

SiZer Analysis for the Comparison of Regression Curves

Cheolwoo Park¹ and Kee-Hoon Kang²

Abstract

In this article we introduce a graphical method for the test of the equality of two regression curves. Our method is based on SiZer (SIgnificant ZERo crossing of the differences) analysis, which is a scale-space visualization tool for statistical inferences. The proposed method does not require any specification of smoothing parameters, it offers a device to compare in a wide range of resolutions, instead. This enables us to find the differences between two curves that are present at each resolution level. The extension of the proposed method to the comparison of more than two regression curves is also done using residual analysis. A broad simulation study is conducted to demonstrate the sample performance of the proposed tool. Applications with two real examples are also included.

Key words: Comparison of multiple curves, Kernel density estimation, Local linear smoothing, SiZer.

1 Introduction

One of the most important problems in statistical inference is the comparison of two or more populations, which is necessary in a variety of contexts. The comparison of several populations can be done by looking at measures of location, measures of dispersion or some other variables, measured in the sample of each population. The comparison of population curves deserves much attention, including densities, regression curves, survival functions and some other characteristic functions of the variables of interest. This comparison can be done either in a parametric or a nonparametric way.

In this paper, we are interested in performing the comparison of several regression curves in a nonparametric context. A statistical challenge in this problem is testing whether there is any statistically significant differences of these curves. Suppose that we have k different samples and $n = \sum_{i=1}^k n_i$ independent observations from the following regression models:

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

where X_{ij} 's are covariates, the 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 of the i th sample and $\sigma_i^2(X_i) = Var(Y_i|X_i)$ is the conditional variance function of the i th sample ($i = 1, \dots, k$).

¹Department of Statistics, University of Georgia, Athens, GA 30602, USA. E-mail : cpark@stat.uga.edu

²Corresponding author. Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, Korea.
E-mail : khkang@hufs.ac.kr

For a motivation of our work we first introduce a real data set which contains monthly expenditures in Dutch guilders of Dutch households on several commodity categories. This data set has been analyzed by Adang and Melenberg (1995), Einmahl and Van Keilegom (2003), Pardo-Fernández et al. (2004), and others. The data set is divided into three groups by the number of members in the household. We model the data using (1.1), with the covariate ‘log of the total monthly expenditure’ and the response variable ‘log of the expenditure on food’ for each different family size.

Put Figure 1 around here.

Figure 1 is an example of local linear smooths with four different smoothing levels. The observed data points are plotted at each panel of Figure 1, with crosses for the first group (two family members), circles for the second group (three family members), and triangles for the third group (four family members), respectively. The local linear kernel estimates are overlaid for each data set, dotted lines for the first, dashed lines for the second, and solid lines for the third group, respectively. The three estimated regression curves with the smallest bandwidth ($h = 0.04$) in Figure 1 (a) seem to be quite different except for the central area. For $h = 0.15$ (Figure 1 (b)), the extent of difference becomes smaller except for the beginning. Note that, however, for the second and the third groups the few observations at the beginning make the local linear fits wavy. As the bandwidth increases (Figures 1 (c) and (d)), most of these estimated curves seem to be similar in a large part of the covariates. As one can see, the choice of smoothing parameter plays an important role in this comparison. In addition, we need to take into account the number of data around the point where the test is being done. Inference based on too small number of data often fails in drawing a significant conclusion. This example will be revisited in Section 5.

Our main concern in this paper is to develop a graphical device for testing the following hypothesis of the equality of mean regression functions

$$H_0 : f_1 = f_2 = \dots = f_k \tag{1.2}$$

versus

$$H_1 : f_i \neq f_j \text{ for some } i \neq j \in \{1, 2, \dots, k\}.$$

The problem of testing the equality of nonparametric regression curves has been widely studied in the literature. Härdle and Marron (1990) compared two regression curves estimated by using kernel methods. A bootstrap method for testing the equality of two regression curves was proposed by Hall and Hart (1990). Delgado (1993), Kulasekera (1995), Kulasekera and Wang (1997), and Neumeyer and Dette (2003) used approaches related to the empirical process. Bowman and Young (1996) introduced the idea of using a reference band for the comparison of two nonparametric curves. Fan (1996), and Fan and Lin (1998) proposed tests based on the adaptive Neyman test and the wavelet thresholding techniques. Some other relevant papers include Munk and Dette (1998), Dette and Neumeyer (2001), and Pardo-Fernández et al. (2004).

Recently, Pardo-Fernández et al. (2004) proposed two types of test statistics for comparing two regression curves. Their approach is based on the estimation of the distribution of the residuals in each population. The calculation is not straightforward and the procedure requires bootstrap approximation to obtain the critical values of the test. As in other nonparametric curve estimation methods, their approach also requires bandwidth selection. This motivates us to consider a graphical device to visualize the differences of two or more regression curves for a wide range of bandwidths.

In this paper we present a graphical device, called SiZer (SIgnificant ZERo crossing of the *differences*), for comparing multiple regression curves. SiZer, which originally stands for SIgnificant ZERo crossing of the *derivatives*, was proposed by Chaudhuri and Marron (1999) as a graphical goodness-of-fit test. It combines the scale-space idea of simultaneously considering a family of smooths (e.g. local linear fits) with the statistical inference that is needed for exploratory data analysis in the presence of noise. It brings an immediate insight into a central scientific issue in exploratory data analysis: *which features observed in a smooth of data are “really” there?* By studying the derivatives of smooths, one can compare sample data to white noise. Also, SiZer avoids the classical problem of bandwidth selection by considering a wide range of bandwidths.

Subsequently, several SiZer tools have been developed and they have been proven to be very powerful in many applications (see, e.g. Park et al., 2006). Hannig and Lee (2005) developed a robust version of SiZer in a regression setting, which can be used for identifying outliers. Hannig and Marron (2006) proposed a new method to reduce unexpected features in the SiZer map. The existing SiZer tools, however, have limitations. For example, they are applicable only to one data set, i.e. one curve. They compare the observed curve with one generated from an assumed (true) model, and test the difference between them. Therefore, it is not possible to directly compare two or more observed curves.

In this paper, a SiZer tool which is capable of comparing multiple curves is developed based on the *differences* of smooths. It gives insightful information about the differences of the curves by combining statistical inference with visualization. The method not only keeps the advantages of the original SiZer tools but also extends their usefulness to a broader range of scientific problems; for example seismic recordings of earthquakes and nuclear explosions, gait analysis, temperature-precipitation patterns, brain potentials evoked by flashes of light, packet/byte counts in Internet traffic, and so on.

This paper is organized as follows. Section 2 describes a SiZer for the comparison of two regression curves. The extension to the comparison of multiple regression curves is done in Section 3. As a byproduct, we develop a SiZer for the comparison of two density functions. Section 4 investigates the finite sample performances of the proposed method via several simulated examples. Applications to real data are illustrated in Section 5. Future work is discussed in Section 6. The quantile for constructing confidence intervals in Section 2 is derived in Section 7.

2 SiZer for the comparison of two regression curves

The original SiZer (Chaudhuri and Marron, 1999) is a visualization method based on nonparametric curve estimates. SiZer analysis enables statistical inference for the discovery of meaningful structure within a data set, while doing exploratory analysis. SiZer addresses the question of which features observed in a smooth are really there, or represent an important underlying structure, and not simply artifacts of the sampling noise.

SiZer is based on scale-space ideas from computer vision, see Lindeberg (1994). Scale-space is a family of kernel smooths indexed by the scale, which is the smoothing parameter or bandwidth h . SiZer considers a wide range of bandwidths which avoids the classical problem of bandwidth selection. Furthermore, the target of a SiZer analysis is shifted from finding features in the “true underlying curve” to inferences about the “smoothed version of the underlying curve”, i.e. the “curve at the given level of resolution”. The idea is that this approach uses all the information that is available in the data at each given scale. The details underlying the statistical inference using SiZer can be found in Chaudhuri and Marron (1999). The essential idea is that SiZer investigates the derivative of smooths and reports the results of a large number of simultaneous hypothesis tests. Chaudhuri and Marron (2000) studied weak convergence of the empirical scale space surface and some related asymptotic results under appropriate regularity conditions.

While the conventional SiZer compares one sample data set to a theoretical model (white noise), our method compares two or more data sets and tests whether they are significantly different. Thus, we do not investigate the derivatives of curves as the original SiZer does, but study the differences of curves. For this reason, in this paper, SiZer stands for SIgnificance of ZERo crossing of the *differences*.

Let us start with the two-sample problem and extend it to multiple samples in Section 3. SiZer applies the local linear fitting method, see e.g. Fan and Gijbels (1996), for obtaining a family of kernel estimates in a regression setting. Precisely, at a particular point x_0 , $\hat{f}_{i,h}(x_0)$ ($i = 1, 2$) are obtained by fitting lines

$$\beta_{i0} + \beta_{i1}(x_0 - X_{ij})$$

to the (X_{ij}, Y_{ij}) , with kernel weighted least squares. Then, $\hat{f}_{i,h}(x_0) = \hat{\beta}_{i0}$ ($i = 1, 2$) where $\hat{\beta}_i = (\hat{\beta}_{i0}, \hat{\beta}_{i1})'$ minimizes

$$\sum_{j=1}^{n_i} \{Y_{ij} - (\beta_{i0} + \beta_{i1}(x_0 - X_{ij}))\}^2 K_h(x_0 - X_{ij}), \quad (2.1)$$

where $K_h(\cdot) = K(\cdot/h)/h$. K is a kernel function, usually a symmetric probability density function. In this paper, we use a Gaussian kernel. Since the solution of (2.1) provides estimates of a regression function for different bandwidths, we can construct the family of smooths parameterized by h and the confidence intervals of the difference of two curves.

In SiZer analysis, we seek confidence intervals for the scale-space version $f_{1,h}(x) - f_{2,h}(x) \equiv E\hat{f}_{1,h}(x) - E\hat{f}_{2,h}(x)$ instead of seeking confidence intervals for $f_1(x) - f_2(x)$ (see Chaudhuri and

Marron, 1999, for discussion on this subject). From this point of view, significance of any difference depends on the scale of resolution, h . Thus, the hypotheses we are testing are

$$H_0 : f_{1,h}(x_0) = f_{2,h}(x_0) \quad \text{vs.} \quad H_1 : f_{1,h}(x_0) \neq f_{2,h}(x_0) \quad (2.2)$$

for a fixed point x_0 .

SiZer visually displays the significance of differences between two regression functions in families of smooths $\{\hat{f}_{i,h}(x), i = 1, 2\}$ over both location x and scale h , using a color map. It is based on confidence intervals for $\hat{f}_{1,h}(x) - \hat{f}_{2,h}(x)$, which will be defined soon, and uses multiple comparison level adjustment. Each pixel shows a color that gives the result of a hypothesis test in (2.2) at the point indexed by the horizontal location x , and by the bandwidth corresponding to the row h . At each (x, h) , if the confidence interval is above (below) 0, meaning that the curves are significantly different, i.e., $f_{1,h}(x) > f_{2,h}(x)$ ($f_{1,h}(x) < f_{2,h}(x)$), then that particular map location is colored black (white, respectively). On the other hand, if the confidence interval contains 0, meaning that the curves are not significantly different, then that map location is given the intermediate gray. Finally, if there are not enough data points to carry out the test, then no decision can be made and the location is colored darker gray. To determine the darker gray areas, based on the definition of Chaudhuri and Marron (1999), we define the estimated effective sample size (ESS), for each (x, h) as

$$\begin{aligned} \text{ESS}_i(x, h) &= \frac{\sum_{j=1}^{n_i} K_h(x - X_{ij})}{K_h(0)} \quad i = 1, 2, \\ \text{ESS}(x, h) &= \min(\text{ESS}_1(x, h), \text{ESS}_2(x, h)). \end{aligned}$$

If $\text{ESS}(x, h) < 5$, then the corresponding pixel is colored darker gray. In order to achieve reasonable computational speed, fast binned implementation of the smoothers and the corresponding hypothesis tests are used, as discussed in Chaudhuri and Marron (1999).

Confidence intervals for $f_{1,h}(x) - f_{2,h}(x)$ are of the form

$$\hat{f}_{1,h}(x) - \hat{f}_{2,h}(x) \pm q \cdot \widehat{SD}(\hat{f}_{1,h}(x) - \hat{f}_{2,h}(x)), \quad (2.3)$$

where q is an appropriate quantile, and \widehat{SD} is the estimated standard deviation, which will be discussed soon. Chaudhuri and Marron (2000) showed that the asymptotic simultaneous level of the test for the entire family of hypotheses is α as long as the true q can be approximated properly. For the approximation of the quantile, Chaudhuri and Marron (1999) suggested several methods including pointwise Gaussian quantiles, number of independent blocks, and bootstrap. Recently, Hannig and Marron (2006) improved the multiple comparison tests using advanced distribution theory. A similar calculation can be done for the comparison of two regression curves. As a result, the quantile for significance level α is defined as

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

where Φ is the standard normal distribution function and g is the number of bins. The “cluster index” θ is given by

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

where $\tilde{\Delta}$ is the distance between the pixels of the SiZer map. We use this quantile in our implementation, and the brief derivation of the cluster index θ is provided in Section 7.

For the estimation of the standard deviation, note that $\hat{f}_{i,h}(x)$ obtained from (2.1) can be written as

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

where

$$W_{i,h}(x, X_{ij}) = \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(x - X_{ij})\} K_h(x - X_{ij})}{\hat{s}_2(x; h)\hat{s}_0(x; h) - \hat{s}_1(x; h)^2} \quad (2.4)$$

and

$$\hat{s}_r(x; h) = \frac{1}{n_i} \sum_{j=1}^{n_i} (x - X_{ij})^r K_h(x - X_{ij}).$$

Then, by independence

$$Var(\hat{f}_{1,h}(x) - \hat{f}_{2,h}(x)) = Var(\hat{f}_{1,h}(x)) + Var(\hat{f}_{2,h}(x)),$$

and

$$\begin{aligned} Var(\hat{f}_{i,h}(x)) &= \frac{1}{n_i^2} \sum_{j=1}^{n_i} (W_{i,h}(x, X_{ij}))^2 Var(Y_{ij}) \\ &= \frac{1}{n_i^2} \sum_{j=1}^{n_i} \sigma_i^2(X_{ij}) (W_{i,h}(x, X_{ij}))^2. \end{aligned}$$

The estimation of $\sigma_i^2(X_{ij})$ can be found in Chaudhuri and Marron (1999).

3 SiZer for the comparison of more than two regression curves

This section is devoted to testing the equality of several regression curves. We consider testing the following scale-space version of the hypotheses in (1.2),

$$H_0 : f_{1,h}(x_0) = f_{2,h}(x_0) = \dots = f_{k,h}(x_0) \quad \text{vs.} \quad H_1 : \text{not } H_0 \quad (3.1)$$

However, the extension of the approach in Section 2 is not straightforward for this testing problem. We borrow an idea from Pardo-Fernández et al. (2004), and compare the residual distributions under the null and alternative hypotheses, respectively. In other words, first we obtain two residual

sets by fitting local linear estimates under the null and alternative hypotheses in (3.1), and then compare their density functions by fitting kernel density estimates. In this way, we convert the comparison of several regression curves into the comparison of two density functions.

For obtaining the residuals, we use a pilot bandwidth h_p , that is different from the bandwidth h used for constructing the SiZer map. If one takes $h_p = h$, then it is much more difficult to draw a SiZer plot due to the various ranges of the residuals. Hence, we treat h and h_p separately, which means the addition of another dimension to the SiZer plot. Our visualization approach for this problem is to draw a series of SiZer plots indexed by the pilot bandwidth h_p .

Let

$$\hat{\sigma}_i^2(x) = n_i^{-1} \sum_{j=1}^{n_i} W_{i,h_p}(x, X_{ij}) Y_{ij}^2 - \hat{f}_{i,h_p}^2(x)$$

be the estimators of the conditional variance function, where the weights $W_{i,h_p}(x, X_{ij})$'s are given in equation (2.4). Let $\hat{f}_{h_p}(\cdot)$ be the local linear estimator of the common scale-spaced regression function $f_{h_p}(\cdot)$ under H_0 , which has the following form:

$$\hat{f}_{h_p}(x) = \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} W_{ij,h_p}(x, X_{ij}) Y_{ij},$$

where the weights $W_{ij,h_p}(x, X_{ij})$'s are similar to those in the equation (2.4) with

$$\hat{s}_r(x; h) = \frac{1}{\sum_{i=1}^k n_i} \sum_{i=1}^k \sum_{j=1}^{n_i} (x - X_{ij})^r K_h(x - X_{ij}).$$

Let $(Y_{ij} - \hat{f}_{i,h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$ be the estimate of the error ε_{ij} from the i th population in model (1.1) and let $(Y_{ij} - \hat{f}_{h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$ be the estimate of the same quantity under the null hypothesis in (3.1). The idea is that if H_0 is true, $(Y_{ij} - \hat{f}_{i,h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$ and $(Y_{ij} - \hat{f}_{h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$ would be quite similar and have the same distribution. Hence, we may check the equality of the regression curves with these two types of residual distributions. Then, we need a SiZer tool for comparing the two densities.

The construction of SiZer for the comparison of two densities is also based on the difference of two functions. Using $(Y_{ij} - \hat{f}_{i,h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$'s and $(Y_{ij} - \hat{f}_{h_p}(X_{ij}))/\hat{\sigma}_i(X_{ij})$'s, we estimate the difference of the two residual density functions, say $g_1 - g_0$, at a point t :

$$\hat{g}_{1,h_p}(t) - \hat{g}_{0,h_p}(t) = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} \left\{ K_{h_p} \left(t - \frac{Y_{ij} - \hat{f}_{i,h_p}(X_{ij})}{\hat{\sigma}_i(X_{ij})} \right) - K_{h_p} \left(t - \frac{Y_{ij} - \hat{f}_{h_p}(X_{ij})}{\hat{\sigma}_i(X_{ij})} \right) \right\} \quad (3.2)$$

where g_1 (g_0) is a probability density function of the residuals under the H_1 (H_0). The idea for constructing the confidence intervals for the difference of the two residual density functions is similar to that of the regression case shown in (2.3).

In our data analysis, we try a wide range of pilot bandwidths h_p to get the residuals. However, if we show all the SiZer plots with the full selection of pilot bandwidths, the complete series of SiZer

plots would be too long. The simultaneous view of all these SiZer plots is hard to comprehend and the information contained in several such plots is often redundant. This motivates us to choose only a subset of SiZer plots. We found three plots are usually enough to convey the needed information. Our choice among the several plots is intended to reflect ‘a wide array of trade-offs’ between undersmoothing and oversmoothing. We first get the range of the covariates in the given data, and divide it into 11 equally spaced values in a logarithmic scale. Then, we obtain residuals and draw SiZer plots for each set of residuals. Finally, we choose three plots that make good representatives. We recommend not choosing the smallest pilot bandwidth since the degree of overfitting may be too high. We always include the second one in our analysis to see the effect of a small pilot bandwidth.

4 Simulation

Section 4.1 shows the simulated examples for comparing two regression curves. In Section 4.2, the comparisons of two density functions and multiple regression curves are provided.

4.1 Comparison of two regression curves

Figure 2 about here.

This section provides six simulated examples. Each example has sample sizes $n_1 = 1000$ and $n_2 = 2000$. X_1 and X_2 are generated independently from a $U(0, 1)$ distribution. The first example has the same constant mean 0 with independent $N(0, 1)$ errors:

$$(i) \quad Y_{ij} = \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, 2,$$

where $\varepsilon_{ij} \sim N(0, 1)$ for $i = 1, 2$. The correct SiZer plot would show no significant difference. Figure 2 (a) displays its SiZer plot. In the top panel, the thin curves display the family of smooths, which are the differences of two local linear smooths, $\hat{f}_{1,h}(x) - \hat{f}_{2,h}(x)$. These differences are located around 0 because both samples have zero constant functions and the same noise distribution. The SiZer map in the lower panel reports the equality test of the two samples by investigating the confidence intervals in (2.3) at each (x, h) . The horizontal locations in the SiZer map are the same as in the top panel, and the vertical locations correspond to the logarithm of bandwidths of the family of smooths shown as thin curves in the top panel. The white dotted curves show effective window widths for each bandwidth, as intervals representing $\pm 2h$. Each pixel shows a color that gives the result of a hypothesis test for the sign of the thin curve, at the point indexed by the horizontal location, and at the bandwidth corresponding to that row. The result shows only gray color, meaning no significant difference, as expected.

For the second example, one sample has mean 2 and the other has mean 0 with both having error distribution $N(0, 1)$:

$$(ii) \quad Y_{1j} = 2 + \varepsilon_{1j}, \quad \text{and} \quad Y_{2j} = \varepsilon_{2j}$$

where $\epsilon_{ij} \sim N(0, 1)$ for $i = 1, 2$. Figure 2 (b) displays its SiZer plot. The upper panel shows the difference of two smooths amounts to approximately 2, which corresponds to the difference of two true regression functions. The SiZer map shows positive differences (black) across all scales since the mean of the first sample is greater than that of the second sample by 2. This shows that SiZer for two samples can successfully detect differences when the data have different mean levels.

The third example studies two different regression functions and has the following regression models:

$$(iii) \quad Y_{1j} = \sin(6\pi X_{1j}) + \epsilon_{1j}, \quad \text{and} \quad Y_{2j} = \epsilon_{2j}$$

where $\epsilon_{ij} \sim N(0, 0.25)$ for $i = 1, 2$. Figure 2 (c) displays its SiZer plot. The difference of the two smooths clearly reveals the sine curves in the top panel, and the SiZer map shows positive (black) and negative (white) differences along the sine curve. From these three simulations, we show that SiZer for the comparison of two regression curves performs well in various settings.

Figure 3 about here.

We make the examples more challenging in the next three simulations by adding errors with different variances. The first error ϵ_{1j} is drawn from $N(0, 0.5)$ and the second error ϵ_{2j} from $N(0, 0.25)$. The following curves are added to the errors:

$$(iv) \quad f_1(x) = f_2(x) = \sin(2\pi x),$$

$$(v) \quad f_1(x) = \sin(2\pi x), \quad f_2(x) = \sin(2\pi x) + x,$$

$$(vi) \quad f_1(x) = \exp(x), \quad f_2(x) = \exp(x) + \sin(2\pi x).$$

Figures 3 (a)-(c) show the SiZer plots of these three examples. No significant features are found in the SiZer map of Figure 3 (a) since they have the same regression curve. When the two regression curves are different, the SiZer maps capture the differences and flag the linear trend (Figure 3 (b)) and the sine wave (Figure 3 (c)) as significant. These three examples show that SiZer for the comparison of two regression curves performs well for differing error variances.

We check the behavior of the proposed method in multiple replications. We repeat the six examples presented above 100 times and average their SiZer maps. Each averaged SiZer map is very close to its corresponding one in Figures 2 or 3. To save the space figures are not included here but available from the authors.

4.2 Comparison of multiple regression curves

Figure 4 about here.

As explained in Section 3, to compare several regression curves we compare density functions of two residual sets, one under H_0 and one under H_1 in (3.1). To get an intuitive idea for comparing

two densities, we simulate three different examples. Each example has sample sizes $n_1 = 1000$ and $n_2 = 2000$. The top panels of Figure 4 show the difference between the two kernel density estimates and the lower panels report the results of testing the equality of the two density functions. In the first example, X_1 and X_2 are separately generated from a $N(0, 1)$ distribution. The family of smooths in the top panel of Figure 4 (a) is located around 0, and the SiZer map in the lower panel shows no features since the two samples are drawn from the same density. The darker gray in the bottom corners of the plots is due to the lack of data points, which can easily happen near the boundaries and at small smoothing levels.

In the second example, X_1 and X_2 are drawn from $N(2, 1)$ and $N(0, 1)$ distributions, respectively. Since they have different centers, the difference of the densities tends to be positive at one side and negative at the other side. Also, note that the range of x axis is wider compared to the first example.

In the third example, X_1 and X_2 are drawn from $N(0, 0.25)$ and $N(0, 1)$, respectively. Since the density of X_1 is more concentrated around the mean in comparison to X_2 , the difference of the two densities tends to be negative at both sides and positive in the middle.

Figure 5 about here.

We saw how SiZer compares two density functions, and move to the simulation for comparing multiple regression curves. We simulate three examples to compare three regression curves. Each example has sample sizes $n_1 = 500$, $n_2 = 1000$, and $n_3 = 1500$. X_1 , X_2 and X_3 are generated from $U(0, 1)$ independently. In the first example, the first sample is drawn from $N(0, 0.25)$, the second from $N(0, 0.5)$, and the third from $N(0, 0.75)$, and the mean regression functions are all zero:

$$(a) \quad Y_{1j} = \varepsilon_{1j}, \quad Y_{2j} = \varepsilon_{2j}, \quad \text{and} \quad Y_{3j} = \varepsilon_{3j}$$

where $\varepsilon_{1j} \sim N(0, 0.25)$, $\varepsilon_{2j} \sim N(0, 0.5)$, and $\varepsilon_{3j} \sim N(0, 0.75)$.

To draw the SiZer plots for comparing the three regression curves, we first obtain two residual sets, one under H_0 and the other under H_1 in (3.1) with 11 different pilot bandwidth h_p as explained in Section 3. Then, with the two residual sets, we construct a SiZer plot based on the difference between two density estimates in (3.2) for each h_p . Figures 5 (a)-(c) show the SiZer plots with the second, the third, and the fourth pilot bandwidths. As expected, the three families of smooths look similar to Figure 4 (a), and the three SiZer maps show no significant features. Different sample sizes and variances do not appear to make any wrong decisions in our method.

Figure 6 about here.

In the second example the error structures remain the same as in the first, but we increase the third mean by 2, that is

$$(b) \quad Y_{1j} = \varepsilon_{1j}, \quad Y_{2j} = \varepsilon_{2j}, \quad \text{and} \quad Y_{3j} = 2 + \varepsilon_{3j}$$

where $\epsilon_{1j} \sim N(0, 0.25)$, $\epsilon_{2j} \sim N(0, 0.5)$, and $\epsilon_{3j} \sim N(0, 0.75)$. Figures 6 (a)-(c) show the SiZer plots of the difference between two residual sets with the second, the fifth, and the eighth pilot bandwidths. Even with the large h_p , the SiZer map clearly flags the significant difference. The SiZer plots look similar to Figure 4 (c). The third sample with the higher mean increases the variance (and possibly shifts the mean as well) of the residuals obtained under H_0 , and this makes the difference between the two densities.

Figure 7 about here.

In the third example, the error structures remain the same as in the first, but we add the sine curve $f_3(x) = 0.4 \sin(6\pi x)$ to the third sample, that is

$$(c) \quad Y_{1j} = \epsilon_{1j}, \quad Y_{2j} = \epsilon_{2j}, \quad \text{and} \quad Y_{3j} = 0.4 \sin(6\pi X_{3j}) + \epsilon_{3j}$$

where $\epsilon_{1j} \sim N(0, 0.25)$, $\epsilon_{2j} \sim N(0, 0.5)$, and $\epsilon_{3j} \sim N(0, 0.75)$. Compared to the second example, the differences of the three curves are not trivial. Figures 7 (a)-(c) show the SiZer plots of the difference between two residual sets with the second, the fourth, and the sixth pilot bandwidths. While the SiZer maps with the second and the fourth h_p 's correctly rejecting H_0 and revealing a few significant features, the map with the sixth h_p shows no significant differences. This happens because the differences between the three curves diminish as h_p increases, as we observed this phenomenon in Figure 1. A bandwidth selection approach might conclude no difference between these three curves, since it selects one particular bandwidth, but our method can correctly detect the difference because it considers a wide range of bandwidths.

These three examples show that the comparison of several regression curves can be successfully done by comparing one residual set under H_0 and the other under H_1 .

We repeated the three examples presented above 100 times and confirmed that each averaged SiZer map was very close to its corresponding one in Figures 5, 6, or 7. We do not report the result to save the space but it is available from the authors.

5 Real examples

This section is devoted to illustrating our procedure applied to real data.

Example 1. The first example revisits the example which has been discussed previously by Hall and Hart (1990). This example was reanalyzed in Munk and Dette (1998). They compared the towns of Coweeta and Lewiston, North Carolina, for the concentration of sulfate in rain as a function of time in a 261 week period between 1979 and 1983. The measurement of these data were taken weekly and the sample sizes are $n_1 = 220$ and $n_2 = 215$. The data actually used in comparing two locations were the natural logarithms of the acid concentration. For a scatterplot of the adjusted data together with kernel regression estimates, see Figure 1 of Hall and Hart (1990). In the analysis,

they found no indication that the error terms were correlated across time using a residual analysis. Here, we compare the sulfate concentration as a function of time in the two towns.

Figure 8 about here.

Since this example compares two towns, a SiZer plot, which is depicted in Figure 8, is constructed based on (2.3). Many little spikes are found in the top panel, but they are not flagged as significant by the SiZer map. Significant features are found at low resolutions (large bandwidth) which suggest that their grand mean is different. Since this SiZer map indicates negative values at those levels the concentration of sulfate in rain at Coweeta is less than that at Lewiston. These results coincide with those of Hall and Hart (1990)[Section 3.4] and Munk and Dette (1998) [Section 4.3]. Our approach offers a more effective visual understanding in a wide range of resolution levels.

Example 2. The second real data example, as introduced in Section 1, was obtained from Data Archive of the Journal of Applied Econometrics, and consists of monthly expenditures in Dutch guilders of Dutch households on several commodity categories, as well as on a number of background variables. We will compare the regression curves for three groups of households: households consisting of 2 members (1575 in total), 3 members (377 in total) and 4 members (292 in total). The data were collected from April 1984 to September 1987, and the average was taken over the 42 months for each household. The autocorrelation plots of the data (not reported), do not indicate any evidence of correlation.

Figure 9 about here.

First, we compare two groups of three pairs and then the three groups altogether. Figure 9 shows SiZer plots for comparing (a) the two and three, (b) the two and four, and (c) the three and four member groups. Since two curves are being compared, SiZer plots are constructed based on (2.3). According to the SiZer maps in the lower panels no difference is found because the other SiZer maps show either gray (not significant) or darker gray (lack of data). A spike and a valley are observed at the beginning in all the families of smooths (with small bandwidths), but they are not flagged as significant since there are not sufficient data around the points where the tests are being performed and the map shows darker gray. As mentioned in Figure 1, there are very few observations at the beginning for the second and the third groups.

Figure 10 about here.

Figure 10 shows SiZer plots for comparing the three groups simultaneously using their residual distributions. The result shows the same conclusion as the comparison of two groups, that is no difference among three groups is found at the second, the fourth, and the sixth pilot bandwidths.

6 Future work

Another approach using ANOVA type statistics can be developed for the comparison of multiple curves. Instead of converting the problem into comparing two densities, this approach compares multiple curves directly at each point. The test statistics for comparing the curves at x can be roughly written as

$$(\text{Constant}) \times \frac{\sum_{i=1}^k (\hat{f}_{i,h}(x) - \hat{f}_h(x))^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \hat{f}_{i,h}(X_{ij}))^2 K_h(x - X_{ij})}$$

where $\hat{f}_{i,h}$ is a local linear fit using i th sample and \hat{f}_h using the combined sample under the null hypothesis. This statistic mimics the ratio of variations from the model and the error in ANOVA. To conduct a test, one needs to find the approximate distribution of this statistic and its degree of freedom. Also, an appropriate multiple adjustment needs to be designed for SiZer. The advantage of this approach over the previous one is that one can compare several curves directly and get the information of their local differences, which reflects the original intention of SiZer. If some differences are found among the curves, multiple pairwise comparisons can be performed as done in ANOVA analysis. We are currently developing this approach.

SiZer is a useful tool to find meaningful structures in the given data, but its usefulness can be limited in the case of dependent data because we assume independent errors. For dependent data, significant features appear in SiZer, which are due to the presence of dependence. The great challenge in time series for applications of SiZer in the trend (of difference between two curves) estimation context is that “trend” and “serial dependence artifacts” cannot be distinguished. For the one curve case, Rondonotti, Marron, and Park (2007) extended the original SiZer to SiZer for time series. This tool finds features of the underlying trend function, while taking into account the dependence structure using the estimated autocovariance function. Another version of SiZer for dependent data is so-called ‘Dependent SiZer’ proposed by Park, Marron, and Rondonotti (2004). Dependent SiZer has a slightly different goal from SiZer for time series. It uses a true autocovariance function of an assumed model instead of estimating it from the observed data. By doing so, a goodness of fit test can be conducted and we can see how different the behavior of the data is from that of the assumed model. For two curves, confidence intervals for $f_{1,h}(x) - f_{2,h}(x)$ have the same form as the independent case in (2.3), but the estimation of the autocovariance function and the choice of the quantile q should be adjusted. Some work was done in Rondonotti, Marron, and Park (2007), but we plan to improve their method and extend it to the comparison of several time series in the future.

7 Appendix

To color the pixels SiZer checks whether the difference of the estimates of the two regression functions

$$\begin{aligned}
\hat{\beta}_{i0} &= c_i^{-1} \left[\sum_{j=1}^{n_i} K_h(x - X_{ij}) Y_{ij} \right] \left[\sum_{j=1}^{n_i} (x - X_{ij})^2 K_h(x - X_{ij}) \right] \\
&- c_i^{-1} \left[\sum_{j=1}^{n_i} (x - X_{ij}) K_h(x - X_{ij}) \right] \left[\sum_{j=1}^{n_i} (x - X_{ij}) K_h(x - X_{ij}) Y_{ij} \right], \tag{7.1} \\
c_i &= \left[\sum_{j=1}^{n_i} K_h(x - X_{ij}) \right] \left[\sum_{j=1}^{n_i} (x - X_{ij})^2 K_h(x - X_{ij}) \right] - \left[\sum_{j=1}^{n_i} (x - X_{ij}) K_h(x - X_{ij}) \right]^2,
\end{aligned}$$

for $i = 1, 2$, is significantly different from 0. By appropriate binning procedure, the points X_{ij} can be converted into the fixed design points $X_l = l\Delta$, $l = 1, \dots, m$, where $\Delta > 0$ is the distance between design points and m is the number of grid points. If x is away from the boundary, it follows from symmetry of the kernel that

$$\sum_{l=1}^m (x - X_l) K_h(x - X_l) \approx 0,$$

which means that the second term in (7.1) disappears.

Let $\tilde{\Delta}$ denote the distance between the pixels of the SiZer map, and $p = \tilde{\Delta}/\Delta$, which denotes the number of data points per SiZer column. For simplicity of notation we can assume that p is a positive integer.

Let g be the number of pixels on each row, and T_1, \dots, T_g denote test statistics of a row in the SiZer map. Then T_k is proportional to $\hat{\beta}_{10} - \hat{\beta}_{20}$ calculated for $x = k\tilde{\Delta} = kp\Delta$. In particular

$$T_k \approx \sum_{q=1}^m W_{kp-q}^h (Y_{1,q} - Y_{2,q}).$$

The form of the W_{kp-q}^h is given in the first term of (7.1) with appropriate binning adjustment. Note that W_{kp-q}^h is proportional to $K_{h/\Delta}(kp - q)$ and thus the weights W_q^h are proportional to the Gaussian kernel with standard deviation h/Δ .

If the null hypothesis of two mean curves being equal is true, then $Y_{1,q} - Y_{2,q}$ are independently distributed Gaussian random variables with mean zero.

The full joint distribution of T_1, \dots, T_g depends on the correlation between them and this corre-

lation is approximated by

$$\begin{aligned}
\rho_k &= \text{corr}(T_l, T_{l+k}) \\
&= \frac{\sum_q W_{q-kp}^h W_q^h}{\sum_q (W_q^h)^2} \\
&\approx \frac{\int K_{h/\Delta}(x - kp) K_{h/\Delta}(x) dx}{\int K_{h/\Delta}(x)^2 dx} \\
&= e^{-(k\tilde{\Delta})^2/(4h^2)}
\end{aligned}$$

where the third line follows by replacing the sums by integral approximations and the last step follows by observing that $p\Delta = \tilde{\Delta}$.

By the Theorem 1 of Hannig and Marron (2006), for each fixed g we can get k step correlation $\rho_{k,g}$ such as

$$\rho_{k,g} = e^{-k^2 C^2 / (4 \log g)},$$

by setting $\tilde{\Delta}/h = C/\sqrt{\log g}$ and

$$\lim_{g \rightarrow \infty} \log(g)(1 - \rho_{k,g}) = \frac{k^2 C^2}{4}.$$

Following the similar arguments in paragraphs after Theorem 1 of Hannig and Marron (2006), we conclude that in the case of SiZer for comparing two curves,

$$P \left[\max_{i=1, \dots, g} T_i \leq x \right] \approx \Phi(x)^{\theta g},$$

where the cluster index θ

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

Acknowledgement

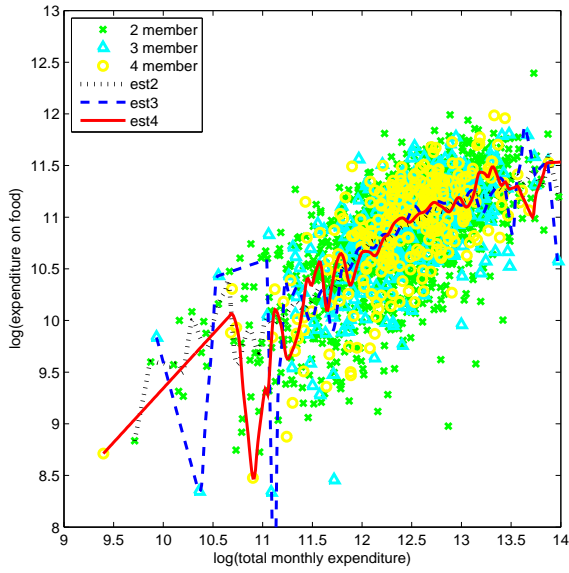
We would like to thank Jan Hannig for helpful discussion. The first author was supported by UGA Faculty Research Grants Program. The second author was supported by Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-015-C00069)

References

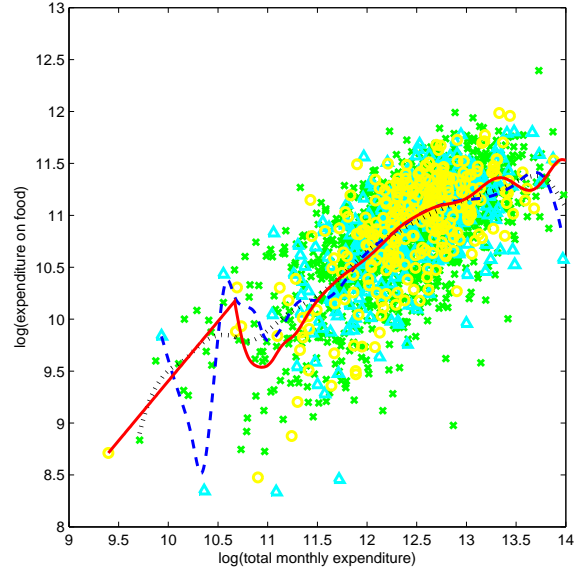
- [1] Adang, P. J. M. and Melenberg, B. (1995). Nonnegativity constraints and intratemporal uncertainty in multi-good life-cycle models. *Journal of Applied Econometrics*, **10**, 1–15.

- [2] Bowman, A. and Young, S. (1996). Graphical Comparison of Nonparametric Curves. *Applied Statistics*, **45**, 83–98.
- [3] Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.
- [4] Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, **28**, 408–428.
- [5] Delgado, M. A. (1993). Testing the equality of nonparametric regression curve. *Statistics & Probability Letters*, **17**, 199–204
- [6] Dette, H. and Neumeier, N. (2001). Nonparametric analysis of covariance. *The Annals of Statistics*, **29**, 1361–1400.
- [7] Einmahl, J. H. J. and Van Keilegom, I. (2003). Goodness-of-fit tests in nonparametric regression. *Discussion Paper*.
- [8] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman’s truncation. *Journal of the American Statistical Association*, **91**, 674–688.
- [9] Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
- [10] Fan, J. and Lin, S. (1998). Test of significance when data are curves. *Journal of the American Statistical Association*, **93**, 1007–1021.
- [11] 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.
- [12] Hannig, J. and Lee, T. (2005). Robust SiZer for exploration of regression structures and outlier detection. *Journal of Computational & Graphical Statistics*, **15**, 101–117 .
- [13] Hannig, J. and Marron, J. S. (2006). Advanced Distribution Theory for SiZer. *Journal of the American Statistical Association*, **101**, 484–499 .
- [14] Härdle W. and Marron, J. S. (1990). Semiparametric comparison of regression curves. *The Annals of Statistics*, **13**, 63–89.
- [15] Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, **90**, 1085–1093.
- [16] Kulasekera, K. B. and Wang, J. (1995). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association*, **92**, 500–511.
- [17] Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer, Boston.

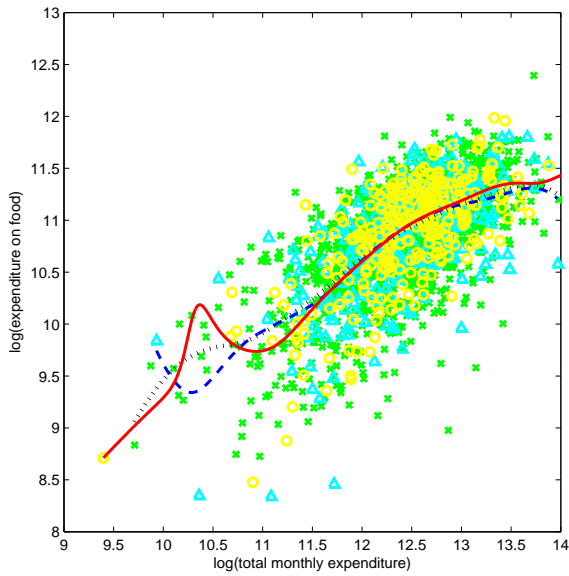
- [18] Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *The Annals of Statistics*, **26**, 2339-2368.
- [19] Neumeyer, N. and Dette, H. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, **31**, 880–920.
- [20] Pardo-Fernández, J. C., Van Keilegom, I. and González-Manteiga, W. (2004). Comparison of regression curves based on the estimation of the error distribution. Discussion Paper DP0416. Available at <http://www.stat.ucl.ac.be/ISpub/ISdp.html>
- [21] 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.
- [22] Park, C., Hernandez-Campos, F., Le, L., Marron, J. S., Park, J., Pipiras, V., Smith, F. D., Smith, R. L., Trovero, M., and Zhu, Z. (2006). Long range dependence analysis of Internet traffic. *Under revision, Technometrics*.
- [23] 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.



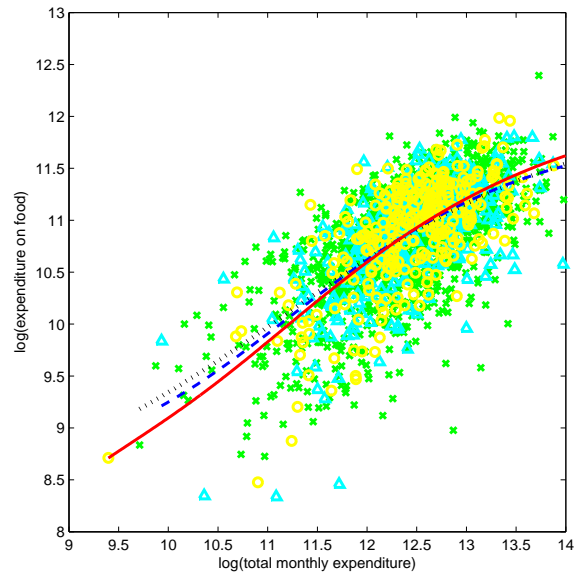
(a) $h = 0.04$



(b) $h = 0.15$



(c) $h = 0.30$



(d) $h = 1.00$

Figure 1: Local linear estimates for each group with different smoothing levels.

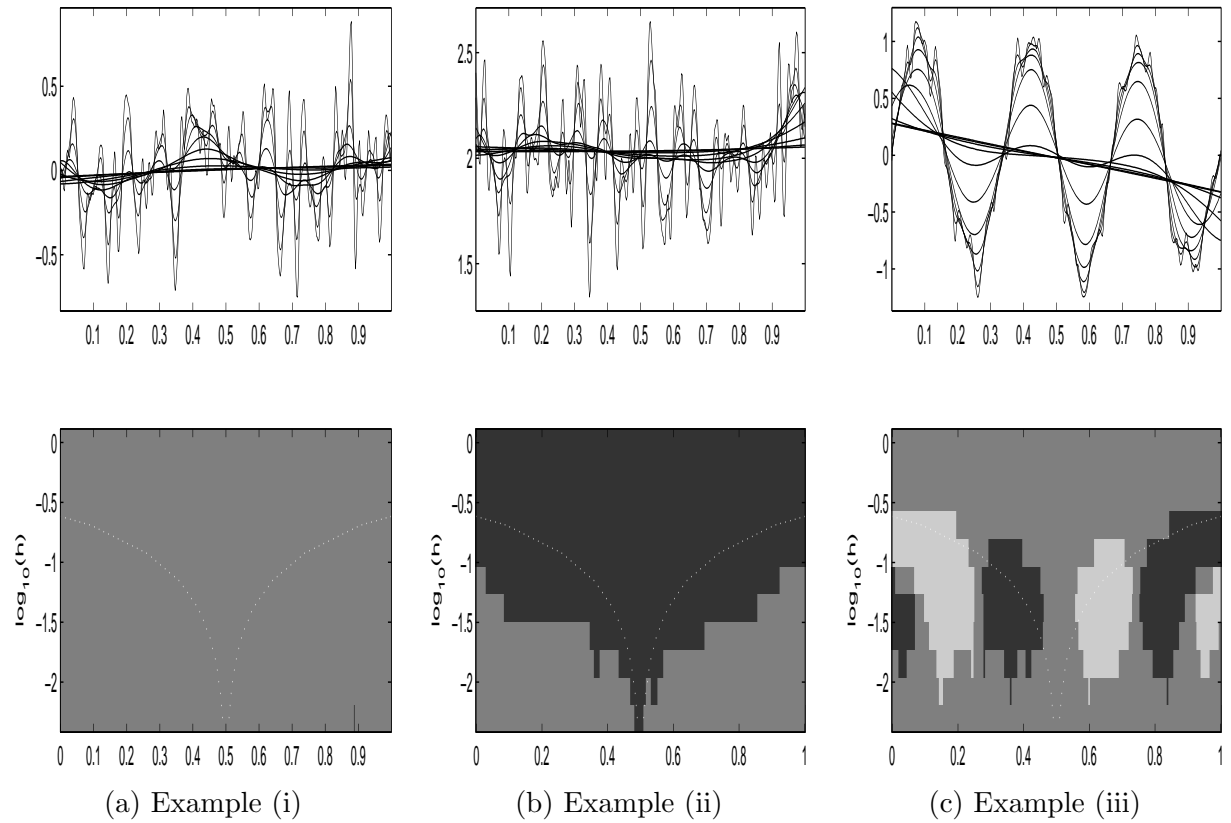


Figure 2: SiZer plots for comparing two regression curves. The two samples are drawn from (a) normal errors with the same mean, (b) normal errors with different means, and (c) normal errors with a sine curve versus a constant mean.

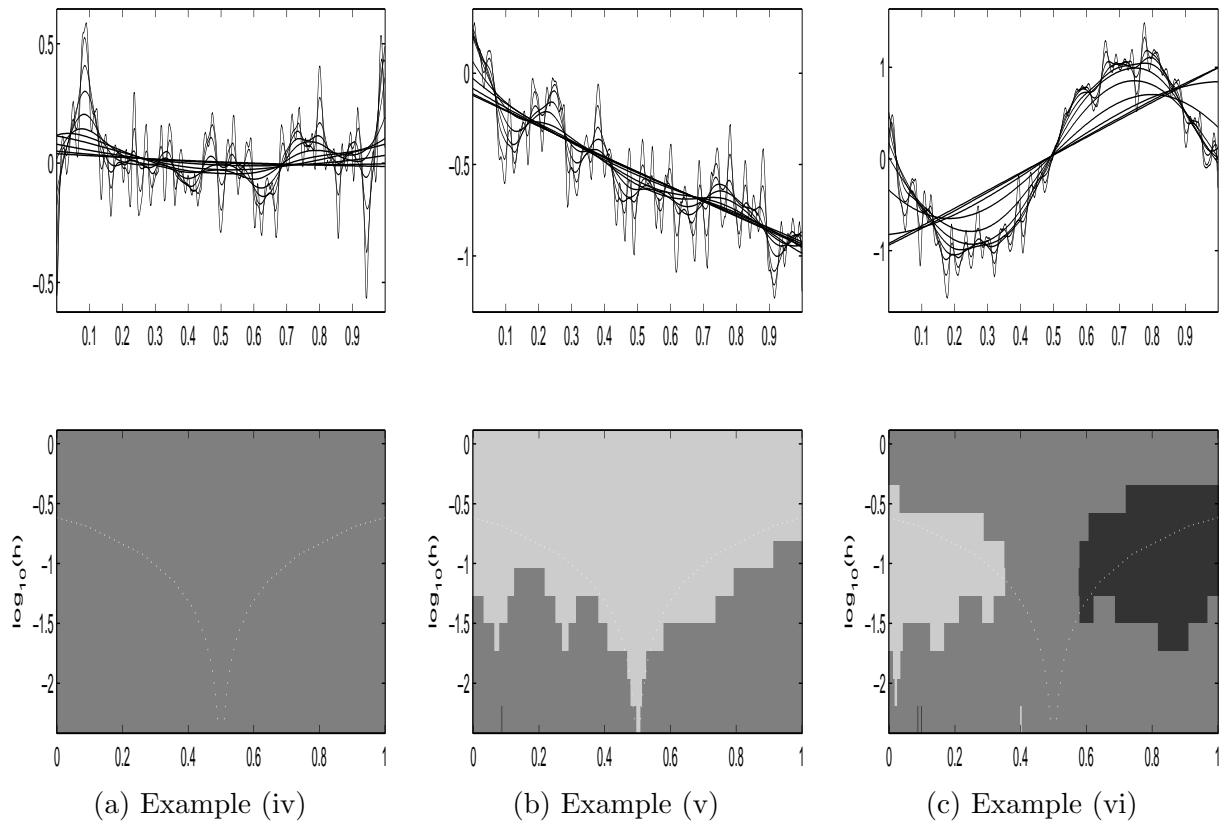


Figure 3: SiZer plots for comparing two regression curves. The two samples have different variances and the two true regression curves are (a) sine curves, (b) sine curves with a linear function versus a zero mean, and (c) exponential functions with a sine curve versus a zero mean.

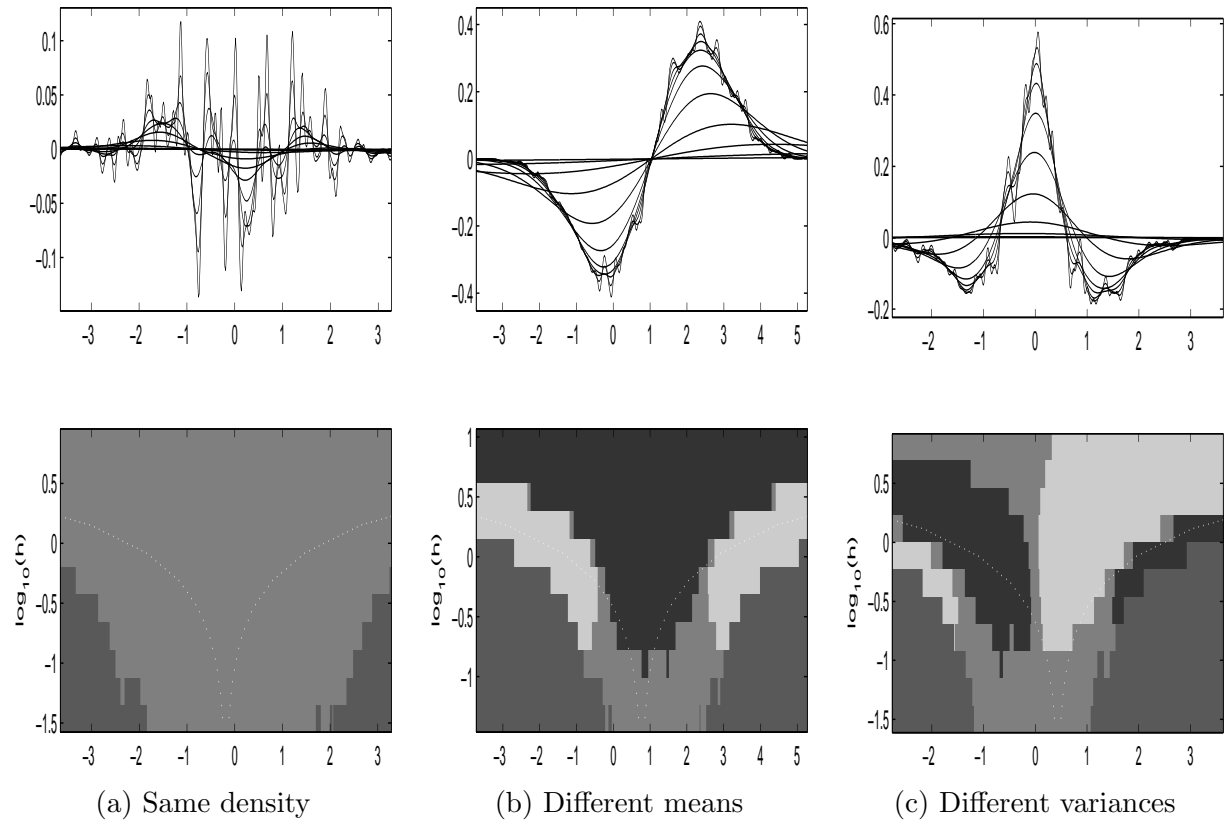


Figure 4: SiZer plots for the comparison of two densities. The two samples are drawn from (a) normal errors with the same mean, (b) normal errors with different means, and (c) normal errors with different variances.

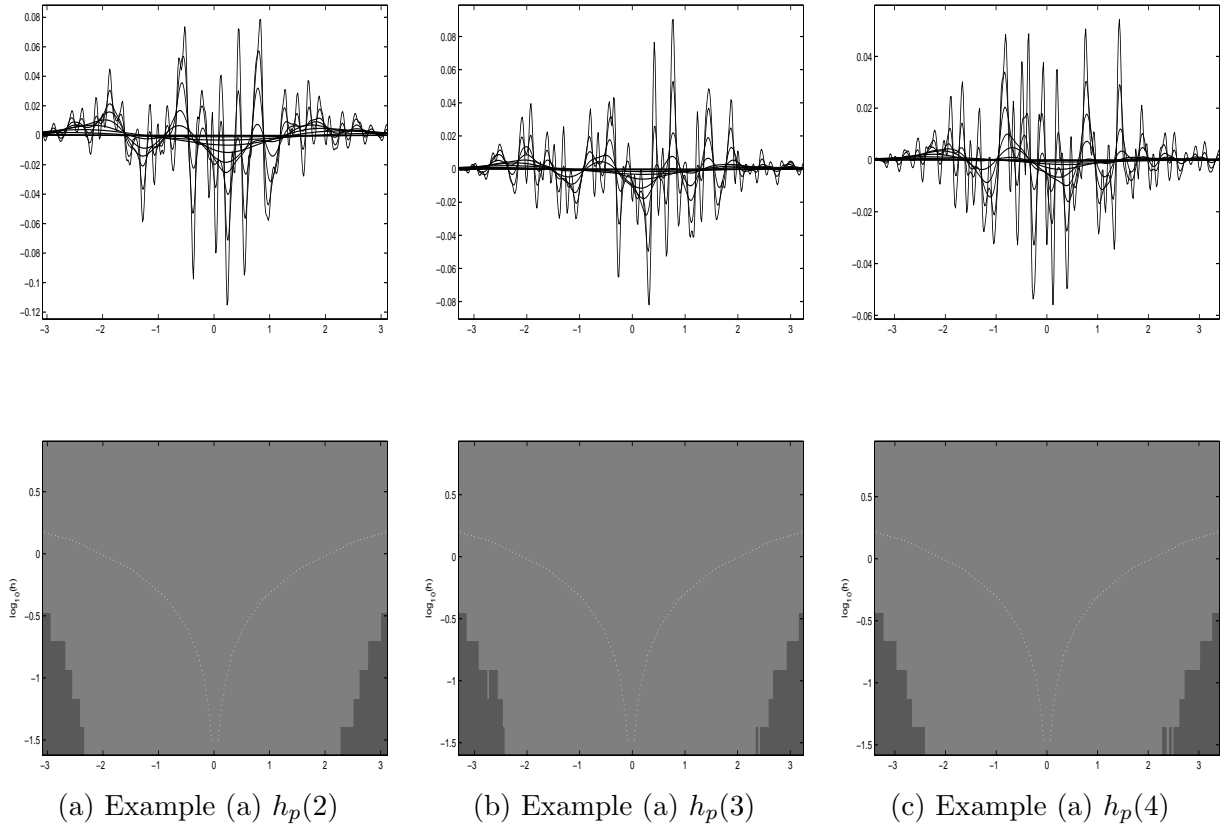


Figure 5: SiZer plots for comparing the densities of two sets of residuals. The three samples are from normal errors with the same regression function and different variances.

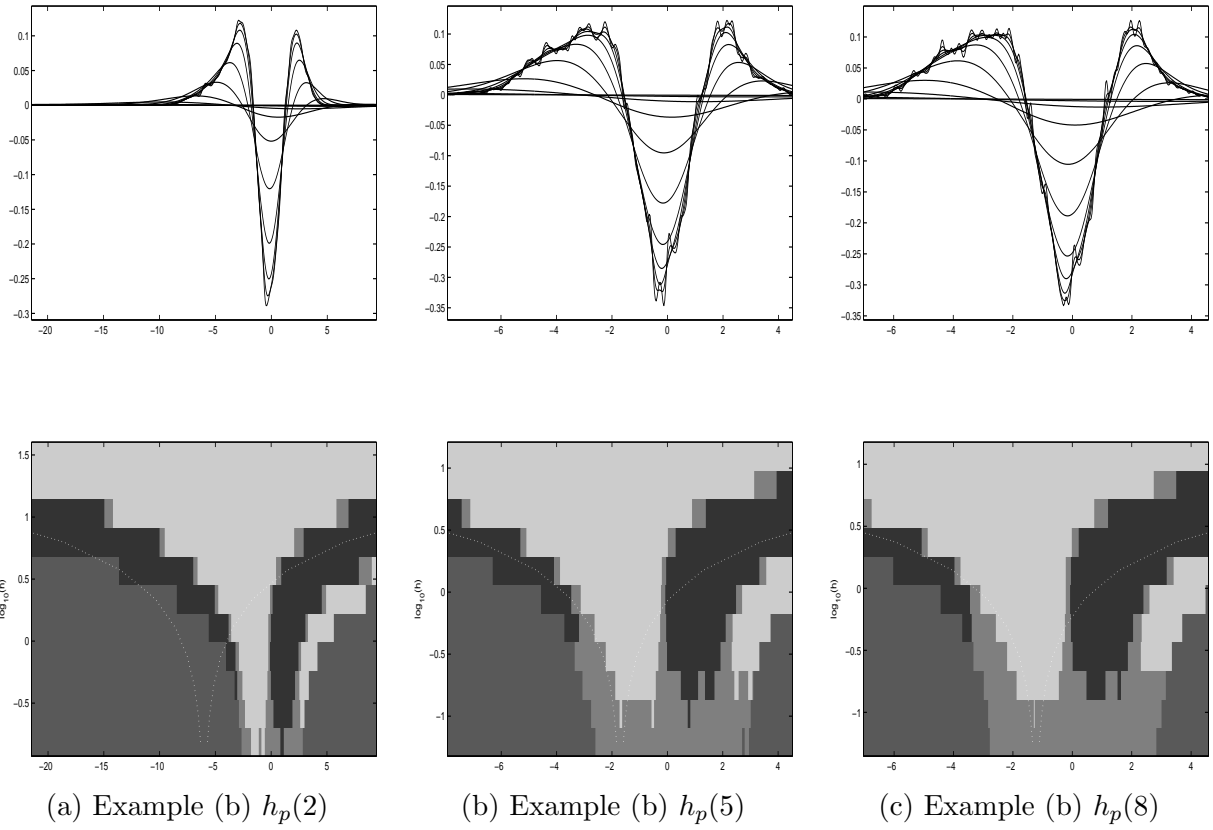


Figure 6: SiZer plots for comparing the densities of two sets of residuals. The three samples are from normal errors with the different mean and different variances.

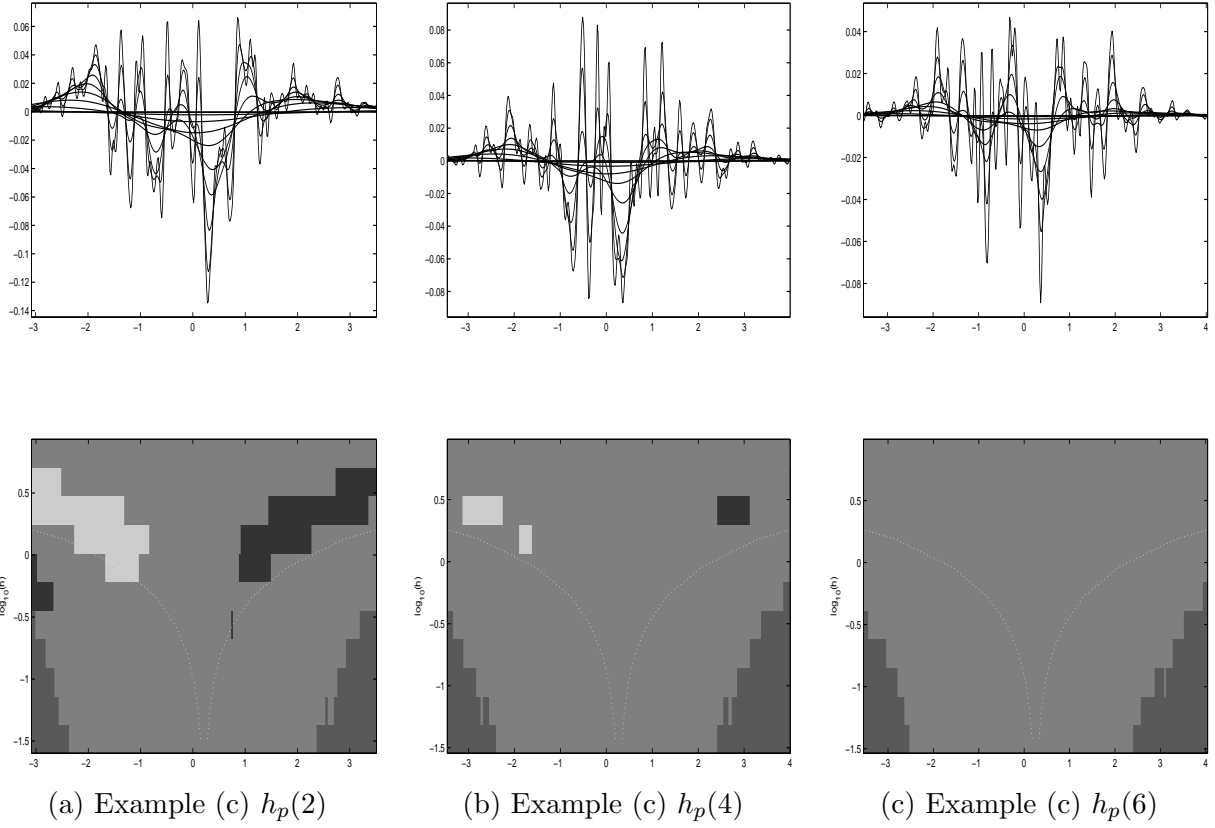


Figure 7: SiZer plots for comparing the densities of two sets of residuals. The three samples are from normal errors with the different regression function and different variances.

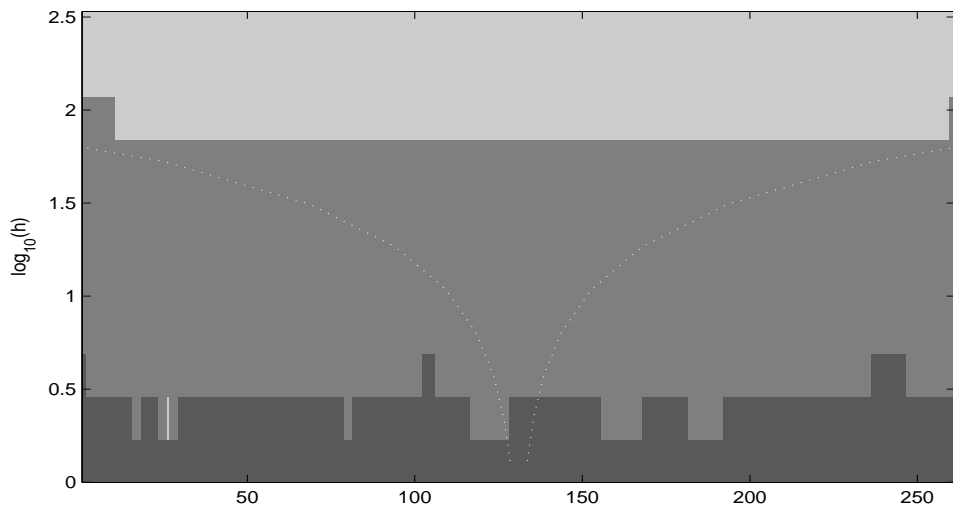
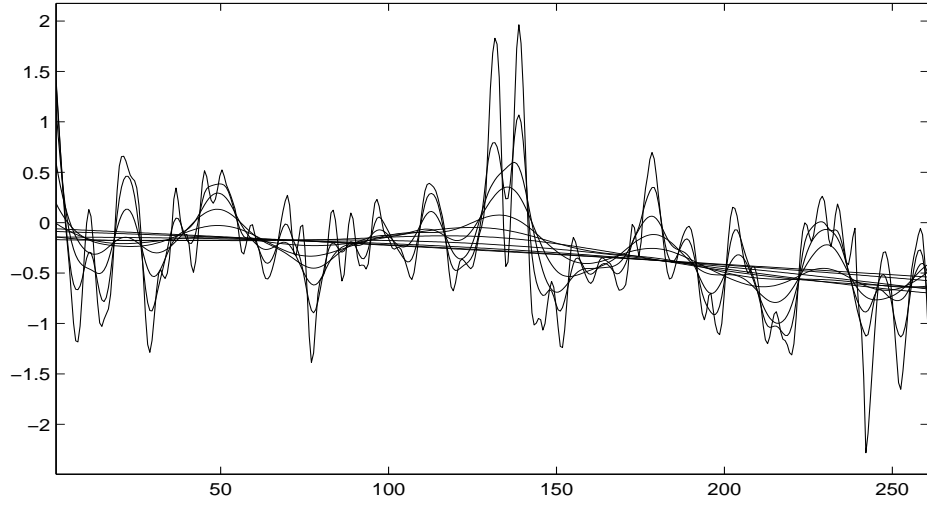


Figure 8: SiZer plots of North Carolina rain data in Example 1.

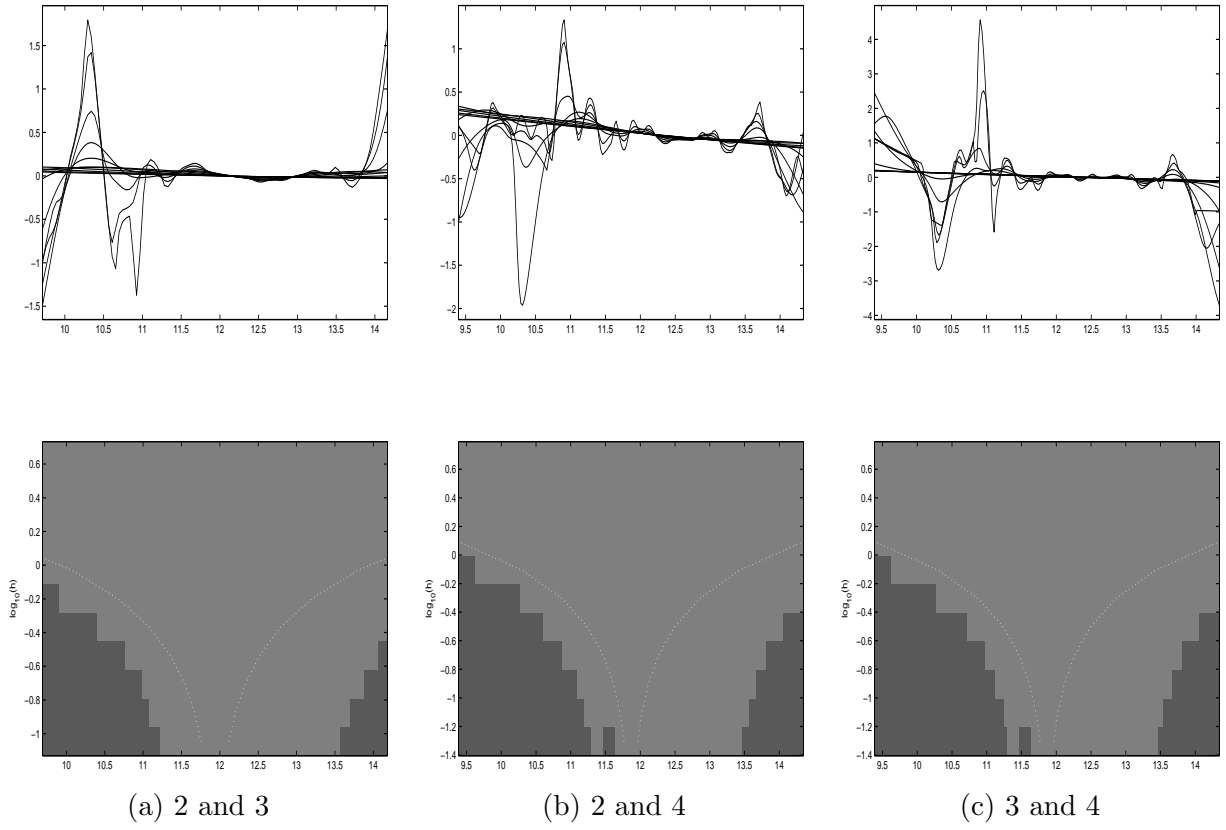


Figure 9: SiZer plots of Dutch households data in Example 2. (a) The two and three, (b) the two and four, and (c) the three and four members groups are compared each other, respectively.

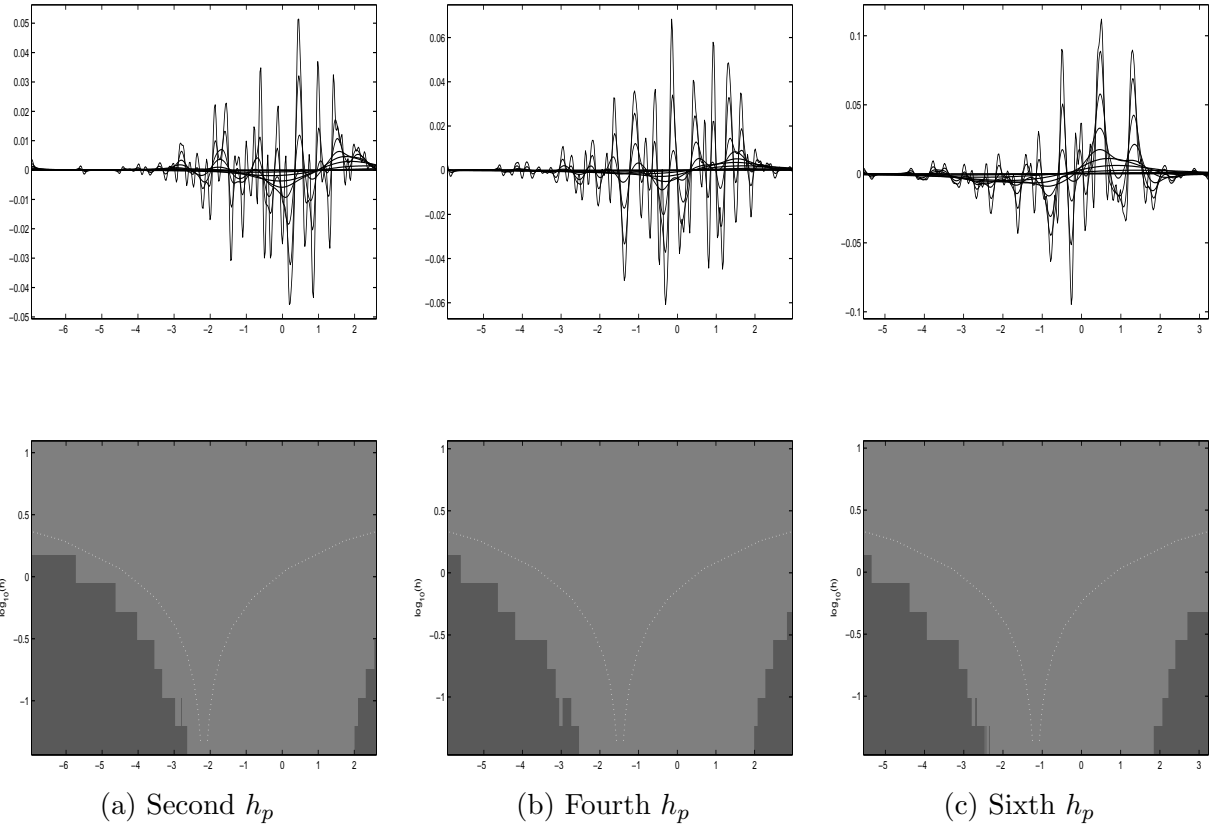


Figure 10: SiZer plots of residuals obtained from Dutch households data.