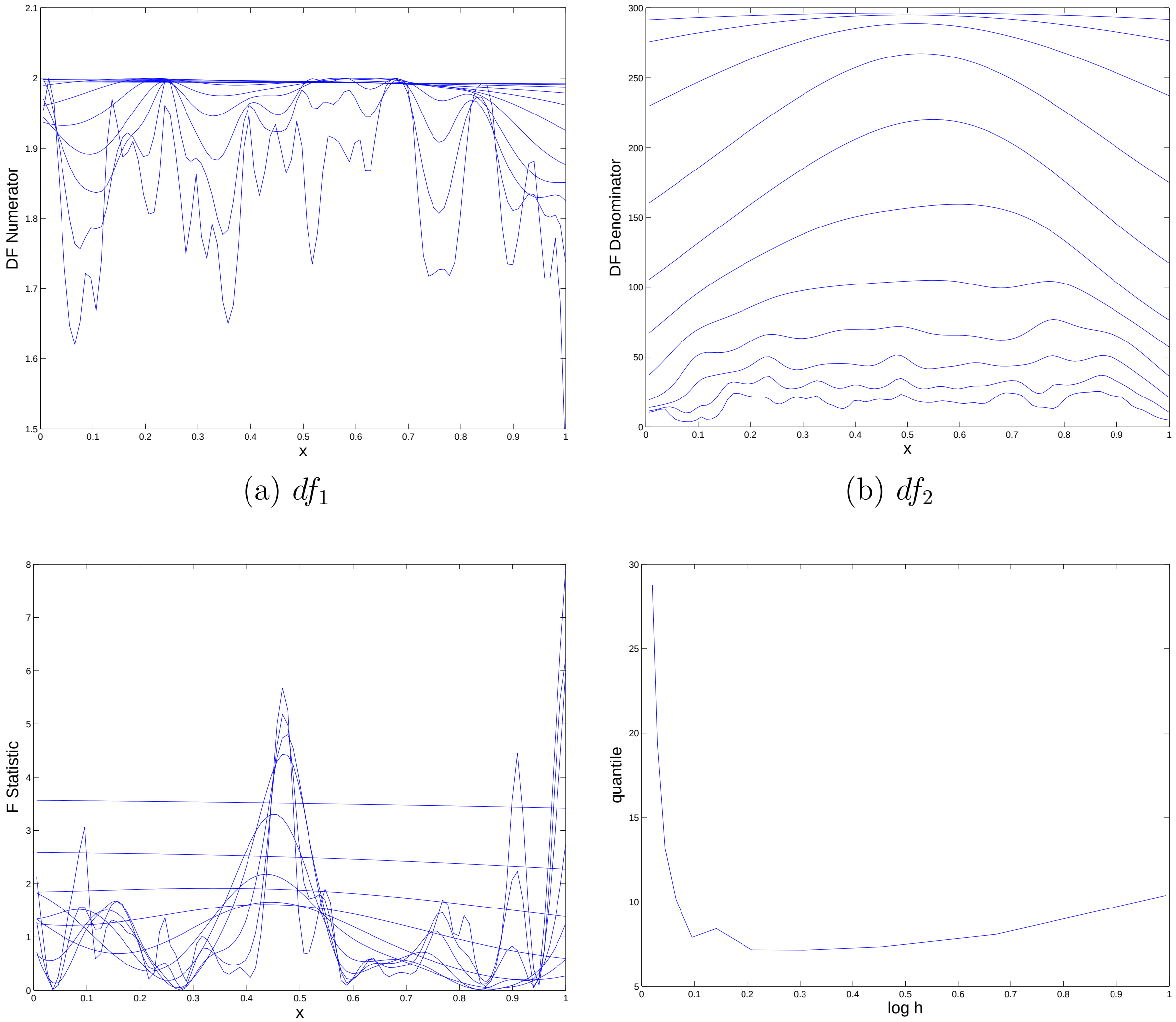
(ii) using the permuted data, calculate $\max_l F_h(x_l)$;

(iii) repeat (i) and (ii) B times;

(iv) using the B repetitions, obtain the empirical distribution of $\max_l F_h(x_l)$ and $100(1-\alpha)\%$ quantile, q_h.

In our numerical study, we use $\alpha = 0.05$ and $B = 1000$. Also, the bandwidths used in our numerical examples are 11 equally spaced values on a logarithmic scale of the range of x.

For illustration purpose, we generate three datasets from $N(0,1)$ with $n_1 = n_2 = n_3 = 100$ to investigate the proposed algorithm more carefully. For simplicity, we use the local constant estimator with $B = 100$. In Figure 1(a), the degrees of freedom of the numerator df_1 are graphed with different bandwidths. For small bandwidths, df_1s are smaller and more wiggly on the grid points $x_1, \ldots, x_g$. On the other hand, df_1 gets closer to 2 regardless of x as h increases. Note that smoothing with the largest bandwidth approximates the simple average of the response variables for each group, which essentially corresponds to ANOVA and its degrees of freedom for the between groups is $k-1=2$ in this case. In Figure 1(b), the degrees of freedom of the denominator df_2 are graphed with different bandwidths. While df_2s are also small and wiggly for small bandwidths, they get close to 300 as h increases. We again note that the degrees of freedom of ANOVA for the within groups corresponds to $n-k=297$ in this case. Figure 1(c) displays the $F_h(x)$ statistic defined in (2.2) for the simulated data. Finally, Figure 1(d) shows the estimated quantile with different h using the permutation idea introduced above. The estimate quantile values are under 10 except for the first a few small bandwidths. The significance of the differences of the three



(a) df_1

(b) df_2

(c) F_h statistic

(d) quantile q_h

Figure 1: We generate three datasets from $N(0,1)$ with $n_1 = n_2 = n_3 = 100$. We visually display (a) df_1 (b) df_2 and (c) $F_h(x)$ in equation (2.2) with the local constant estimator, against the grid points x for different h using permutation with $B = 100$. The plots show how the degrees of freedom and the test statistic change along with location x and scale h. Note that the curves in (a), (b), and (c) become smoother with greater average df_1, df_2, or F_h as the bandwidth h increases. We also draw the quantile (d) q_h vs. $\log h$ for this simulated setting. Statistical significance is determined by comparing the $F_h(x)$ and q_h for each x and h.

curves are determined by comparing $F_h(x)$ and q_h each other at a particular location and scale (x, h). It can be seen that the value of $F_h(x)$ is always less than q_h at any location for a given smoothing level, and thus no significant features would be found for this simulation. Given that the three curves are generated from the same model, this result demonstrates the accuracy of the proposed SiZer tool. More simulated examples will be illustrated in Section 3.

SiZer summarizes the result of the series of tests at (x, h) as a colored map called a *SiZer map*. The variation of colors in a SiZer map provides the statistical evidence for the differences in the curves for different scales. At each (x, h), if the test statistic is equal to or above the quantile value q_h, which means that the curves are significantly different one another, then the pixel is colored white. On the other hand, if the test statistic is less than q_h, which means that the curves are not significantly different, then that particular map location is given gray. There is one more color in a SiZer map when there are not sufficient data points for statistical decision. In that case, the test is not conducted and the location is colored darker gray. To determine these gray areas, we define the estimated effective sample size (ESS, Chaudhuri and Marron (1999)) for each (x, h) as

$$ESS(x, h) = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} K_h(x - X_{ij})}{K_h(0)}.$$

If ESS$(x, h) < 5k$, then the corresponding pixel is colored darker gray.

If the null hypothesis in (2.1) is rejected, one would be interested in pairwise comparisons of the curves. This can be done by SiZer for comparing two curves developed in Park and Kang (2008).

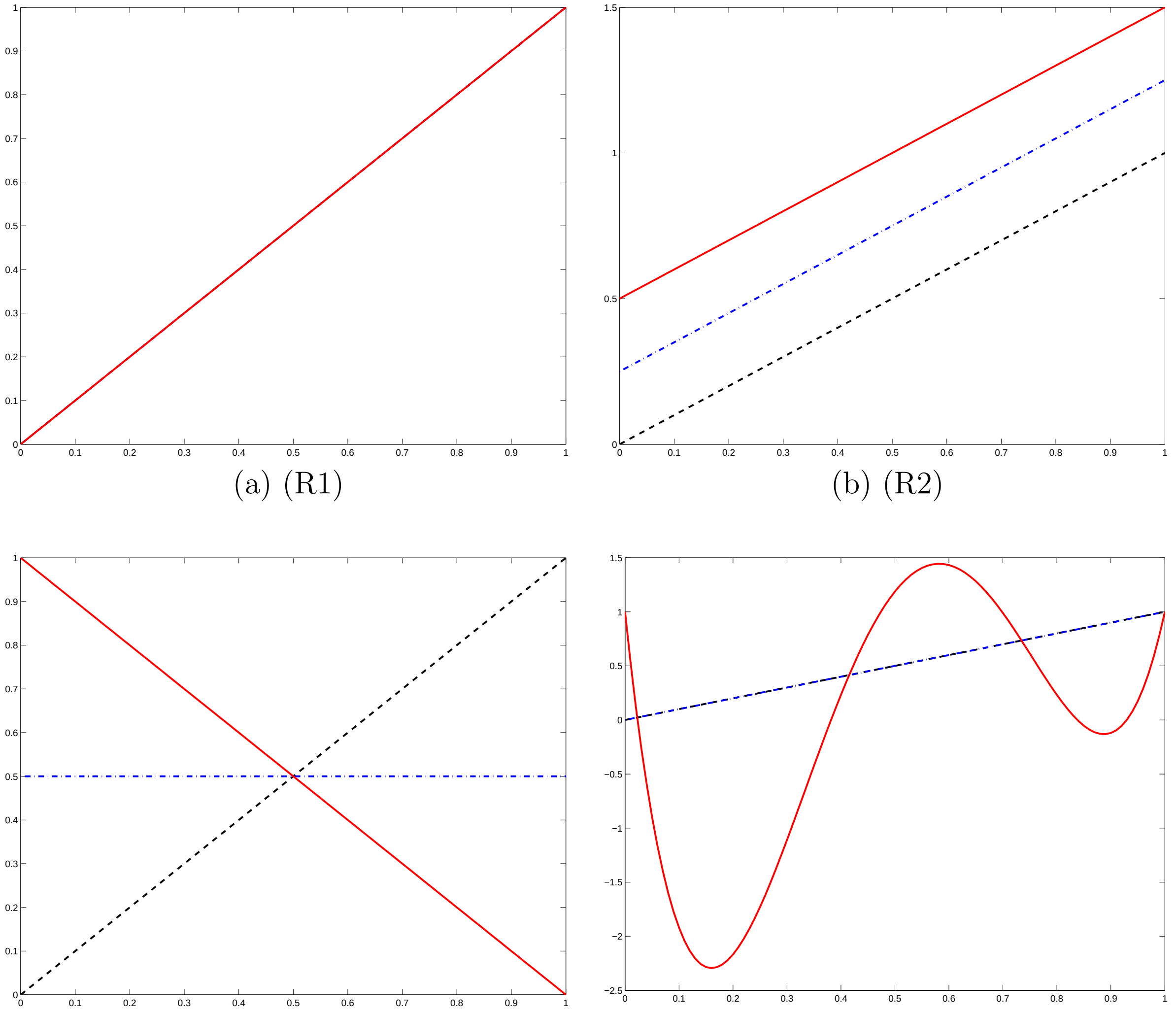
3 Simulation

This section demonstrates the finite sample performance of the proposed scale-space tool under various simulation settings. The simulated data are generated by the model in (1.1) with three samples, i.e., $k = 3$. The regression functions f_1, f_2, and f_3 are chosen from the following four groups of the functions:

(R1) $f_1(x) = f_2(x) = f_3(x) = x$

(R2) $f_1(x) = x$, $f_2(x) = x + 0.25$, $f_3(x) = x + 0.5$

(R3) $f_1(x) = x$, $f_2(x) = 0.5$, $f_3(x) = 1 - x$



(a) (R1)

(b) (R2)

(c) (R3)

(d) (R4)

Figure 2: Four different regression functions used in the simulation.

(R4) $f_1(x) = f_2(x) = x$, $f_3(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$.

These four groups are displayed in Figure 2.

The variance functions σ_1^2, σ_2^2, and σ_3^2 are also chosen from the following four groups of the functions:

(V1) $\sigma_1^2(x) = \sigma_2^2(x) = \sigma_3^2(x) = 0.5$

(V2) $\sigma_1^2(x) = \sigma_2^2(x) = \sigma_3^2(x) = 0.5(0.5 + 2x)$

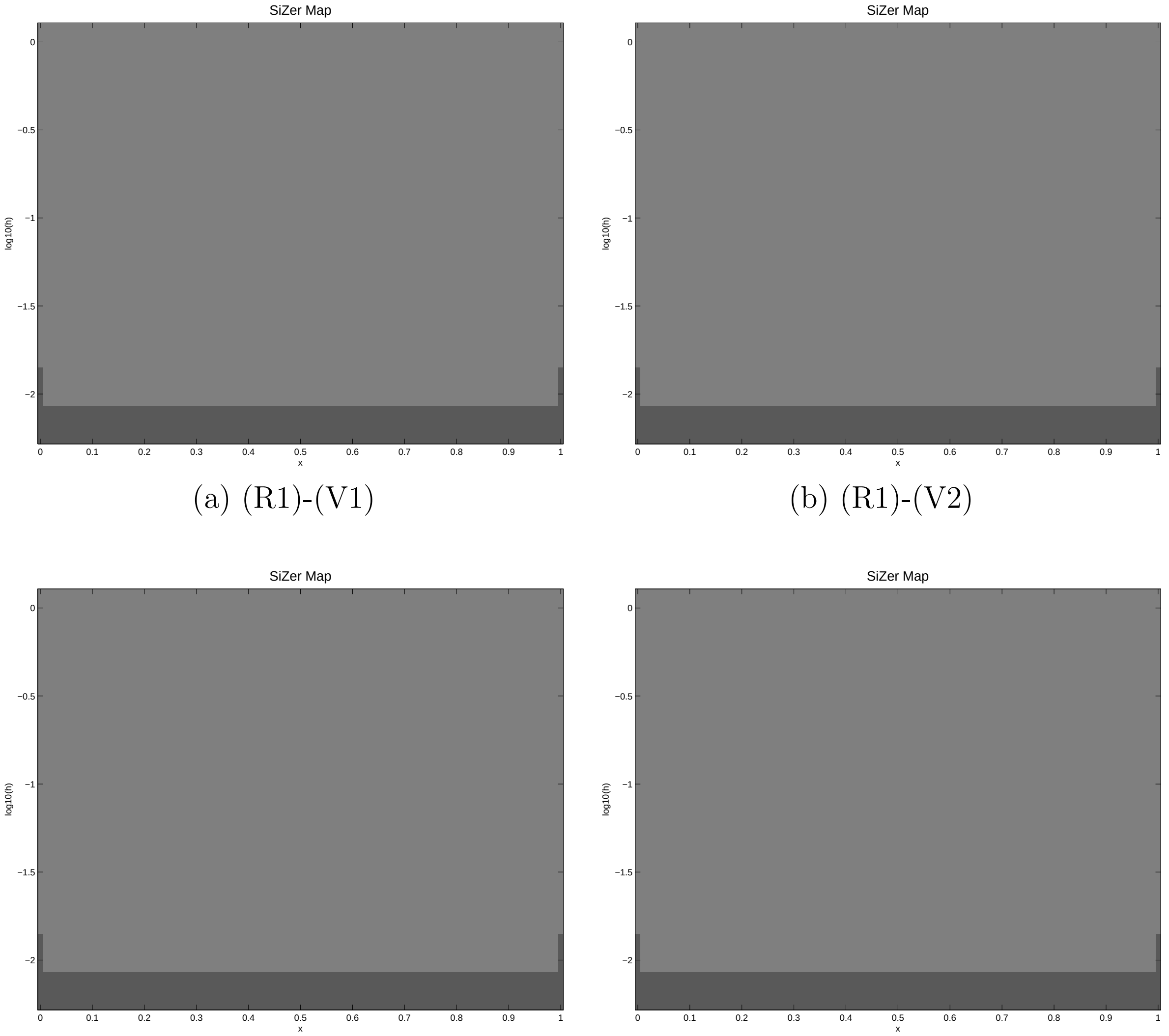
(V3) $\sigma_1^2(x) = \sigma_2^2(x) = \sigma_3^2(x) = 0.5(2.5 - 2x)$

(V4) $\sigma_1^2(x) = \sigma_2^2(x) = \sigma_3^2(x) = 0.5(-4x^2 + 4x + .5)$.

Pardo-Fernández et al. (2007) considered the first three regression functions (R1)-(R3) with the constant variance (V1), and Park et al. (2010) used the fourth case (R4) of the regression functions with the four variance functions (V1)-(V4). In the model, X_{1j}, X_{2j} and X_{3j} are generated from the uniform distribution on $(0, 1)$ independently. Also, each example has the sample sizes $n_1 = 300$, $n_2 = 400$, and $n_3 = 500$. We repeat each combination of regression and variance functions for 100 times with $B = 1000$ and report the average SiZer map over 100 repetitions for each combination, which graphically presents the testing results of comparing three regression curves. At each iteration, each pixel in a SiZer map takes one of the three values: 1 for indecisive, 2 for insignificant, and 3 for significant feature. An average SiZer map is created by taking the mean of the 100 values at each pixel.

In Figure 3, the four average SiZer maps are depicted for the case of the regression function (R1) with four different variance functions. Almost all pixels are colored gray in the four maps, which provides strong evidence of no significant difference across the entire locations and scales. This is the correct decision because the three regression functions share the same linear trend in Figure 2(a). We also conclude that the SiZer inference is rather insensitive to different types of variance functions in this example.

From the average SiZer maps in Figure 4, we can observe that the significant features (white) are consistently found at middle and large scales in the four maps, which correctly reveals the overall difference of the three curves in Figure 2(b). It is noted, however, that the SiZer map for (V2) (and (V3)) fails to flag significant features at the end (beginning, respectively). It suggests



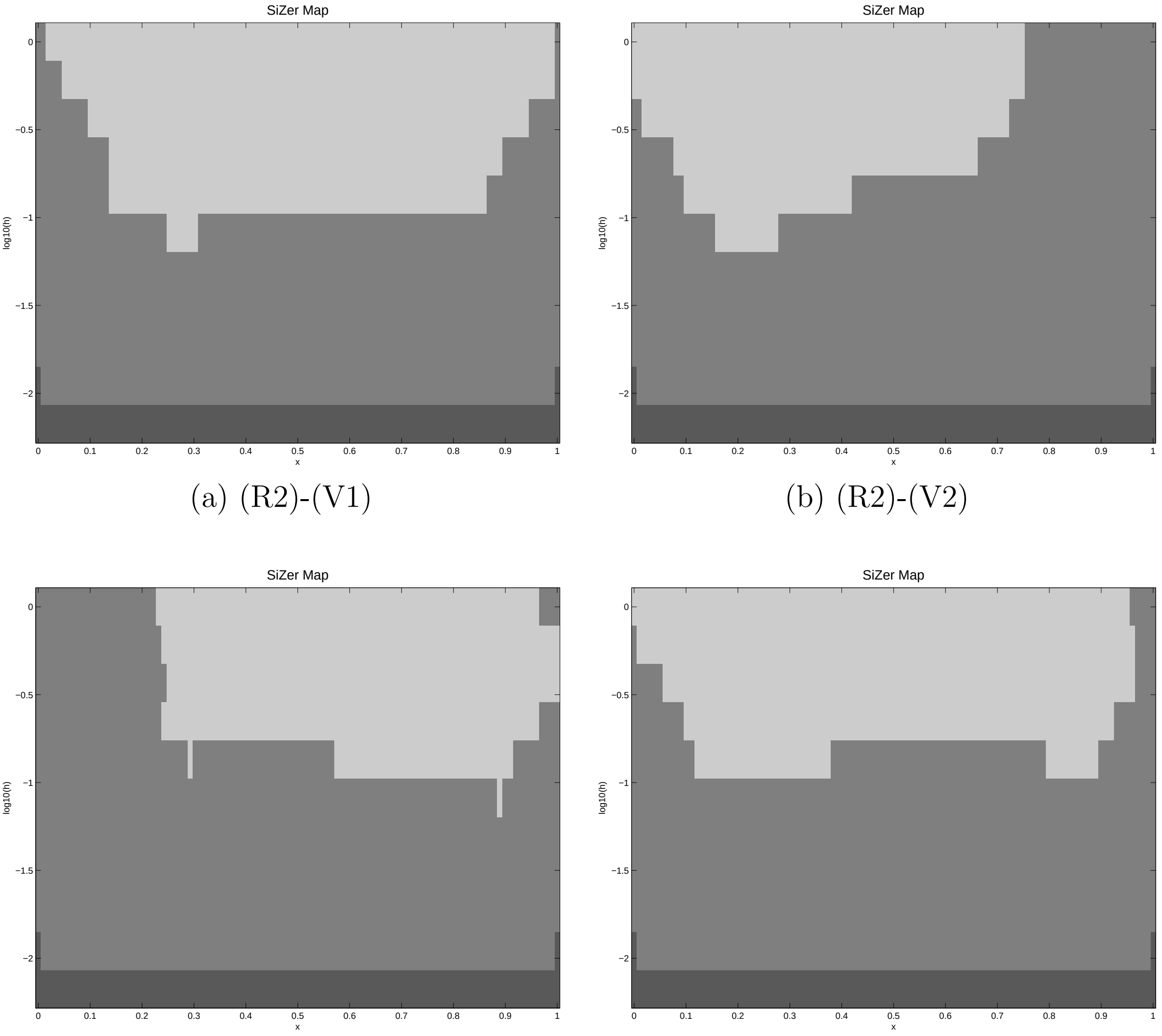
(a) (R1)-(V1)

(b) (R1)-(V2)

(c) (R1)-(V3)

(d) (R1)-(V4)

Figure 3: The average SiZer maps for the regression function (R1) with four different variance functions (V1)-(V4). Almost all pixels are colored gray in the four maps, which indicates no significant different among the three curves.



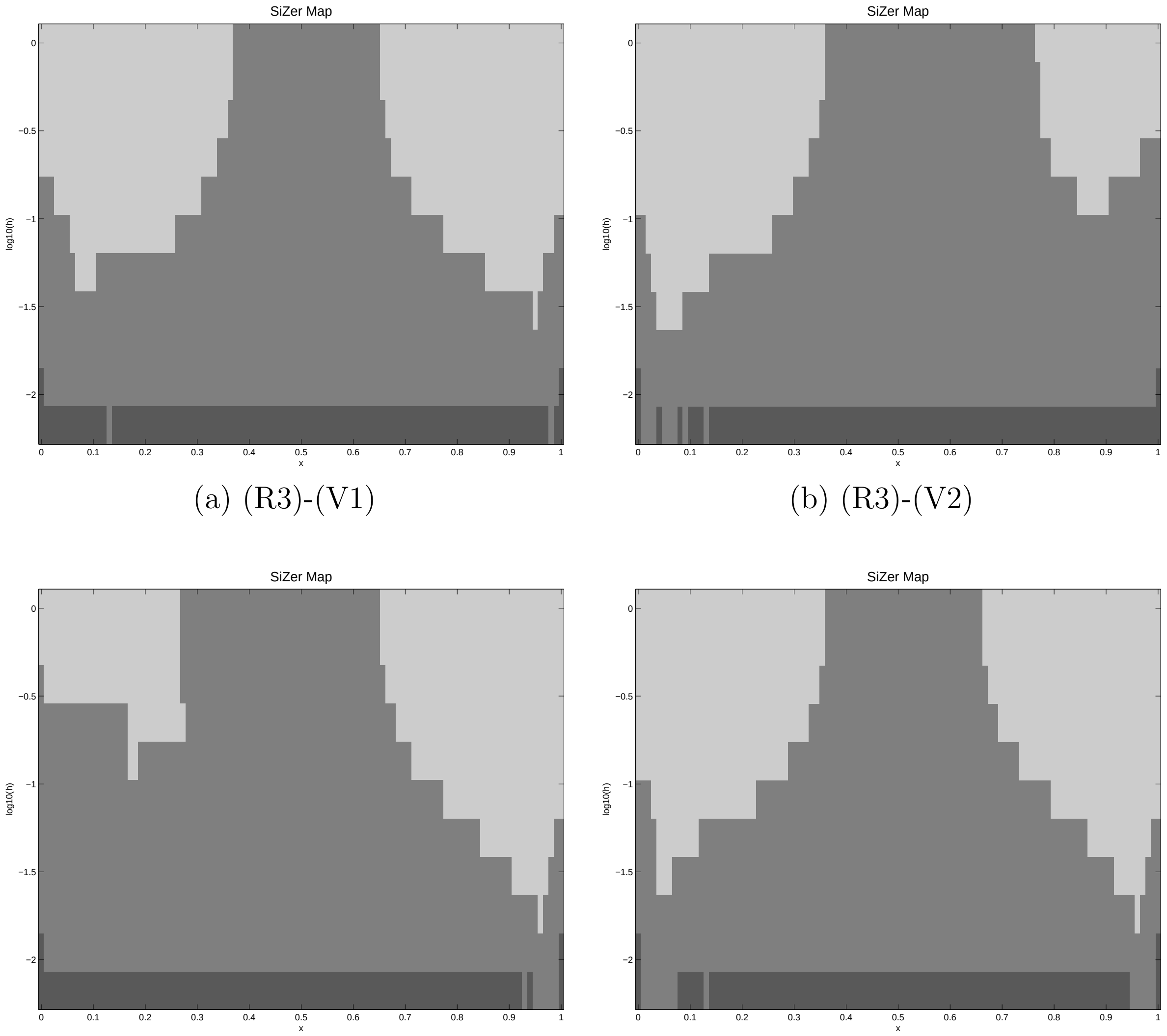
(a) (R2)-(V1)

(b) (R2)-(V2)

(c) (R2)-(V3)

(d) (R2)-(V4)

Figure 4: The average SiZer maps for the regression function (R2) with four different variance functions (V1)-(V4). The significant features (white) are consistently found at middle and large scales in the four maps.



(a) (R3)-(V1)

(b) (R3)-(V2)

(c) (R3)-(V3)

(d) (R3)-(V4)

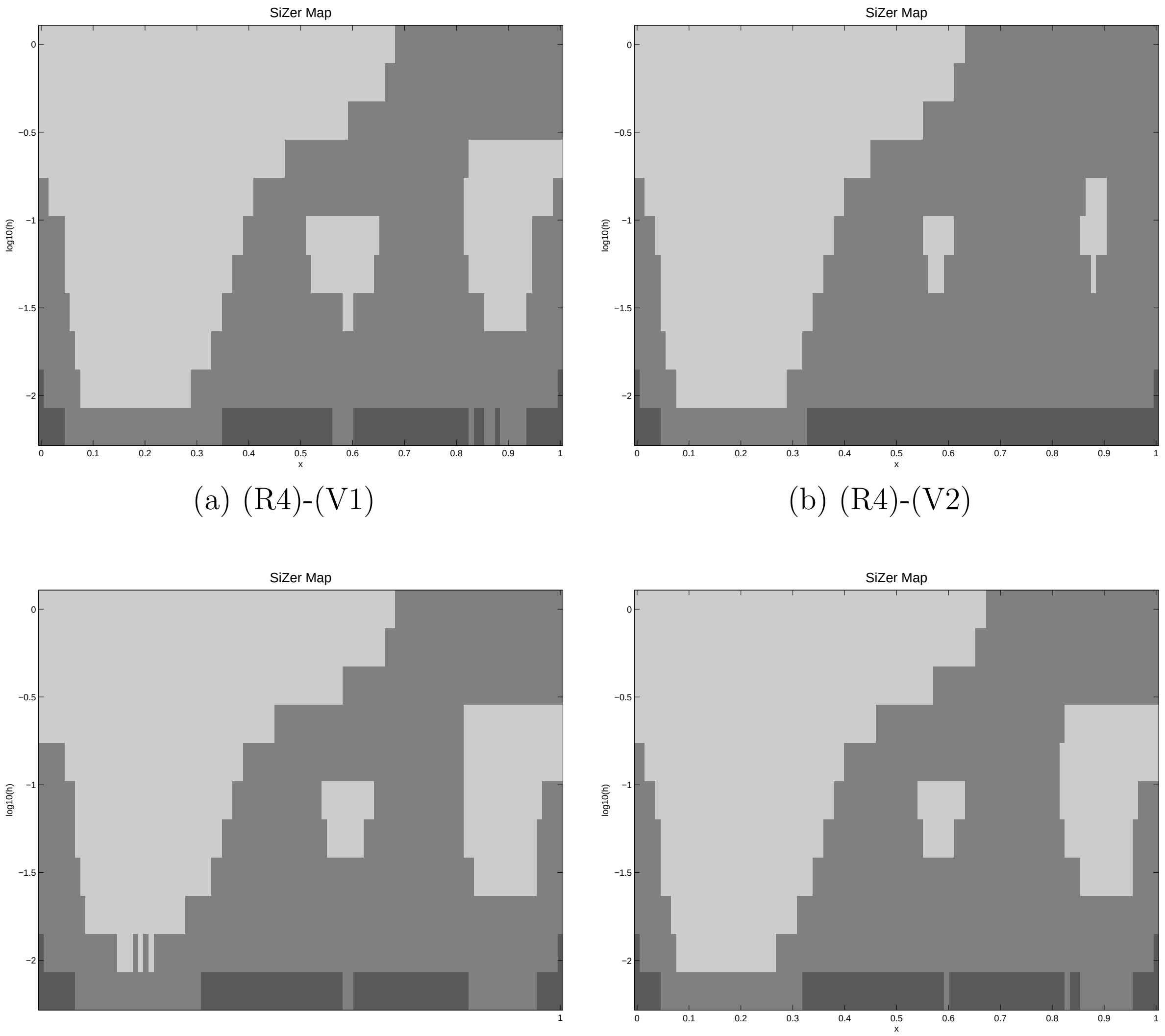
Figure 5: The average SiZer maps for the regression function (R3) with four different variance functions (V1)-(V4). The significant feature are found in most of the locations except for the center where the three curves intersect.

that some significant features can be possibly omitted in a SiZer map when both regression and variance functions are linear.

The average SiZer maps in Figure 5 illustrate that, for four types of the variance functions, the three regression functions significantly differ to one another in most of the locations except for the center where the three curves in Figure 2(c) intersect. For the linear variance functions (V2) and (V3), we note a similar phenomenon observed in Figure 4; less significant features are declared compared to the other two variance cases.

Pardo-Fernández et al. (2007), who considered the first three regression functions (R1)-(R3) with the constant variance (V1), also concluded that there is no difference for (R1) and global differences exist for both (R2) and (R3). However, it is difficult to directly compare the performance of the proposed approach and theirs because SiZer analyzes the data at multiple scales instead of working with a single bandwidth and attempts to find the locations where the difference occurs instead of conducting a test for overall equality. Furthermore, SiZer also can be helpful for drawing a single conclusion about overall equality of the curves in practice. If a SiZer map shows no (or few isolated and spurious) significant features, it can be concluded that there is no evidence of overall inequality, see for example (R1). On the other hand, a fair amount of significant features at multiple scales would support overall inequality, see for example (R2) and (R3). In addition, the proposed SiZer provides additional information about locations and scales of the differences. Park and Kang (2008) applied the idea of Pardo-Fernández et al. (2007) to SiZer inference and compared multiple curves based on the residuals. They compared the densities of two residual sets, one of which is obtained from the pooled data under the null hypothesis and the other from k separate groups under the alternative hypothesis. If the two densities tend to be similar to (different from) each other, it would suggest overall equality (inequality, respectively). However, because it does not directly compare the original regression curves, it is not possible to find the exact locations of their differences.

Figure 6 displays the four average SiZer maps for (R4) in Figure 2(d). From the maps, we can correctly infer that there is a big difference among the three curves in the first half of the region because significant features are detected at most of the scales. In the second half, two middle-sized features are flagged as significant around $x = 0.6$ and 0.9 at small and middle scales because the differences among the curves are not large enough to be caught at large scales. We again note



(a) (R4)-(V1)

(b) (R4)-(V2)

(c) (R4)-(V3)

(d) (R4)-(V4)

Figure 6: The average SiZer maps for the regression function (R4) with four different variance functions (V1)-(V4). A large significant feature is detected in the first half of the region at most of the scales. In the second half, two middle-sized features are flagged as significant around $x = 0.6$ and 0.9 at small and middle scales.

slightly inferior performance particularly with the positively linear variance function because the same trends exist in both regression and variance functions and they seem to be confounded each other.

In what follows we assess the accuracy of SiZer inference by calculating type I error and power for each simulation setting. Since SiZer conducts multiple tests at various locations and scales, the conventional concepts of type I error and power are not applicable. To circumvent this difficulty, we adopt the approach used in Rondonotti et al. (2007) and Hannig et al. (2013), in which they measured how often SiZer maps created from the observed data (called observed SiZer) are in agreement with the *true* ones without noise (called oracle SiZer).

Type I error and power for SiZer analysis are calculated as follows. For the case of (R1) where no difference should be found,

$$\text{type I error} = \frac{\#\text{ (pixels in an observed SiZer map that are flagged as significant)}}{\#\text{ (total pixels)}},$$

and the power is not calculated. In a map, darker gray regions are excluded because no statistical decision is made. For (R2)-(R4), we create two SiZer maps with the true regression functions f_1, f_2, and f_3 (oracle SiZer) and with the observed data (observed SiZer). Then, we compare the two SiZer maps each other pixel by pixel, and define

$$\text{type I error} = \frac{\#\text{ (observed SiZer=significant, oracle SiZer=insignificant)}}{\#\text{ (total pixels)}},$$

and

$$\text{power} = 1 - \frac{\#\text{ (observed SiZer=insignificant, oracle SiZer=significant)}}{\#\text{ (total pixels)}}.$$

Table 1 shows the mean and standard error of the 100 type I errors and the powers for each simulation setting. It can be seen that the type I errors are below 0.004 for all the cases. This suggests that the proposed SiZer inference makes few mistakes in its decision of overall equality. The power ranges between 0.5090 to 0.7312, which implies that there are some pixels that are not detected by the observed SiZer maps. From our numerical experiments, however, an observed SiZer rarely misses an important difference among the curves although it might not detect a few significant pixels around the feature. We observe that the power is slightly higher for (R4) when the regression and variance have different types of trends. We also note that the standard errors are small for all the cases, which indicates the stable performance of the proposed SiZer.

Table 1: Type I error and Power

	Type I	Error	Power	
Settings	Mean	S.E.	Mean	S.E.
(R1)-(V1)	0.0025	0.0010	-	-
(R1)-(V2)	0.0025	0.0010	-	-
(R1)-(V3)	0.0031	0.0012	-	-
(R1)-(V4)	0.0025	0.0011	-	-
(R2)-(V1)	0	0	0.6280	0.0040
(R2)-(V2)	0	0	0.5653	0.0062
(R2)-(V3)	0	0	0.5718	0.0054
(R2)-(V4)	0	0	0.6195	0.0042
(R3)-(V1)	0	0	0.5621	0.0031
(R3)-(V2)	0	0	0.5119	0.0036
(R3)-(V3)	0	0	0.5090	0.0042
(R3)-(V4)	0	0	0.5572	0.0031
(R4)-(V1)	0.0004	0.0002	0.7312	0.0028
(R4)-(V2)	0.0002	0.0001	0.6562	0.0038
(R4)-(V3)	0.0004	0.0002	0.7038	0.0026
(R4)-(V4)	0.0003	0.0002	0.7126	0.0027

4 Real Data Analysis

As mentioned in Section 1 we analyze two real datasets in this section using the proposed scale-space tool for comparing multiple curves.

The first example, the monthly household expenditures on several commodities in Dutch guilders, has been analyzed by several authors including Adang and Melenberg (1995); Einmahl and Van Keilegom (2006); Pardo-Fernández et al. (2007); Park and Kang (2008). We obtain the dataset from Data Archive of the Journal of Applied Econometrics. The dataset is divided into three groups by the number of members in the household: two ($n_1 = 1575$), three ($n_2 = 377$) or four members ($n_3 = 292$). We are interested in testing the equality of mean regression functions for each group, where the response variable is the logarithm of the expenditure on food and the covariate is the logarithm of the total monthly expenditure for each different family size. In an economic sense, it is similar to testing whether the Engel's law has a different meaning according to the household size.

The first top three panels of Figure 7 display the observed data points and their smoothed curves indexed by different bandwidths. These kernel estimates look similar at large scales but show some differences at small scales particularly for the left regions from $x = 10$ to $x = 11$. According to the SiZer map in the lower panel there are no differences among the three groups because the pixels are colored either gray (insignificant) or darker gray (no decision). The test results confirm that the central areas of the three groups are similar to each other. However, the differences observed at the beginning in the kernel estimates with small bandwidths cannot be determined because there are few data available around the regions especially for the second and the third groups. Therefore, for each different family size, there is no evidence for the overall inequality of the relationships. This conclusion is consistent with that of Park and Kang (2008).

The second example, taken from the Belgian part of the European Community Household Panel, concerns the working population in Belgium for the year 1994. This data set was used by Nolan and Whelan (1996), Whelan et al. (2000), Whelan et al. (2003) and Verbeek (2004), etc. The purpose was to explore the relationship between persistent income poverty and life-style deprivation, and finding out factors which can explain the wage differential. In our analysis, we focus on the relationship between wage (gross hourly wage rate in euro) on a log scale and years of experience with five education level groups from low (1) to high (5). The sample sizes for each

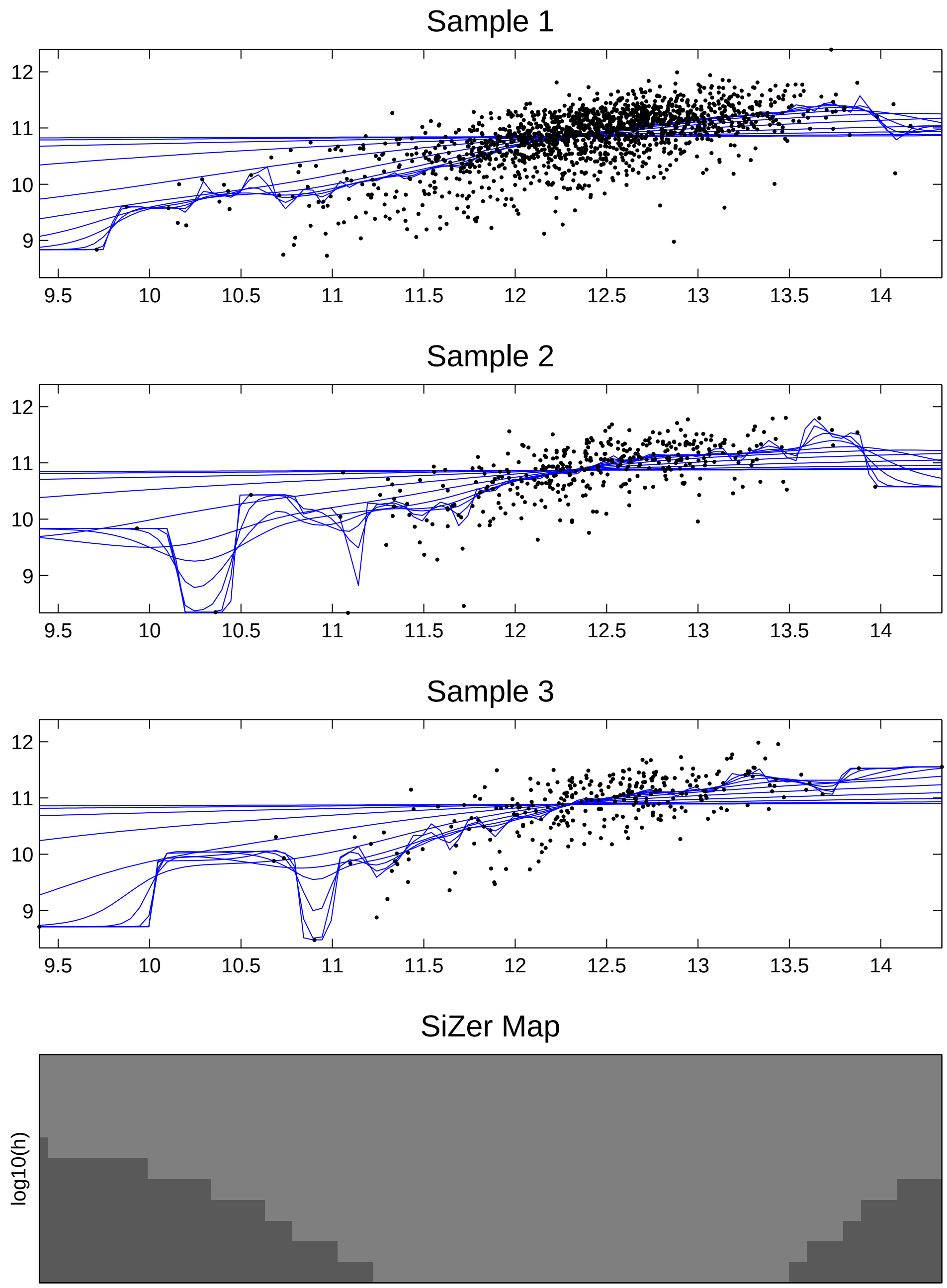


Figure 7: SiZer plots for the Dutch household data. The first top three plots display the data points and kernel estimates with different bandwidths for each sample and the bottom shows the corresponding SiZer map.

group are $n_1 = 99$, $n_2 = 265$, $n_3 = 420$, $n_4 = 356$ and $n_5 = 332$, respectively.

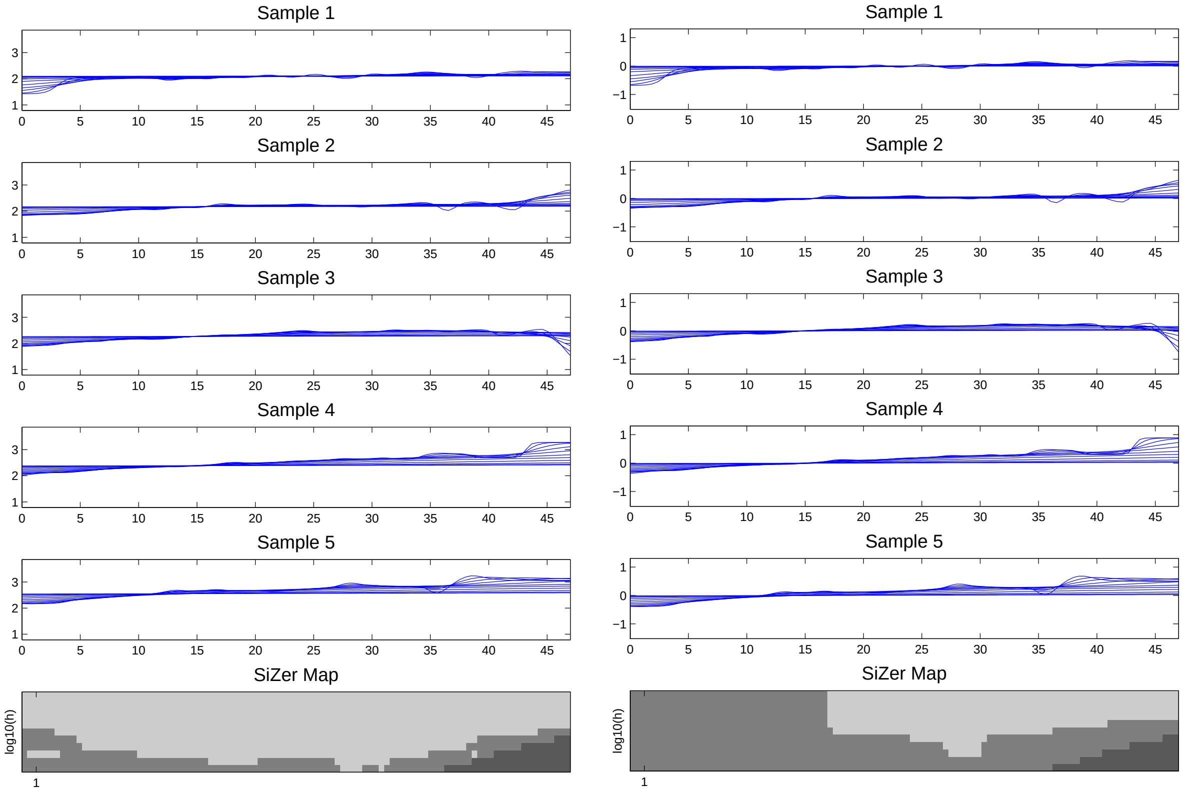
The SiZer plots in Figure 8(a) indicate that there are significant differences, flagged as white in the map, among the five education levels. Because these features are found in all locations and most of scales, and also the wage seems to have a higher mean as the education level increases from the kernel estimates, the mean differences can be regarded as the major driving force. Note that the darker gray colors at the bottom right corner imply that there are not sufficient data points around those regions (many years of experience) for the education levels 4 and 5 to make a statistically meaningful decision at small scales. When there exists an overall mean difference, it is sometimes difficult to identify other trends in the data. Hence, we redraw SiZer plots after subtracting the overall mean of each group in Figure 8(b). For the centered data, the features from $x = 17$ to the end at middle and large scales remain. This difference seems to come from an increasing trend starting around 20 years of experience for the levels 3, 4, and 5.

In Figure 9, we expand our investigation for sets of subgroups using the centered data. Figure 9(a) displays SiZer plots using only the first three levels. The SiZer map using the centered data flags the features from $x = 20$ to the end at middle and large scales. This difference also can be similarly found in Figure 8(b). In the comparisons among the education levels 3, 4, and 5 in Figure 9(b), the kernel estimates hint a stronger increasing trend for the education levels 4 and 5 compared to the level 3. The SiZer map supports it as significant features are found at large scales in the second half of the locations.

To sum up, SiZer analysis discovers that the overall mean wage tends to be higher as the education levels increases, and there exist increasing trends starting around 20 years of experience for the education level from 3 to 5. We point out, however, that the SiZer inferences for subgroup comparisons possibly cause another issue of multiple testing adjustments among SiZer maps. We suggest this improvement as future work.

5 Appendix: derivation of the test statistic

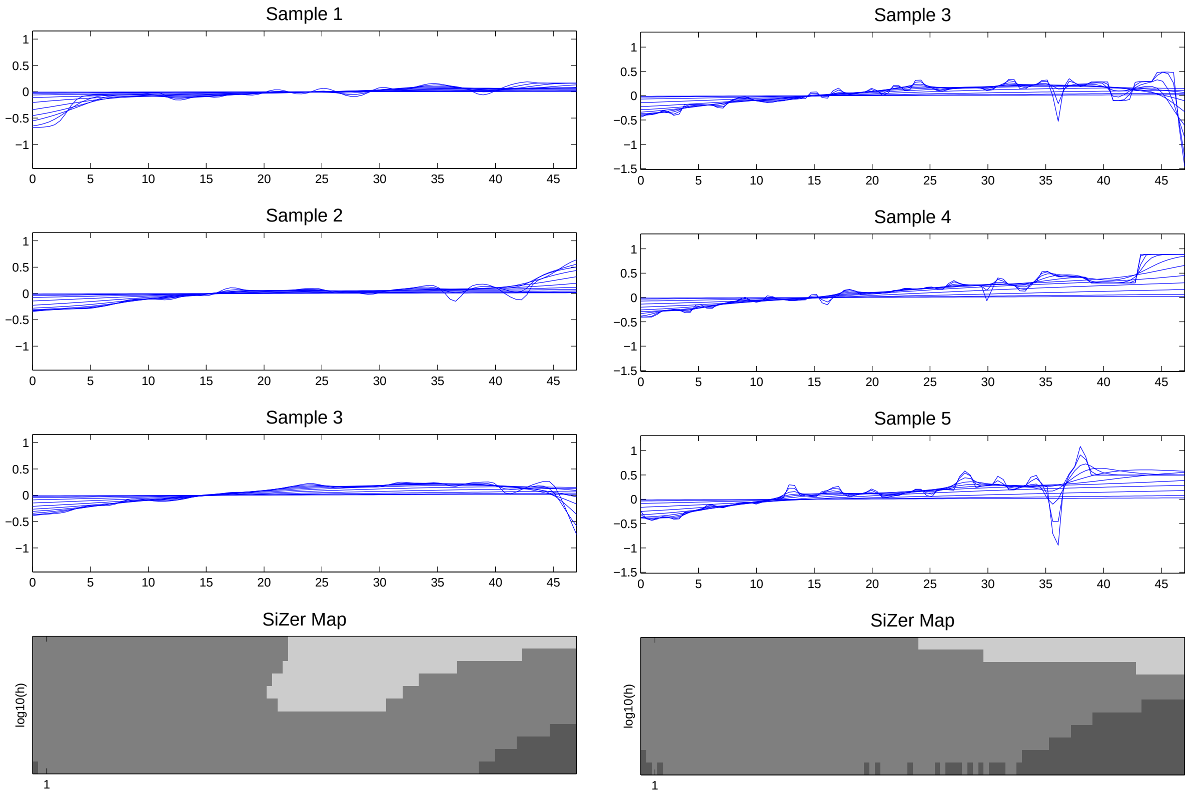
This section provides the derivation of the scale factors and degrees of freedom in the test statistic (2.2) using both local constant and local linear estimators. The expectations and variances are conditionally calculated given $X_{ij} = x_{ij}$.



(a) Raw data

(b) Centered data

Figure 8: SiZer plots for the Belgian wage data. All five groups are compared using (a) raw data and (b) centered data. The five groups represent education levels from 1 (low) to 5 (high).



(a) Education levels 1, 2, 3

(b) Education levels 3, 4, 5

Figure 9: SiZer plots for the centered Belgian wage data: (a) education levels 1, 2, and 3 and (b) education levels 3, 4, and 5 from the top.

5.1 Local constant estimator

Let us consider the numerator of (2.2) first. Let

$$\begin{aligned} T(x) &= \sum_{i=1}^{k} (\hat{f}_{i,h}(x) - \hat{f}_h(x))^2 \\ &= \sum_{i=1}^{k} (\hat{f}_{i,h}(x) - \sum_{l=1}^{k} r_l(x)\hat{f}_{l,h}(x))^2, \end{aligned} \tag{5.1}$$

where

$$r_l(x) = \frac{\sum_{j=1}^{n_l} K_h(x - X_{lj})}{\sum_{l=1}^{k}\sum_{j=1}^{n_l} K_h(x - X_{lj})}.$$

Let

$$a_i^2(x) = \frac{\sigma_i^2(x)}{n_i h}\int K^2(u)du.$$

Then, the expected value and the variance of (5.1) can be represented as

$$\begin{aligned} ET(x) &= \sum_{i=1}^{k} a_i^2(x)(1 - 2r_i(x) + kr_i(x)^2) \\ VarT(x) &= 2(ET(x))^2 - 4\sum_{1\le i<j\le k}\big\{a_i^2(x)a_j^2(x) \\ &\quad \times \big(1 - 2(r_i(x) + r_j(x)) - (r_i(x) - r_j(x))^2 + k(r_i^2(x) + r_j^2(x))\big)\big\}. \end{aligned}$$

Then, the degrees of freedom (df_1) and the scale factor (c_1) of the numerator can be obtained by calculating

$$df_1 = \frac{2(ET(x))^2}{VarT(x)}, \;\; c_1 = \frac{VarT(x)}{2ET(x)}.$$

Next, we consider the denominator of (2.2). Let

$$\begin{aligned} T_i(x) &= \sum_{j=1}^{n_i} \big(Y_{ij} - \hat{f}_{i,h}(X_{ij})\big)^2 K_h(x - X_{ij}) \\ &= \sum_{j=1}^{n_i} w_{ij}(x)\big(Y_{ij} - \sum_{l=1}^{n_i} r_i(j,l)Y_{il}\big)^2, \end{aligned} \tag{5.2}$$

where $w_{ij}(x) = K_h(x - X_{ij})$ and

$$r_i(j,l) = \frac{K_h(X_{ij} - X_{il})}{\sum_{l=1}^{n_i} K_h(X_{ij} - X_{il})}.$$

If we set

$$\tilde{r}_i(j,l) = \begin{cases} r_i(j,l), & \text{if } j \neq l \\ r_i(j,j) - 1, & \text{if } j = l, \end{cases}$$

(5.2) can be rewritten as

$$T_i(x) = \sum_{j=1}^{n_i} w_{ij}(x)\big(\sum_{l=1}^{n_i} \tilde{r}_i(j,l)Y_{il}\big)^2.$$

If we assume that Y_{ij}s are independently distributed with $N(0, \sigma_i^2(X_{ij}))$, then it can be shown that

$$\begin{aligned} ET_i(x) &= \sum_{j=1}^{n_i}\sum_{l=1}^{n_i} w_{ij}(x)\big[\tilde{r}_i(j,l)\big]^2\sigma_i^2(X_{il}) \\ VarT_i(x) &= 2\big(ET_i(x)\big)^2 - 4\sum\sum_{1\le j<l\le n_i}\sum\sum_{1\le p<q\le n_i}\big\{\sigma_i^2(X_{il})\sigma_i^2(X_{ij})w_{ip}(x)w_{iq}(x) \\ &\qquad \times\big(\tilde{r}_i(p,l)\tilde{r}_i(q,j) - \tilde{r}_i(p,j)\tilde{r}_i(q,l)\big)^2\big\}. \quad (5.3) \end{aligned}$$

Then, the denominator of (2.2) corresponds to $T(x) = \sum_{i=1}^{k} T_i$ and one can get its expected value and variance by

$$ET(x) = \sum_{i=1}^{k} ET_i(x) \quad \text{and} \quad VarT(x) = \sum_{i=1}^{k} VarT_i(x).$$

Again, the degrees of freedom and the scale factor of the distribution of the denominator can be obtained by calculating

$$df_2 = \frac{2(ET(x))^2}{VarT(x)}, \quad c_2 = \frac{VarT(x)}{2ET(x)}.$$

When we have non-central parameters, i.e., when Y_{ij}s are independently distributed with $N(f_{i,h}(X_{ij}), \sigma_i^2(X_{ij}))$, let $Y'_{ij} = Y_{ij} - f_{i,h}(X_{ij})$ and replace Y_{ij} by Y'_{ij} in (5.2). Then, we have the same formula in (5.3). In order to get Y'_{ij}, we need to plug in the estimate $\hat{f}_{i,h}$.

5.2 Local linear estimator

For the local linear estimator, exact derivation of the variance of the numerator is rather complicated. However, we note in (2.2) that we only need

$$c_1 \cdot df_1 = \frac{(ET(x))^2}{ET(x)}.$$

The expectation of numerator is given as

$$\begin{aligned} E(T(x)) &= E\left(\sum_{i=1}^{k}(\hat{f}_{i,h}(x)-\hat{f}_h(x))^2\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(K_{ij}^2(x)+k\tilde{K}_{ij}^2(x)-2K_{ij}(x)\tilde{K}_{ij}(x)\right)\sigma_i^2(x_{ij}). \end{aligned}$$

In the denominator of (2.2),

$$\begin{aligned} T_i(x) &= \sum_{j=1}^{n_i}\big(Y_{ij}-\hat{f}_{i,h}(x_{ij})\big)^2 K_h(x-X_{ij}) \\ &= \sum_{j=1}^{n_i} w_{ij}(x)\big(Y_{ij}-\sum_{l=1}^{n_i}K_{il}(x_{ij})Y_{il}\big)^2, \end{aligned} \tag{5.4}$$

where $w_{ij}(x)=K_h(x-x_{ij})$. If we set

$$\tilde{r}_i(j,l)=\begin{cases} K_{il}(x_{ij}), & \text{if } j\neq l \\ K_{ij}(x_{ij})-1, & \text{if } j=l. \end{cases}$$

(5.4) can be rewritten as

$$T_i(x)=\sum_{j=1}^{n_i} w_{ij}(x)\big(\sum_{l=1}^{n_i}\tilde{r}_i(j,l)Y_{il}\big)^2.$$

If we assume that Y_{ij}s are independently distributed with $N(0,\sigma_i^2(x_{ij}))$, then it can be shown that

$$ET_i(x) = \sum_{j=1}^{n_i}\sum_{l=1}^{n_i} w_{ij}(x)\big[\tilde{r}_i(j,l)\big]^2\sigma_i^2(x_{il}),$$

and

$$ET(x)=\sum_{i=1}^{k}ET_i(x).$$

Again,

$$c_2\cdot df_2=\frac{(ET(x))^2}{ET(x)}.$$

Acknowledgments

The authors would like to thank two referees and Associate Editor for their helpful comments. The second author was was supported in part by the National Science Foundation under Grant No.

1007543 and 1016441. The third author was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0076781).

References

Adang, P. J. M. and Melenberg, B. (1995). Nonnegativity constraints and intratemporal uncertainty in multi-good life-cycle models. *Journal of Applied Econometrics*, 10:1–15.

Bowman, A. and Young, S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics*, 45:83–98.

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

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, 17:199–204.

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.

Einmahl, J. H. J. and Van Keilegom, I. (2006). Goodness-of-fit tests in nonparametric regression. Discussion Paper No. 2006-79, CentER.

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 Oigard, 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.

Hall, P. and Hart, J. D. (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association*, 85:1039–1049.

Hannig, J. and Lee, T. (2006). Robust sizer for exploration of regression structures and outlier detection. *Journal of Computational & Graphical Statistics*, 15:101–117.

Hannig, J., Lee, T., and Park, C. (2013). Metrics for sizer map comparison. *Stat*, 2:49–60.

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

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *Annals of Statistics*, 13:63–89.

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

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, 90:1085–1093.

Kulasekera, K. B. and Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association*, 92:500–511.

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

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

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

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Neumeyer, N. and Dette, H. (2003). Noparametric comparison of regression curves: an empirical process approach. *Annals of Statistics*, 31:880–920.

Nolan, B. and Whelan, C. T. (1996). Measuring poverty using income and deprivation indicators: Alternative approaches. *Journal of European Social Policy*, 6:225–240.

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

Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, 17:1115–1137.

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. (2009a). Improved sizer for time series. *Statistica Sinica*, 19:1511–1530.

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. (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. *To appear in Journal of Computational and Graphical Statistics.*

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. (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.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114.

Sørbye, S., 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.

Verbeek, M. (2004). *A Guide to Modern Econometrics., 2nd ed.* John Wiley & Sons, West Sussex.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyã Series A*, 26:359–372.

Whelan, C. T., Layte, R., and Maître, B. (2003). Persistent income poverty and deprivation in the European Union: An analysis of the first three waves of the European community household panel. *Journal of Social Policy*, 32:1–18.

Whelan, C. T., Layte, R., Maître, B., and Nolan, B. (2000). Poverty dynamics: An analysis of the 1994 and 1995 waves of the ECHP. *European Societies*, 2:505–531.