Testing significance of peaks in kernel density estimator by SiZer map

Aleksandra Baszczyńska[1]

Abstract

In kernel density estimation the researcher needs two parameters of kernel method: the kernel function and smoothing parameter called as bandwidth. The special care is required in choosing the last one. Too small value of bandwidth results in spurious peaks in the density estimator. Too large value makes it oversmoothed.

In paper, a useful technique known as SiZer map is presented. This technique helps in determining whether peaks in density estimator are significant or not. The density kernel estimator is viewed thought the different level of smoothing. The SiZer map can be used by non-experts and speeds the procedure of deciding which features are signals and which are noise. The procedure of testing the hypothesis about significance of this type is described. The applications of SiZer map is illustrated by analysis of carbon dioxide emission in countries made by density function estimation.

Keywords: *kernel density estimation, SiZer map, testing hypothesis*

JEL Classification: C12, C13

1. Introduction

Density estimation is one of the mostly used way of identifying and describing the structure of data on the basis of the random sample. Nonparametric methods, especially kernel density estimation, becomes more and more popular in the analysis of, among others, economic variables (Li and Racine, 2007). In the process of density function estimation by kernel method, the researcher has to determine two parameters of the method: kernel function and smoothing parameter. Some kernel functions are presented in literature but the influence of this parameter on the results of density estimation is regarded not to be significant. The smoothing parameter, known as the bandwidth, which determines the level of smoothing in the process of estimation, plays an important role in resulting estimator. So, the ways of choosing the appropriate value of smoothing parameter in the process of estimation are taken into regard in, for example, in Silverman (1996). The classical approach to kernel density estimator means regarding one value of smoothing parameter in kernel density estimation that results in a single estimated function. Even when a good choice of smoothing parameter is made, misleading impression can be created due to the bumps of the estimator. The problem of assessing if these bumps are "really there" and avoiding spurious noise should be regarded

[1] University of Łódź, Department of Statistical Methods, 90-214 Łódź, Rewolucji 1905 r. 41, Poland, albasz@uni.lodz.pl.

in the data structure analysis. In technical analysis this problem means determining which structure is signal and which is noise.

The SiZer map is a graphical tool used in analyzing the visible feature representing important underlying structures through different levels of smoothing what means that the estimation of kernel density function is made and analyzed for different values of bandwidths. The idea of considering a family of smooths can be found in scale space theory in computer science. Chaudhuri and Marron (2000) explored this problem in a statistical point of view.

The bump in the structure of curve like density function is characterized by going up one side and going down the other. The bump is a zero crossing of the derivative and it is statistically significant when the derivative estimate is significantly positive to the left and statistically negative to the right. The name of SiZer map stems from assessing the SIgnificant ZERo crossing of the derivative. Compering with the classical approach there are two main differences. Firstly, SiZer studies a very wide range of bandwidths instead of looking at just one. Secondly, instead of focusing on a "true underlying curve" in classical, SiZer "has" looking at the true curve viewed at varying bandwidths what can lead to recovering the significant aspects of the underlying function for different levels of smoothing. Benefits are evident - it speeds up the process of deciding which features are "really there" and makes this type of inference readily do-able by non-experts.

2. Kernel method

Kernel method can be applied in different areas: in density estimation, regression estimation, classification and pattern recognition.

In density function estimation, kernel method, known as Parzen-Rosenblat method, is one of the mostly used procedures in assessing the characteristic features of random variable. A comprehensive review of kernel density methods can be found in Silverman (1986) and Li and Racine (2007). Kernel density estimator is defined in the following way (Rosenblatt 1956; Parzen 1962):

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{1}$$

where $X_1, X_2, ..., X_n$ is the random *n*-element sample, h is the smoothing parameter, $K(\cdot)$ is the kernel function.

Kernel functions, which are in most cases density functions, are presented, among others, in Domański and Pruska (2000). The most widely used is Gaussian kernel which is density

function of normal standardized distribution. When this kernel is used in kernel density estimation, the number of zero crossing of the derivative estimate is always a decreasing function of smoothing parameter h. Because of this feature just Gaussian kernel is used in SiZer map.

In classical approach of kernel density estimation the researcher has to make a decision which value of smoothing parameter h is appropriate in particular estimation. Smoothing parameter controls, like in other nonparametric curve estimators (for example histogram), the level of smoothness. Small value of h leads to jagged estimate, while big value tends to produce over smoothed estimator. In literature some procedures indicating this value are presented, such as Silverman's rule of thumb, cross-validation, plug-in method and their modifications. In SiZer map the smoothing parameter range, instead one value like in classical approach, is taken into consideration.

3. Testing hypotheses in SiZer map

In SiZer map we have the possibility of regarding not only one density kernel estimator constructed for a particular kernel function and particular value of smoothing parameter but the family of density estimators with Gaussian kernel function and the range of smoothing parameter. The family of smooth curves is the following:

$$\left\{\hat{f}_h(x): h \in [h_{\min}, h_{\max}]\right\} \tag{2}$$

where: $h_{\min} = 2B$, B is the binwidth, $h_{\max} = x_{\max} - x_{\min}$.

The case of $h \in (0, \infty)$ is also regarded.

The family (2) represents different structures of the curve under different levels of smoothing and can be called as scale space surface. While $E\left(\hat{f}_h(x)\right)$ is the true curve viewed at different scales of resolution.

When a peak is observed, before the peak the sign of derivative is positive, at the point of maximum the derivative is equal to 0, after peak the derivative is negative. When a valley is observed before the valley the sign of derivative is negative, at the point of minimum the derivative is equal to 0, after valley the derivative is positive. Hence, peaks and valleys are determined by zero crossing of the derivative.

In SiZer map Gaussian kernel function is used:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},$$

because in kernel density estimation with this kernel function, the number of zero crossings of derivative (number of peaks) decreases monotonically with the increase of the bandwidth (Silverman 1986). Chaudhuri and Marron (2000) show that in kernel regression with Gaussian kernel function, the number of zero crossings of the mth order derivative decreases monotonically with the increase of the bandwidth.

In SiZer the following hypotheses are regarded:

$$H_0^{h,x} : \frac{\partial^m E\left(\hat{f}_h(x)\right)}{\partial x^m} = 0, \tag{3}$$

$$H_1^{h,x} : \frac{\partial^m E\left(\hat{f}_h(x)\right)}{\partial x^m} \neq 0. \tag{4}$$

If $H_0^{h,x}$ is rejected, there is an evidence that $\frac{\partial^m E\left(\hat{f}_h(x)\right)}{\partial x^m}$ is positive or negative, according to the sign of $\frac{\partial^m \hat{f}_h(x)}{\partial x^m}$ (Chaudhuri and Marron, 2000). The test is done independently at each location in the scale space.

In the calculation of the quantile q the following fact is used: if two locations u_1 and u_2 are sufficiently far apart, relative to h then $\hat{f}_h(u_1)$ and $\hat{f}_h(u_2)$ are independent which implies that $\hat{f}_h'(u_1)$ and $\hat{f}_h'(u_2)$ are independent. The simultaneous confidence limit problem is then approximated by m independent confidence intervals. The estimate for m is calculated through an $ESS(x,h)$ estimated effective sample size:

$$ESS(x,h) = \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}{\frac{1}{h}K(0)}. \tag{5}$$

When kernel is uniform $ESS(x,h)$ is simply the number of data points in the window of width h centered at x. For Gaussian kernel the data points are downweighted according to the height of the kernel function. Next m is chosen to be the number of independent blocks (m confidence intervals) of average size available from a dataset of size n:

$$m(h) = \frac{n}{avg_x ESS(x,h)}. \tag{6}$$

The $ESS(x,h)$ can also be used to indicate where the smooth is based on sparse data by highlighting the regions where $ESS(x,h) \leq n_0$. Chaudhuri and Marron (1999) suggested that

$n_0 = 5$. Therefore the calculation of block size $m(h)$ is modified to avoid problems with small $ESS(x,h)$ to:

$$m(h) = \frac{n}{avg_{x \in D_h} ESS(x,h)}, \tag{7}$$

where $D_h = \{x : ESS(x,h) \geq 5\}$, is the set of locations where the data are "dense".

Assuming independence of $m(h)$ blocks of data the approximate simultaneous quantile for a $(1-\alpha)100\%$ confidence interval is:

$$q(h) = \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m}}{2}\right). \tag{8}$$

For the derivative estimate $\hat{f}_h'(x)$ the confidence limits, depending on h, can be constructed:

$$\left\{\hat{f}_h'(x) - q\,\hat{sd}\left(\hat{f}_h'(x)\right), \hat{f}_h'(x) + q\,\hat{sd}\left(\hat{f}_h'(x)\right)\right\}, \tag{9}$$

where: q is appropriate quantile, and calculation of $\hat{sd}\left(\hat{f}_h'(x)\right)$ is based on the fact that the derivative estimator $\hat{f}_h'(x)$ is an average of the derivative kernel functions:

$$\hat{\text{var}}\left(\hat{f}_h'(x)\right) = \hat{\text{var}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}K'\left(\frac{X_i - x}{h}\right)\right) =$$
$$= \frac{1}{n}s^2\left(\frac{1}{h^2}K'\left(\frac{X_1 - x}{h}\right),...,\frac{1}{h^2}K'\left(\frac{X_n - x}{h}\right)\right),$$

where $s^2(k_1,...,k_n)$ is the sample variance of $k_1,...,k_n$.

On the vertical axis in the SiZer map is x and on the horizontal axis is h. From the SiZer map it is possible to present the information, for given x and h, about the positivity and negativity of the derivative of $f_h(x) = \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{u-x}{h}\right) f(u)du$. The following color codes are used:

1. blue, $\hat{f}_h(x)$ is significantly increasing, (zero is greater than the upper confidence limit),

2. red, $\hat{f}_h(x)$ is significantly decreasing (zero is less than the lower confidence limit),

3. purple, $\hat{f}_h(x)$ is not significantly increasing or decreasing (zero within confidence limits),

4. grey, indicates regions where the data are too spare to make statements about significance, the effective sample size is less than 5.

In SiZer map the $\log_{10}$ scale is used for h in the display (it gives smooths that are more equally spaced). The dotted white curves show effective window widths for each bandwidth, as intervals representing $\pm 2h$ (± 2 standard deviations of the Gaussian kernel).

There is a variation of SiZer map named SiCon map (Significant CONvexity), where statistical inference is made taking into account second derivative and regions of statistically significant curvature are highlighted (special color code is used: cyan – significant concavity, downward curvature; orange – significant convexity, upward curvature; green – no significant curvature).

4. Application of SiZer map

In literature there are examples of using the Sizer map in analysis of economic data (Zambom and Dias, 2012), medical data (Skrovseth, Bellika, Godtliebsen, 2012) or geochemical data (Rudge, 2008).

The application of SiZer map is illustrated in the analysis of the carbon dioxide emission in countries in the world. The data was downloaded from the data bank (http://data.worldbank.org/topic/environment [25.02.2014]). Total carbon dioxide emission (in thousand metric tons) is available for 214 countries in the world for 1960-2010. The last year was taken into account in the research. Samples of sizes 10, 30 and 50 countries were chosen and on the basis of these samples the SiZer maps are obtained using the codes in Matlab. Figure 1 shows the results where the kernel density estimator for different values of smoothing parameters is presented (top) and the SiZer map (bottom) for sample size 10.

In the SiZer map blue shows regions of significant positive $\hat{f}_h(x)$, red regions of significantly negative $\hat{f}_h(x)$, purple regions where $\hat{f}_h(x)$ is not significantly increasing or decreasing and grey regions where it is not possible to make inference. For large values of bandwidth the density function significantly increases, then there is a region where SiZer is unable to distinguish and then there is a region where the density function significantly decreases. The SiZer map results in grey region for small values of bandwidth, it means that it is not possible to separate signal and noise. This situation is closed connected with the sample size. For such small sample size the process of estimating the density function is rather difficult.

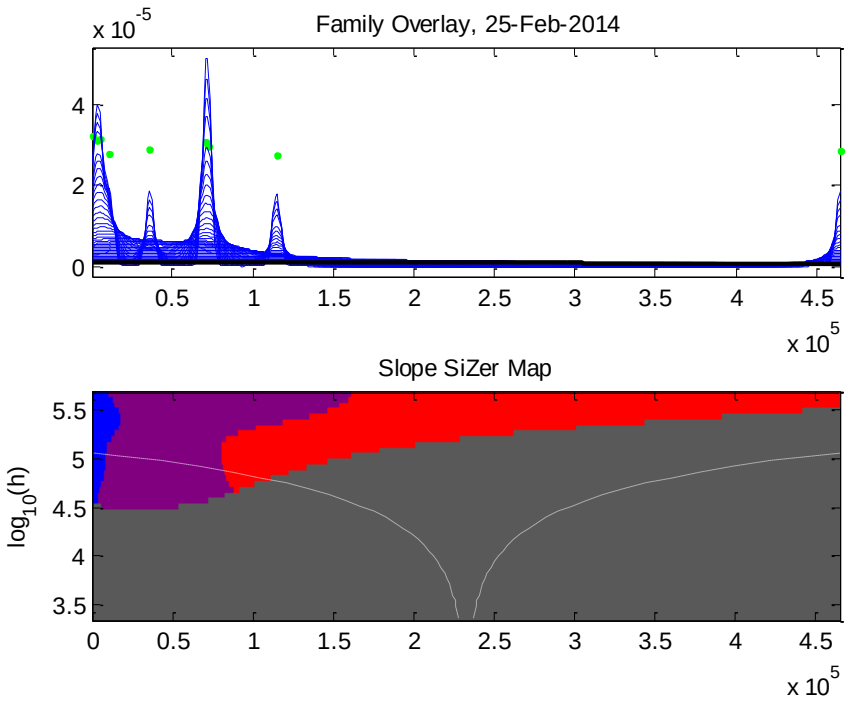

Fig. 1. SiZer map for n = 10.

Figure 2-3 presents SiZer map for bigger sample sizes. It should be noted that when sample size is increasing, the grey region becomes smaller.

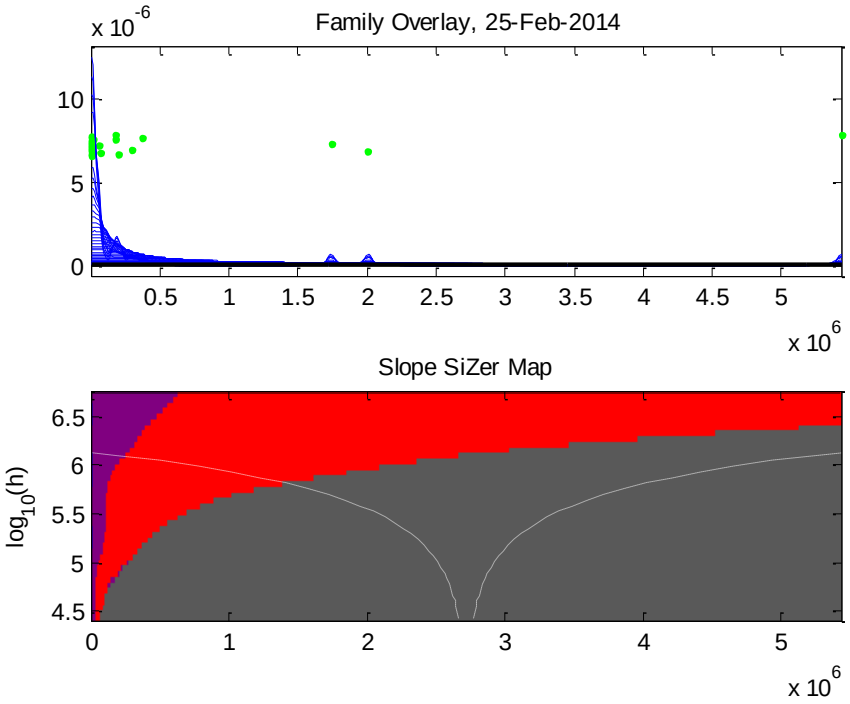

Fig. 2. SiZer map for n = 30.

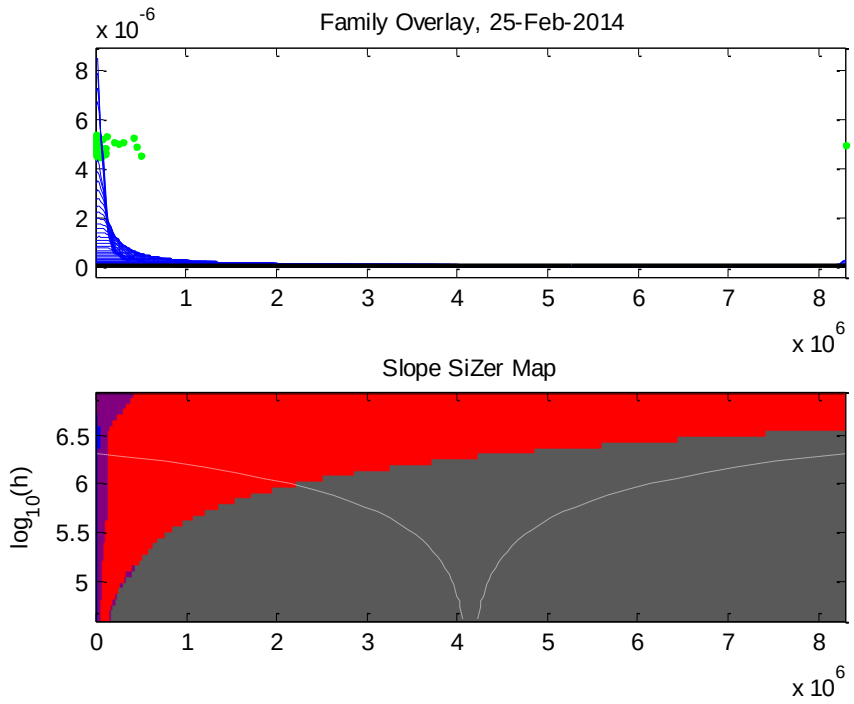

Fig. 3. SiZer map for n = 50.

Conclusion

The SiZer map is very useful technique in determining structure of the data. It can be treated as nonclassical method because of its multiple results. Taking into account not only one value of smoothing parameter like in classical approach but the range of values, broadens the researcher's point of view. But the special issue should be underlined: the sample size. Too small sample size unables detailed analysis of structure of date. Further research should be made to determine the influence of the sample size on the results of SiZer map.

Acknowledgements

This work was supported by the project number DEC-2011/01/B/HS4/02746 from the National Science Centre.

References

Chaudhuri, P., & Marron, S. (1999). SiZer for exploration of structure of curves. *JASA*, *94*, 807-823.

Chaudhuri, P., & Marron, S. (2000). Scale space view of curve estimation. *The Annals of Statistics*, *28*, 402-428.

Domański, Cz., & Pruska, K. (2000). *Nieklasyczne metody statystyczne*. PWE, Warszawa.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics. Theory and practice*, Princeton University Press, Princeton and Oxford.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 3.

Rosenblatt, M. (1956). Remarks on some nonparametric estimation of a density function, *Ann. Math. Statist.*, 27.

Rudge, J. (2008). Finding peaks in geochemical distributions: A re-examination of the helium-continental crust correlation, *Earth and Planetary Science Letters*, 274, 179-188.

Silverman, B.(1996). *Density estimation for statistics and data analysis*, Chapman and Hall, London

Skrovseth, S., Bellika, J., & Godtliebsen, F. (2012). Causality in scale space as an approach to change detection. Retrived from http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0052253&representation=PDF.

Turner, L. (2003). Exploring structure of curves using SiZer. Retrived from http://www.stat.ubc.ca/~webmaste/howto/statsoftware/misc/sizer/paper.pdf.

Zambom, A. Z., & Dias, R. (2012). A review of kernel density estimation with applications to econometrics. Retrieved from http://arxiv.org/pdf/1212.2812.pdf.