A Heuristic Method for The Selection of The Regularization Parameter in Kernel Regression

Helder Fernando Pedrosa e Sousa, Francisco Lage Calheiros

***Abstract*— A new method and criterion are proposed to find the best regularization or smoothing parameter, *h*, in nonparametric regression by the kernel method on a sample obtained at points equally spaced. The new heuristic criterion is based in the skewness coefficient of the regression estimator derivatives in order of *h* on all data points, and allows to obtain a comprehensive tool to construct a practical method to control the smoothness of the curve, without the usual and more elaborated criterion based in convergence over the estimator that have, for example, conditions on kernel functions. With a simple calculation for a distance, it is possible to obtain the regularization parameter, *h*. The selection of the regularization parameter must be clearly different if it is aimed to minimize the error or to calculate the zeros of the regression function.**

***Index Terms*—Heuristic method, Kernel regression, Nonparametric Regression, Smoothing parameter selection**

I. Introduction

The original idea of kernel estimation was initiated in an attempt to estimate the density function, $d_n(x,h)$, of a random variable *X*,

$$d_n(x,h) = \hat{f}(x) = (1/nh)\sum_{i=1}^{n} K\left((x-x_i)/h\right), \qquad (1)$$

from a sample $(x_1,\ ,x_n)$, of size *n*, using classical histograms [1]-[3]. The use of the kernels, $K(x)$ as core functions depends of a regularization parameter, *h*, which for higher and for smaller values of *h*, corresponds to a narrower or wider curve, respectively.

The density estimation is then an average of density distributions with equal weights. Several kernel functions are used, some of the simpler ones are: Uniform, Epachinikov, Biweight, triangular, and Gaussian, among others. In this work, the Gaussian function is used,

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

due to its properties, in particular, to the fact that it is continuously and infinitely differentiable.

By using the kernel density estimator of the joint probability function of the independent variable *X* and the dependent variable *Y*, applied to a set of points $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ of dimension *n*,

$$\hat{f}(x,y) = \\ = \left(1/nh^2\right)\sum_{i=1}^{n} K\left((x-x_i)/h\right)K\left((y-y_i)/h\right), \qquad (3)$$

and applying it to the empiric regression estimator,

$$y(x) = E\left(Y|X=x\right) = \\ = \int y f_{Y|X=x}(y)\,dy = \int y f(x,y)/f(x)\,dy, \qquad (4)$$

where $f(x,y)$ is the joint function probability, $f_{Y|X=x}(y)$ the *Y* probability density function conditioned by $X=x$, and $f(x)$ the marginal density function of *X*, it will be possible to obtain the Nadaraya-Watson regression estimator [4], NW or Kernel regression estimator,

$$\hat{m}_h(x) = \hat{y}(x) = \\ = \sum_{k=1}^{n} y_i K\left((x-x_k)/h\right) \Big/ \sum_{k=1}^{n} K\left((x-x_k)/h\right). \qquad (4)$$

The kernel method applied to regression, or to density function, always presents some difficulties, because there is not a simple method to find the best *h*, and no evolution was found in this matter on the last decade. Some known methods, are for example, the cross-validation or the Plug-in method [5]-[9], which are based on calculations of values depending on the properties of the kernel functions, as well as on the unknown specific density function. Other approaches have been made defining new asymmetric or not all positive kernel functions [10]-[18].

The problem of sample size on the regression estimator is important when, as in most of the cases, the smoothing parameters should be local, i.e., individualized for each kernel, and when there is a smaller sample size, an individualized selection of these parameters, or another kernel function construction becomes necessary. Moreover, a sample size that can give enough information to estimate all necessary regularity parameters must be considered.

This work presents a new heuristic method to find the optimal smoothing or regularization parameter *h* for samples obtained at constant intervals.

II. Looking for an Estimation of the Optimal Smoothing or Regularization Parameter

The criterion presented is based in the study of the behavior of derivative in order of *h* in all observed points. It is applicable to a set of *n* points, $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, where the dependent variable *Y* was obtained with some error in constant intervals of valor of the independent variable *X*, i.e., $x_{i+1}-x_i = a$, where *a* is a positive real constant and

Manuscript received October 23, 2016

Helder Fernando Pedrosa e Sousa, Mathematics Department, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

Francisco Lage Calheiros, Civil Engineering Department, Faculty of Engineering, University of Porto, Porto, Portugal

$j=0,\dots,n-1$. This type of data allow the use of symmetric kernel and constant smoothing parameter *h* in all kernel functions simplifying the regression estimator use. There are many science areas that use this type of data, where the sample is collected in constants intervals of time or space, giving applicability of the presented method.

The regression model is $y_i=f(x_i)+\varepsilon_i$, $i=1,\dots,n$ with the error ε_i as a normal distribution with zero mean, μ and some positive standard deviation σ and $f(x)$ is the unknow function that relates *y* with *x*.

The two samples used in this work were generated using $f(x)$, as one of even degree and another of odd degree polynomial functions, both with multiple zero points, because they are not so favorable to obtain good regression curves. When $f(x)$ has only one type of monotony, or the sample error is lower, the method will work better that shown in examples 1 and 2 below.

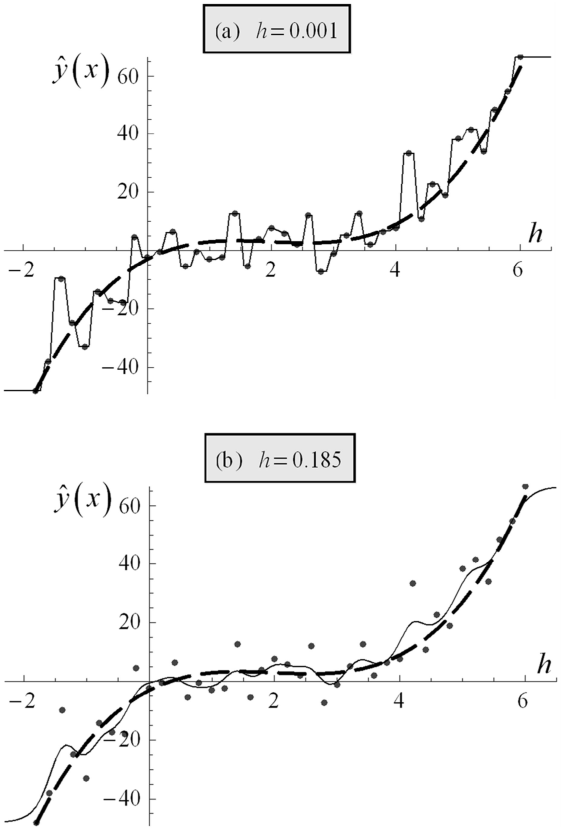
A. *Example 1*

$n=40$, $f(x)=(x-2)^3-x+5$ with $\varepsilon_i\sim N(0,8)$, and sample values of *x* axis from -1.8 to 6 equally spaced.

B. *Example 2*

$n=100$, $f(x)=(x-2)^2-5$ and $\varepsilon_i\sim N(0,8)$, and sample values of *x* axis from -1.92 to 6 equally spaced.

For data of example 1, regression was calculated for different values of *h*, which are shown in Fig. 1.



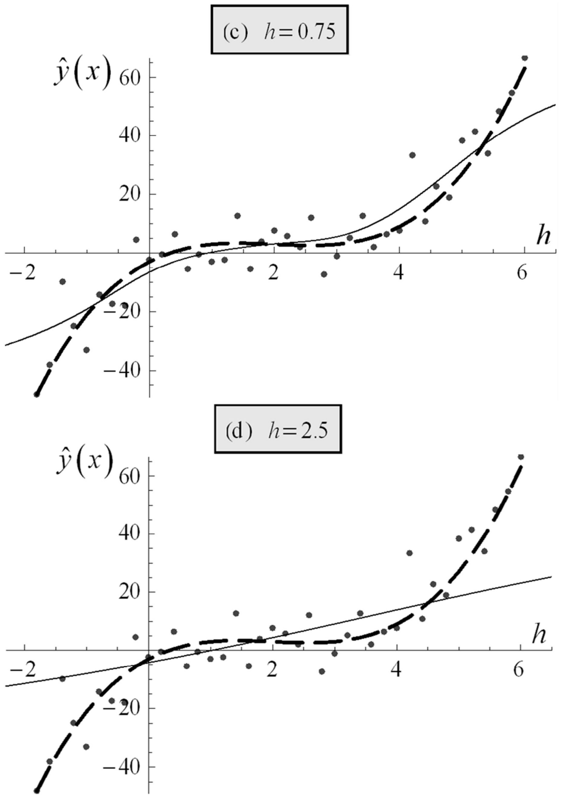


Fig. 1. Different steps of nonparametric regression for data of example 1, where the dashed line corresponds to the function plot, $f(x)$, and the continuous line to the regression estimation for the regularization parameter, *h*, on the four indicated cases, (a) *h*=0.001, (b) *h*=0.185, (c) *h*=0.75, (d) *h*=2.5.

As it can be observed in Fig. 1 (a), the "discrete" behavior to parameters *h* near zero makes the regression a step function, the curve becomes as close to points (x_i,y_i), as intended, simply by decreasing the value of *h*, and that is why the least squares method and kernel regression goes in some sense against one another. The curve is locally characterized by the individually of the points, and it does not take into account the group's behavior. When the parameter has very high values, Fig. 1 (c) and (d), points behave as a single group in the sense that the curve will slowly converge at all points to the value which characterizes the variable *y*: the average of the observed *y* values. The limit curve, as *h* goes to infinity, is a horizontal line as it can easily be observed, by calculation of the limits on (5). The question of calculating a value of *h*, which is optimal, Fig. 1 (b), lies in understanding the process of moving from the "discrete" to the "group" zone and how can the balance between these two areas be measured. The optimal *h* that will be obtained, will be a value that achieves the equilibrium between a smooth, and close enough curve of the observed points.

It can be easily observed that the dominant behavior of the evolution of the curve is in the vicinity of each point. The spaces between the points follow the influence of their nearest points, as *h* increases, drawing the complete curve. As a result, the derivative of the regression curve $\partial\hat{y}(x,h)/\partial x$ at each point, x_i, $i=1,\dots,n$, of the sample (for Gaussian kernels),

$$\left.\frac{\partial \hat{y}(x,h)}{\partial x}\right|_{x=x_i} = \frac{\sum_{i=1}^{n}\left(\frac{\hat{y}(x_j,h)-y_i}{h}\right)\left(\frac{x_j-x_i}{h}\right)K\left(\frac{x_j-x_i}{h}\right)}{\sum_{k=1}^{n}K\left(\frac{x_j-x_k}{h}\right)}, \qquad (5)$$

measures the smoothing of the regression curve, in three phases, "discrete", "transition" and "group" zones:

A. *"Discrete" Zone*

For a small *h*, the slope at observed points extend beyond the vicinity of each of the sample value of *x*, with an abrupt variation of the curve in the middle of each interval between points. In this interval, the slopes are approximately zero in all *x* values observed, and estimated $\hat{y}(x_i)$ values near y_i with $i=1,\dots,n$.

B. *"Transition" Zone*

The optimum *h*, *h**, can be found in this zone. The slopes (from the regression line in all the sample values *x*) evolve to a state where no rapid changes in the curve exist, i.e., a smooth regression curve is obtained. It should be noted that the most important property in this area lies in the fact that the slopes reached its maximum. From here, values will continuously decrease. In this zone, as in the above zone, the estimated values of *y* in the range of the abscissa of the points remain close enough to the *y* corresponding observed values.

C. *"Group" zone*

After a situation of equilibrium, all regression curve slopes, at each point, x_i, $i=1,\dots,n$, evolve to zero but, unlike the first area, the estimated values of *y* evolve to the observed average, moving away, of the observed values, except in regard to the more central ones.

Also, for all values of *h*, regressions using extrapolations are not valid. In fact, the interior points evolve faster in the value of its derivatives than the outer points, stays near the derived optimum value before decreasing, as *h* increases. Throughout the evolution of *h*, from near zero to infinity, the outer points cannot even reach the expected slope, converging back to zero before they get there.

The derivative of the regression curve at each observed sample value of *x*, is therefore a good tool to control the smoothness of the regression curve, and hence to find a way to reach the optimum h^*.

Let $\mathbf{D}(h)$ be the vector of all derivatives of the kernel regression estimator, $\hat{y}(x,h)$ on each of the observed value of the variable *X*, $\mathbf{x}=(x_i,\dots,x_n)$,

$$\mathbf{D}(h)=\mathbf{D}(h\,|\,\mathbf{x},\mathbf{y})=\left(\left.\frac{\partial \hat{y}(x,h)}{\partial x}\right|_{x=x_1},\dots,\left.\frac{\partial \hat{y}(x,h)}{\partial x}\right|_{x=x_n}\right), \quad (6)$$

where $\mathbf{y}=(y_i,\dots,y_n)$.

Fig. 2 presents, for the example 1, the skewness of the vector $\mathbf{D}(h)$ in function of *h*, and two function of errors, $\varepsilon_s(h)$ and $\varepsilon_f(h)$, that are respectively the mean square error between the estimator of the observed values and of the function $f(x)$, used to generate the sample,

$$\varepsilon_s(h)=\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}(x_i)-y_i\right)^2, \qquad (7)$$

and,

$$\varepsilon_f(h)=\frac{1}{n}\sum_{j=1}^{n}\left(\hat{y}(x_i)-f(x_i)\right)^2. \qquad (8)$$

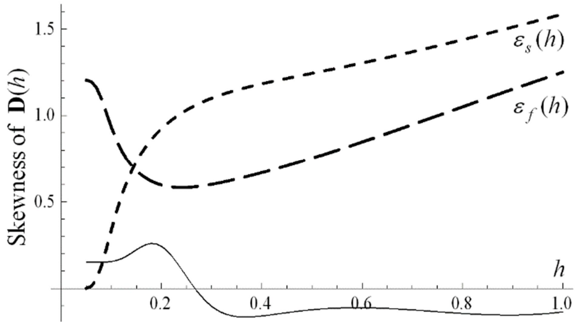


Fig. 2. Skewness of the vector $\mathbf{D}(h)$ in function of *h*, for example 1. The long and small dashed line corresponds to the two function of errors, $\varepsilon_s(h)$ and $\varepsilon_f(h)$, of the mean square error between the estimator and the observed values, and of the function $f(x)$, respectively.

Fig. 2 shows the first and absolute maximum of skewness function near the minimum of function $\varepsilon_f(h)$. The skewness coefficient of a variable *X* is the third moment of *X* divided by the third power of the second moment of *X*.

The curve of $\mathbf{D}(h)$ variance is similar for any sample and it can be observed that the maximum value represents the change from the "discrete" zone to the "transition" zone, slowly decreasing thereafter to zero. The curves of skewness and kurtosis have values without much variation in the same "discrete" zone, where the variance grows quickly to its maximum, and then starts to present different behaviors in the "transition" area. Already in the "group" zone, the skewness slowly converges to zero and to some positive value for the kurtosis. What is expected to find for optimal *h* is a smooth regression curve, which is represented by a vector $\mathbf{D}(h)$ with different consecutive component values but not very distant from each other. It is therefore necessary to balance these components, representing slopes, between asymmetry / information and symmetry / kurtosis, as they should represent the maximum sample information, which is reflected in having the maximum absolute asymmetry, with the smaller possible kurtosis.

Thus, the proposed criterion is that the optimal value of h is the one within the "transition" zone that gives $\mathbf{D}(h)$ a greater asymmetry (skewness) in the absolute value, to the lowest kurtosis. The "transition" zone can be located in order to stay after the value of h, that maximizes the variance of $\mathbf{D}(h)$, and for almost cases the first absolute maximum of the skewness corresponds to the optimal h.

Fig. 3 presents, for the example 2, the skewness of the vector $\mathbf{D}(h)$ in function of h, and two function of errors, $\varepsilon_s(h)$ and $\varepsilon_f(h)$. In this sample, with a larger n, the first absolute maximum of the skewness function is located before the minimum of function $\varepsilon_f(h)$, resulting of the higher error used generating data, and the $\mathbf{D}(h)$ skewness function favors the collection of information on sample data, producing a more discrete regression as it can observed in Fig. 4. If data was generated with less error, the optimal h would be more near the minimum of $\varepsilon_f(h)$.

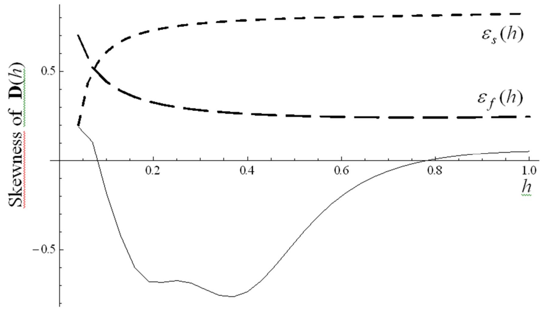


Fig. 3. Skewness of the vector $\mathbf{D}(h)$ in function of h, for example 2. The long and small dashed line corresponds to the two function of errors, $\varepsilon_s(h)$ and $\varepsilon_f(h)$, of the mean square error between the estimator and the observed values, and of the function $f(x)$, respectively.

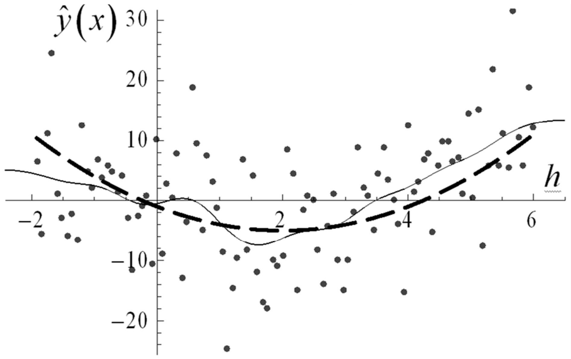


Fig. 4. Nonparametric regression for data of example 2, where the dashed line corresponds to the function plot, $f(x)$, and the continuous line to the regression estimation for the regularization parameter, h=0.4.

III. Heuristic method

In conclusion, this work presents a heuristic method to obtain an optimal regularity parameter, h* for kernel regression on a sample with equally spaced values.

Let $\hat{y}(x)$ be the kernel estimator applied to a set of points $\{(x_1,y_1),\dots,(x_n,y_n)\}$, where $x_{i+1}-x_i=a$, and a is a positive real constant and $j=0,\dots,n-1$. Also, h is the regularization parameter and $K(x)$ the kernel function differentiable associated with the estimator. The optimum value of h, h* is the value that maximizes the absolute value of the function that gives the skewness coefficient of the vector,

$$\mathbf{D}(h)=\mathbf{D}(h\,|\,\mathbf{x},\mathbf{y})=\left(\left.\frac{\partial\hat{y}(x,h)}{\partial x}\right|_{x=x_1},\dots,\left.\frac{\partial\hat{y}(x,h)}{\partial x}\right|_{x=x_n}\right),\quad(10)$$

for values of h in the “transition” zone that are greater than the maximum of variance of the same vector $\mathbf{D}(h)$.

References

[1] M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” *Ann. Math. Stat.*, vol. 27, no. 3, 1956, pp. 832-837.

[2] E. Parzen, “On estimation of a probability density function and mode,” *Ann. Math. Stat.*, vol. 33, no. 3, Sep. 1962, pp. 1065-1076.

[3] W. Wertz, *Statistical Density Estimation: A Survey*, No. 13, Vandenhoeck & Ruprecht, 1978.

[4] E. A. Nadaraya, “On estimating regression,” *Theor. Probab. Appl.*, vol. 9, no. 1, 1964, pp. 141-142.

[5] W. Härdle, and J. S. Marron, “Optimal bandwidth Selection in nonparametric regression function estimation,” *Ann. Stat.*, vol. 13, no. 4, Dec. 1985, pp. 1465-1481.

[6] P. Hall, and J. S. Marron, “Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation,” *Probab. Theor. Relat. Field*, vol. 74, no. 4, Apr. 1987, pp. 567-581.

[7] W. Härdle, *Smooting Technique with Implementation in S, ser.*, Springer Series in Statistics. New York: Springer-Verlag, 1991.

[8] J. Fan, P. Hall, M. Martin, and P. Patil, “On local smoothing of nonparametric curve estimators,” *J. Am. Stat. Assoc.*, vol. 91, no. 433, Mar. 1996, pp. 258-266.

[9] A. Tsybakov, *Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats*, New York: Springer Science+Business Media, 2009.

[10] P. Green, and B. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, CRC Press, Boca Raton, 1993.

[11] M. Samiuddin, and G. M. El-Sayyad, “On nonparametric kernel density estimates,” *Biometrika*, vol. 77, no. 4, Dec. 1990, pp. 865-874.

[12] Holiday, D. “Near optimal weights in nonparametric regression under some common restrictions,” *Stat. Probab. Lett.*, vol. 22, no. 1, Jan. 1995, pp. 33-42.

[13] S. X. Chen, “Beta kernel estimators for density functions,” *Comput. Stat. Data Anal.*, vol. 31, no. 2, Aug. 1999, pp. 131-145.

[14] S. X. Chen, “Probability density function estimation using gamma kernels,” *Ann. Inst. Stat. Math.*, vol. 52, no. 3, Sep. 2000, pp. 471-480.

[15] K. M. Abadir, and S. Lawford, “Optimal asymmetric kernels,” *Econ. Lett.*, vol. 83, no. 1, Apr. 2004, pp. 61-68.

[16] W. Härdle, M. Muller, S. Sperlich, and A. Werwatz, *Nonparametric and Semiparametric Models*, ser., Springer Series in Statistics. Berlin; New York: Springer-Verlag Berlin Heidelberg, 2004.

[17] O. Scaillet, “Density estimation using inverse and reciprocal inverse Gaussian kernels,” J. Nonparametr. Stat., vol. 16, no. 1-2, Feb. 2004, pp. 217-226.

[18] M. Fernandes, E. F. Mendes, and O. Scaillet, “Testing for symmetry and conditional symmetry using asymmetric kernels,” *Ann. Inst. Stat. Math.*, vol. 67, no. 4, Aug. 2015, pp. 649-671.