VERSITA

TATRA MOUNTAINS
Mathematical Publications

DOI: 10.2478/v10127-011-0035-z
Tatra Mt. Math. Publ. **50** (2011), 39–50

k-NEAREST NEIGHBOUR KERNEL DENSITY ESTIMATION, THE CHOICE OF OPTIMAL k

Jan Orava

ABSTRACT. The k-nearest neighbour kernel density estimation method is a special type of the kernel density estimation method with the local choice of the bandwidth. An advantage of this estimator is that smoothing varies according to the number of observations in a particular region. The crucial problem is how to estimate the value of the parameter k. In the paper we discuss the problem of choosing the parameter k in a way that minimizes the value of the asymptotic mean integrated square error (AMISE). We define the class of the modified cosine densities that meet the requirements given by the AMISE. The results are compared in a simulation study.

1. Introduction

The integral $\int f(x)\,\mathrm{d}x$ denotes the Riemann integral over $\mathcal{R}$ if it is not specified otherwise.

Let $X_1, X_2 \ldots, X_n$ be independent, identically distributed random variables with bounded continuous density $f(x)$. The k-nearest neigbour density estimate (in the rest of article we will call it KNN) proposed in [4] is given by

$$\hat{f}_{KNN}(x,k) = \frac{1}{nr_n}\sum_{i=1}^{n} \mathsf{K}\left(\frac{x - X_i}{r_n}\right), \tag{1}$$

where $r_n = r_n(x)$ is a Euclidean distance between x and the kth nearest neighbour of x among X_j's,

$$r_n(x) = \min\big(k, \{|x - X_j|, \text{where } j = 1, \ldots, n\}\big), \tag{2}$$

where $\min(k, A)$ is the kth smallest member of the set A; K is a kernel function which satisfies

$$\int \mathsf{K}(x)\,\mathrm{d}x = 1,$$

2010 Mathematics Subject Classification: 62G07.
Keywords: kernel density estimation, k-nearest neighbour, modified cosine density.

and $k = \{k(n)\}$ is a sequence of positive integers with

$$k \to \infty, \frac{k}{n} \to 0 \qquad \text{as } n \to \infty.$$

The kernel function is usually chosen as a non-negative density function which is symmetric about 0. This implies that the estimate of density using the global kernel density estimation method defined in [7] will be the density itself. However, this is not valid for KNN method. The integral of KNN density estimate is usually very close to 1, but it is not 1. The choice of kernel function does not have a great influence on the final quality of the result. Kernel functions were closely studied, e.g., in [7]. In this paper only Epanechnik kernel will be used

$$\mathsf{K}(x) = \begin{cases} \frac{3}{4}(x^2 - 1) & \text{if} \quad x \in [-1, 1], \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In the following theorems we will use notation

$$\beta_2(g) = \int g^2(x)\,\mathrm{d}x, \tag{4}$$

$$\mu_2(g) = \int x^2 g(x)(d)x. \tag{5}$$

In the past, there has been extensive research on the properties of kernel estimates. Several articles have been published on asymptotic properties, the rate of convergence and consistency of the estimate (for example, see [1], [2], [3] and [6]). The main object of this paper is to derive a practical method that could be used for choosing the parameter k. The main idea is based on the paper [4], where the asymptotic mean integrated square error is expressed. Minimizing this error for a certain reference density will lead to the universal formula that can be used for estimating the value of the parameter k.

Theorem 1 (Mack and Rosenblatt)**.** *Let f be the bounded density function. The kernel function $\mathsf{K}(x)$ is assumed to be bounded with*

$$\int |x|^2 |\mathsf{K}(x)|\,\mathrm{d}x < \infty, \quad \int |x| \mathsf{K}(x)\,\mathrm{d}x = 0. \tag{6}$$

Furthermore, let x be a point with $f(x) > 0$ and f be continuously differentiable up to the second order in a neighbourhood of x. Then the asymptotic variance and the asymptotic bias of the KNN estimate can be expressed as

$$\widehat{\mathrm{Var}}\big(\hat{f}_{KNN}(x)\big) = \frac{2}{k}\beta_2(\mathsf{K}) f^2(x) + o\left(\frac{1}{k}\right). \tag{7}$$

$$\widehat{\mathrm{Bias}}\left(\hat{f}_{KNN}(x)\right) = \frac{1}{2^3}\left(\frac{k}{n}\right)^2 \mu_2(\mathsf{K}) \frac{f''(x)}{f^2(x)} + o\left(\left(\frac{k}{n}\right)^2 + \frac{1}{k}\right). \tag{8}$$

P r o o f. See [4]. □

Since the asymptotic mean integrated square error (AMISE) is

$$\mathrm{AMISE}(\hat{f}) = \int \widehat{\mathrm{Var}}\big(\hat{f}(x)\big) + \int \widehat{\mathrm{Bias}}^{\,2}(\hat{f}),$$

then using the notations (4) and (5) we get

$$\mathrm{AMISE}(\hat{f}_{KNN}) = \frac{2}{k}\beta_2(\mathsf{K})\beta_2(f) + \frac{1}{64}\left(\frac{k}{n}\right)^4 \mu_2^2(\mathsf{K})\beta_2\left(\frac{f''}{f^2}\right). \tag{9}$$

The value of k that minimizes $\mathrm{AMISE}(\hat{f}_{KNN})$ can be expressed as

$$\begin{aligned} k_{\mathrm{AMISE}} &= \arg\min_{k=2,\ldots,n} \mathrm{AMISE}(\hat{f}_{KNN}) \\ &= \mathrm{round}\left(2n^{\frac{4}{5}}\left(\frac{\beta_2(\mathsf{K})\beta_2(f)}{\mu_2^2(\mathsf{K})\beta_2\left(\frac{f''}{f^2}\right)}\right)^{\frac{1}{5}}\right). \end{aligned} \tag{10}$$

The value of k_{AMISE} will be called k AMISE optimal. The proof can be done easily by deriving $\mathrm{AMISE}(\hat{f}_{KNN})$ with respect to k and setting it equal to 0.

Since the expression for k_{AMISE} depends on an unknown density function, it cannot be used in practice. Our goal is to substitute the unknown density f with a reference density that will allow us to estimate the value of k_{AMISE}.

The value of the functional $\beta_2(f)$ can be easily computed for commonly used densities, but the problem is to express the value of $\beta_2\big(\frac{f''}{f^2}\big)$. For commonly used densities the value of this functional goes to zero, thus k_{AMISE} goes to infinity.

We will attempt to develop a new type of density function that will give us a non-zero value of $\beta_2\big(\frac{f''}{f^2}\big)$. The idea is to use the cosine density function defined by

$$f_{\cos}(x) = \begin{cases} \frac{1}{2}\cos(x) & \text{if} \quad |x| < \frac{\pi}{2}, \\ 0 & \text{otherwise}, \end{cases} \tag{11}$$

and cut out the edges of the function.

The question is how big a part of the edges should be cut out and how to transform the new function into a density function. The solution of this problem will be suggested in the next section of the article.

2. Class of modified cosine densities

The class of modified cosine densities is defined by $\big\{g_l(x)\big\}$, $l = 2, 3, \ldots, \infty$, where

$$g_l(x) = \begin{cases} \frac{1}{2}\frac{D}{\sigma}\sin^{-1}\left(\frac{\pi}{2}\frac{l}{l+1}\right)\cos\left(\frac{D}{\sigma}x\right) & \text{if} \quad |x| < m, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where

$$m = \frac{\pi}{2}\frac{l}{l+1}\frac{\sigma}{D},$$

parameters D and σ are positive integers. It can be easily showed that g_l is positive function with $\int g_l(x)\mathrm{d}x = 1$ for any positive l, D and σ, this implies that $g_l(x)$ is a density function.

Theorem 2. *Let g_l be a cosine modified density function defined in* (12)*. If the value of parameter D is*

$$D = \left(\frac{\pi^2}{8}\frac{l^2}{(l+1)^2} - 1 + \frac{\pi}{2}\frac{l}{l+1}\arctan\left(\frac{\pi}{2}\frac{l}{l+1}\right)\right)^{\frac{1}{2}},$$

then the variance of $g_l(x)$ is given by $\mathrm{Var}(\mathrm{g}_l) = \sigma^2$.

P r o o f. The proof can be obtained by substituting D in expression $g_l(x)$ in

$$\mathrm{Var}\big(g_l(x)\big) = \int_{-m}^{m} x^2 g_l(x)\,\mathrm{d}x.$$

□

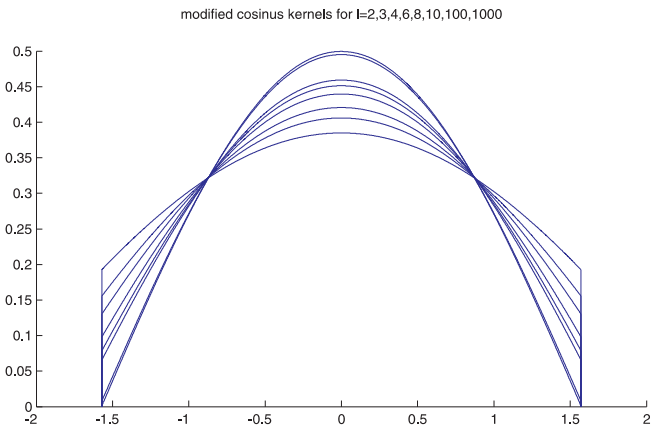

Figure 1. Graphs of modified cosine densities with $m = \pi/2$ for different values of the parameter l.

The modified cosine density can be taken as cosine density with cut edges. The parameter l says how big a part was cut out. When l goes to infinity then g_l becomes identical to cosine kernel defined in 11.

The situation is illustrated in Figure 1. The flattest graph is the graph of g_2, the sharpest graph corresponds to g_{1000}. We can see that the graph of g_{1000} resembles a cosine density, because only small part of the edges has been cut out.

Now we can compute the values of functionals for the modified cosine density, that we need for estimating k_{AMISE},

$$\beta_2(\mathsf{g}_l) = \frac{1}{8}\frac{D}{\sigma}\frac{\pi l + 2(l+1)\sin\left(\frac{\pi}{2}\frac{l}{l+1}\right)\cos\left(\frac{\pi}{2}\frac{l}{l+1}\right)}{(l+1)\left(1-\cos^2\left(\frac{\pi}{2}\frac{l}{l+1}\right)\right)}, \tag{13}$$

$$\beta_2\left(\frac{\mathsf{g}_l''}{\mathsf{g}_l^2}\right) = 8\frac{D}{\sigma}\sin^3\left(\frac{\pi}{2}\frac{l}{l+1}\right)\cos^{-1}\left(\frac{\pi}{2}\frac{l}{l+1}\right), \tag{14}$$

$$\frac{\beta_2(\mathsf{g}_l)}{\beta_2\left(\frac{\mathsf{g}_l''}{\mathsf{g}_l^2}\right)} = \frac{1}{64}\frac{\pi l + (l+1)\sin\left(\pi\frac{l}{l+1}\right)}{l+1}\frac{\cos\left(\frac{\pi}{2}\frac{l}{l+1}\right)}{\sin^5\left(\frac{\pi}{2}\frac{l}{l+1}\right)}. \tag{15}$$

We can see that the ratio (15) does not depend on the parameter D and even more interestingly, it does not depend on the value of variance σ^2. This means that k AMISE optimal value of density from the modified cosine class does not depend on the variance of unknown density.

Theorem 3. *Assume that all conditions of Theorem 1 are satisfied. Let* K *be Epanechnik kernel defined in* (3) *and the density* f *be chosen from a modified cosine densities class, then the estimation of* k *AMISE optimal value is given by*

$$\hat{k}_{\mathrm{AMISE}} = \mathrm{round}\left(n^{\frac{4}{5}}C(l)\right), \tag{16}$$

where

$$C(l) = \left(\frac{15}{2}\frac{\pi l + (l+1)\sin\left(\pi\frac{l}{l+1}\right)}{l+1}\frac{\cos\left(\frac{\pi}{2}\frac{l}{l+1}\right)}{\sin^5\left(\frac{\pi}{2}\frac{l}{l+1}\right)}\right)^{\frac{1}{5}}. \tag{17}$$

P r o o f. First of all we compute values of functionals

$$\beta_2(\mathsf{K}) \quad \text{and} \quad \mu_2(\mathsf{K})$$

for Epanechnik kernel

$$\beta_2(\mathsf{K}) = \int_{-1}^{1}\mathsf{K}^2(x)\,\mathrm{d}x = \int_{-1}^{1}\left(\frac{3}{4}\left(x^2-1\right)\right)^2\mathrm{d}x = \frac{3}{5},$$

$$\mu_2(\mathsf{K}) = \int_{-1}^{1}x^2\mathsf{K}(x)\,\mathrm{d}x = \int_{-1}^{1}x^2\frac{3}{4}\left(x^2-1\right)\mathrm{d}x = \frac{1}{5}.$$

Then by substituting (15) in (10) we get

$$\begin{aligned}\hat{k}_{\mathrm{AMISE}} &= n^{\frac{4}{5}}\left(\frac{1}{2}\frac{\pi l+(l+1)\sin\left(\pi\frac{l}{l+1}\right)}{l+1}\frac{\cos\left(\frac{\pi}{2}\frac{l}{l+1}\right)}{\sin^5\left(\frac{\pi}{2}\frac{l}{l+1}\right)}\right)^{\frac{1}{5}}\left(\frac{\beta_2(\mathsf{K})}{\mu_2^2(\mathsf{K})}\right)^{\frac{1}{5}}\\ &= n^{\frac{4}{5}}\left(\frac{15}{2}\frac{\pi l+(l+1)\sin\left(\pi\frac{l}{l+1}\right)}{l+1}\frac{\cos\left(\frac{\pi}{2}\frac{l}{l+1}\right)}{\sin^5\left(\frac{\pi}{2}\frac{l}{l+1}\right)}\right)^{\frac{1}{5}}\\ &= \mathrm{round}\left(n^{\frac{4}{5}}C(l)\right).\end{aligned}$$

□

Table 2 illustrates the behavior of functional $C(l)$. We can see that $C(l)$ is a decreasing step function and that for increasing l the rate of decrease is decreasing. It means a big increase of l causes only a small decrease of $C(l)$.

TABLE 1. Values of functional $C(l)$ for different l.

l	1	2	3	4	5	6	7
$C(l)$	2,38	1,87	1,67	1,56	1,48	1,43	1,38
l	8	9	10	10^2	10^3	10^4	10^5
$C(l)$	1,35	1,31	1,29	0,82	0,52	0,33	0,21

In Theorem 3 it was proved that estimation of k AMISE optimal depends only on the size of a random sample n and on the parameter l. Thus we can use $\hat{k}_{\mathrm{AMISE}}$ when estimating unknown density function. Since the value n is known, l is the only value that has to be estimated. The parameter l represents a member of the class of modified cosine densities.

2.1. Measuring the quality of result

The quality of estimated densities will be measured by integrated square error (ISE)

$$\mathrm{ISE}\big(\hat{f}_{KNN}(x,k)\big)=\int\big(\hat{f}_{KNN}(x,k)-f(x)\big)^2\,\mathrm{d}x. \tag{18}$$

For a better presentation we will use a natural logarithm of ISE in graphs. The value k that minimizes ISE is given by

$$k_{\mathrm{opt}}=\arg\min_{k=2,\ldots,n}\mathrm{ISE}\big(\hat{f}_{KNN}(x,k)\big).$$

The parameter k-optimal will be called k-optimal value. ISE is minimized for $k=2,\ldots,n$. If $k=1$ then the denominator in (1) is equal to zero, for $x=X_j$ we get $r_1(X_j)=|X_j-X_i|=0$ as $i=j$. We do not take the value $k=1$ into account.

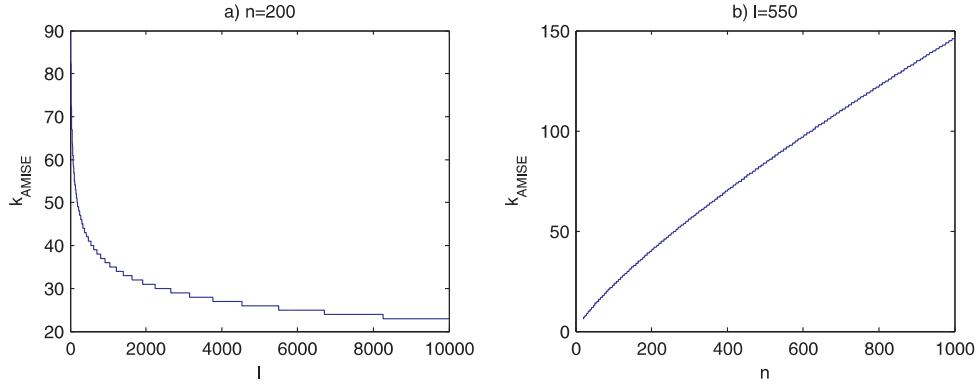

FIGURE 2. Demonstration of dependence $\hat{k}_{\mathrm{AMISE}}$ on the parameter l (a) size of the random sample is $n = 200$) and on the size of the random sample n (b) the parametr $l = 550$).

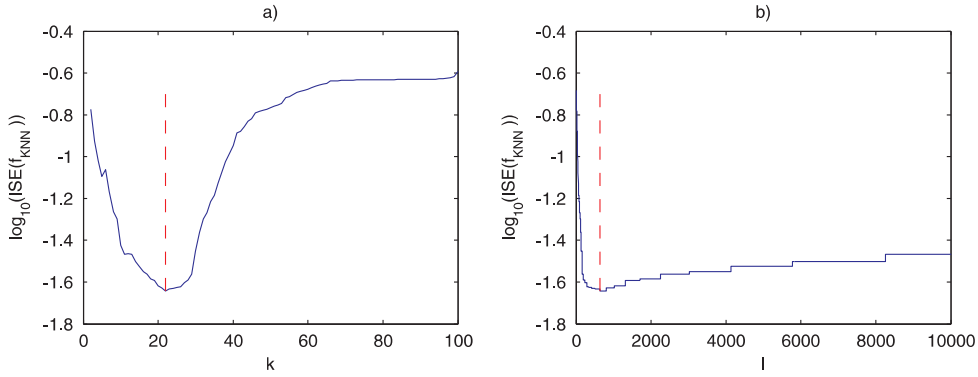

FIGURE 3. Example of dependence of functional $\log_{10}(\mathrm{ISE}\hat{f}_{KNN})$ on the parameter k (part a), resp. on the parameter l (part b), for simulated data from a distribution with the density function f_6 defined in Section 3.

In Figure 3 we can see the example of functional log ISE for a random sample of the size $n = 100$ from a distribution with the density function f_6 (defined in next Section 3).

The part a) illustrates dependence ISE on k. The red dashed line marks the minimum of ISE. In our case the functional ISE is flat in the neighbourhood of the minimum. If we manage to estimate k in this region, we get the result with ISE that is close to the minimum value.

The part b) illustrates the dependence of ISE on l. The red dashed line marks the minimum of functional ISE. We can see the functional is again flat in the neighbourhood of the minimum, especially on the right side of the minimum, but in this case it is much flatter than in the part a).

It appears that if we transform the problem of estimating k into the problem of estimating l, we can get better results making the same error when estimating an unknown parameter. We need to realize this hypothesis is based only on the observation of the functional ISE behavior. We will try to verify this hypothesis in a simulation study in the next section of this article.

3. Simulation study

The simulation study is divided into two parts. In the first part we will compare the quality of results of KNN estimate for different densities and for different values of the parameter l. The main goal of the first part will be to propose a concrete value of parameter l that will yield acceptable results for all tested densities.

The second part of the simulation study will be dedicated to the comparison of KNN method with two other kernel density methods.

In the simulation study we will use simulated data from six different distributions with normal mixture densities:

1. standard normal $f_1 \sim N(0,1)$,
2. skewed data $f_2 \sim \frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{4}, \frac{4}{9}\right) + \frac{3}{5}N\left(\frac{13}{12}, \frac{25}{81}\right)$,
3. kurtotic unimodal $f_3 \sim \frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \frac{1}{100}\right)$,
4. asymetric bimodal $f_4 \sim \frac{4}{5}N(0,1) + \frac{1}{5}N(2, \frac{1}{25})$,
5. symetric trimodal $f_5 \sim \frac{9}{20}N\left(-\frac{7}{4}, 1\right) + \frac{9}{20}N\left(\frac{7}{4}, 1\right) + \frac{1}{10}N\left(0, \frac{1}{25}\right)$,
6. asymetric trimodal $f_5 \sim \frac{3}{10}N\left(-2, \frac{1}{4}\right) + \frac{3}{10}N\left(\frac{7}{4}, \frac{1}{5}\right) + \frac{2}{5}N(0,2)$.

All densities are continuous with infinite support. Different unimodal, bimodal and trimodal densities with high and low peaks were chosen.

3.1. Choice of optimal value of l

The Section 2 demonstrated that k AMISE optimal depends only on the reference density function through the parameter l. Our hypothesis is that we can choose one value of l that will be used while estimating all kinds of densities and that can give us satisfactory results.

Random samples of size $n = 50, \ldots, 300$ were simulated from the distributions with densities $f_1, \ldots, f_6$. The results of ISE were compared for $l = 2, 3, \ldots, n$ for six different random samples, where each random sample belongs to different distribution. Then the value of l was chosen that on average yielded the best results for all densities. This step was repeated 1000 times.

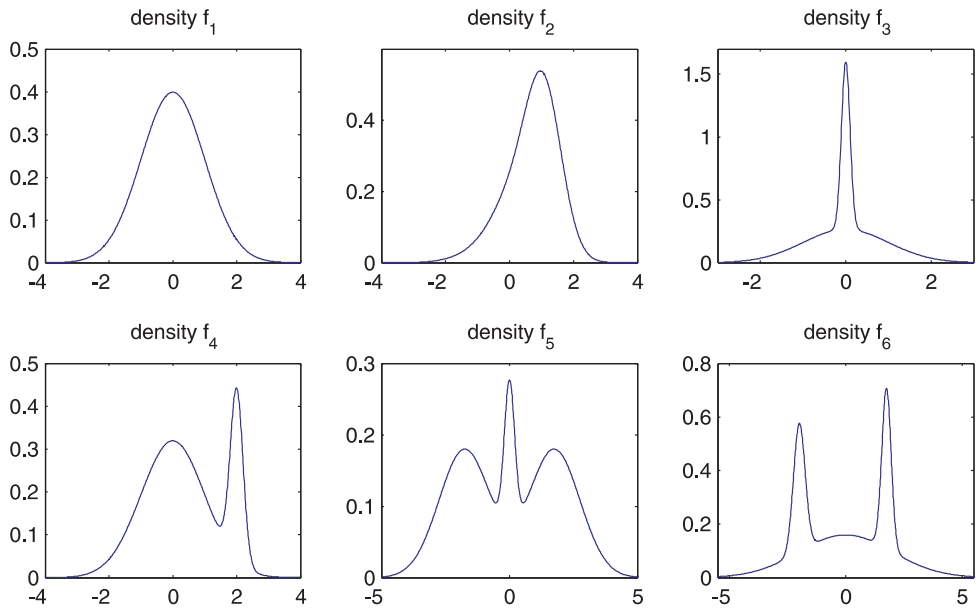

FIGURE 4. Graphs of densities used in the simulation study.

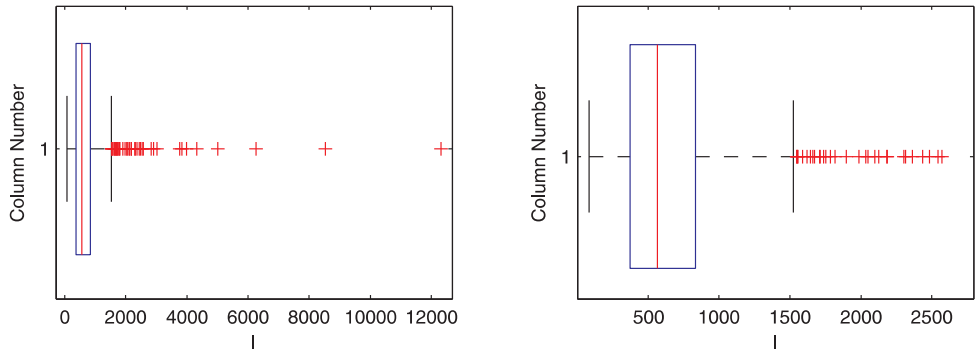

FIGURE 5. Box plots of values l, that minimized ISE for simulated data with $f_1, \ldots, f_6$ densities.

The box plot on the left in Figure 5 shows all 1000 estimated values of l that minimized AMISE in the certain step of the simulation. The box plot on the right is an enlarged box plot on the left side without extreme values.

We propose to take a median of estimates of l as optimal value of l, so

$$l_{\text{opt}} = 550, \quad \text{then} \quad \hat{k}_{\text{AMISE}} = \text{round}\big(0,587 \cdot n^{\frac{4}{5}}\big). \tag{19}$$

We will test this result in the second part of the simulation study.

3.2. Comparison of KNN estimate with other estimation methods

In this section we will compare the quality of results of KNN method using k_{opt}, KNN method using $\hat{k}_{\text{AMISE}}$, KNN method using $\hat{k}_{\text{CVML}}$ chosen by cross validation maximum likelihood method (for details see [5]) and the kernel density method using the global smoothing parameter h estimated by the least squared cross validation method (described in [7]).

Data were simulated from distributions with densities $f_1, \ldots, f_6$; four different sizes of random samples $n = [50, 75, 100, 300]$ were chosen. The simulation was repeated 1000 times.

Figure 6 shows the total overview of simulation results. Each box plot presents results for simulated data from distribution with certain density f_i and a random sample size n. Four different methods mentioned above were compared. We can see that only the ISE has decreasing tendencies for increasing n of all methods. When comparing KNN methods, the best results were achieved using k_{opt}, and not $\hat{k}_{\text{AMISE}}$. Surprisingly, the data driven method of the estimation $\hat{k}_{\text{CVML}}$ yielded worse results than the choice of $\hat{k}_{\text{AMISE}}$ with fixed l in all cases.

As k_{opt} cannot be used in practice, the $\hat{k}_{\text{AMISE}}$ seems to be the best estimated choice of the parameter k. Let us compare the quality of KNN method with the classical kernel estimation method. The result varies for different densities. Clearly, the classical kernel density estimation has better results for the densities f_1 and f_2. These two densities are simple unimodal densities, so the classical estimate method provides sufficient results. Estimating more variate densities with high peaks f_3 and f_6 yields a converse result. The results for densities f_4 and f_5 are quite similar and it is difficult to say which method yields to better results in the sense of ISE.

On Figure 7 we can see an overview of all results for simulated data with all six densities.

4. Conclusion

In this paper a method for estimating the value of k was proposed and tested in a simulation study.

It is difficult to decide whether this method can be used in the practice or not. We can only say that the KNN method has comparable ISE to the classical method. There are big differences for various densities. In general, the KNN

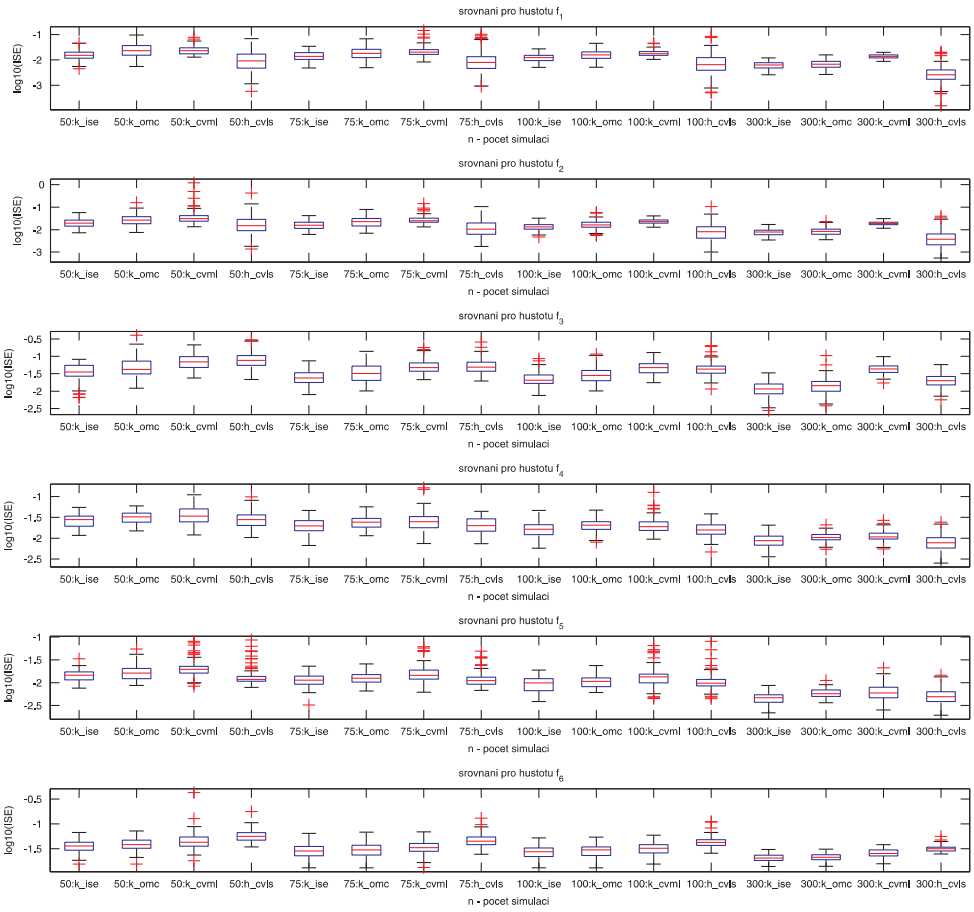

FIGURE 6. The box plots of $\log \mathrm{ISE}(\hat{f}_i)$, for $i = 1, \ldots, 6$. The notation k_ise represents the KNN estimation using k_{opt}, k_mcd represents KNN estimation using $\hat{k}_{\mathrm{AMISE}}$, k_cvml represents KNN estimation using $\hat{k}_{\mathrm{CVML}}$ and h_cvls represents the classical kernel density estimate using h estimated by the least square cross validation method. First four box plots correspond to the same random sample of size $n = 50$, second four box plots correspond to the same random sample of size $n = 75$, third four box plots correspond to the samples of the size $n = 100$ and the last four have $n = 300$.

method gives better results for multimodal densities, densities with high peaks etc. On the other hand, we can see that for simple densities these two KNN methods give worse results. The simulation study also showed that for increasing n the ISE of the estimate has decreasing tendencies for all compared methods.

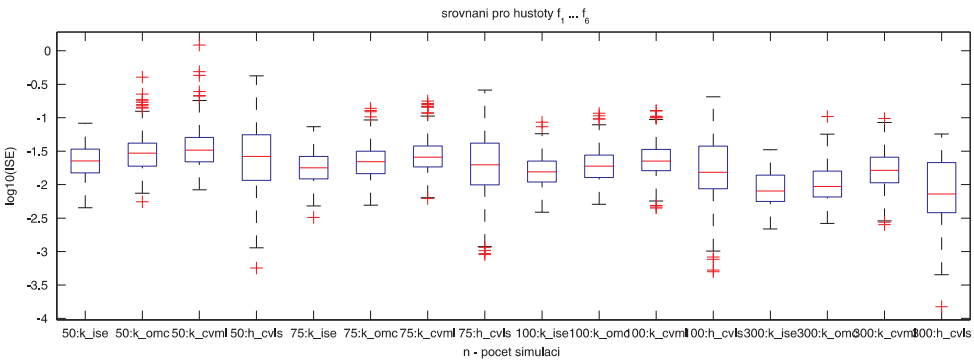

FIGURE 7. Total box plots of $\log \mathrm{ISE}(\hat{f}_i)$, for $i = 1, \ldots, 6$. Notation is the same like in the Figure 6.

The conclusion is that for simulated densities the KNN method gave comparable error. The advantage of estimating the value of $\hat{k}_{\mathrm{AMISE}}$ is that it is not demanding in terms of a computing capacity. The KNN method uses a simple expression that depends only on the size of a random sample.

REFERENCES

[1] BOWMAN, A. W.: *An alternative method cross-validation for the smoothing of density estimates*, Biometrika **71** (1984), 353–360.
[2] GYÖRFI, L.: *On the rate of convergence of nearest neighbor rules*, IEEE Trans. Inform. Theory **24** (1978), 509–512.
[3] GYÖRFI, L.: *The rate of convergence of k_n-nn regression estimates and classification rules*, IEEE Trans. Inform. Theory **27** (1981), 362–364.
[4] MACK, Y. P.—ROSENBLATT, M.: *Multivariate k-nearest neighbour density estimates*, J. Multivariate Anal. **9** (1979), 1–15.
[5] ORAVA, J.: *K-nearest neighbour kernel density estimation, the choice of optimal k*, Biometrika (accepted).
[6] VAN RYZIN, J.: *On strong consistency of density estimates*, Ann. Math. Statist. **40** (1969), 1765–1772.
[7] WAND, M. P.—JONES, M. C.: *Kernel Smothing*. Chapman and Hall, London, 1995.

Received August 14, 2011

Masaryk University
Department of Mathematics and Statistics
Kotlářská 2
CZ–611-37 Brno
CZECH REPUBLIC
E-mail: orava@mail.muni.cz