The Priestley-Chao Estimator of Conditional Density with Uniformly Distributed Random Design

Kateřina Konečná [1] | *Brno University of Technology, Brno, Czech Republic*

Abstract

The present paper is focused on non-parametric estimation of conditional density. Conditional density can be regarded as a generalization of regression thus the kernel estimator of conditional density can be derived from the kernel estimator of the regression function. We concentrate on the Priestley-Chao estimator of conditional density with a random design presented by a uniformly distributed unconditional variable. The statistical properties of such an estimator are given. As the smoothing parameters have the most significant influence on the quality of the final estimate, the leave-one-out maximum likelihood method is proposed for their detection. Its performance is compared with the cross-validation method and with two alternatives of the reference rule method. The theoretical part is complemented by a simulation study.[2]

Keywords

Priestley-Chao estimator of conditional density, random design, uniform marginal density, bandwidth selection, maximum likelihood method, reference rule method

JEL code

C14

INTRODUCTION

Kernel smoothing is still a popular non-parametric procedure, in theory as well as in practice. There are numerous monographs concerned with the kernel smoothing approach, e.g., Wand and Jones (1994). Computational implementations in MATLAB were developed by Horová et al. (2012). The present paper focuses on the kernel conditional density estimation. Several estimator types can be found in the literature with the Nadaraya-Watson one being probably best known (see Rosenblatt, 1969). The local linear estimator of conditional density was suggested by Fan et al. (1996) for its better statistical properties and boundary effects.

Conditional density can be regarded as a generalization of regression, which models the conditional mean while conditional density models the whole distribution. This is the reason why a kernel regression estimator can be generalized to a kernel conditional density estimator. The present paper extends the

[1] Institute of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Žižkova 17, 602 00 Brno, Czech Republic. E-mail: konecna.k@fce.vutbr.cz.

[2] This article is based on contribution at the conference *Robust 2018*.

Priestley-Chao regression estimator (for detailed information see Priestley and Chao, 1972) to estimate even conditional densities.

Each kernel estimator depends on the smoothing parameters called bandwidths, values which significantly influence the final estimation. This is the reason why so much importance is given to their selection. There are many methods discussed in the literature, most of them suggested for the Nadaraya-Watson estimator, with only a few of them for the local linear estimator.

Introduced by Fan and Yim (2004), Hansen (2004) and Hall et al. (2004) and based on minimizing the Integrated Squared Error, cross-validation is a method typical of bandwidth selection. Bashtannyk and Hyndman (2001) suggested a reference rule method for normal underlying conditional density and for two marginal density choices – normal and uniform. Some methods extend the methods suggested for kernel regression. The iterative method proposed by Konečná and Horová (2014), for one, is motivated by the iterative method developed for kernel density estimation and for kernel regression (for detailed information, see Horová and Zelinka (2007), Horová et al. (2012), Koláček and Horová (2012)). Other examples include the bootstrap method by Bashtannyk and Hyndman (2001) and Fan and Yim (2004) as well as the fast dual-tree based algorithms using a maximum likelihood criterion (see Holmes et al., 2012).

Kernel conditional density estimation is still employed in practice: Takeuchi et al. (2009) show its application in medicine (the relative change in spinal bone mineral density is explored as a function of the age of adolescents), Jeon and Taylor (2012) are interested in 1-to-72-hours ahead wind-power prediction from which the management of wind farms and electricity systems can profit. Another application, forecasting electricity smart meter data, helping consumers to analyze and to minimize their electricity consumption and enabling new pricing strategies for suppliers, is introduced by Arora and Taylor (2016).

As mentioned above, papers are focused primarily on the Nadaraya-Watson or the local linear estimator. The present paper suggests the Priestley-Chao estimator for the uniformly distributed design, based on the estimator suggested for the equally spaced design (see Konečná, 2017). The leave-one-out maximum likelihood method follows the one proposed by Konečná (2018).

The paper is organized as follows: Section 1 deals with the Priestley-Chao estimator of conditional density and its statistical properties. The optimal values of the smoothing parameters are derived, and the leave-one-out maximum likelihood method for their practical estimation proposed in Section 2. This method is complemented by the cross-validation method and by two alternatives of the reference rule method. A simulation study in Section 3 then presents the performance of the methods by a simulation study. The proofs of the statistical properties can be found in the Appendix.

1 THE PRIESTLEY-CHAO ESTIMATOR OF CONDITIONAL DENSITY

The conditional density $f(y|x)$ models the probability of a random variable Y given a random variable X, represented by a fixed observation $X = x$. Let $\{(X_i,Y_i), i = 1,\ldots,n\}$ be an observed data sample of a pair of real random variables (X,Y). The kernel estimate of conditional density generally takes the form:

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

where $w_i(x)$ is a weight function, and K is a real, symmetric, nonnegative kernel function satisfying:

$$\int_R K(x)\,dx = 1, \int_R xK(x)\,dx = 0, \int_R x^2 K(x)\,dx = \beta_2(K) \neq 0. \tag{2}$$

The present paper uses the Gaussian kernel. The smoothing parameters $h_x > 0$, $h_y > 0$ control the smoothness of the estimate. The estimate of conditional density is also influenced by the estimator type (1).

Our focus is on the Priestley-Chao estimator, originally proposed for the kernel regression estimation (Priestley and Chao, 1972). Konečná (2017) dealt with the Priestley-Chao estimator for conditional density with the fixed design, i.e., the fixed values $x_i = \frac{i}{n}, i = 1, \ldots, n$ of the design variable X were assumed.

Next, we are concerned with the estimator for the random design specified by a uniformly distributed variable X on the interval [0,1]. The estimator can easily be extended for the design variable X on the interval $[a, b]$, $a < b$. The statistical properties of the estimator will be given and methods for bandwidth detection proposed.

Let X be a uniformly distributed random variable with the marginal density function:

$$g(x) = \begin{cases} 1, & x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

As the focused weight function in Formula (1) is $w_i^{PC}(x) = \frac{1}{n} K_{h_x}(x - X_i)$, the Priestley-Chao estimator takes the form:

$$\hat{f}_{PC}(y|x; h_x, h_y) = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) K_{h_y}(y - Y_i). \tag{3}$$

The Priestley-Chao estimator of the regression function is expressed by the conditional mean of the Formula (3):

$$\hat{m}_{PC}(x; h_x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) Y_i.$$

Theorem 1 *Let X be a uniformly distributed random variable on the interval* [0,1], *Y be a random variable with density $f(y|x)$ being at least twice continuously differentiable, and $K(x)$ be a real, symmetric, nonnegative kernel function satisfying (2). For $x \in [h_x, 1 - h_x]$, $h_x \to 0$, $h_y \to 0$ and $nh_xh_y \to \infty$ as $n \to \infty$, the asymptotic bias* (AB) *and the asymptotic variance* (AV) *are given by:*

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

$$\mathrm{AV}\{\hat{f}_{PC}(y|x; h_x, h_y)\} = \frac{1}{nh_xh_y} R^2(K) f(y|x),$$

where: $R_2(K) = \int_R K_2(u)\, du$.

Proof. The proof can be found in the Appendix.

The local quality of the estimate at the point $[x,y]$ is given by the mean squared error (MSE) which is the simple decomposition to the variance (V) and the squared bias (SB). Considering the main terms only, the asymptotic MSE (AMSE) is obtained as:

$$\begin{aligned} &\mathrm{AMSE}\{\hat{f}_{PC}(y|x; h_x, h_y)\} = AV\{\hat{f}_{PC}(y|x; h_x, h_y)\} + ASB\{\hat{f}_{PC}(y|x; h_x, h_y)\} \\ &= \frac{1}{nh_xh_y} R^2(K) f(y|x) + \left(\frac{1}{2} h_x^2 \beta_2(K) \frac{\partial^2 f(y|x)}{\partial x^2} + \frac{1}{2} h_y^2 \beta_2(K) \frac{\partial^2 f(y|x)}{\partial y^2} \right)^2. \end{aligned} \tag{4}$$

The statistical properties of the Formula (3), particularly the global quality measure expressed by the asymptotic mean integrated squared error (AMISE), are necessary for assessing the quality of the estimate and the theoretical values of the smoothing parameters. AMISE is obtained by integrating (4) weighted by the marginal density $g(x)$ as:

$$\mathrm{AMISE}\{\hat{f}_{PC}(\cdot \mid \cdot; h_x, h_y)\} = \iint \mathrm{AMSE}\{\hat{f}_{PC}(y|x; h_x, h_y)\} g(x)\, dx\, dy.$$

The following form of the AMISE is more succinct for further processing:

$$\mathrm{AMISE}\{\hat{f}(\cdot \mid \cdot; h_x, h_y)\} = \frac{1}{n h_x h_y} c_1 + c_2 h_x^4 + c_3 h_y^4 + c_4 h_x^2 h_y^2, \tag{5}$$

where the constants c_1, c_2, c_3, c_4 are given by:

$$\begin{aligned} c_1 &= R^2(K), \\ c_2 &= \frac{1}{4}\beta_2^2(K) \iint \left(\frac{\partial^2 f(y|x)}{\partial x^2}\right)^2 dx\, dy, \\ c_3 &= \frac{1}{4}\beta_2^2(K) \iint \left(\frac{\partial^2 f(y|x)}{\partial y^2}\right)^2 dx\, dy, \\ c_4 &= \frac{1}{2}\beta_2^2(K) \iint \frac{\partial^2 f(y|x)}{\partial x^2} \frac{\partial^2 f(y|x)}{\partial y^2}\, dx\, dy. \end{aligned} \tag{6}$$

Remark. Note that all the integrals with respect to *x* are computed over the support of the *X* variable, i.e., over the interval [0,1]. The integrals with respect to y are considered over *R*.

2 METHODS FOR BANDWIDTH SELECTION

The values of the smoothing parameters have an essential significance for the final estimate of conditional density. First, the optimal widths of the smoothing parameters are derived as the values minimizing the AMISE. As the optimal bandwidths depend on the true conditional and marginal density function, it is necessary to develop a data-driven method for their estimation. In this section, the leave-one-out maximum likelihood method, the cross-validation method, and two alternatives of the reference rule method are suggested for their detection.

2.1 Optimal values of the smoothing parameters

The optimal values of the smoothing parameters are given as the values which minimize AMISE given by (5). By differentiating (5) with respect to h_x and h_y and setting the derivatives to 0, we obtain the following system of non-linear equations:

$$\begin{aligned} -\frac{1}{n h_x^2 h_y} c_1 + 4 c_2 h_x^3 + 2 c_4 h_x h_y^2 &= 0, \\ -\frac{1}{n h_x h_y^2} c_1 + 4 c_3 h_y^3 + 2 c_4 h_x^2 h_y &= 0. \end{aligned} \tag{7}$$

Solving system (7), the optimal bandwidths are given by:

$$\begin{aligned} h_x^* &= n^{-1/6} c_1^{1/6} \left(4 \left(\frac{c_2^5}{c_3}\right)^{1/4} + 2 c_4 \left(\frac{c_2}{c_3}\right)^{3/4}\right)^{-1/6}, \\ h_y^* &= \left(\frac{c_2}{c_3}\right)^{1/4} h_x^*. \end{aligned} \tag{8}$$

Both h_x^* and h_y^* are of order $n^{-1/6}$ while the order of AMISE is $n^{-2/3}$.

2.2 The leave-one-out maximum likelihood method

As mentioned above, with a real dataset, a data-driven method is needed for bandwidth selection. We will modify the maximum likelihood method, which is a standard statistical procedure for estimating

unknown parameters. This method was originally proposed for kernel density estimation by Leiva-Murillo and Artes-Rodriguez (2012), and their approach is generalized to include the Priestley-Chao estimator of conditional density.

Since the objective function:

$$L(h_x, h_y) = \prod_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(X_j - X_i) K_{h_y}(Y_j - Y_i) \tag{9}$$

is considered for all n observations, the optimization problem $L \to \max$ has a trivial solution. If $i = j$ in (9), the objective function (9) increases to infinity for $h_x \to 0$ and $h_y \to 0$. Of course, this is not the desired behaviour because, with very small values of bandwidths, the final estimate tends to be undersmoothed.

This problem can be solved by leaving out one observation and employing the modified objective function:

$$L^*(h_x, h_y) = \prod_{j=1}^{n} \frac{1}{n} \sum_{i=1, i \neq j}^{n} K_{h_x}(X_j - X_i) K_{h_y}(Y_j - Y_i)\,. \tag{10}$$

If the natural logarithm of the likelihood function L^* given by (10) is taken into account, the values of the smoothing parameters maximize:

$$l^*(h_x, h_y) = \sum_{j=1}^{n} \ln\left(\frac{1}{n} \sum_{i=1, i \neq j}^{n} K_{h_x}(X_j - X_i) K_{h_y}(Y_j - Y_i) \right),$$

and are developed as:

$$(\hat{h}_x, \hat{h}_y) = \arg \max_{(h_x, h_y)} l^*(h_x, h_y).$$

2.3 The leave-one-out cross-validation method

The cross-validation method is a standard procedure for bandwidth selection in kernel smoothing. Introduced by Fan and Yim (2004), Hansen (2004) and Hall et al. (2004), the method is associated with the global quality measure of the estimator, with the integrated squared error (ISE). With $\hat{f}_{PC,-i}(Y_i|X_i; h_x, h_y)$, being the estimate at the point (X_i, Y_i) using the points $\{(X_j, Y_j), j \neq i\}$, the cross-validation function:

$$\mathrm{CV}(h_x, h_y) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\substack{j=1, \\ j \neq i}}^{n} K_{h_x\sqrt{2}}(X_i - X_j) K_{h_y\sqrt{2}}(Y_i - Y_j) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{PC,-i}(Y_i|X_i; h_x, h_y),$$

is the proper estimator of the ISE.

The values of the smoothing parameters are given by:

$$(\hat{h}_x, \hat{h}_y) = \arg \min_{(h_x, h_y)} \mathrm{CV}(h_x, h_y).$$

2.4 The reference rule method

The reference rule method was originally proposed for the Nadaraya-Watson estimator by Bashtannyk and Hyndman (2001). They assumed a normally distributed random variable $Y|(X = x)$ with linear conditional mean and constant or linear standard deviation. Additionally, they distinguished two possibilities for the marginal distribution, considering uniform and truncated normal marginal densities.

Our approach via the Priestley-Chao estimator corresponds to the choice of a uniform marginal density. We assume that the conditional distribution is normal with the mean $m(x)$ and standard deviation $\sigma(x)$. Hence, the conditional density of $Y|(X = x)$ is:

$$f(y|x) = \frac{1}{\sigma(x)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-m(x)}{\sigma(x)}\right)^2\right\}. \tag{11}$$

Two different situations are considered:
(a) According to Bashtannyk and Hyndman (2001), the model with the linear conditional mean $m(x) = p_0 + p_1 x$ and the linear standard deviation $\sigma(x) = q_0 + q_1 x$ is assumed. The values of the constants $c_1, \dots, c_4$ are given by the expressions:

$$\begin{aligned} c_1 &= R^2(K), \\ c_2 &= \frac{3}{512}\beta_2^2(K)\frac{wz}{q_1\sqrt{\pi}}, \\ c_3 &= \frac{3}{128}\beta_2^2(K)\frac{z}{q_1\sqrt{\pi}}, \\ c_4 &= \frac{3}{128}\beta_2^2(K)\frac{z(2p_1 - 3q_1^2)}{q_1\sqrt{\pi}}, \end{aligned} \tag{12}$$

where $z = \frac{(q_0+q_1)^4 - q_0^4}{(q_0+q_1)^4 q_0^4}$ and $w = 19q_1^4 + 4p_1^4 + 28p_1^2q_1^2, p_1 \neq 0$.
The values of the smoothing parameters are obtained by substituting (12) into (8).
(b) The model with the quadratic conditional mean $m(x) = p_0 + p_1 x + p_2 x^2$ and the constant standard deviation σ is suggested. The constants $c_1, \dots, c_4$ are given by:

$$\begin{aligned} c_1 &= R^2(K), \\ c_2 &= \frac{3}{160}\beta_2^2(K)\frac{1}{\sigma^5\sqrt{\pi}}(16p_2^4 + 40p_1p_2^3 + 40p_1^2p_2^2 + 20p_1^3p_2 + 5p_1^4 + 10p_2^2\sigma^2), \\ c_3 &= \frac{3}{32}\beta_2^2(K)\frac{1}{\sigma^5\sqrt{\pi}}, \\ c_4 &= \frac{1}{16}\beta_2^2(K)\frac{1}{\sigma^5\sqrt{\pi}}(4p_2 + 6p_1p_2 + 3p_1^2). \end{aligned} \tag{13}$$

The values of the smoothing parameters are obtained by substituting the terms (13) into (8).
Remark. The expressions (13) are obtained by differentiating (11) twice and substituting them into (6). As the computations of the integrals in (6) include many auxiliary derivations, only a sketch of them is presented.

A conditional random variable $Y|(X = x) \sim N(m(x), \sigma^2)$ with the density function $f(y|x)$ is assumed. The following equality:

$$f^2(y|x) = \frac{1}{2\pi\sigma^2}\exp\left\{-\left(\frac{y-m(x)}{\sigma}\right)^2\right\} = \frac{1}{2\sigma\sqrt{\pi}}\frac{1}{\frac{\sigma}{\sqrt{\pi}}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y-m(x)}{\frac{\sigma}{\sqrt{2}}}\right)^2\right\}$$

$$= \frac{1}{2\sigma\sqrt{\pi}} f_1(y|x),$$

where $f_1(y|x)$ is a density function of a conditional random variable $Y_1|(X = x) \sim N(m(x), \frac{1}{2}\sigma^2)$, was used. For evaluating the integrals in (6), the below auxiliary expression was derived:

$$\iint_A x^k y^n f_1(y|x)\,dx\,dy = K_{0,n} + \sum_{i=1}^{2n+1} \frac{K_{i,n}}{k+i}, k = 0, \dots, 4,$$

where $A = R \cdot [0,1]$ is the domain of integration,

$$K_{0,n} = \begin{cases} 0 & n = 0, 1, \\ \frac{1}{2}\sigma^2 \frac{1}{k+1} & n = 2, \\ \frac{3}{2}\sigma^2 \sum_{i=1}^{n} \frac{p_{i-1}}{k+i} & n = 3, \\ \frac{3}{4}\sigma^4 \frac{1}{k+1} + 3\sigma^2 \sum_{i=1}^{n+1} \frac{K_{i,n-2}}{k+i} & n = 4, \end{cases}$$

$$K_{i,n} = \begin{cases} \sum_{j=1}^{\min\left\{n+2-\left\lceil\frac{i}{2}\right\rceil, \left\lceil\frac{i}{2}\right\rceil\right\}} \binom{n}{\left\lceil\frac{i}{2}\right\rceil + j - 2} \binom{\left\lceil\frac{i}{2}\right\rceil + j - 2}{\left\lceil\frac{i}{2}\right\rceil - j} p_0^{n-\left\lceil\frac{i}{2}\right\rceil - j + 2} p_1^{2j-2} p_2^{\left\lceil\frac{i}{2}\right\rceil - j} & \text{for } i \text{ odd}, \\ \sum_{j=1}^{\min\left\{n+1-\frac{i}{2}, \frac{i}{2}\right\}} \binom{n}{\frac{i}{2} + j - 1} \binom{\frac{i}{2} + j - 1}{\frac{i}{2} - j} p_0^{n-\frac{i}{2}-j+1} p_1^{2j-1} p_2^{\frac{i}{2}-j} & \text{for } i \text{ even}, \end{cases}$$

and $\lceil \cdot \rceil$ denotes the ceiling function.

3 SIMULATION STUDY

In this section, a simulation study comparing four methods for bandwidth estimation is conducted. The considered methods are the maximum likelihood method (ML), the cross-validation method (CV), the reference rule method with linear conditional mean and linear standard deviation (REF1), and the reference rule method with quadratic conditional mean and constant standard deviation (REF2). Two models are involved in the simulation study. To demonstrate the adaptability of the methods to various shapes of the regression function or conditional density, a changing shape of the conditional mean is presented in the first model. In the second model, a bimodal and non-symmetric conditional distribution is chosen as a mixture of two normal densities. The models are defined as:

$$M_1: Y_i = \sin(\pi X_i^2) + X_i + \varepsilon_i, X_i \sim \text{unif}(0, 2), \ \varepsilon_i \sim \text{N}(0, 0.5^2), \text{i} = 1, \ldots, \text{n},$$

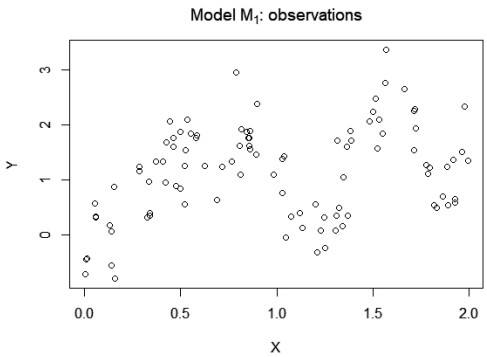
$$M_2: Y_i = \exp\{X_i\} + \varepsilon_i, X_i \sim \text{unif}(0, 3), \varepsilon_i | X_i \sim W_i T_i + (1 - W_i) U_i, T_i \sim N(2X_i, 1),$$

$$U_i \sim N(-3X_i, 2), P(W_i = 0) = P(W_i = 1) = 0.5, i = 1, \ldots, n.$$

In both simulation studies, one hundred observations ($n = 100$) were generated. The described methods for bandwidth selection are compared from several points of view – the estimates of the smoothing parameters and the quality measure of the estimate. The quality of the final estimate was measured by an estimate of the integrated squared error (ISE) given by:

$$\widehat{\text{ISE}}\{\hat{f}_{PC}(\cdot \mid \cdot; h_x, h_y)\} = \frac{\Delta}{n} \sum_{j=1}^{N} \sum_{i=1}^{n} \left(\hat{f}_{PC}(y_j | X_i; h_x, h_y) - f(y_j | X_i)\right)^2,$$

Figure 1 A scatterplot of one hundred observations of model M_1



Source: The author's own construction

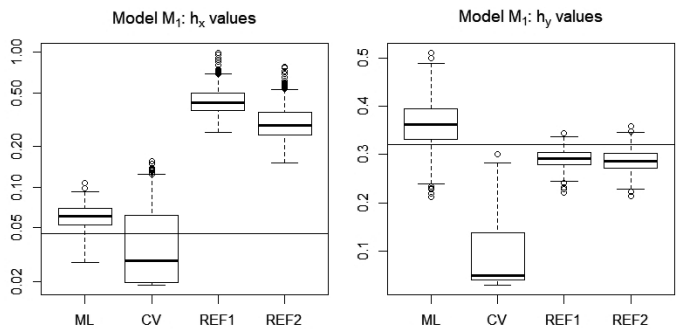
where $\boldsymbol{y} = (y_1, \ldots, y_N)$ is a vector of equally spaced values over the sample space of Y and Δ is the distance between two consecutive values of y, i.e. $\Delta = y_{j+1} - y_j$, $j = 1, \ldots, N - 1$. In both simulation studies, the number of $\boldsymbol{y}$ values was set to $N = 100$.

For both models, five hundred repetitions have been made to obtain the described characteristics. The results are displayed in boxplots and supplemented by numerical values in the text.

First, the results for the model M_1 are summarized. A scatterplot of one set of the sample values of model M_1 is displayed in Figure 1.

Boxplots of the estimates of the smoothing parameters h_x and h_y are displayed in Figure 2. It can be seen, that the ML and CV methods lead to good values for h_x (medians are 0.0613 for ML and 0.0288 for CV), the REF1 and REF2 methods produce highly variable bandwidths exceeding the optimal value $h_x^* = 0.0455$.

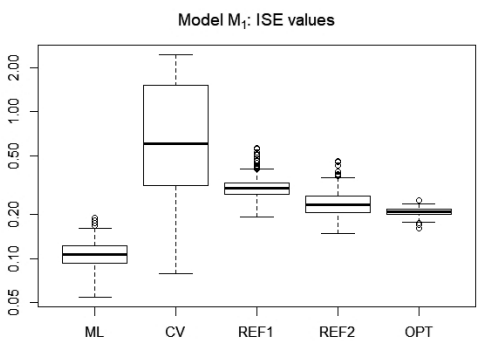
Figure 2 Estimates of the smoothing parameters h_x and h_y along with the optimal values (horizontal lines) for the ML, CV, REF1, and REF2 methods in model M_1



Note: The log-scale of the vertical axis in the left-hand-side panel.
Source: The author's own construction

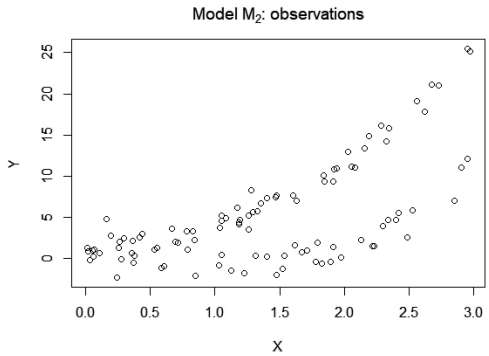
Considering the estimates of the smoothing parameter h_y, the values estimated using REF1 resemble those with REF2. Their medians 0.2927 and 0.2863 (in this order) are close to the optimal value $h_y^* = 0.3213$, and the estimates are characterized by a low variability (their standard deviations are close to 0.02). The ML method gives slightly higher values than the optimum h_y^*, the median of the values being 0.3631. The CV method tends to produce values well under the optimal value which results in a much undersmoothed estimate of conditional density.

Figure 3 Estimates of the ISE values (expressed in the log-scale) for the ML, CV, REF1, and REF2 method and for the optimal bandwidth choice (OPT) in model M_1



Source: The author's own construction

Figure 4 A scatterplot of 100 observations by model M_2



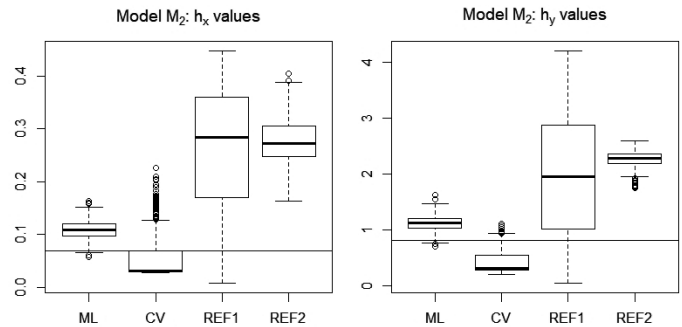
Source: The author's own construction

The ML method results in an ISE that is smaller than any other methods considered as well as the optimal bandwidth choice (OPT). Both reference rule methods produce ISE values slightly higher than OPT while the values of the CV method are well above the optimal bandwidth choice.

Now, we focus on the results of model M_2. The scatterplot of a sample generated by model M_2 is shown in Figure 4.

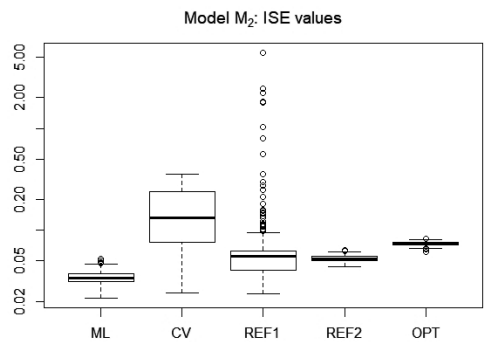
Boxplots of bandwidth estimates are shown in Figure 5. Both reference rule methods produce values of h_x and h_y well above the optimal values, which results in an overvalued final estimate with

Figure 5 Estimates of the smoothing parameters h_x and h_y and the optimal values (horizontal lines) for the ML, CV, REF1, and REF2 method in model M_2



Source: The author's own construction

Figure 6 Estimates of the ISE values (expressed in the log-scale) for the ML, CV, REF1, and REF2 methods and for the optimal bandwidths (OPT) in model M_2



Source: The author's own construction

worse capability to adapt to bimodal conditional density. On the other hand, the low values of both smoothing parameters obtained by the CV method lead to the undersmoothing of conditional density and abundance of useless information in the data. The ML method performs the best in this simulation study.

The ML method is also suitable in terms of the ISE (see Figure 6). The medians of the ISE values obtained by the reference rule methods (0.0555 for REF1 and 0.0527 for REF2) do not reach the median (0.0742) of the ISE values for optimal bandwidths, but the REF1 method suffers from the large variability (standard deviation is 0.3143). The CV method provides highly variable ISE values exceeding the OPT values.

CONCLUSION

The presented paper generalizes the Priestley-Chao estimator from the restrictive fixed design to the uniformly distributed random design variable X. The statistical properties of this estimator are derived, and the methods for bandwidth selection are suggested.

The leave-one-out maximum likelihood method, a modification of the classical likelihood approach, is proposed for bandwidth detection. This method is complemented by the cross-validation method and the reference rule method. The original approach of the reference rule method was extended to a normally distributed conditional variable with quadratic mean and constant standard deviation.

The performance of the suggested methods is presented using two simulation studies focusing on the bandwidth estimates and the quality measure estimates. The results show that the cross-validation method tends to undersmooth significantly. The reference rule method assuming the quadratic conditional mean produces results similar to or better than the reference rule with a linear conditional mean and linear standard deviation, but none of these two methods outperforms the ML method. On the other hand, the results of these two references are better than those of the CV method, even in the cases of the underlying conditional density not resembling the conditional density assumed by the reference model.

The ML method can adapt well not only to the changing shape of the conditional mean and conditional normal distribution but also to a bimodal or an asymmetric distribution. The method always results in an ISE that is smaller than the optimal bandwidth choice. It also detects the bandwidths which decrease ISE estimates, but does not underestimate the parameter h_x as the optimal case usually does. The simulation study shows that the proposed maximum likelihood method is a reasonable tool for bandwidth selection.

ACKNOWLEDGMENT

The research was supported by the project of specific university research at Brno University of Technology FAST-S-18-5184.

References

ARORA, S. AND TAYLOR, J. W. Forecasting electricity smart meter data using conditional kernel density estimation. *Omega*, 2016, 59 (Part A), pp. 47–59.

BASHTANNYK, D. M. AND HYNDMAN R. J. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 2001, 36(3), pp. 279–298.

FAN, J., YAO Q., TONG H. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 1996, 83(1), pp. 189–206.

FAN, J. AND YIM T. H. A crossvalidation method for estimating conditional densities. *Biometrika*, 2004, 91(4), pp. 819–834.

HALL, P., RACINE, J., LI, Q. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 2004, 99(468), pp. 1015–1026.

HANSEN, B. E. *Nonparametric conditional density estimation.* Unpublished manuscript, 2004.

HOLMES, M. P., GRAY, A. G., ISBELL, C. L. Fast Nonparametric Conditional Density Estimation. *ArXiv e-prints*, 2012.

HOROVÁ, I., KOLÁČEK, J., ZELINKA, J. *Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing.* Singapore: World Scientific Publishing Co. Pte. Ltd., 2012.

HOROVÁ, I. AND ZELINKA, J. Contribution to the bandwidth choice for kernel density estimates. *Computational Statistics*, 2007, 22(1), pp. 31–47.

JEON, J. AND TAYLOR, J. W. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 2012, 107(497), pp. 66–79.

KOLÁČEK, J. AND HOROVÁ, I. Iterative bandwidth method for kernel regression. *Journal of Statistics: Advances in Theory and Applications*, 2012, 8, pp. 93–101.

KONEČNÁ, K. Priestley-Chao estimator of conditional density. *Mathematics, Information Technologies and Applied Sciences 2017, post-conference proceedings of extended versions of selected papers*, 2017, pp. 151–163.

KONEČNÁ, K. The leave-one-out maximum likelihood method for the Priestley-Chao estimator of conditional density. *Proceedings, 17th Conference on Applied Mathematics – APLIMAT 2018*, 2018, pp. 577–589.

KONEČNÁ, K., HOROVÁ, I., KOLÁČEK, J. Conditional Density Estimations. In: SKIADAS, C. H. *Theoretical and Applied Issues in Statistics and Demography*, Athens: International Society for the Advancement of Science and Technology (ISAST), 2014, pp. 15–31.

LEIVA-MURILLO, J. M. AND ARTES-RODRIGUEZ, A. Algorithms for maximum-likelihood bandwidth selection in kernel density estimators. *Pattern Recognition Letters*, 2012, 33(13), pp. 1717–1724.

PRIESTLEY, M. B. AND CHAO, M. T. Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1972, 34(3), pp. 385–392.

ROSENBLATT, M. Conditional probability density and regression estimators. *Multivariate analysis II*, 1969, 25, pp. 25–31.

TAKEUCHI, I., NOMURA, K., KANAMORI, T. Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, 2009, 21(2), pp. 533–559.

WAND, M. P. AND JONES M. C. *Kernel smoothing.* 1st Ed. London: Crc Press, 1994.

APPENDIX

Here, a detailed proof of **Theorem 1** can be found.

Proof. All the computations are based on Taylor expansions with higher-order terms ignored. First, the expectation (E) of the Formula (3) is derived:

$$\mathrm{E}\{\hat{f}_{PC}(y|x;h_x,h_y)\} = n\frac{1}{n}\mathrm{E}\left\{\mathrm{K}_{\mathrm{h_x}}(\mathrm{x}-\mathrm{X_i})\mathrm{K}_{\mathrm{h_y}}(\mathrm{y}-\mathrm{Y_i})\right\}$$

$$= \iint K_{h_x}(x-u)K_{h_y}(y-v)f(v|u)g(u)\,du\,dv$$

$$= f(y|x) + \frac{1}{2}h_x^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2} + O(h_x^4) + O\left(h_y^4\right) + O\left(h_x^2h_y^2\right).$$

Then, the asymptotic bias is given as:

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

The variance of the estimator is derived by the well-known law of total variance. Let X and Y be random variables. Then, the following equality holds:

$$\mathrm{var}\{Y\} = \mathrm{E}\left\{\mathrm{var}_{Y|X}\{Y|X\}\right\} + \mathrm{var}\left\{\mathrm{E}_{Y|X}\{Y|X\}\right\}. \tag{14}$$

First, by (14), the variance of the i-th term of the Formula (3) is derived.

The conditional expectation of the estimator's i-th term $\hat{f}_{PC}(y|x;h_x,h_y)$ can be written as:

$$E_{f(y|X_i)}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)|X_i\right\}$$
$$= \frac{1}{n}K_{h_x}(x-X_i)\left(f(y|X_i) + \frac{1}{2}h_y^2\beta_2(K)\frac{\partial^2 f(y|X_i)}{\partial y^2} + O\left(h_y^4\right)\right). \tag{15}$$

The conditional expectation of the squared i-th term in (3) is given by:

$$\mathrm{E}_{f(y|X_i)}\left\{\frac{1}{n^2}K_{h_x}^2(x-X_i)K_{h_y}^2(y-Y_i)|X_i\right\}$$
$$= \frac{1}{n^2h_y}K_{h_x}^2(x-X_i)\left(R(K)f(y|X_i) + \frac{1}{2}h_y^2G(K)\frac{\partial^2 f(y|X_i)}{\partial y^2} + O\left(h_y^4\right)\right), \tag{16}$$

where $\mathrm{G(K)} = \int u^2 K^2(u)du$. By subtracting the second power of (15) from (16), we have:

$$\begin{aligned}&\mathrm{var}_{f(y|X_i)}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)|X_i\right\}\\&=\frac{1}{n^2}K_{h_x}^2(x-X_i)\left(\frac{1}{h_y}R(K)f(y|X_i)-f^2(y|X_i)+\frac{1}{2}h_yG(K)\frac{\partial^2 f(y|X_i)}{\partial y^2}+O\left(h_y^2\right)\right).\end{aligned}\tag{17}$$

Finally, by applying the expectation to (17), the first term of (14) is acquired:

$$\begin{aligned}&\mathrm{E}\left\{\mathrm{var}_{f(y|X_i)}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)\right\}\right\}\\&=\frac{1}{n^2h_xh_y}R^2(K)f(y|x)-\frac{1}{n^2h_x}R(K)f^2(y|x)+\frac{h_y}{2n^2h_x}R(K)G(K)\frac{\partial^2 f(y|x)}{\partial y^2}+O\left(h_y^2\right).\end{aligned}\tag{18}$$

The expected value of (16) and its square are used to derive the variance of the conditional Formula (16).

$$\begin{aligned}&\mathrm{E}\left\{\mathrm{E}_{f(y|X_i)}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)\right\}\right\}\\&=\frac{1}{n}f(y|x)+\frac{h_x^2}{2n}\beta_2(K)\frac{\partial^2 f(y|x)}{\partial x^2}+\frac{h_y^2}{2n}\beta_2(K)\frac{\partial^2 f(y|x)}{\partial y^2}+O\left(h_y^4\right),\end{aligned}\tag{19}$$

$$\mathrm{E}\left\{\mathrm{E}_{f(y|X_i)}^2\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)\right\}\right\}=\frac{1}{n^2h_x}R(K)f^2(y|x)+O\left(\frac{h_x}{n^2}\right)+O\left(\frac{h_y}{n^2}\right).\tag{20}$$

Thus, the variance of (16) is obtained by subtracting (19) squared from (20):

$$\begin{aligned}&\mathrm{var}\left\{\mathrm{E}_{f(y|X_i)}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)\right\}\right\}\\&=\frac{1}{n^2h_x}R(K)f^2(y|x)-\frac{1}{n^2}f^2(y|x)+O\left(\frac{h_x}{n^2}\right)+O\left(\frac{h_y^2}{n^2}\right).\end{aligned}\tag{21}$$

As the expression (21) is the desired second term of (14), the variance of the *i*-th term of the Priestley-Chao estimator is obtained by summing up (18) and (21):

$$\mathrm{var}\left\{\frac{1}{n}K_{h_x}(x-X_i)K_{h_y}(y-Y_i)\right\}=\frac{1}{n^2h_xh_y}R^2(K)f(y|x)+O\left(\frac{1}{n^2}\right)+O\left(\frac{h_x}{n^2}\right)+O\left(\frac{h_y}{n^2}\right).$$

It can be easily shown that the equation:

$$\mathrm{cov}\left\{\frac{1}{n}K_{h_x}(x-X_1)K_{h_y}(y-Y_1),\frac{1}{n}K_{h_x}(x-X_2)K_{h_y}(y-Y_2)\right\}=0$$

holds, i.e., the two terms of the Priestley-Chao estimator are not correlated. Then, the variance of the Formula (3) is given by:

$$\mathrm{var}\{\hat{f}_{PC}(y|x;h_x,h_y)\}=\frac{1}{nh_xh_y}R^2(K)f(y|x)+O\left(\frac{1}{n}\right)+O\left(\frac{h_x}{n}\right)+O\left(\frac{h_y}{n}\right).$$

By taking into account only the leading terms of bias and variance, the theorem is proven.