Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

THE LEAVE-ONE-OUT MAXIMUM LIKELIHOOD METHOD FOR THE PRIESTLEY-CHAO ESTIMATOR OF CONDITIONAL DENSITY

KONEČNÁ Kateřina (CZ)

Abstract. The contribution is focused on a kernel estimation of conditional density. Kernel smoothing is still popular non-parametric method, in theory as well as in practice. The Priestley-Chao estimator of conditional density is introduced and the statistical properties of the estimator are given. The smoothing parameters called bandwidths play a significant role in kernel smoothing. This is the reason for suggesting the methods for their estimation. The typical approach - the cross-validation method - is supplemented with the leave-one-out maximum log-likelihood method. The performance of the suggested methods is compared via a simulation study and an application on a real data set.

Keywords: kernel smoothing, conditional density, bandwidths, Priestley-Chao estimator, leave-one-out maximum likelihood method, cross-validation method

Mathematics subject classification: Primary 62G07; Secondary 62G20, 62G08

1 Introduction

A conditional density provides a comprehensive information about the data set. It can be regarded as a generalization of regression: while regression models only the conditional expectation, conditional density models also an uncertainty. Thus, the conditional density estimator provides except the estimation of the distribution in fixed points of the independent variable also the estimation of a regression function.

The estimators of the conditional density depend on the smoothing parameters called bandwidths. The bandwidths control the smoothness in the both directions - in the direction of the independent and dependent variable. This is the reason why so much importance is given to their detection. The optimal values of the smoothing parameters generally depend on the true marginal and conditional density, that is why they can be used only for the simulated data. In the case of the real data sets, their values need to be estimated by any data-driven method, there are several methods for bandwidths selection in the literature.

Some methods proceed from the methods suggested for kernel density estimation or kernel regression, because the conditional density estimator is a combination of the density estimator and the estimator of a regression function. Thus, the bandwidths can be estimated separately - at first, by methods used for the bandwidth estimation of kernel density, for example the cross-validation method ([4]), the reference rule method ([17], [18]) or the maximum smoothing principle ([20], [19]) could be used. Further, the methods used in kernel regression can be applied for the estimation of the smoothing parameter h_x, for example the cross-validation method ([5]) or the method of penalizing functions (see [15], [10]).

The other methods for bandwidth selection are usually a generalization of the methods used in kernel density estimation and/or kernel regression. The most popular method is the cross-validation method, introduced by [3], which is based on a minimization of the global measure of the quality of the estimate. An iterative method is based on the minimization of the proper relation between the Asymptotic Integrated Square Bias (AISB) and Asymptotic Integrated Variance (AIV). The method is described in [12] and it is inspired by the iterative method for kernel density estimations (see [6] and [7]) and for kernel regression ([9]).

The expressions for the bandwidths estimations are given by a reference rule method, which assume uniform or normal marginal density and normal conditional density with linear mean and linear variance. This method was proposed by Bashtannyk and Hyndman in [1]. Further, a bootstrap method (see [1], [3]) and a method of penalizing functions (for detailed information see [1]) could be mentioned.

All the mentioned methods were derived for the widest used estimator of conditional density - for the Nadaraya-Watson estimator. In this contribution, the Priestley-Chao estimator of conditional density is focused on, the proposed methods - the cross-validation methods and the leave-one-out maximum log-likelihood method - will be derived for this estimator.

2 The Priestley-Chao estimator of conditional density and its statistical properties

In this section, the construction of the estimator is focused on and the statistical properties of the estimator are given. As the most of the methods for bandwidth detection is based on the minimization of a global measure of the quality of the estimator, the derivation of the statistical properties is very important.

In kernel smoothing generally, the elemental building block is the kernel function.

Definition 2.1 *[21] Let K be a real valued function satisfying:*

1. $K \in \mathrm{Lip}[-1,1]$, *i.e.* $|K(x) - K(y)| \leq L|x - y|$, $\forall x, y \in [-1,1]$, $L > 0$,
2. $\mathrm{supp}(K) = [-1,1]$,
3. *moment conditions:*

$$\int_{-1}^{1} K(x)\,\mathrm{d}x = 1,\ \int_{-1}^{1} xK(x)\,\mathrm{d}x = 0,\ \int_{-1}^{1} x^2 K(x)\,\mathrm{d}x = \beta_2(K) \neq 0.$$

Such a function K is called a kernel of order 2.

There are several examples of the kernel function - the Epanechnikov kernel, the uniform, the quartic, the triangular kernel, etc. Probably the best known kernel function is the Gaussian kernel, represented by the density of the random variable with standardized normal distribution. The disadvantage of this kernel is in its unconstrained support. On the other hand, this kernel is used in practice very often, especially because of its simple manipulation and computational aspect.

The estimation of conditional density with one independent variable is focused on, this approach can be generalized for multidimensional independent vector. In the most cases, the independent variable is considered to be a random variable X with a marginal density $g(x)$. Then, we talk about a model with random design.

Let (X, Y) be a random vector and $(X_1, Y_1),\ (X_2, Y_2),\ \ldots,\ (X_n, Y_n)$ its observations. The kernel estimation of the distribution of a random variable Y given by a random variable X characterized by a fixed point x is generally given by the expression

$$\hat{f}\left(y|x\right) = \sum_{i=1}^{n} w_i(x) K_{h_y}\left(y - Y_i\right),$$

where $w_i(x)$ is a weight function at the point x. In kernel conditional density estimates, there are several types of the estimates corresponding to the proper weight function.

Probably the widest used and best known estimator is the Nadaraya-Watson estimator of conditional density ([16]). Its name comes from the kernel estimator of the regression function - the Nadaraya-Watson estimator - due to the similarity between the weight functions of both estimators. The weight function takes the form

$$w_i^{NW}(x) = \frac{K_{h_x}\left(x - X_i\right)}{\sum\limits_{i=1}^{n} K_{h_x}\left(x - X_i\right)}.$$

The local linear estimator is another widely used estimator. The estimator is suggested in [2], its weight function of the estimator is expressed as

$$w_i^{LL}(x) = \frac{K_{h_x}\left(x - X_i\right)\left(\hat{s}_2(x) - \left(x - X_i\right)\hat{s}_1(x)\right)}{\hat{s}_0(x)\hat{s}_2(x) - \hat{s}_1^2(x)},$$

$\hat{s}_r(x) = \frac{1}{n}\sum\limits_{i=1}^{n}\left(x - X_i\right)^r K_{h_x}\left(x - X_i\right)$ is the auxiliary function. The local linear estimator is more difficult, especially as statistical properties and the methods for bandwidths selection are concerned. On the other hand, the estimator is distinguished by better bias property and better boundary effects.

These two estimators mostly suppose a non-uniform distributed random design variable X. In practice, there are many situations when the conditional distribution has to be estimated, but the conditioning variable X is made by equally spaced values of X. Analysis of time series can be such an example in which the variable X is comprised by the time occasions. This is the reason for proposing a new type of estimator - the Priestley-Chao estimator of conditional density.

The Priestley-Chao estimator was originally suggested for the kernel estimation of a regression function. The original paper by Priestley and Chao ([14]) is followed and the generalization of the original estimator to the estimator of conditional density is made.

Suppose a fixed design, i.e. suppose the equally spaced design variable X with fixed values $x_i,\ i = 1,\ 2,\ \ldots,\ n$ and $\delta = x_{i+1} - x_i$. Although the fixed design in the form $x_i = \frac{i}{n},\ i = 1,\ 2,\ \ldots,\ n$ is usually assumed, the design points can not be restricted only on the interval $[0, 1]$ but generally on $[a, b]\,,\ a < b$.

The Priestley-Chao estimator of conditional density is defined as

$$\hat{f}_{PC}\left(y|x\right) = \delta \sum_{i=1}^{n} K_{h_x}\left(x - x_i\right) K_{h_y}\left(y - Y_i\right). \tag{1}$$

The similarity of the Priestley-Chao estimator of conditional density with the Priestley-Chao estimator of regression function is following. The conditional expectation of the estimator (1), denoted by $\widehat{m}_{PC}(x)$, takes the form

$$\widehat{m}_{PC}(x) = \delta \sum_{i=1}^{n} K_{h_x}\left(x - x_i\right) Y_i, \tag{2}$$

the estimator (2) is exactly the Priestley-Chao estimator of regression function, introduced in [14].

Now, the statistical properties of the estimator (1) are focused on.

Theorem 2.1 *[11] Let x be a fixed design, Y a random variable with conditional density $f\left(y|x\right)$ being at least twice continuously differentiable, and K be a kernel function satisfying Definition 2.1. For $h_x \to 0$, $h_y \to 0$ and $nh_xh_y \to \infty$ as $n \to \infty$, asymptotic bias (AB) and asymptotic variance (AV) of the Priestley-Chao estimator are given by the expressions*

$$\mathrm{AB}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = \frac{1}{2}h_x^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2}, \tag{3}$$

$$\mathrm{AV}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = \frac{\delta}{h_xh_y}R^2(K)f(y|x), \tag{4}$$

where $R(K) = \int K^2(u)\,\mathrm{d}u$.

Sketch of the proof. At first, the property

$$\mathrm{E}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = n\delta\mathrm{E}\left\{K_{h_x}\left(x - x_i\right) K_{h_y}\left(y - Y_i\right)\right\}$$

along with the symmetry of the kernel function, the Taylor expansion of the conditional density and $\mathcal{O}$-notation are used for deriving the expression (3).

A well known law of the total variance is the fundamental idea for deriving (4). Let U and V be a random variables, then variance of V is stated by

$$\mathrm{var}\left\{V\right\} = \mathrm{E}\left\{\mathrm{var}_{V|U}\left\{V|U\right\}\right\} + \mathrm{var}\left\{\mathrm{E}_{V|U}\left\{V|U\right\}\right\}.$$

Thus, the variance of i-th term of the estimator (1) is derived. Finally, using the stochastically independence of all the terms in (1), the final variance is expressed as the sum of the variances.

The complete proof can be found in [11]. □

For assessing the local measure of the quality of the estimator (1) at the point $[x, y]$, the Asymptotic Mean Square Error (AMSE) should be used. AMSE is defined as a sum of the the Asymptotic Variance and the Asymptotic Square Bias by the expression

$$\mathrm{AMSE}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = \frac{\delta}{h_xh_y}R^2(K)f(y|x) + \left(\frac{1}{2}h_x^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2.$$

As it was said earlier, the values of the smoothing parameters play a very important role in kernel smoothing. A lot of methods are based on a minimization of the global measure of the quality of the estimate, that is why this characteristics is needed. The global quality of the estimator (1) is given by the Asymptotic Mean Integrated Square Error (AMISE) in the form

$$\mathrm{AMISE}\left\{\hat{f}_{PC}\left(\cdot|\cdot\right)\right\} = \frac{\delta}{h_x h_y} c_1 + c_2 h_x^4 + c_3 h_y^4 + c_4 h_x^2 h_y^2, \tag{5}$$

where $c_1,\ c_2,\ c_3$ and c_4 are constants defined below by the integrals over the ranges of x and Y:

$$\begin{aligned}
c_1 &= \int R^2(K)\,\mathrm{d}x,\\
c_2 &= \frac{1}{4}\beta_2^2(K)\iint\left(\frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 \mathrm{d}x\,\mathrm{d}y,\\
c_3 &= \frac{1}{4}\beta_2^2(K)\iint\left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2 \mathrm{d}x\,\mathrm{d}y,\\
c_4 &= \frac{1}{2}\beta_2^2(K)\iint \frac{\partial^2 f(y|x)}{\partial x^2}\frac{\partial^2 f(y|x)}{\partial y^2}\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

3 Methods for bandwidths detection

The smoothing parameters influence the final estimation of conditional density significantly. While choosing too great values of the bandwidths, the estimate will tend to oversmooth, on the other hand, the estimate will be undersmoothed in the choice of the small values. Our aim is to estimate the proper values of the smoothing parameters to get a satisfactory estimate.

As the values of the smoothing parameters depend on the unknown conditional density, their optimal values of them can be computed only in the case of the simulated data. While having the real data set, a data-driven method has to be used for the bandwidths detection.

At first, the attention is paid to the optimal values of the smoothing parameters. Further, the cross-validation method and the leave-one-out maximum log-likelihood method are described.

3.1 The optimal values of the smoothing parameters

The optimal values of the smoothing parameters can be used in the cases when the true conditional density is known. Their knowledge is used especially while implementing new methods for the bandwidths estimations to determine the suitability of the method.

The optimal values of the bandwidths are such values which minimize a global measure of the quality of the estimator. They are derived by differentiating AMISE with respect to both smoothing parameters, and by solving the system of two non-linear equations given by setting the derivatives to zero.

The optimal values of the smoothing parameters, denoted by h_x^* and h_y^*, are given by

$$\begin{aligned}
h_x^* &= \delta^{1/6} c_1^{1/6}\left(4\left(\frac{c_2^5}{c_3}\right)^{1/4} + 2c_4\left(\frac{c_2}{c_3}\right)^{3/4}\right)^{-1/6}\\
h_y^* &= \left(\frac{c_2}{c_3}\right)^{1/4} h_x^*.
\end{aligned}$$

3.2 The cross-validation method

The cross-validation method is a standard procedure widely used for estimating the smoothing parameters, not only in the kernel estimates of conditional density, but in kernel smoothing generally.

The main idea of the method is based on a minimization of the cross-validation function (6). The cross-validation function is associated with the global quality of the estimator - ISE (the Integrated Square Error) in this case. In this contribution, the leave-one-out cross-validation method is concerned, i.e. we use the estimation at the points (x_i, Y_i) using the points $\{(x_j, Y_j)\,,\ j \neq i\}$.

The leave-one-out cross-validation function is given in the form

$$\mathrm{CV}\left(h_x, h_y\right) = \delta^2 \sum_i \sum_{j \neq i} h_x h_y K_{h_x\sqrt{2}}\left(x_i - x_j\right) K_{h_y\sqrt{2}}\left(Y_i - Y_j\right) - 2\delta \sum_i \hat{f}_{PC}\left(Y_i | x_i\right). \tag{6}$$

The values of the smoothing parameters using the leave-one-out cross-validation method are given by

$$\left(h_x^{CV}, h_y^{CV}\right) = \underset{(h_x, h_y)}{\arg\min}\, \mathrm{CV}\left(h_x, h_y\right).$$

3.3 The leave-one-out log-likelihood method

The maximum likelihood method is a standard statistical procedure for selecting unknown parameters. In general, having an underlying density is supposed. In this case, the underlying unknown conditional density is substituted by the Priestley-Chao estimator (1), represented by given observations and unknown bandwidths.

This approach was proposed by Leiva-Murillo and Artes-Rodríguez for kernel density estimations (see the paper [13]). In this paper, their idea is followed and extended to conditional density estimations.

Assume the objective function

$$\mathcal{L}\left(h_x, h_y | \boldsymbol{x}, \boldsymbol{Y}\right) = \prod_{j=1}^{n} \hat{f}_{PC}\left(Y_j | x_j\right) \tag{7}$$

with the independent and identically distributed observations $(x_i, Y_i)\,,\ i = 1,\ 2,\ \ldots,\ n$ considered as the fixed parameters, and the smoothing parameters h_x, h_y as the unknown parameters.

In the case that all n observation is considered, the maximizing problem (7) has a trivial degenerate solution. In (7), if $i = j$, then

$$K_{h_x}\left(x_j - x_i\right) = K_{h_x}(0) \quad \text{and} \quad K_{h_x}\left(x_j - x_i\right) = K_{h_x}(0).$$

Letting $h_x \to 0$ and $h_y \to 0$, then $\mathcal{L} \to \infty$, i.e. the objective function becomes unbounded and does not have a finite maximum. This is an undesirable property because of estimating the values of bandwidths tending to zero, thus the final estimate is undersmoothed, usually with abundance of information and high variability. This is a motivation for implementation of a modification of the classical maximum-likelihood approach.

The modification of the classical maximum likelihood approach lies in the leaving one observation out - here is the name of the method, the leave-one-out maximum likelihood. Thus, the modified likelihood function is expressed by

$$\mathcal{L}^*\left(h_x, h_y|\boldsymbol{x}, \boldsymbol{Y}\right) = \prod_{\substack{j=1\\ j\neq i}}^{n} \hat{f}_{PC}\left(Y_j|x_j\right). \tag{8}$$

As maximizing (8) is equivalent to maximizing the functional

$$\ell^*\left(h_x, h_y|\boldsymbol{x}, \boldsymbol{Y}\right) = \ln \prod_{\substack{j=1\\ j\neq i}}^{n} \hat{f}_{PC}\left(Y_j|x_j\right) = \sum_{\substack{j=1\\ j\neq i}}^{n} \ln \hat{f}_{PC}\left(Y_j|x_j\right), \tag{9}$$

the maximizing problem (9) will be used rather than (8), especially due to the computational aspects and computational time.

The estimations of the smoothing parameters are given by

$$\left(\hat{h}_x^{\ell^*}, \hat{h}_y^{\ell^*}\right) = \underset{(h_x,h_y)}{\arg\max}\, \ell^*\left(h_x, h_y|\boldsymbol{x}, \boldsymbol{Y}\right).$$

4 Simulation study

In this section, we conduct two simulation studies comparing the results given by the the cross-validation method and by the proposed leave-one-out maximum log-likelihood method. The simulation study involves two models defined as

$$M_1 : Y_i = \mathrm{e}^{x_i} + \varepsilon_i, \quad x_i = \frac{i}{n}, \quad i = 1\, \ldots, 100, \quad \varepsilon_i \sim N(0, 0.5^2)$$
$$M_2 : Y_i = \sin\left(3\pi x_i^2\right) + \varepsilon_i, \quad x_i = \frac{i}{n}, \quad i = 1\, \ldots, 100, \quad \varepsilon_i \sim N(1, 1)$$

At first, one hundred observations are generated from each model to apply the Priestley-Chao estimator for conditional density detection. For both simulation studies, an exactly given grid of 100 times 100 points is considered to construct an estimation and to measure the error term. The x grid is formed by the observations x_i, the y grid is formed by the exact equidistant points at the range of Y values.

The methods for bandwidth detection are performed from several points of view, the accuracy of the estimates of the smoothing parameters to the optimal bandwidths is assessed as well as the measure of the quality estimation is focused on. The measure of the quality of the estimate is given by the Integrated Square Error

$$\mathrm{ISE}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = \iint \left\{\hat{f}_{PC}\left(y|x\right) - f\left(y|x\right)\right\}^2 \mathrm{d}x\,\mathrm{d}y.$$

Due to computational aspect, its estimation is used:

$$\widehat{\mathrm{ISE}}\left\{\hat{f}_{PC}\left(y|x\right)\right\} = \frac{\Delta}{n}\sum_{j=1}^{N}\sum_{i=1}^{n}\left(\hat{f}_{PC}\left(y_j|x_i\right) - f\left(y_j|x_i\right)\right)^2,$$

where $\boldsymbol{y} = (y_i,\, \ldots, y_N)$ is a vector of equally spaced values over the sample space of Y and Δ is the distance between two consecutive values of $\boldsymbol{y}$.

The observations were generated two hundred times in total. The values of the smoothing parameters, the ISE estimation and the computational time were received from every repetition of the simulation. The results of the estimates using the proposed methods are compared with the theoretical values and they are presented by boxplots.

The estimations of the smoothing parameters for the model M_1 are displayed in Fig. 1. The CV method gives more stable results than the LOO-MLln method in the estimation of h_x, on the other hand the median is slightly closer to the optimal value h_x^* (vertical line) while the LOO-MLln method is used. Both methods tend to underestimate the parameter h_x. As the smoothing parameter h_y is concerned, the CV method gives the underestimated values too, the median is located slightly above the optimal value h_y^* while using the LOO-MLln method.

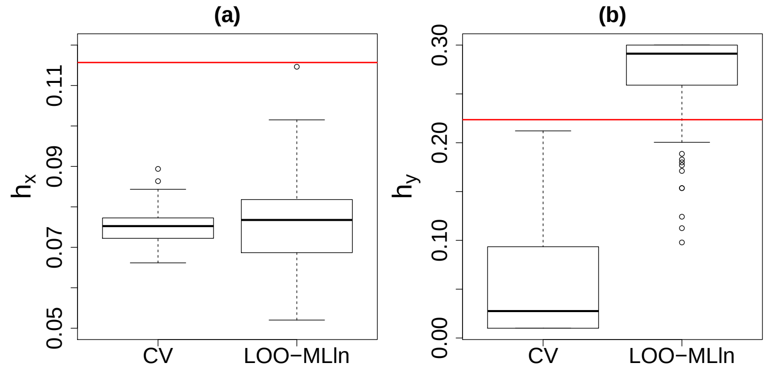

Fig. 1. The model M_1: The estimations of the smoothing parameters (a) h_x and (b) h_y for the CV and the LOO-MLln method. The red vertical line represents the optimal value of the smoothing parameter (a) $h_x^* = 0.116$, (b) $h_y^* = 0.224$.

In the model M_1, the estimations of the ISE values and the computational times are displayed in Fig. 2. The ISE values given by the CV and the LOO-MLln method are compared with the ISE values for the estimations constructed for the optimal values of the smoothing parameters (OPT).

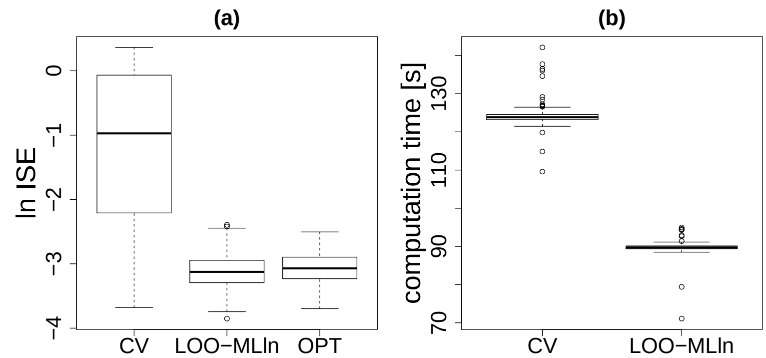

Fig. 2. The model M_1: The estimations of (a) the natural logarithms of the ISE values for the CV, the LOO-MLln method and OPT, (b) computational time.

The LOO-MLln method gives the estimations of the natural logarithms of the ISE error estimation very close to the optimal ISE. Other advantage of the LOO-MLln method lies in its computational difficulty - the computation using the LOO-MLln method takes about three quarters of the CV computational time.

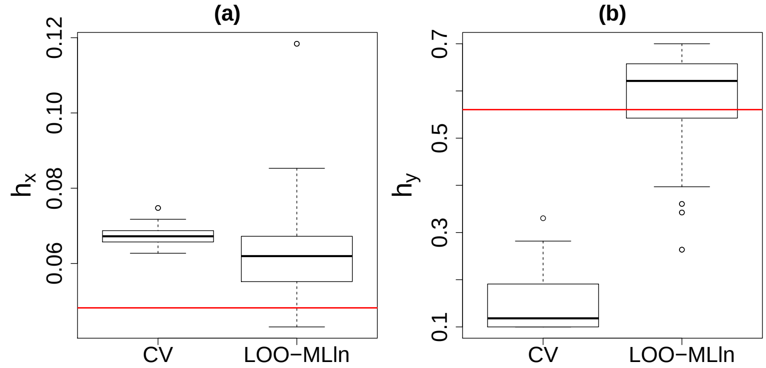

Fig. 3. The model M_2: The estimations of the smoothing parameters (a) h_x and (b) h_y for the CV and the LOO-MLln methods in the model M_2. The red vertical line represents the optimal value of the smoothing parameter (a) $h_x^* = 0.048$, (b) $h_y^* = 0.56$.

The estimations of the smoothing parameters for the model M_2 are displayed in Fig. 3. Both methods give slightly overestimated values of the smoothing parameter h_x, the LOO-MLln estimations are characterized by the higher variance with the median closer to the optimal value h_x^*. The estimations of h_y are very undervalued for the CV method, the LOO-MLln gives estimations well reflecting the optimal value h_y^*.

The estimations of the ISE values and the computational times are displayed for the model M_2 in Fig. 4. As it can be seen in Fig. 4 (a), the LOO-MLln method gives the ISE estimations very close to the optimal ISE values, while the estimations by the CV method are high and variable. The LOO-MLln method is distinguished by the lower computational time.

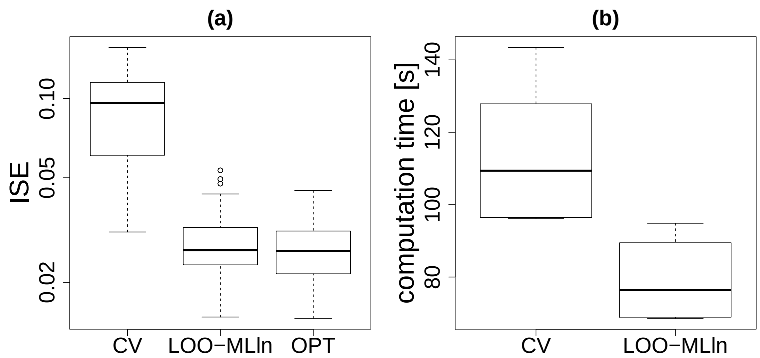

Fig. 4. The model M_2: The estimations of (a) the ISE values for the CV, the LOO-MLln method and OPT, (b) computational time.

5 Application to a temperature data

In this chapter, the comparison among the results given by the proposed leave-one-out maximum log-likelihood method and the cross-validation method is focused on. Both methods are implemented on a real data set provided by Berkeley Earth hosted in Kaggle [8].

The average annual temperatures in the Czech Republic during 1753 – 2012 are focused on. Early data was collected by technicians using mercury thermometers, while in the 1980's, there was a move to electronic thermometers that are said to have a cooling bias.

Our aim is to compare the estimations of the smoothing parameters by the cross-validation method (CV) and the leave-one-out log-likelihood method (LOO-MLln) and their influence on the final estimate of conditional density. Also the computation time and the conditional mean estimation are in our interest.

There are two variables in the data set:

- time factor (year), independent variable,
- the average annual temperature (measured in $^\circ$ C), dependent variable.

The temperature data is illustrated in Fig. 5. The results of the methods are presented in Tab. 1 and in Fig. 6 and Fig. 7.

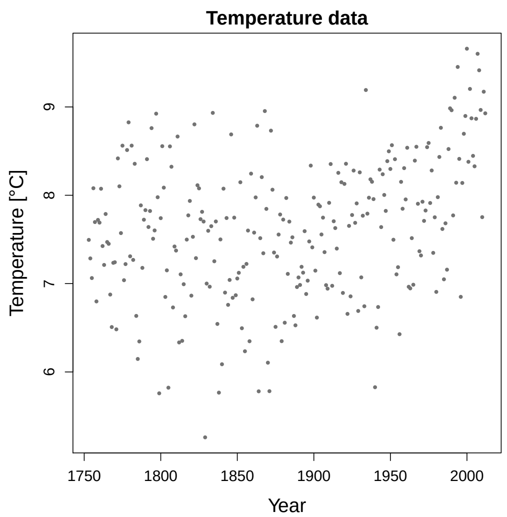

Fig. 5. A scatterplot of the temperature data.

At first, the estimates of the values of the smoothing parameters as well as the computational times were focused on. The results are brought in Tab. 1. Here, the most significant advantage, the computational time, of the LOO-MLln method over the CV method can be seen. The computational time for the LOO-MLln method is almost three times lower than the computational time for the CV method.

method	$\hat{h}_x$	$\hat{h}_y$	computational time $[s]$
CV	8.45	0.00656	1021
LOO-MLln	10.2	0.34	285

Tab. 1. Estimates of the smoothing parameters and computational times for methods used for estimating the temperature data.

Considering the estimates of the smoothing parameters, the differences among their estimations are distinguished by a significant impact on the final visual presentation of the estimations of conditional density. While the estimation using the CV method is characterized by the high-variable estimate

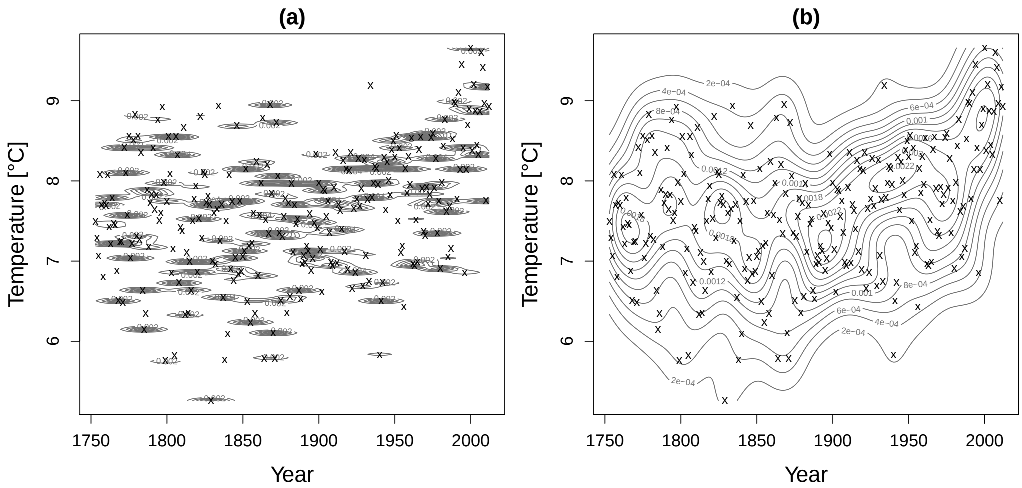

Fig. 6. Contour plots for conditional density estimations for two compared methods: (a) CV, (b) LOO-MLln.

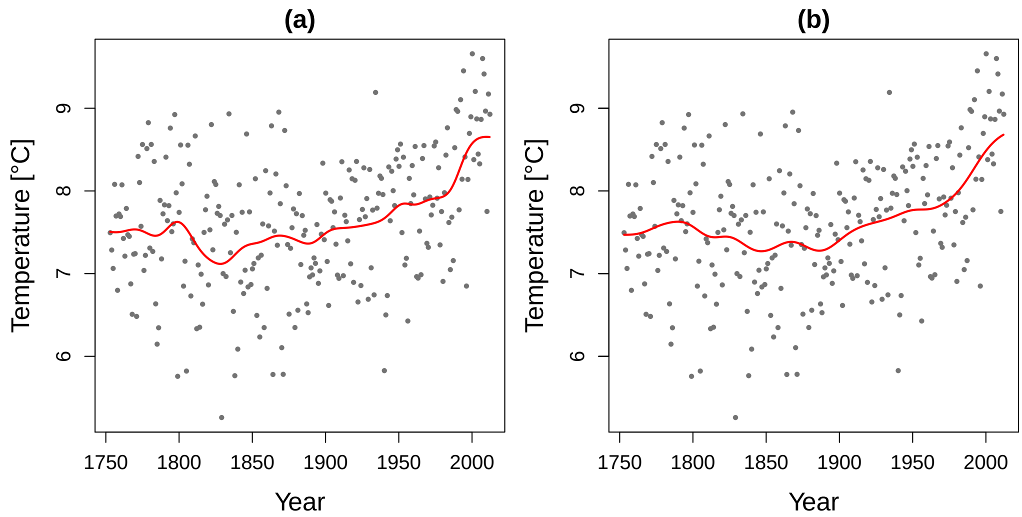

Fig. 7. Estimations of conditional expectation for two compared methods: (a) CV, (b) LOO-MLln.

with many peaks and unrecognizable distribution while time changing, the estimation using the LOO-MLln method is not so much fluctuating, it is much smoother with evident distribution.

The estimates of the conditional mean are displayed for each method in Fig. 7. Despite the very different conditional distribution, not so distinctive differences among the estimated conditional means can be seen. The estimation using the CV method is characterized by lower smoothness.

6 Conclusion

A new type of the kernel estimator of conditional density - the Priestley-Chao estimator - was focused on. Despite the "classical" approach consisted in the random design assumption, the fixed design is assumed in the Priestley-Chao estimator. The statistical properties including the local and the global measures of the quality of the estimate are included. As the smoothing parameters play a significant role in kernel smoothing, the optimal values of the smoothing parameters are supplemented with the data-driven methods. The typical approach, the cross-validation method, is mentioned and a new method, the leave-one-out log-likelihood method, is proposed.

The suggested methods are compared via a simulation study. Two simulation studies are conducted, the emphasis is placed on the comparisons of the estimations of the smoothing parameters to the theoretical values and of the estimations of ISE with its theoretical values. The computational time for both methods is also included.

The results show the unequivocal advantage of the maximum likelihood approach. Except for the lower computational time, the proposed method gives even more precise and stable estimations of the smoothing parameters. The estimations of ISE approach to the ISE values computed for the estimates with the optimal values of the smoothing parameters.

The CV and LOO-MLln are also applied to a real data set. The temperature data in the Czech Republic during 1753 – 2012 is focused on. The results show the differences between the estimations of conditional density. While the cross-validation method results in very undersmoothed estimates with very complicatedly recognizable distribution, the LOO-MLln gives satisfactory results.

The results show that the proposed maximum log-likelihood method can be a reasonable and reliable tool for bandwidth selection.

Acknowledgement

The research was supported by the project 22738 Institutional support for development of research organization (Brno University of Technology).

References

[1] BASHTANNYK, D. M., HYNDMAN, R. J., Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, vol. 36, no. 3, 2001: pp. 279–298, ISSN 0167-9473.

[2] FAN, J., YAO, Q., TONG, H., Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, vol. 83, no. 1, 1996: pp. 189–206, ISSN 00063444.

[3] FAN, J., YIM, T. H., A crossvalidation method for estimating conditional densities. *Biometrika*, vol. 91, no. 4, 2004: pp. 819–834, ISSN 0006-3444.

[4] HALL, P., Cross-validation in density estimation. *Biometrika*, vol. 69, no. 2, 1982: pp. 383–390, ISSN 0006-3444.

[5] HÄRDLE, W., *Applied nonparametric regression*. Cambridge university press, 1990.

[6] HOROVÁ, I., KOLÁČEK, J., VOPATOVÁ, K., Full bandwidth matrix selectors for gradient kernel density estimate. *Computational Statistics & Data Analysis*, vol. 57, no. 1, 2013: pp. 364–376, ISSN 0167-9473.

[7] HOROVÁ, I., ZELINKA, J., Contribution to the bandwidth choice for kernel density estimates. *Computational Statistics*, vol. 22, no. 1, 2007: pp. 31–47, ISSN 1613-9658.

[8] KAGGLE, *Climate change: Earth surface temperature data*, https://www.kaggle.com/berkeleyearth/datasets, accessed: 2017-09-13.

[9] KOLÁČEK, J., HOROVÁ, I., Iterative bandwidth method for kernel regression. *Journal of Statistics: Advances in Theory and Applications*, vol. 8, 2012: pp. 91–103, ISSN 0975-126.

[10] KOLÁČEK, J., Problems of automatic data-driven bandwidth selectors for nonparametric regression. *Journal of Electrical Engineering*, vol. 53, 2002: pp. 48–51, ISSN 1335-3632.

[11] KONEČNÁ, K., Priestley-chao estimator of conditional density, In *Mathematics, Information Technologies and Applied Sciences 2017, post-conference proceedings of extended versions of selected papers*, Brno: University of Defence, 2017, ISBN 978-80-7582-026-6, pp. 151–163.

[12] KONEČNÁ, K., HOROVÁ, I., Conditional Density Estimations. *International Society for the Advancement of Science and Technology (ISAST)*, Athens, 2014: pp. 15–31, ISBN 978-618-81257-7-3.

[13] LEIVA-MURILLO, J. M., ARTÉS-RODRÍGUEZ, A., Algorithms for maximum-likelihood bandwidth selection in kernel density estimators. *Pattern Recognition Letters*, vol. 33, no. 13, 2012: pp. 1717–1724, ISSN 0167-8655.

[14] PRIESTLEY, M. B., CHAO, M. T., Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 3, 1972: pp. 385–392, ISSN 00359246.

[15] RICE, J., Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 1984: pp. 1215–1230.

[16] ROSENBLATT, M., Conditional probability density and regression estimators. *Multivariate analysis II*, vol. 25, 1969: p. 31.

[17] SILVERMAN, B. W., Algorithm as 176: Kernel density estimation using the fast fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, no. 1, 1982: pp. 93–99, ISSN 00359254, 14679876.

[18] SILVERMAN, B. W., *Density estimation for statistics and data analysis*. CRC press, 1986.

[19] TERRELL, G. R., The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, vol. 85, no. 410, 1990: pp. 470–477, ISSN 01621459.

[20] TERRELL, G. R., SCOTT, D. W., Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, vol. 80, no. 389, 1985: pp. 209–214, ISSN 01621459.

[21] WAND, M. P., JONES, M. C., *Kernel smoothing*. Crc Press, 1994, ISBN 9780412552700.

Current address

Konečná Kateřina, Mgr.
Institute of Mathematics and Descriptive Geometry
Faculty of Civil Engineering
Brno University of Technology
Žižkova 17, 602 00 Brno, Czech Republic
E-mail: konecna.k@fce.vutbr.cz