MASARYK UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS AND STATISTICS

Habilitation Thesis

BRNO 2014 **JAN KOLÁČEK**

MASARYK UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS AND STATISTICS

Theory and Practice of Kernel Smoothing

Habilitation Thesis

Jan Koláček

Brno 2014

Contents

Abstrakt

Habilitační práce je souborem článků [2 – 10] publikovaných v mezinárodních časopisech, z nichž čtyři jsou evidovány v databázi Web of Science. Většina těchto článků má spoluautory, jimiž jsou Ivana Horová, Kamila Vopatová, R.J. Karunamuni a další, přičemž podíl všech autorů na společných článcích je rovnocenný. Práce také odkazuje na knihu [1], která vyšla v roce 2012 v nakladatelství *World Scientific* a je shrnutím získaných poznatků a praktickou aplikací našich výsledků v jazyce Matlab.

Oblastí našeho výzkumu je teorie jádrového vyhlazování, které zaznamenalo v posledních dvaceti letech nebývalý rozmach. V současnosti patří jádrové vyhlazování ke standardním neparametrickým technikám používaných při zpracování a modelování dat. Základy teorie jádrového vyhlazování jsou popsány v monografiích [48, 74, 79]. Řídícím faktorem při jádrovém vyhlazování je vyhlazovací parametr, který se v jednorozměrném případě nazývá šířka vyhlazovacího okna, ve vícerozměrném případě jej nazýváme vyhlazovací matice. V našem výzkumu jsme se tedy zaměřili především na volbu tohoto vyhlazovacího parametru.

V případě jádrových odhadů regresní funkce byly navrženy dvě nové metody. První předpokládá cyklický plán, kdy se data periodicky opakují, a byla publikována v [10]. Druhá metoda byla představena v článku [7] a její statistické vlastnosti byly odvozeny v článku [3], který byl loni přijat k publikaci. V souvislosti s odhady regresní funkce byly také studovány hraniční efekty (viz [25]) a dále problematika jádrových odhadů regresní funkce pro korelovaná data (viz [4]).

Neméně zajímavým tématem v této oblasti je problematika hraničních efektů, které při jádrových odhadech nastávají. Zaměřili jsme se zejména na hraniční efekty při jádrových odhadech distribuční funkce. V článku [9] jsme se zabývali potlačením těchto efektů při odhadech ROC křivky. V článku [8] jsme studovali vliv a potlačení efektů při odhadech rizikové funkce. Dále jsme se také zabývali využitím jádrových odhadů ve financích, konkrétně při odhadování indexů a křivek, které popisují kvalitu skóringových modelů (viz [5]).

Velmi významnou část našeho výzkumu tvoří zobecnění principů jádrových odhadů v jednorozměrném případě na vícerozměrný prostor. Zaměřili jsme se nejprve na jádrové odhady hustoty. V článku [6] byla představena iterační metoda pro hledání optimální vyhlazovací matice, zejména její grafická interpretace ve speciálním případě dvourozměrného prostoru a za předpokladu diagonální matice. Statistické vlastnosti a zobecnění na plnou matici pro tuto metodu byly odvozeny v článku [2].

Preface

The thesis is a collection of articles [2 – 10]. Four of them have been published in international journals indexed by Web of Science. The paper [3] was accepted in December 2013. The thesis also refers to the book [1] which is a summary of all results in our research area.

Our main research interest lies in the theory of kernel smoothing. Kernel methods are well-known and intensively used by the community of non-parametricians because they are a useful tool for local weighting. Kernel estimators combine two main advantages: simple expression and ease of implementation.

It is well known that the most important factor in kernel estimation is a choice of smoothing parameters. This choice is particularly important because of its role in controlling both the amount and the direction of smoothing. This problem has been widely discussed in many monographs and papers.

The following overview starts with a motivation of the theory of kernel smoothing and then briefly describes the main contributions of the book [1] and the papers [2 – 10]. In order to make the presentation more compact, the thesis consists of the author's selected papers in the area. In References one can find the list of other related publications of the author [11 – 28].

Pronouncement

Almost all papers included in this thesis have co-authors, namely I. Horová, K. Vopatová, R. J. Karunamuni, J. Zelinka, M. Řezáč and D. Lajdová. In all cases, the contributions of all authors were equivalent, since the results were based on common discussions. Formally, the author's contribution to the paper [10] was 100%, the author's contribution to the papers [3, 5, 7, 8, 9] was 50% and the author's contribution to the monograph [1] and the papers [2, 4, 6] was 33%.

Acknowledgement

I wish to thank all the co-authors for their friendly and always very helpful collaboration. I would like to express my gratitude to my colleague Prof. Ivana Horová for our numerous interesting discussions. And most importantly, I would like to thank my wife Veronika. Her support, encouragement, patience and love were the bedrock upon which the past eight years of my life have been built.

1 Introduction

Kernel smoothing belongs to a general category of techniques for nonparametric curve estimations including nonparametric regression, nonparametric density estimators and nonparametric hazard functions. These estimations depend on a smoothing parameter called a bandwidth which controls the smoothness of the estimate and on a kernel which plays a role of weight function. As far as the kernel function is concerned, a key parameter is its order which is related both to the number of its vanishing moments and to the number of existing derivatives for the underlying curve to be estimated. As concerns a bandwidth choice – it is the crucial problem in the kernel smoothing and this is the main topic of our research.

The first part of our research includes a methodology for nonparametric regression analysis, complemented with practical applications. In nonparametric regression estimation, a critical and inevitable step is to choose the smoothing parameter (bandwidth) to control the smoothness of the curve estimate. The smoothing parameter considerably affects the features of the estimated curve. Although in practice one can try several bandwidths and choose a bandwidth subjectively, automatic (data-driven) selection procedures could be useful for many situations; see [73] for more examples. Several automatic bandwidth selectors were proposed and studied in [37], [50], [49], [38] with the references included. It is well recognized that these bandwidth estimates are the subject to large sample variation. The kernel estimates based on the bandwidths selected by these procedures could have very different appearances. Due to the large sample variation, classical bandwidth selectors might not be very useful in practice. This fact has motivated us in our research to find new methods for bandwidth selection which give much more stable bandwidth estimates.

In connection with the kernel regression analysis we have to mention one essential fact. The regression model assumes no correlation in measurements. In the case of independent observations the literature on bandwidth selection methods is quite extensive. Nevertheless, if an autocorrelation structure of errors occurs in data, then classical bandwidth selectors have not always provided applicable results (see [35]). Many real data sets (especially time series) show the autocorrelation. This has led us to study possibilities for overcoming the effect of dependence on the bandwidth selection.

The next part of our research is focused on the studying of boundary effects in kernel estimation. In practical processing we encounter data which are bounded in some interval. The quality of the estimate in the boundary region is affected since the "effective" window does not belong to this interval, therefore the finite equivalent of the moment conditions on the kernel

function does not apply any more. This phenomenon is called the *boundary effect*. Although there is a vast literature on boundary correction in density estimation context, the boundary effects problem in the cumulative distribution function and the regression function context has been less studied. Thus, we have focused our research to these areas of kernel smoothing.

As we have already mentioned, kernel smoothing is widely used in many statistical research areas. One of them is focused on studying discrimination measures used to determine the quality of models at separating in a binary classification system. There are many possible ways to measure the performance of the classification rules. It is often very helpful to be given a method for displaying and summarizing performance over a wide range of conditions. This aim is fulfilled, e.g., by the ROC (Receiver Operating Characteristic) curve, Information value curve, Lift, Kolmogorov-Smirnov statistics and others. There are many problems in the estimation of these curves in practice and the kernel smoothing approach seems to be very helpful. Thus, our research has been directed also to this area.

The important part of our research is devoted to the extension of the univariate kernel density estimate to the multivariate setting. As we have already explained, the typical question, motivated by the origins of this research area, asks to determine the optimal smoothing parameter (matrix). Some "classical" methods in the multivariate case were developed and widely discussed in papers [31], [42], [41], [68], [39]. Tarn Duong's PhD thesis ([39]) provided a comprehensive survey of bandwidth matrix selection methods for kernel density estimation. Papers [32], [40] investigated general density derivative estimators, i.e., kernel estimators of multivariate density derivatives using general (or unconstrained) bandwidth matrix selectors. We have followed mentioned papers and we have proposed a new data-driven bandwidth matrix selection method. Ideas similar to this method have been applied to kernel estimates of multivariate regression functions.

We would like to emphasize a great interest and usefulness of all mentioned problems in many fields of applied sciences (environmetrics, chemometrics, biometrics, medicine, econometrics, ...). Thus our works deal not only with the theoretical background of the considered problems but also with the application to real data. For example, see [4] where the utility of the proposed method was illustrated through an application to the time series of ozone data. For applications of smoothing methods in medicine see [14]. The wide range of applications in finance can be found in [5, 16, 17, 20, 21, 18]. The use of some proposed methods for modeling in environmetrics was described in [22]. See the list in "Other Publications of the Author" at the end of the thesis for more references.

Author's Contribution

Our interest is focused on an outstanding open problem of the optimal bandwidth matrix selection in the multivariate case. Although there exist several classical approaches, it is problematic to implement them in practice because of their computational difficulty. Our results concerning this problem are described in Section 6. The author considers these results to be the most valuable part of the thesis since they can potentially constitute a significant step towards a more effective computable solution of the problem.

In Section 3 we overview our results concerning two other related problems. The main part describes results concerning optimal bandwidth selection for the univariate kernel regression and the remaining part deals with the problem of autocorrelated data in kernel regression.

Our investigations of boundary effects in kernel smoothing (Section 4) serve as a supporting ground for new techniques in reliability assessment (Section 5), and the results obtained there could be beneficial for applications in other research areas.

Finally, Section 7 presents a monograph, where all results of our research are summarized. An integral part of the book is a special toolbox in MATLAB. The toolbox is described in the book in detail and provides a practical implementation of presented methods.

2 Assumptions and notations

In this section, we introduce a definition of the kernel and show notations and general assumptions used in our research.

Definition 1. *Let ν, k be nonnegative integers, $0 \le \nu < k$. Let K be a real valued function satisfying $K \in S_{\nu,k}$,*
where

$$S_{\nu,k} = \left\{ \begin{array}{l} K \ \in Lip[-1,1], \ support(K) = [-1,1] \\ \\ \int\limits_{-1}^{1} x^j K(x)dx = \left\{ \begin{array}{ll} 0, & 0 \le j < k, j \ne \nu \\ (-1)^\nu \nu!, & j = \nu \\ \beta_k \ne 0, & j = k. \end{array} \right. \end{array} \right. \tag{1}$$

Such a function is called a kernel of order k. The integral conditions are often called moment conditions.

A commonly used kernel function is the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Nevertheless, this kernel has an unbounded support and thus it does not belong to the class $S_{\nu,k}$.

2.1 The univariate case

Let us consider a univariate function f (a density function or a regression function) which should be estimated. We present a short overview of the notation and assumptions used in our research.

$(N1)$ The positive number h is a smoothing parameter called also a *bandwidth*. The bandwidth h is depending on n, $h = h(n)$: $\{h(n)\}_{n=1}^{\infty}$ is a nonrandom sequence of positive numbers.

(N_2) $K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right)$, $K \in S_{0,k}$, k is even, $h > 0$.

(N_3) $V(\rho) = \int\limits_{\mathbb{R}} \rho^2(x)dx$ for any square integrable scalar valued function ρ.

(A_1) $K \in S_{0,k} \cap C^{\nu}[-1,1]$, $K^{(j)}(-1) = K^{(j)}(1) = 0$, $j = 0, 1, \ldots, \nu$, $\nu \in \mathbb{N}$, i.e., $K^{(\nu)} \in S_{\nu,k+\nu}$ (see [46, 60]).

(A_2) $f \in C^{k_0}$, $\nu + k \leq k_0$, $f^{(\nu+k)}$ is square integrable.

(A_3) $\lim\limits_{n\to\infty} h = 0$, $\lim\limits_{n\to\infty} nh^{2\nu+1} = \infty$.

2.2 The multivariate case

This part is devoted to the extension of assumptions for the univariate case to the multivariate setting. Let us consider a d-dimensional space as the domain of the estimated function f.

(N_1) $\mathcal{H}$ denotes a class of $d \times d$ symmetric positive definite matrices.

(N_2) $V(g) = \int\limits_{\mathbb{R}^d} g(\mathbf{x})g^T(\mathbf{x})d\mathbf{x}$ for any square integrable vector valued function g.

(A_1) The kernel function K satisfies the moment conditions $\int K(\mathbf{x})d\mathbf{x} = 1$, $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = \mathbf{0}$, $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \beta_2\mathbf{I}_d$, $\mathbf{I}_d$ is the $d \times d$ identity matrix.

(A_2) $\mathbf{H} \in \mathcal{H}$, $\mathbf{H} = \mathbf{H}_n$ is a sequence of bandwidth matrices such that $n^{-1/2}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1})^j$, $j = 0, 1, \ldots, \nu$, $\nu \in \mathbb{N}$, and entries of $\mathbf{H}$ approach zero ($(\mathbf{H}^{-1})^0$ is considered as equal to 1).

(A_3) Each partial derivative of f of order $j+2$, $j = 0, 1, \ldots, \nu$, is continuous and square integrable.

3 Kernel estimation of a regression function

One of our research interest includes the methodology for nonparametric regression analysis, combined with practical applications.

The aim of regression analysis is to produce a reasonable analysis of an unknown regression function m. By reducing the observational errors it allows the interpretation to concentrate on important details of the mean dependence of Y on X. Kernel regression estimates are one of the most popular nonparametric estimates.

Let us consider a standard regression model of the form

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2}$$

where m is an unknown regression function, $Y_1, \ldots, Y_n$ are observable data variables with respect to the design points $x_1, \ldots, x_n$. The residuals $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed random variables for which

$$\mathrm{E}(\varepsilon_i) = 0, \qquad \mathrm{var}(\varepsilon_i) = \sigma^2 > 0, \qquad i = 1, \ldots, n.$$

We suppose the *fixed equally spaced design*, i.e., design variables are not random and $x_i = i/n,\ i = 1, \ldots, n$. In the case of *random design*, where the design points $X_1, \ldots, X_n$ are random variables with the same density f, all considerations are similar to the fixed design. A more detailed description of the random design can be found, e.g., in [79].

The most popular regression estimator was proposed by Nadaraya and Watson ([64] and [80]) and it is defined as

$$\widehat{m}_{NW}(x, h) = \frac{\sum_{i=1}^{n} K_h(x_i - x) Y_i}{\sum_{i=1}^{n} K_h(x_i - x)}. \tag{3}$$

In order to complete the overview of commonly used nonparametric methods for estimating $m(x)$ we mention these estimators:

- *local – linear estimator* ([76, 36])

$$\widehat{m}_{LL}(x, h) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{\hat{s}_2(x, h) - \hat{s}_1(x, h)(x_i - x)\} K_h(x_i - x) Y_i}{\hat{s}_2(x, h)\hat{s}_0(x, h) - \hat{s}_1(x, h)^2}, \tag{4}$$

  where

$$\hat{s}_r(x, h) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^r K_h(x_i - x), \qquad r = 0, 1, 2,$$

- *Priestley – Chao estimator* ([66])

$$\widehat{m}_{PCH}(x,h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_i - x)Y_i, \tag{5}$$

- *Gasser – Müller estimator* ([44])

$$\widehat{m}_{GM}(x,h) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(t-x)dt, \tag{6}$$

where

$$s_i = \frac{x_i + x_{i+1}}{2} = \frac{2i+1}{2n},\ i = 1,\dots,n-1, \quad s_0 = 0, \quad s_n = 1.$$

One can see from these formulas that kernel estimators can be generally expressed as

$$\widehat{m}(x,h) = \sum_{i=1}^{n} W_i^{(j)}(x,h)Y_i, \tag{7}$$

where weights $W_i^{(j)}(x,h)$, $j \in \{NW, LL, PCH, GM\}$ correspond to weights of estimators $\widehat{m}_{NW}$, $\widehat{m}_{LL}$, $\widehat{m}_{PCH}$ and $\widehat{m}_{GM}$ defined above.

In the univariate case, these estimators depend on a bandwidth, which is a smoothing parameter controlling the smoothness of an estimated curve and a kernel which is considered as a weight function.

3.1 Choosing the shape of the kernel

The choice of the kernel does not influence the asymptotic behavior of the estimate so significantly as the bandwidth does. We assume $K \in S_{0,k}$ and under the additional assumption that k is even, $k > 0$. More detailed procedures for choosing the optimal kernel are described in [1].

3.2 Choosing the optimal bandwidth

The choice of the smoothing parameter is a crucial problem in the kernel regression. The literature on bandwidth selection is quite extensive, e.g., monographs [79, 74, 75], papers [48, 33, 34, 67, 77, 37, 38, 58, 10].

Although in practice one can try several bandwidths and choose a bandwidth subjectively, automatic (data-driven) selection procedures could be

useful for many situations; see [73] for more examples. Most of these procedures are based on estimating Average Mean Square Error. They are asymptotically equivalent and asymptotically unbiased (see [48, 33, 34]). However, in simulation studies ([58]), it is often observed that most selectors are biased toward undersmoothing and yield smaller bandwidths more frequently than predicted by asymptotic results.

As a part of our research we developed two methods for the optimal bandwidth selections.

3.2.1 Plug-in method

In the simulation study of [33], it was observed that standard criterions give smaller bandwidths more frequently than predicted by the asymptotic theorems. [33] provided an explanation for the cause and suggested a procedure to overcome the difficulty. By applying the procedure, we have introduced a method for bandwidth selection which gives much more stable bandwidth estimates (see [10]). As a result, we have obtained a type of plug-in method.

Our ideas are based on an assumption of a "cyclic design", that is, we suppose m to be a smooth periodic function and the estimate is obtained by applying the kernel on the extended series $\widetilde{Y}_i$, $i = -n+1, -n+2, \ldots, 2n$, where generally $\widetilde{Y}_{j+ln} = Y_j$ for $j = 1, \ldots, n$ and $l \in \mathbb{Z}$. Similarly $x_i = i/n$, $i = -n+1, -n+2, \ldots, 2n$.

The main result of the paper [10] is the plug-in estimator of the optimal bandwidth h

$$\hat{h}_{\mathrm{PI}} = \left(\frac{\hat{\sigma}^2 V(K)(k!)^2}{2kn\beta_k^2 \widehat{A}_k} \right)^{\frac{1}{2k+1}} . \tag{8}$$

We would like to point out the computational aspect of the proposed estimator. It has preferable properties compared to the classical methods because there is no problem of minimization of any error function. Also, the sample size which is necessary for computing the estimation is far less than for classical methods. On the other hand, a minor disadvantage could be the fact that we need a "starting" approximation of the unknown parameter h. We would also like to specify the proposed method was developed for a rather limited case: the cyclic design.

3.2.2 Iterative method

Successful approaches to the bandwidth selection in kernel density estimation can be transferred to the case of kernel regression. The iterative method for the kernel density was developed and widely discussed in [54]. The ideas

of this paper were extended to the regression case. The obtained selector was introduced in [7] and its statistical properties were derived in [3]. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated square bias

$$\mathrm{AIV}\,\{\widehat{m}(\cdot, h_{opt})\} - 2k\,\mathrm{AISB}\{\,\widehat{m}(\cdot, h_{opt})\} = 0, \tag{9}$$

where

$$\mathrm{AIV}\,\{\widehat{m}(\cdot, h_{opt})\} = \frac{\sigma^2 V(K)}{nh}$$

and

$$\mathrm{AISB}\{\,\widehat{m}(\cdot, h_{opt})\} = \frac{1}{n}\sum_{i=1}^{n}(E\widehat{m}(x_i, h) - m(x_i))^2.$$

The main idea consists in finding a fixed point of the equation

$$h = \frac{\hat{\sigma}^2 V(K)}{2knh\widehat{\mathrm{AISB}}\,\{\widehat{m}(\cdot, h)\}}. \tag{10}$$

We use Steffensen's iterative method with the starting approximation $\hat{h}_0 = 2/n$. This approach leads to an iterative quadratically convergent process (see [54]).

3.3 Kernel regression for correlated data

As mentioned above, the literature on bandwidth selection methods is quite extensive in the case of independent observations. Nevertheless, if an autocorrelation structure of errors occurs in data, then classical bandwidth selectors have not always provided applicable results (see [35]). There exist several possibilities for overcoming the effect of dependence on the bandwidth selection.

In the paper [4] we used the results of [35] and [10] and developed a new flexible plug-in approach for estimating the optimal smoothing parameter. The utility of the method was illustrated through a simulation study and application to the time series of ozone data obtained from the Vernadsky station in Antarctica.

4 Boundary effects in kernel estimation

In practical processing we encounter data which are bounded in some interval. The quality of the estimate in the boundary region is affected since the "effective" window $[x-h, x+h]$ does not belong to this interval, so the finite equivalent of the moment conditions on the kernel function does not apply any more. This phenomenon is called the *boundary effect*. There are several methods to cope with boundary effects. One of them is based on the construction of special *boundary kernels*. Their construction was described in details for instance in [63] or [51]. These kernels can be used successfully in kernel regression but their use in density or distribution function estimates gives often inappropriate results.

Although there is a vast literature on the boundary correction in density estimation context, the boundary effects problem in distribution function and regression function context has been less studied. Thus we focused our research on these areas of kernel smoothing.

4.1 Boundary effects in kernel regression

If the support of the true regression curve is bounded then most nonparametric methods give estimates that are severely biased in regions near the endpoints. To be specific, the bias of $\widehat{m}(x)$ is of order $O(h)$ rather than $O(h^2)$ for $x \in [0,h] \cup [1-h,1]$. This boundary problem affects the global performance visually and also in terms of a slower rate of convergence in the usual asymptotic analysis. It has been recognized as a serious problem and many works are devoted to reducing the effects.

[44, 45, 46] and [63] discussed boundary kernel methods. Another approach to the boundary problems are reflection methods which generally consist in reflecting data about the boundary points and then estimating the regression function. These methods were discussed, e.g., in [69, 47]. The reflection principles used in kernel density estimation can be also adapted to kernel regression. The regression estimator with the assumption of the "cyclic" model described in [10] can be also considered as the special case of a reflection technique. A short comparative study of methods for boundary effects eliminating was given in [25].

4.2 Boundary effects in kernel estimation of a distribution function

We have focused also on the boundary correction in kernel estimation of a cumulative distribution function (CDF) which is important for other applica-

tions – especially for kernel estimation of ROC curves and hazard functions.

In the paper [9], we developed a new kernel type estimator of the ROC curve that removes boundary effects near the end points of the support. The estimator is based on a new boundary corrected kernel estimator of distribution functions and it is based on ideas of [56, 57], developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is a type of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We have derived expressions for the bias and variance of the proposed estimator. Furthermore, the proposed estimator has been compared with the "classical estimator" using simulation studies.

Using similar ideas as in [9] we have developed a new kernel estimator of the hazard function. The method was proposed in [8] and successfully removes boundary effects and performs considerably better than classical estimators.

5 Kernel estimation and reliability assessment

The following part of our research is focused on studying discrimination measures used for detecting the quality of models at separating in a binary classification system. There are many possible ways of measuring the performance of the classification rules. It is often very helpful to know a way of displaying and summarizing performance over a wide range of conditions. This aim is fulfilled by the ROC (Receiver Operating Characteristic) curve. It is a single curve summarizing the distribution functions of the scores of two classes.

In our research, we have followed the financial sphere, where the discrimination power of scoring models is evaluated. However, most of all studied indices have wide application in many other areas, where models with binary output are used, like biology, medicine, engineering and so on.

References on this topic are quite extensive, see, e.g., [72, 29, 78]. In [5], we summarized the most important quality measures and gave some alternatives to them. All of the mentioned indices are based on the density or on the distribution function, therefore one can suggest the technique of kernel smoothing for estimation. More detailed studies on all indices can be also found, e.g., in [20, 21]. Finally, a new conservative approach to quality assessment was proposed in [18].

6 Multivariate kernel density estimation

An important part of our research is devoted to the extension of the univariate kernel density estimate to the multivariate setting.

Let a d-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be drawn from a density f. The kernel density estimator $\hat{f}$ at the point $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$\hat{f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \tag{11}$$

where K is a kernel function, which is often taken to be a d-variate symmetric probability function, $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix and $K_{\mathbf{H}}$ is the scaled kernel function

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

with $|\mathbf{H}|$ the determinant of the matrix $\mathbf{H}$.

In a univariate case, kernel estimates depend on a bandwidth, which is a smoothing parameter controlling smoothness of an estimated curve and a kernel which is considered as a weight function. The choice of the smoothing parameter is a crucial problem in the kernel density estimation. The literature on bandwidth selection is quite extensive, e.g., monographs [79], [74], [75], papers [61], [65], [71], [55], [30]. As far as the kernel estimate of density derivatives is concerned, this problem has received significantly less attention. In paper [50], an adaptation of the least squares cross-validation method was proposed for the bandwidth choice in the kernel density derivative estimation. In paper [52], the automatic procedure of simultaneous choice of the bandwidth, the kernel and its order for kernel density and its derivative estimates was proposed. But this procedure can be only applied in case that the explicit minimum of the Asymptotic Mean Integrated Square Error of the estimate is available. It is known that this minimum exists only for $d = 2$ and the diagonal matrix H. In paper [6], the basic formula for the corresponding procedure was given.

The need for nonparametric density estimates for recovering the structure in multivariate data is greater since a parametric modelling is more difficult than in the univariate case. The extension of the univariate kernel methodology is not without problems. The most general smoothing parameterization of the kernel estimator in d dimensions requires the specification entries of $d \times d$ positive definite bandwidth matrix. The multivariate kernel density estimator we have dealt with is a direct extension of the univariate estimator (see, e.g., [79]).

Successful approaches to the univariate bandwidth selection can be transferred to the multivariate settings. The least squares cross-validation and plug-in methods in the multivariate case were developed and widely discussed in papers [31], [42], [41], [68], [39]. Some papers (e.g., [23], [6], [19]) were focused on constrained parameterization of the bandwidth matrix such as a diagonal matrix. It is a well-known fact that a visualization is an important component of the nonparametric data analysis. In paper [6], this effective strategy was used to clarify the process of the bandwidth matrix choice using bivariate functional surfaces. The paper [53] brought a short communication on a kernel gradient estimator. Tarn Duong's PhD thesis ([39]) provided a comprehensive survey of bandwidth matrix selection methods for kernel density estimation. The papers [32], [40] investigated general density derivative estimators, i.e., kernel estimators of multivariate density derivatives using general (or unconstrained) bandwidth matrix selectors. They defined the kernel estimator of the multivariate density derivative and provided results for the Mean Integrated Square Error convergence asymptotically and for finite samples. Moreover, the relationship between the convergence rate and the bandwidth matrix was established here. They also developed estimates for the class of normal mixture densities.

We have followed the mentioned papers and in [2] we proposed a new data-driven bandwidth matrix selection method. This method is based on an optimally balanced relation between the integrated variance and the integrated squared bias, see [54]. Similar ideas have been applied to kernel estimates of regression functions (see [7] or [3]). We have discussed the statistical properties and relative rates of convergence of the proposed method as well.

7 The monograph

The knowledge obtained in our research in kernel smoothing theory has resulted in writing a monograph [1]. The book provides a brief comprehensive overview of statistical theory. We do not concentrate on details since there exists a number of excellent monographs developing statistical theory ([79, 48, 62, 74, 75, 70] *etc.*). Instead, the emphasis is given to the implementation of presented methods in MATLAB. All created programs are included into a special toolbox which is an integral part of the book. This toolbox contains many MATLAB scripts useful for kernel smoothing of density, distribution function, regression function, hazard function, multivariate density and also for kernel estimation and reliability assessment. The toolbox can be downloaded from the public web page (see [59]).

The toolbox is divided into six parts according to the chapters of the book. All scripts are included in a user interface and it is easy to manipulate with this interface. Each chapter of the book contains a detailed help for the related part of the toolbox.

The monograph is intended for newcomers to the field of smoothing techniques and would be also appropriate for a wide audience: advanced graduate and PhD students, researchers from both the statistical science and interface disciplines.

8 Conclusion and further research

The previous text summarizes all our results in kernel smoothing which belongs to a general category of techniques for nonparametric curve estimations. We have studied several parts of kernel smoothing theory. The most interesting theoretical results were obtained in multivariate kernel estimating and in the choosing of the optimal smoothing parameter.

We have also paid attention to the use of our results in many fields of applied sciences like environmetrics, biometrics, medicine or econometrics. Thus our works deal not only with the theoretical background of the considered problems but also with the application to real data.

In the further research we would like to aim at extending our previous results to modeling for functional data sets. The functional data set can be defined as the observation of the random variable which takes values in an infinite dimensional space (or functional space). Thus the analysis of functional data seems to be a natural extension of our ideas. For more about functional data analysis see, e.g., [43].

References

Publications Included in the Thesis

[1] I. Horová, J. Koláček, and J. Zelinka, *Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing*. Singapore: World Scientific Publishing Co. Pte. Ltd., 2012.

[2] I. Horová, J. Koláček, and K. Vopatová, "Full bandwidth matrix selectors for gradient kernel density estimate," *Computational Statistics & Data Analysis*, vol. 57, no. 1, pp. 364–376, 2013.

[3] J. Koláček and I. Horová, "Selection of bandwidth for kernel regression," *Communications in Statistics - Theory and Methods*. to appear.

[4] I. Horová, J. Koláček, and D. Lajdová, "Kernel regression model for total ozone data," *Journal of Environmental Statistics*, vol. 4, no. 2, pp. 1–12, 2013.

[5] M. Řezáč and J. Koláček, "Lift-based quality indexes for credit scoring models as an alternative to gini and ks," *Journal of Statistics: Advances in Theory and Applications*, vol. 7, no. 1, pp. 1–23, 2012.

[6] I. Horová, J. Koláček, and K. Vopatová, "Visualization and bandwidth matrix choice," *Communications in Statistics – Theory and Methods*, vol. 41, no. 4, pp. 759–777, 2012.

[7] J. Koláček and I. Horová, "Iterative bandwidth method for kernel regression," *Journal of Statistics: Advances in Theory and Applications*, vol. 8, no. 2, pp. 91–103, 2012.

[8] J. Koláček and R. J. Karunamuni, "A generalized reflection method for kernel distribution and hazard functions estimation," *Journal of Applied Probability and Statistics*, vol. 6, no. 2, pp. 73–85, 2011.

[9] J. Koláček and R. J. Karunamuni, "On boundary correction in kernel estimation of ROC curves," *Austrian Journal of Statistics*, vol. 38, no. 1, pp. 17–32, 2009.

[10] J. Koláček, "Plug-in method for nonparametric regression," *Computational Statistics*, vol. 23, no. 1, pp. 63–78, 2008.

Other Publications of the Author

[11] K. Vopatová, I. Horová, and J. Koláček, "Bandwidth matrix selectors for multivariate kernel density estimation," in *Theoretical and Applied Issues in Statistics and Demography*, pp. 123–130, Barcelona: International Society for the Advancement of Science and Technology (ISAST), 2013.

[12] K. Konečná, I. Horová, and J. Koláček, "Conditional density estimations," in *Theoretical and Applied Issues in Statistics and Demography*, pp. 39–45, Barcelona: International Society for the Advancement of Science and Technology (ISAST), 2013.

[13] D. Lajdová, J. Koláček, and I. Horová, "Kernel regression model with correlated errors," in *Theoretical and Applied Issues in Statistics and Demography*, pp. 81–88, Barcelona: International Society for the Advancement of Science and Technology (ISAST), 2013.

[14] M. Trhlík, R. Soumarová, P. Bartoš, M. Těžká, J. Koláček, K. Vopatová, I. Horová, and P. Šupíková, "Neoadjuvant chemotherapy for primary advanced ovarian cancer," in *The International Journal of Gynecological Cancer – October 2012, vol 22, issue 8, supplement 3, E517*, 2013.

[15] I. Horová, J. Koláček, K. Vopatová, and J. Zelinka, "Contribution to bandwidth matrix choice for multivariate kernel density estimate," in *ISI 2011, Proceedings of the 58th World Statistics Congress*, ISI 2011, 2011.

[16] M. Řezáč and J. Koláček, "Adjusted empirical estimate of information value for credit scoring models," in *PROCEEDINGS ASMDA 2011*, (Rome), pp. 1162–1169, Edizioni ETS, 2011.

[17] J. Koláček and M. M. Řezáč, "Quality measures for predictive scoring models," in *PROCEEDINGS ASMDA 2011* (C. H. S. Raimondo Manca, ed.), (Rome, Italy), pp. 720–727, Edizioni ETS, 2011.

[18] J. Koláček and M. Řezáč, "A conservative approach to assessment of discriminatory models," in *Workshop of the Jaroslav Hájek Center and Financial Mathematics in Practice I, Book of short papers* (I. H. Jiří Zelinka, ed.), (Brno), pp. 30–36, Masaryk University, 2011.

[19] K. Vopatová, I. Horová, and J. Jan Koláček, "Bandwidth matrix choice for bivariate kernel density derivative," in *Proceedings of the 25th International Workshop on Statistical Modelling*, (Glasgow (UK)), pp. 561–564, 2010.

[20] J. Koláček and M. Řezáč, "Assessment of scoring models using information value," in *19th International Conference on Computational Statistics, Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, (Paris), pp. 1191–1198, SpringerLink, 2010.

[21] M. Řezáč and J. Koláček, "On aspects of quality indexes for scoring models," in *19th International Conference on Computational Statistics, Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, (Paris), pp. 1517–1524, SpringerLink, 2010.

[22] I. Horová, J. Koláček, J. Zelinka, and A. H. El-Shaarawi, "Smooth estimates of distribution functions with application in environmental studies," in *Advanced topics on mathematical biology and ecology*, (Mexico), pp. 122–127, WSEAS Press, 2008.

[23] I. Horová, J. Koláček, J. Zelinka, and K. Vopatová, "Bandwidth choice for kernel density estimates.," in *Proceedings IASC*, (Yokohama), pp. 542–551, IASC, 2008.

[24] J. Koláček, "An improved estimator for removing boundary bias in kernel cumulative distribution function estimation," in *Proceedings in Computational Statistics COMPSTAT'08*, (Porto), pp. 549–556, Physica-Verlag, 2008.

[25] J. Koláček and J. Poměnková, "A comparative study of boundary effects for kernel smoothing," *Austrian Journal of Statistics*, vol. 35, no. 2, pp. 281–289, 2006.

[26] J. Koláček, "Use of fourier transformation for kernel smoothing," in *Proceedings in Computational Statistics COMPSTAT'04*, pp. 1329 – 1336, Springer, 2004.

[27] J. Koláček, "Some stabilized bandwidth selectors for nonparametric regression," *Journal of Electrical Engineering*, vol. 54, no. 12, pp. 65–68, 2003.

[28] J. Koláček, "Problems of automatic data-driven bandwidth selectors for nonparametric regression," *Journal of Electrical Engineering*, vol. 53, no. 12, pp. 48–51, 2002.

Other References

[29] R. Anderson. *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press, 2007.

[30] R. Cao, A. Cuevas, and W. González Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, 17(2):153–176, 1994.

[31] J. E. Chacón and T. Duong. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19(2):375–398, 2010.

[32] J. E. Chacón, T. Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21(2):807–840, 2011.

[33] S. Chiu. Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. *Biometrika*, 77(1):222–226, 1990.

[34] S. Chiu. Some stabilized bandwidth selectors for nonparametric regression. *Annals of Statistics*, 19(3):1528–1546, 1991.

[35] C. K. Chu and J. S. Marron. Choosing a kernel regression estimator. *Statistical Science*, 6(4):404–419, 1991.

[36] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[37] P. Craven and G. Wahba. Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1979.

[38] B. Droge. Some comments on cross-validation. Technical Report 1994-7, Humboldt Universitaet Berlin, 1996.

[39] T. Duong. *Bandwidth selectors for multivariate kernel density estimation*. PhD thesis, School of Mathematics and Statistics, University of Western Australia, oct 2004.

[40] T. Duong, A. Cowling, I. Koch, and M. P. Wand. Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis*, 52(9):4225–4242, 2008.

[41] T. Duong and M. Hazelton. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis*, 93(2):417–433, 2005.

[42] T. Duong and M. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.

[43] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

[44] T. Gasser and H.-G. Müller. Kernel estimation of regression functions. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, volume 757 of *Lecture Notes in Mathematics*, pages 23–68. Springer Berlin / Heidelberg, 1979.

[45] T. Gasser, H.-G. Müller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):238–252, 1985.

[46] B. Granovsky and H.-G. Müller. Optimizing kernel methods - a unifying variational principle. *International Statistical Review*, 59(3):373–388, 1991.

[47] P. Hall and T. E. Wehrly. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415):pp. 665–672, 1991.

[48] W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1st edition, 1990.

[49] W. Härdle, P. Hall, and J. Marron. How far are automatically chosen regression smoothing parameters from their optimum. *Journal of the American Statistical Association*, 83(401):86–95, 1988.

[50] W. Härdle, J. S. Marron, and M. P. Wand. Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):223–232, 1990.

[51] I. Horová. Boundary kernels. In *Summer schools MATLAB 94, 95*, pages 17–24. Brno: Masaryk University, 1997.

[52] I. Horová, P. Vieu, and J. Zelinka. Optimal choice of nonparametric estimates of a density and of its derivatives. *Statistics & Decisions*, 20(4):355–378, 2002.

[53] I. Horová and K. Vopatová. Kernel gradient estimate. In F. Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, pages 177–182. Springer-Verlag Berlin Heidelberg, 2011.

[54] I. Horová and J. Zelinka. Contribution to the bandwidth choice for kernel density estimates. *Computational Statistics*, 22(1):31–47, 2007.

[55] M. C. Jones and R. F. Kappenman. On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, 19(4):337–349, 1991.

[56] R. Karunamuni and T. Alberts. A generalized reflection method of boundary correction in kernel density estimation. *Canad. J. Statist.*, 33:497–509, 2005b.

[57] R. Karunamuni and S. Zhang. Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters*, 78:497–507, 2008.

[58] J. Koláček. *Kernel Estimation of the Regression Function (in Czech)*. PhD thesis, Masaryk University, Brno, feb 2005.

[59] J. Koláček and J. Zelinka. MATLAB toolbox, 2012.

[60] J. S. Marron and D. Nolan. Canonical kernels for density-estimation. *Statistics & Probability Letters*, 7(3):195–199, 1988.

[61] J. S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(4):653–671, 1994.

[62] H.-G. Müller. *Nonparametric regression analysis of longitudinal data*. Springer, New York, 1988.

[63] H.-G. Müller. Smooth optimum kernel estimators near endpoints. *Biometrika*, 78(3):521–530, 1991.

[64] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

[65] B. Park and J. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

[66] M. B. Priestley and M. T. Chao. Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):385–392, 1972.

[67] J. Rice. Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12(4):1215–1230, 1984.

[68] S. Sain, K. Baggerly, and D. Scott. Cross-validation of multivariate densities. *Journal of the American Statistical Association*, 89(427):807–817, 1994.

[69] E. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory end Methods*, 14(5):1123–1136, 1985.

[70] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. Wiley, 1992.

[71] D. W. Scott and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.

[72] N. Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Wiley and SAS Business Series. Wiley, 2006.

[73] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47:1–52, 1985.

[74] B. W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[75] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, New York, 1996.

[76] C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

[77] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 36(2):111–147, 1974.

[78] L. Thomas. *Consumer credit models: pricing, profit, and portfolios*. Oxford University Press, 2009.

[79] M. Wand and M. Jones. *Kernel smoothing*. Chapman and Hall, London, 1995.

[80] G. S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Reprints of articles

Computational Statistics and Data Analysis 57 (2013) 364–376

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Full bandwidth matrix selectors for gradient kernel density estimate

Ivana Horová [a,*], Jan Koláček [a], Kamila Vopatová [b]

[a] Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic
[b] Department of Econometrics, University of Defence, Brno, Czech Republic

ARTICLE INFO

Article history:
Received 4 July 2011
Received in revised form 2 July 2012
Accepted 5 July 2012
Available online 10 July 2012

Keywords:
Asymptotic mean integrated square error
Multivariate kernel density
Unconstrained bandwidth matrix

ABSTRACT

The most important factor in multivariate kernel density estimation is a choice of a bandwidth matrix. This choice is particularly important, because of its role in controlling both the amount and the direction of multivariate smoothing. Considerable attention has been paid to constrained parameterization of the bandwidth matrix such as a diagonal matrix or a pre-transformation of the data. A general multivariate kernel density derivative estimator has been investigated. Data-driven selectors of full bandwidth matrices for a density and its gradient are considered. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated squared bias. The analysis of statistical properties shows the rationale of the proposed method. In order to compare this method with cross-validation and plug-in methods the relative rate of convergence is determined. The utility of the method is illustrated through a simulation study and real data applications.

1. Introduction

Kernel density estimates are one of the most popular nonparametric estimates. In a univariate case, these estimates depend on a bandwidth, which is a smoothing parameter controlling smoothness of an estimated curve and a kernel which is considered as a weight function. The choice of the smoothing parameter is a crucial problem in the kernel density estimation. The literature on bandwidth selection is quite extensive, e.g., monographs Wand and Jones (1995), Silverman (1986) and Simonoff (1996), papers Marron and Ruppert (1994), Park and Marron (1990), Scott and Terrell (1987), Jones and Kappenman (1991) and Cao et al. (1994). As far as the kernel estimate of density derivatives is concerned, this problem has received significantly less attention. In paper Härdle et al. (1990), an adaptation of the least squares cross-validation method is proposed for the bandwidth choice in the kernel density derivative estimation. In paper Horová et al. (2002), the automatic procedure of simultaneous choice of the bandwidth, the kernel and its order for kernel density and its derivative estimates was proposed. But this procedure can be only applied in case that the explicit minimum of the Asymptotic Mean Integrated Square Error of the estimate is available. It is known that this minimum exists only for $d = 2$ and the diagonal matrix H. In paper Horová et al. (2012), the basic formula for the corresponding procedure is given.

The need for nonparametric density estimates for recovering structure in multivariate data is greater since a parametric modeling is more difficult than in the univariate case. The extension of the univariate kernel methodology is not without its problems. The most general smoothing parameterization of the kernel estimator in d dimensions requires the specification entries of $d \times d$ positive definite bandwidth matrix. The multivariate kernel density estimator we are going to deal with is a direct extension of the univariate estimator (see, e.g., Wand and Jones (1995)).

Successful approaches to the univariate bandwidth selection can be transferred to the multivariate settings. The least squares cross-validation and plug-in methods in the multivariate case have been developed and widely discussed in papers

* Correspondence to: Department of Mathematics and Statistics, Kotlářská 2, 61137, Brno, Czech Republic. Tel.: +420 549494429; fax: +420 549491421.
E-mail addresses: horova@math.muni.cz (I. Horová), kolacek@math.muni.cz (J. Koláček), 63985@mail.muni.cz (K. Vopatová).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.006

Chacón and Duong (2010), Duong and Hazelton (2005b,a), Sain et al. (1994) and Duong (2004). Some papers (e.g., Horová et al. (2008, 2012) and Vopatová et al. (2010)) have been focused on constrained parameterization of the bandwidth matrix such as a diagonal matrix. It is well-known fact that a visualization is an important component of the nonparametric data analysis. In paper Horová et al. (2012), this effective strategy was used to clarify the process of the bandwidth matrix choice using bivariate functional surfaces. The paper Horová and Vopatová (2011) brings a short communication on a kernel gradient estimator. Tarn Duong's PhD thesis (Duong, 2004) provides a comprehensive survey of bandwidth matrix selection methods for kernel density estimation. Papers Chacón et al. (2011) and Duong et al. (2008) investigated general density derivative estimators, i.e., kernel estimators of multivariate density derivatives using general (or unconstrained) bandwidth matrix selectors. They defined the kernel estimator of the multivariate density derivative and provided results for the Mean Integrated Square Error convergence asymptotically and for finite samples. Moreover, the relationship between the convergence rate and the bandwidth matrix has been established here. They also developed estimates for the class of normal mixture densities.

The paper is organized as follows: In Section 2 we describe kernel estimates of a density and its gradient and give a form of the Mean Integrated Square Error and the exact MISE calculation for a d-variate normal kernel as well. The next sections are devoted to a data-driven bandwidth matrix selection method. This method is based on an optimally balanced relation between the integrated variance and the integrated squared bias, see Horová and Zelinka (2007a). Similar ideas were applied to kernel estimates of hazard functions (see Horová et al. (2006) or Horová and Zelinka (2007b)). It seems that the basic idea can be also extended to a kernel regression and we are going to investigate this possibility. We discuss the statistical properties and relative rates of convergence of the proposed method as well. Section 5 brings a simulation study and in the last section the developed theory is applied to real data sets.

2. Estimates of a density and its gradient

Let a d-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be drawn from a density f. The kernel density estimator $\hat{f}$ at the point $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$\hat{f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \tag{1}$$

where K is a kernel function, which is often taken to be a d-variate symmetric probability function, $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix and $K_{\mathbf{H}}$ is the scaled kernel function

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

with $|\mathbf{H}|$ the determinant of the matrix $\mathbf{H}$.

The kernel estimator of the gradient Df at the point $\mathbf{x} \in \mathbb{R}^d$ is

$$\widehat{Df}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} DK_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \tag{2}$$

where $DK_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2}\mathbf{H}^{-1/2} DK(\mathbf{H}^{-1/2}\mathbf{x})$ and DK is the column vector of the partial derivatives of K.

Since we aim to investigate both density itself and its gradient in a similar way, we introduce the notation

$$\widehat{D^r f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} D^r K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \quad r = 0, 1, \tag{3}$$

where $D^0 f = f$, $D^1 f = Df$.

We make some additional assumptions and notations:

- (A_1) The kernel function K satisfies the moment conditions $\int K(\mathbf{x})d\mathbf{x} = 1$, $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = \mathbf{0}$, $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \beta_2 \mathbf{I}_d$, $\mathbf{I}_d$ is the $d \times d$ identity matrix.
- (A_2) $\mathbf{H} = \mathbf{H}_n$ is a sequence of bandwidth matrices such that $n^{-1/2}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1})^r$, $r = 0, 1$, and entries of $\mathbf{H}$ approach zero ($(\mathbf{H}^{-1})^0$ is considered as equal to 1).
- (A_3) Each partial density derivative of order $r + 2$, $r = 0, 1$, is continuous and square integrable.
- (N_1) $\mathcal{H}$ is a class of $d \times d$ symmetric positive definite matrices.
- (N_2) $V(\rho) = \int_{\mathbb{R}} \rho^2(x)dx$ for any square integrable scalar valued function ρ.
- (N_3) $V(g) = \int_{\mathbb{R}^d} g(\mathbf{x})g^T(\mathbf{x})d\mathbf{x}$ for any square integrable vector valued function g. In the rest of the text, $\int$ stands for $\int_{\mathbb{R}^d}$ unless it is stated otherwise.
- (N_4) $DD^T = D^2$ is a Hessian operator. Expressions like $DD^T = D^2$ involve "multiplications" of differentials in the sense that

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} = \frac{\partial^2}{\partial x_i \partial x_j}.$$

This means that $(D^2)^m$, $m \in \mathbb{N}$, is a matrix of the $2m$-th order partial differential operators.

(N_5) vec$\mathbf{H}$ is a $d^2 \times 1$ vector obtained by stacking columns of $\mathbf{H}$.
(N_6) Let $d^* = d(d+1)/2$, vech$\mathbf{H}$ is $d^* \times 1$ a vector-half obtained from vec$\mathbf{H}$ by eliminating each of the above diagonal entries.
(N_7) The matrix $\mathbf{D}_d$ of size $d^2 \times d^*$ of ones and zeros such that

$$\mathbf{D}_d \text{vech}\mathbf{H} = \text{vec}\mathbf{H}$$

is called the *duplication matrix* of order d.
(N_8) $\mathbf{J}_d$ denotes $d \times d$ matrix of ones.

The quality of the estimate $\widehat{D^r f}$ can be expressed in terms of the Mean Integrated Square Error

$$\text{MISE}_r\{\widehat{D^r f}(\cdot, \mathbf{H})\} = E \int \|\widehat{D^r f}(\mathbf{x}, \mathbf{H}) - D^r f(\mathbf{x})\|^2 d\mathbf{x},$$

with $\|\cdot\|$ standing for the Euclidean norm, i.e., $\|\mathbf{v}\|^2 = \mathbf{v}^T\mathbf{v} = \text{tr}(\mathbf{v}\mathbf{v}^T)$. For the sake of simplicity we write the argument of MISE_r as $\mathbf{H}$. This error function can be also expressed as the standard decomposition

$$\text{MISE}_r(\mathbf{H}) = \text{IV}_r(\mathbf{H}) + \text{ISB}_r(\mathbf{H}),$$

where $\text{IV}_r(\mathbf{H}) = \int \text{Var}\{\widehat{D^r f}(\mathbf{x}, \mathbf{H})\} d\mathbf{x}$ is the integrated variance and

$$\begin{aligned}\text{ISB}_r(\mathbf{H}) &= \int \|E\widehat{D^r f}(\mathbf{x}, \mathbf{H}) - D^r f(\mathbf{x})\|^2 d\mathbf{x} \\ &= \int \left\| \int K(\mathbf{z}) D^r f(\mathbf{x} - \mathbf{H}^{1/2}\mathbf{z}) d\mathbf{z} - D^r f(\mathbf{x}) \right\|^2 d\mathbf{x} \\ &= \int \|(K_{\mathbf{H}} * D^r f)(\mathbf{x}) - D^r f(\mathbf{x})\|^2 d\mathbf{x}\end{aligned}$$

is the integrated square bias (the symbol $*$ denotes convolution).

Since MISE_r is not mathematically tractable, we employ the Asymptotic Mean Integrated Square Error. The AMISE_r theorem has been proved (e.g., in Duong et al. (2008)) and reads as follows:

Theorem 1. *Let assumptions (A_1) –(A_3) be satisfied. Then*

$$\text{MISE}_r(\mathbf{H}) \simeq \text{AMISE}_r(\mathbf{H}),$$

where

$$\text{AMISE}_r(\mathbf{H}) = \underbrace{n^{-1}|\mathbf{H}|^{-1/2}\text{tr}\left\{(\mathbf{H}^{-1})^r V(D^r K)\right\}}_{\text{AIV}_r} + \underbrace{\frac{\beta_2^2}{4}\text{vech}^T\mathbf{H}\boldsymbol{\Psi}_{4+2r}\text{vech}\mathbf{H}}_{\text{AISB}_r}. \tag{4}$$

The term $\boldsymbol{\Psi}_{4+2r}$ involves higher order derivatives of f and its subscript $4+2r$, $r = 0, 1$, indicates the order of derivatives used. It is a $d^* \times d^*$ symmetric matrix.

It can be shown that

$$\int \|\{\text{tr}(\mathbf{H}D^2)D^r\} f(\mathbf{x})\|^2 d\mathbf{x} = \text{vech}^T\mathbf{H}\,\boldsymbol{\Psi}_{4+2r}\text{vech}\mathbf{H}.$$

Then (4) can be rewritten as

$$\text{AMISE}_r(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}\text{tr}(\mathbf{H}^{-1})^r V(D^r K) + \frac{\beta_2^2}{4}\int \|\{\text{tr}(\mathbf{H}D^2)D^r\} f(\mathbf{x})\|^2 d\mathbf{x}, \quad r = 0, 1. \tag{5}$$

Let $K = \phi_{\mathbf{I}}$ be the d-variate normal kernel and suppose that f is the normal mixture density $f(\mathbf{x}) = \sum_{l=1}^k w_l \phi_{\boldsymbol{\Sigma}_l}(\mathbf{x} - \boldsymbol{\mu}_l)$, where for each $l = 1, \ldots, k$, $\phi_{\boldsymbol{\Sigma}_l}$ is the d-variate $N(\mathbf{0}, \boldsymbol{\Sigma}_l)$ normal density and $\mathbf{w} = (w_1, \ldots, w_k)^T$ is a vector of positive numbers summing to one.

In this case, the exact formula for MISE_r was derived in Chacón et al. (2011). For $r = 0, 1$ it takes the form

$$\text{MISE}_r(\mathbf{H}) = 2^{-r}n^{-1}(4\pi)^{-d/2}|\mathbf{H}|^{-1/2}(\text{tr}\mathbf{H}^{-1})^r + \mathbf{w}^T\left\{(1-n^{-1})\boldsymbol{\Omega}_2 - 2\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_0\right\}\mathbf{w}, \tag{6}$$

where

$$(\boldsymbol{\Omega}_c)_{ij} = (-1)^r \phi_{c\mathbf{H}+\boldsymbol{\Sigma}_{ij}}(\boldsymbol{\mu}_{ij})\left\{\boldsymbol{\mu}_{ij}^T(c\mathbf{H} + \boldsymbol{\Sigma}_{ij})^{-2}\boldsymbol{\mu}_{ij} - 2\text{tr}(c\mathbf{H} + \boldsymbol{\Sigma}_{ij})^{-1}\right\}^r$$

with $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j$, $\boldsymbol{\mu}_{ij} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$.

3. Bandwidth matrix selection

The most important factor in multivariate kernel density estimates is the bandwidth matrix $\mathbf{H}$. Because of its role in controlling both the amount and the direction of smoothing this choice is particularly important.

Let $\mathbf{H}_{(A)MISE,r}$ stand for a bandwidth matrix minimizing (A)MISE$_r$, i.e.,

$$\mathbf{H}_{MISE,r} = \arg\min_{\mathbf{H}\in\mathcal{H}} \mathrm{MISE}_r(\mathbf{H})$$

and

$$\mathbf{H}_{AMISE,r} = \arg\min_{\mathbf{H}\in\mathcal{H}} \mathrm{AMISE}_r(\mathbf{H}).$$

As it has been mentioned in former works (see, e.g., Duong and Hazelton (2005a,b)), the discrepancy between $\mathbf{H}_{MISE,r}$ and $\mathbf{H}_{AMISE,r}$ is asymptotically negligible in comparison with the random variation in the bandwidth matrix selectors that we consider. The problems of estimating $\mathbf{H}_{MISE,r}$ and $\mathbf{H}_{AMISE,r}$ are equivalent for most practical purposes.

If we denote $D_{\mathbf{H}} = \frac{\partial}{\partial \mathrm{vech}\mathbf{H}}$, then using matrix differential calculus yields

$$D_{\mathbf{H}}\mathrm{AMISE}_r(\mathbf{H}) = -(2n)^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\left\{(\mathbf{H}^{-1})^r V(D^r K)\right\}\mathbf{D}_d^T\mathrm{vec}\mathbf{H}^{-1} + n^{-1}|\mathbf{H}|^{-1/2} r\left(-\mathbf{D}_d^T\mathrm{vec}(\mathbf{H}^{-1}V(D^r K)\mathbf{H}^{-1})\right) + \frac{\beta_2^2}{2}\boldsymbol{\Psi}_{4+2r}\mathrm{vech}\mathbf{H}.$$

Unfortunately, there is no explicit solution for the equation

$$D_{\mathbf{H}}\mathrm{AMISE}_r(\mathbf{H}) = \mathbf{0} \tag{7}$$

(with an exception of $d = 2$, $r = 0$ and a diagonal bandwidth matrix $\mathbf{H}$, see, e.g., Wand and Jones (1995)). But nevertheless the following lemma holds.

Lemma 2.

$$\mathrm{AIV}_r(\mathbf{H}_{AMISE,r}) = \frac{4}{d+2r}\mathrm{AISB}_r(\mathbf{H}_{AMISE,r}). \tag{8}$$

Proof. See Complements for the proof. □

It can be shown (Chacón et al., 2011) that

$$\mathbf{H}_{AMISE,r} = \mathbf{C}_{0,r}n^{-2/(d+2r+4)} = O(n^{-2/(d+2r+4)}\mathbf{J}_d)$$

and then AMISE$_r(\mathbf{H}_{AMISE,r})$ is of order $n^{-4/(d+2r+4)}$.

Since $\mathbf{H}_{AMISE,r}$ resp. $\mathbf{H}_{MISE,r}$ cannot be found in practice, the data-driven methods for selection of $\mathbf{H}$ have been proposed in papers Chacón and Duong (2010), Duong (2004), Duong and Hazelton (2005b), Sain et al. (1994) and Wand and Jones (1994) etc.. The performance of bandwidth matrix selectors can be assessed by its relative rate of convergence. We generalize the definition for the relative rate of convergence for the univariate case to the multivariate one.

Let $\widehat{\mathbf{H}}_r$ be a data-driven bandwidth matrix selector. We say that $\widehat{\mathbf{H}}_r$ converges to $\mathbf{H}_{AMISE,r}$ with relative rate $n^{-\alpha}$ if

$$\mathrm{vech}(\widehat{\mathbf{H}}_r - \mathbf{H}_{AMISE,r}) = O_p(\mathbf{J}_{d^*}n^{-\alpha})\mathrm{vech}\mathbf{H}_{AMISE,r}. \tag{9}$$

This definition was introduced by Duong (2004).

Now, we remind cross-validation methods $CV_r(\mathbf{H})$ (Duong and Hazelton, 2005b; Chacón and Duong, 2012) which aim to estimate MISE$_r$. $CV_r(\mathbf{H})$ is an unbiased estimate of MISE$_r(\mathbf{H}) - \mathrm{tr}V(D^r f)$ and

$$CV_r(\mathbf{H}) = (-1)^r\mathrm{tr}\left\{\frac{1}{n^2}\sum_{i,j=1}^{n} D^{2r}(K_{\mathbf{H}} * K_{\mathbf{H}})(\mathbf{X}_i - \mathbf{X}_j) - \frac{2}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n} D^{2r}K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j)\right\}, \tag{10}$$

$$\widehat{\mathbf{H}}_{CV_r} = \arg\min_{\mathbf{H}\in\mathcal{H}} CV_r(\mathbf{H}).$$

It can be shown that the relative rate of convergence to $\mathbf{H}_{MISE,r}$ is $n^{-d/(2d+4r+8)}$ (Chacón and Duong, 2012) and to $\mathbf{H}_{AMISE,r}$ is $n^{-\min\{d,4\}/(2d+4r+8)}$ (see Duong and Hazelton (2005b) for $r = 0$).

Plug-in methods for the bandwidth matrix selection were generalized to the multivariate case in Wand and Jones (1994). The idea consists of estimating the unknown matrix $\boldsymbol{\Psi}_{4+2r}$. The relative rate of convergence to $\mathbf{H}_{MISE,r}$ and $\mathbf{H}_{AMISE,r}$ is the same $n^{-2/(d+2r+6)}$ when $d \geq 2$ (see, e.g., Chacón (2010) and Chacón and Duong (2012)).

In papers Horová et al. (2008, 2012), a special method for bandwidth matrix selection for a bivariate density for the case of diagonal bandwidth matrix has been developed and the rationale of this method has been explained. This method is based on formula (8). As concerns the bandwidth matrix selection for the kernel gradient estimator, the aforementioned method was extended to this case in Vopatová et al. (2010) and Horová and Vopatová (2011). Because the problem of the bandwidth matrix choice both for density itself and its gradient are closely related one to each other, we address the problem of these choices together.

4. Proposed method and its statistical properties

As mentioned, our method is based on Eq. (8) in the sense that a solution of $D_{\mathbf{H}}\mathrm{AMISE}_r(\mathbf{H}) = \mathbf{0}$ is equivalent to solving Eq. (8). But $\mathrm{AISB}_r(\mathbf{H})$ depends on the unknown density. Thus we adapt the similar idea as in the univariate case (Horová and Zelinka (2007a)) and use a suitable estimate of AISB_r.

Eq. (8) can be rewritten as

$$(d+2r)n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\left\{(\mathbf{H}^{-1})^r V(D^r K)\right\} - \beta_2^2 \int \|\{\mathrm{tr}(\mathbf{H}D^2)D^r\}f(\mathbf{x})\|^2 d\mathbf{x} = 0. \tag{11}$$

Let us denote

$$\Lambda(\mathbf{z}) = (K*K*K*K - 2K*K*K + K*K)(\mathbf{z}),$$
$$\Lambda_{\mathbf{H}}(\mathbf{z}) = |\mathbf{H}|^{-1/2}\Lambda(\mathbf{H}^{-1/2}\mathbf{z}).$$

Then the estimate of $\mathrm{AISB}_r(\mathbf{H})$ can be considered as

$$\overline{\mathrm{AISB}}_r(\mathbf{H}) = \int \|(K_{\mathbf{H}} * \widehat{D^r f})(\mathbf{x},\mathbf{H}) - \widehat{D^r f}(\mathbf{x},\mathbf{H})\|^2 d\mathbf{x}.$$

This estimate involves non-stochastic terms, therefore, according to Taylor (1989), Jones and Kappenman (1991) and Jones et al. (1991), we eliminated these terms and propose an (asymptotically unbiased) estimate

$$\widehat{\mathrm{AISB}}_r(\mathbf{H}) = \mathrm{tr}\left\{\frac{(-1)^r}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j)\right\}.$$

Now, instead of Eq. (11) we aim to solve the equation

$$(d+2r)n^{-1}|\mathbf{H}|^{-1/2}\mathrm{tr}\left\{(\mathbf{H}^{-1})^r V(D^r K)\right\} - 4\mathrm{tr}\left\{\frac{(-1)^r}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n} D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_i - \mathbf{X}_j)\right\} = 0. \tag{12}$$

Remark 1. The bandwidth matrix selection method based on Eq. (12) is called the *Iterative method* (IT method) and the bandwidth estimate is denoted $\widehat{\mathbf{H}}_{\mathrm{IT}_r}$.

Remark 2. In the following we assume that K is the standard normal density $\phi_{\mathbf{I}}$. Thus $\Lambda(\mathbf{z}) = \phi_{4\mathbf{I}}(\mathbf{z}) - 2\phi_{3\mathbf{I}}(\mathbf{z}) + \phi_{2\mathbf{I}}(\mathbf{z})$ and $\beta_2 = 1$. We are going to discuss statistical properties of the Iterative method which will show its rationality.

Let $\Gamma_r(\mathbf{H})$ stand for the left hand side of (11) and $\widehat{\Gamma}_r(\mathbf{H})$ for the left hand side of (12).

Theorem 3. *Let the assumptions* (A_1) *–*(A_3) *be satisfied and* $K = \phi_{\mathbf{I}}$. *Then*

$$E(\widehat{\Gamma}_r(\mathbf{H})) = \Gamma_r(\mathbf{H}) + o(\|\mathrm{vec}\mathbf{H}\|^{5/2}),$$
$$\mathrm{Var}(\widehat{\Gamma}_r(\mathbf{H})) = 32n^{-2}|\mathbf{H}|^{-1/2}\|\mathrm{vec}\mathbf{H}\|^{-2r}V(\mathrm{vec}D^{2r}\Lambda)V(f) + o(n^{-2}|\mathbf{H}|^{-1/2}\|\mathrm{vec}\mathbf{H}\|^{-2r}).$$

Proof. For the proof see Complements. □

As far as the convergence rate of the IT method is concerned, we are inspired with AMSE lemma (Duong, 2004; Duong and Hazelton, 2005a). The following theorem takes place.

Theorem 4. *Let the assumptions* (A_1) *–*(A_3) *be satisfied and* $K = \phi_{\mathbf{I}}$. *Then*

$$\mathrm{MSE}\{\mathrm{vech}\widehat{\mathbf{H}}_{\mathrm{IT}_r}\} = O\left(n^{-\min\{d,4\}/(d+2r+4)}\mathbf{J}_{d^*}\right) \times \mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE},r}.$$

Proof. Proof of theorem can be found in Complements. □

Corollary 5. *The convergence rate to* $\mathbf{H}_{\mathrm{AMISE},r}$ *is* $n^{-\min\{d,4\}/(2d+4r+8)}$ *for the IT method.*

Remark 3. For the r-th derivative the cross-validation method is of order $n^{-\min\{d,4\}/(2d+4r+8)}$ and the plug-in method is of order $n^{-2/(d+2r+6)}$ (with respect to $\mathbf{H}_{\mathrm{AMISE},r}$).

5. Computational aspects and simulations

Eq. (12) can be rewritten as

$$|\widehat{\mathbf{H}}_{\mathrm{IT}_r}|^{1/2} 4 \operatorname{tr}\left\{\frac{(-1)^r}{n}\sum_{\substack{i,j=1\\ i\neq j}}^{n} D^{2r} \Lambda_{\widehat{\mathbf{H}}_{\mathrm{IT}_r}}(\mathbf{X}_i - \mathbf{X}_j)\right\} = (d+2r)\operatorname{tr}\left\{(\widehat{\mathbf{H}}_{\mathrm{IT}_r}^{-1})^r V\left(D^r K\right)\right\}.$$

This equation represents a nonlinear equation for d^* unknown entries of $\widehat{\mathbf{H}}_{\mathrm{IT}_r}$. In order to find all these entries we need additional $d^* - 1$ equations. Below, we describe a possibility of obtaining these equations.

We adopt a similar idea as in the case of the diagonal matrix (see also Terrell (1990), Scott (1992), Duong et al. (2008) and Horová and Vopatová (2011)). We explain this approach for the case $d = 2$ with the matrix

$$\widehat{\mathbf{H}}_{\mathrm{IT}_r} = \begin{pmatrix} \hat{h}_{11,r} & \hat{h}_{12,r} \\ \hat{h}_{12,r} & \hat{h}_{22,r} \end{pmatrix}.$$

Let $\widehat{\boldsymbol{\Sigma}}$ be a sample covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_{11}^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_{22}^2 \end{pmatrix}.$$

The initial estimates of entries of $\widehat{\mathbf{H}}_{\mathrm{IT}_r}$ can be chosen as

$$\hat{h}_{11,r} = \hat{h}_{1,r}^2 = (\hat{\sigma}_{11}^2)^{(12+r)/12} n^{(r-4)/12},$$
$$\hat{h}_{22,r} = \hat{h}_{2,r}^2 = (\hat{\sigma}_{22}^2)^{(12+r)/12} n^{(r-4)/12},$$
$$\hat{h}_{12,r} = \operatorname{sign}\hat{\sigma}_{12}|\hat{\sigma}_{12}|^{(12+r)/12} n^{(r-4)/12}.$$

For details see Horová and Vopatová (2011).

Hence

$$\hat{h}_{22,r} = \left(\frac{\hat{\sigma}_{22}^2}{\hat{\sigma}_{11}^2}\right)^{(12+r)/12} \hat{h}_{11,r}, \tag{13}$$

$$\hat{h}_{12,r}^2 = \left(\frac{\hat{\sigma}_{12}^2}{\hat{\sigma}_{11}^2}\right)^{(12+r)/12} \hat{h}_{11,r} \tag{14}$$

and further

$$|\widehat{\mathbf{H}}_{\mathrm{IT}_r}| = \hat{h}_{11,r}^2 \left((\hat{\sigma}_{11}\hat{\sigma}_{22})^{(12+r)/6} - \hat{\sigma}_{12}^{(12+r)/6}\right) \Big/ \hat{\sigma}_{11}^{(12+r)/3}$$
$$= \hat{h}_{11,r}^2 S(\hat{\sigma}_{ij}).$$

Thus we arrive at the equation for the unknown $\hat{h}_{11,r}$

$$4\hat{h}_{11,r}\sqrt{S(\hat{\sigma}_{ij})}\operatorname{tr}\left\{\frac{(-1)^r}{n}\sum_{\substack{i,j=1\\ i\neq j}}^{n} D^{2r} \Lambda_{\widehat{\mathbf{H}}_{\mathrm{IT}_r}}(\mathbf{X}_i - \mathbf{X}_j)\right\} = (d+2r)\operatorname{tr}\left\{(\widehat{\mathbf{H}}_{\mathrm{IT}_r}^{-1})^r V\left(D^r K\right)\right\}. \tag{15}$$

This approach is very important for computational aspects of solving Eq. (12). Putting Eqs. (13)–(15) forms one nonlinear equation for the unknown $\hat{h}_{11,r}$ and it can be solved by means of an appropriate iterative numerical method. This procedure gives the name of the proposed method. Evidently, this approach is computationally much faster than a general minimization process.

To test the effectiveness of our estimator, we simulated its performance against the least squares cross-validation method. All simulations and computations were done in MATLAB. The simulation is based on 100 replications of 6 bivariate normal mixture densities, labeled *A*–*F*. Means and covariance matrices of these distributions were generated randomly. Table 1 brings the list of the normal mixture densities. Densities *A* and *B* are unimodal, *C* and *D* are bimodal and *E* and *F* are trimodal. Their contour plots are displayed in Fig. 1.

The sample size of $n = 100$ was used in all replications. We calculated the Integrated Square Error (ISE)

$$\mathrm{ISE}_r\{\widehat{D^r f}(\cdot, \mathbf{H})\} = \int \|\widehat{D^r f}(\mathbf{x}, \mathbf{H}) - D^r f(\mathbf{x})\|^2 d\mathbf{x}$$

for each estimated density and its derivative over all 100 replications. The logarithm of results is displayed in Tables 2 and 3 and in Fig. 2. Here "ITER" denotes the results for our proposed method, "LSCV" stands for the results of the Least Squares Cross-validation method (10) and "MISE" is a tag for the results obtained by minimizing (6).

Finally, we compared computational times of all methods. Results are listed in Table 4.

Table 1
Normal mixture densities.

Density	Formula $N(\mathrm{vec}^T \boldsymbol{\mu}, \mathrm{vec}^T \boldsymbol{\Sigma})$
A	$N\Big((-0.2686, -1.7905), (7.9294, -10.0673; -10.0673, 22.1150)\Big)$
B	$N\Big((-0.6847, 2.6963), (16.9022, 9.8173; 9.8173, 6.0090)\Big)$
C	$\frac{1}{2}N\Big((0.3151, -1.6877), (0.1783, -0.1821; -0.1821, 1.0116)\Big) + \frac{1}{2}N\Big((1.1768, 0.3731), (0.2414, -0.8834; -0.8834, 4.2934)\Big)$
D	$\frac{1}{2}N\Big((1.8569, 0.1897), (1.5023, -0.9259; -0.9259, 0.8553)\Big) + \frac{1}{2}N\Big((0.3349, -0.2397), (2.3050, 0.8895; 0.8895, 1.2977)\Big)$
E	$\frac{1}{3}N\Big((0.0564, -0.9041), (0.9648, -0.8582; -0.8582, 0.9332)\Big) + \frac{1}{3}N\Big((-0.7769, 1.6001), (2.8197, -1.4269; -1.4269, 0.9398)\Big)$ $+ \frac{1}{3}N\Big((1.0132, 0.4508), (3.9982, -3.7291; -3.7291, 5.5409)\Big)$
F	$\frac{1}{3}N\Big((2.2337, -2.9718), (0.6336, -0.9279; -0.9279, 3.1289)\Big) + \frac{1}{3}N\Big((-4.3854, 0.5678), (2.1399, -0.6208; -0.6208, 0.7967)\Big)$ $+ \frac{1}{3}N\Big((1.5513, 2.2186), (1.1207, 0.8044; 0.8044, 1.0428)\Big)$

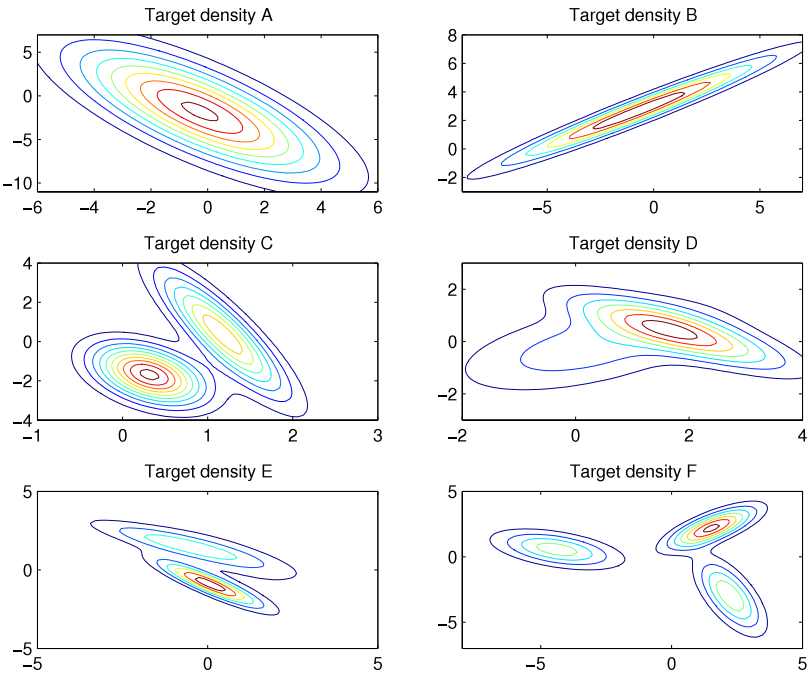


Fig. 1. Contour plots for target densities.

Table 2
Logarithm of ISE_0 for bandwidth matrices.

Target density		*A*	*B*	*C*	*D*	*E*	*F*
ITER	Mean	−7.562	−6.345	−4.319	−4.918	−4.779	−5.103
	Std	0.459	0.448	0.264	0.274	0.203	0.180
LSCV	Mean	−7.110	−5.781	−4.332	−4.957	−4.917	−5.138
	Std	0.531	0.610	0.407	0.518	0.385	0.325
MISE	Mean	−7.865	−4.256	−4.168	−3.521	−2.763	−3.903
	Std	0.397	0.418	0.188	0.340	0.237	0.164

6. Application to real data

An important question arising in application to real data is which observed features – such as a local extremes – are really there. Chaudhuri and Marron (1999) introduced the SiZer (Significant Zero) method for finding structure in smooth data. Duong et al. (2008) proposed a framework for feature significance in *d*-dimensional data which combines kernel density derivative estimators and hypothesis tests for modal regions. Distributional properties are given for the gradient and curvature estimators, and pointwise tests extend the two-dimensional feature significance ideas of Godtliebsen et al. (2002).

Table 3
Logarithm of ISE_1 for bandwidth matrices.

Target density		A	B	C	D	E	F
ITER	Mean	−7.618	−4.005	−0.888	−2.698	−1.991	−3.203
	Std	0.289	0.405	0.055	0.099	0.030	0.032
LSCV	Mean	−5.364	−0.210	0.503	−1.214	−0.501	−1.544
	Std	2.638	2.960	2.364	2.437	2.373	1.914
MISE	Mean	−7.939	−4.314	−1.813	−3.544	−2.732	−3.864
	Std	0.391	0.443	0.311	0.359	0.241	0.172

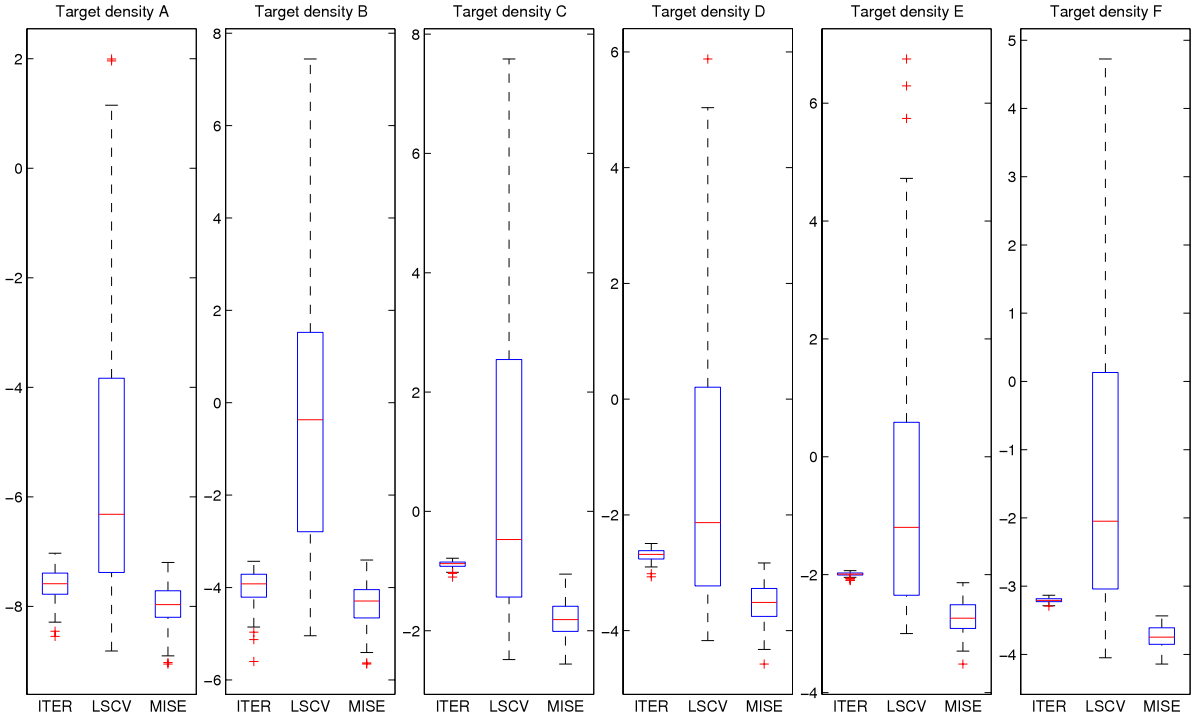


Fig. 2. Box plots for log(ISE).

Table 4
Average computational times (in seconds).

Target density	r	A	B	C	D	E	F
ITER	0	0.0826	0.0685	0.0596	0.0801	0.0754	0.0591
	1	0.8295	0.8201	0.8542	0.8605	0.8538	0.8786
LSCV	0	0.5486	0.5732	0.5182	0.4844	0.5004	0.5004
	1	1.7936	1.6483	1.3113	1.3128	1.6495	1.5581
MISE	0	0.1927	0.1982	0.7126	0.5540	1.8881	2.4000
	1	0.5236	0.3112	1.2653	1.3452	2.3089	4.1172

We started with the well-known 'Old Faithful' data set (Simonoff, 1996), which contains characteristics of 222 eruptions of the 'Old Faithful Geyser' in Yellowstone National Park, USA, during August 1978 and August 1979. Kernel density and first derivative estimates using the standard normal kernel based on the following bandwidth matrices obtained by the IT method

$$\widehat{\mathbf{H}}_{\mathrm{IT}_0} = \begin{pmatrix} 0.0703 & 0.7281 \\ 0.7281 & 9.801 \end{pmatrix}, \qquad \widehat{\mathbf{H}}_{\mathrm{IT}_1} = \begin{pmatrix} 0.2388 & 3.006 \\ 3.006 & 50.24 \end{pmatrix}$$

are displayed in Fig. 3. The intersections of $\partial f/\partial x_1 = 0$ and $\partial f/\partial x_2 = 0$ show the existence of extremes.

The second data set is taken from UNICEF—"The State of the World's Children 2003". It contains 72 pairs of observations for countries with a GNI less than 1000 US dollars per capita in 2001. X_1 variable describes the under-five mortality rate, i.e., the probability of dying between birth and exactly five years of age expressed per 1000 live births, and X_2 is a life expectancy at birth, i.e., the number of years newborn children would live if subject to the mortality risks prevailing for

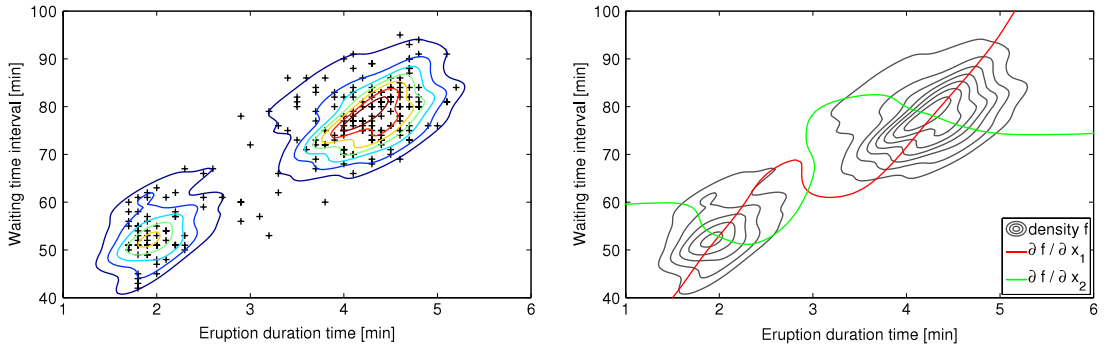


Fig. 3. 'Old Faithful' data contour plots—estimated density $\hat{f}$ (left) and estimated partial derivatives $\widehat{\partial f}/\partial x_1 = 0$, $\widehat{\partial f}/\partial x_2 = 0$ (right).

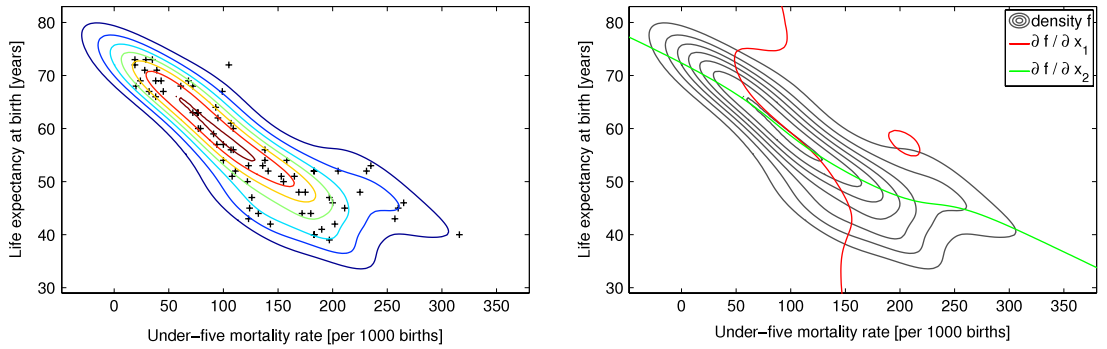


Fig. 4. 'UNICEF Children' data contour plots—estimated density $\hat{f}$ (left) and estimated partial derivatives $\widehat{\partial f}/\partial x_1 = 0$, $\widehat{\partial f}/\partial x_2 = 0$ (right).

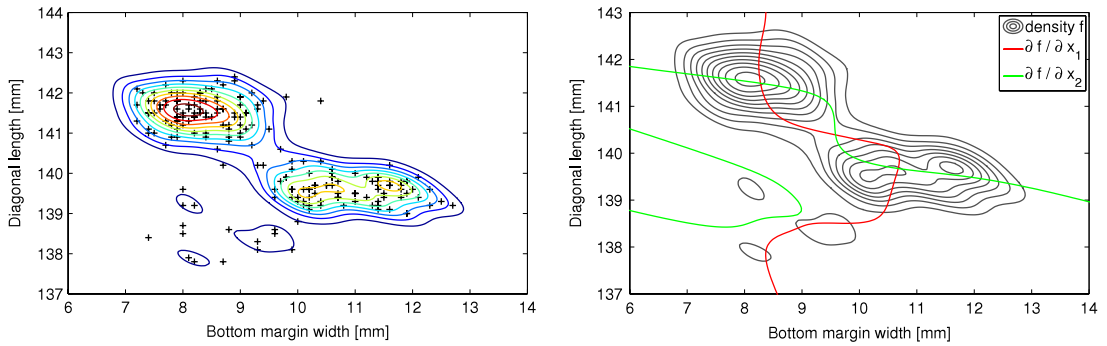


Fig. 5. Swiss bank notes data contour plots—estimated density $\hat{f}$ (left) and estimated partial derivatives $\widehat{\partial f}/\partial x_1 = 0$, $\widehat{\partial f}/\partial x_2 = 0$ (right).

the cross-section of population at the time of their birth (UNICEF, 2003). These data have also been analyzed in Duong and Hazelton (2005b).

Bandwidth matrices for the estimated density $\hat{f}$ and its gradient $\widehat{Df}$ are

$$\widehat{\mathbf{H}}_{\mathrm{IT}_0} = \begin{pmatrix} 1112.0 & -138.3 \\ -138.3 & 24.20 \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{H}}_{\mathrm{IT}_1} = \begin{pmatrix} 2426 & -253.7 \\ -253.7 & 38.38 \end{pmatrix},$$

respectively. Fig. 4 illustrates the use of the iterative bandwidth matrices for the 'UNICEF Children' data set.

We also analyzed a Swiss bank notes data set from Simonoff (1996). It contains measurements of the bottom margin and diagonal length of 100 real Swiss bank notes and 100 forged Swiss bank notes. Contour plots in Fig. 5 represent kernel estimates of the joint distribution of the bottom margin and diagonal length of the bills using bandwidth matrices

$$\widehat{\mathbf{H}}_{\mathrm{IT}_0} = \begin{pmatrix} 0.1227 & -0.0610 \\ -0.0610 & 0.0781 \end{pmatrix}, \qquad \widehat{\mathbf{H}}_{\mathrm{IT}_1} = \begin{pmatrix} 0.6740 & -0.3159 \\ -0.3159 & 0.4129 \end{pmatrix}.$$

The bills with longer diagonal and shorter bottom margin correspond to real bills.

The density estimate shows a bimodal structure for the forged bills (bottom right part of the plot) and it seems that the gradient estimate does not match this structure. The elements of the bandwidth matrix for the gradient estimate are bigger

in magnitude than the ones of the bandwidth matrix for density estimate, as expected from the theory. Three bumps in the tails are too small and the gradient estimator is not able to distinguish them.

7. Conclusion

We restricted ourselves on the use of the standard normal kernel. This kernel satisfies smoothness conditions and provides easy computations of convolutions. Due to these facts it was possible to compare the IT method with the LSCV method.

The simulation study and application to real data show that the IT method provides a sufficiently reliable way of estimating arbitrary density and its gradient. The IT method is also easy implementable and seems to be less time consuming (see Horová and Zelinka (2007a) for $d = 1$, see also Table 4 for $d = 2$).

Further assessment of the practical performance and an extension to a curvature density estimate would be very important further research. Although the theoretical comparison also involves PI methods, they are not included in the simulation study. This would be an interesting task for further research.

8. Complements

We start with introducing some facts on matrix differential calculus and on the Gaussian density (see Magnus and Neudecker (1979, 1999) and Aldershof et al. (1995)).

Let $\mathbf{A}$, $\mathbf{B}$ be $d \times d$ matrices and $r = 0, 1$:

1°. $\mathrm{tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{vec}^T\mathbf{A}\mathrm{vec}\mathbf{B}$

2°. $D_{\mathbf{H}}|\mathbf{H}|^{-1/2} = -\frac{1}{2}|\mathbf{H}|^{-1/2}\mathbf{D}_d^T\mathrm{vec}\mathbf{H}^{-1}$

3°. $D_{\mathbf{H}}\mathrm{tr}(\mathbf{H}^{-1}\mathbf{A}) = -\mathbf{D}_d^T\mathrm{vec}(\mathbf{H}^{-1}\mathbf{A}\mathbf{H}^{-1})$

4°. $\int \phi_{c\mathbf{I}}(\mathbf{z})\{\mathrm{tr}(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = c\{\mathrm{tr}(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})$

$\int \phi_{c\mathbf{I}}(\mathbf{z})\{\mathrm{tr}^2(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = 3c^2\{\mathrm{tr}^2(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})$

$\int \phi_{c\mathbf{I}}(\mathbf{z})\{\mathrm{tr}^k(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)\mathrm{tr}(\mathbf{H}^{1/2}D\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = \mathbf{0}$, $k \in \mathbb{N}_0$

5°. $\Lambda(\mathbf{z}) = \phi_{4\mathbf{I}}(\mathbf{z}) - 2\phi_{3\mathbf{I}}(\mathbf{z}) + \phi_{2\mathbf{I}}(\mathbf{z})$,

then using 4° yields

$\int \Lambda(\mathbf{z})d\mathbf{z} = 0$

$\int \Lambda(\mathbf{z})\{\mathrm{tr}(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = \mathbf{0}$

$\int \Lambda(\mathbf{z})\{\mathrm{tr}^2(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = 6\{\mathrm{tr}^2(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})$

$\int \Lambda(\mathbf{z})\{\mathrm{tr}^k(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)\mathrm{tr}(\mathbf{H}^{1/2}D\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z} = \mathbf{0}$, $k \in \mathbb{N}_0$

6°. $\int D^k f(\mathbf{x})[D^k f(\mathbf{x})]^T d\mathbf{x} = (-1)^k \int D^{2k} f(\mathbf{x})f(\mathbf{x})d\mathbf{x}$, $k \in \mathbb{N}$

7°. Taylor expansion in the form (for $r = 0,\ 1$)

$$D^{2r}f(\mathbf{x} - \mathbf{H}^{1/2}\mathbf{z}) = D^{2r}f(\mathbf{x}) - \{\mathbf{z}^T\mathbf{H}^{1/2}DD^{2r}\}f(\mathbf{x}) + \frac{1}{2!}\{(\mathbf{z}^T\mathbf{H}^{1/2}D)^2D^{2r}\}f(\mathbf{x}) + \cdots + \frac{(-1)^k}{k!}\{(\mathbf{z}^T\mathbf{H}^{1/2}D)^kD^{2r}\}f(\mathbf{x}) + o(\|\mathbf{H}^{1/2}\mathbf{z}\|^k\mathbf{J}_{d^r}).$$

Sketch of the proof of Lemma 2:

Proof. Consider Eq. (7) and multiply it from the left by $\frac{1}{2}\mathrm{vech}^T\mathbf{H}$.

Then

$$(4n)^{-1}|\mathbf{H}|^{-1/2}\mathrm{vech}^T\mathbf{H}\mathrm{tr}\left\{(\mathbf{H}^{-1})^rV(D^rK)\right\}\mathbf{D}_d^T\mathrm{vec}\mathbf{H}^{-1} + (2n)^{-1}|\mathbf{H}|^{-1/2}r\mathrm{vech}^T\mathbf{H}\left(\mathbf{D}_d^T\mathrm{vec}(\mathbf{H}^{-1}V(D^rK))\mathbf{H}^{-1}\right) = \frac{\beta_2^2}{4}\mathrm{vech}^T\mathbf{H}\boldsymbol{\Psi}_{4+2r}\mathrm{vech}\mathbf{H}.$$

The right hand side of this equation is AISB_r. Further, if we use the facts on matrix calculus, we arrive at formula (8). □

We only present a sketch of proofs of theorems. Detailed proofs are available on request from the first author.

Sketch of the proof of Theorem 3:

Proof. In order to show the validity of the relation for the expected value of $\widehat{\Gamma}_r(\mathbf{H})$, we evaluate $E(\widehat{\mathrm{AISB}}_r(\mathbf{H}))$ and start with

$$\begin{aligned} E\ \mathrm{tr}\left\{D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{X}_2)\right\} &= \mathrm{tr}\iint D^{2r}\Lambda_{\mathbf{H}}(\mathbf{x} - \mathbf{y})f(\mathbf{x})f(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= \mathrm{tr}\iint \Lambda_{\mathbf{H}}(\mathbf{x} - \mathbf{y})f(\mathbf{x})D^{2r}f(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= \mathrm{tr}\iint \Lambda(\mathbf{z})D^{2r}f(\mathbf{x} - \mathbf{H}^{1/2}\mathbf{z})f(\mathbf{x})d\mathbf{z}d\mathbf{x}. \end{aligned}$$

Taylor expansion, defined in 7°, and using 5° yields

$$
\begin{aligned}
&= \operatorname{tr}\iint \Lambda(\mathbf{z})\left(\sum_{i=0}^{5}\frac{(-1)^i}{i!}\{(\mathbf{z}^T\mathbf{H}^{1/2}D)^i D^{2r}\}f(\mathbf{x}) + o(\|\mathbf{H}^{1/2}\mathbf{z}\|^5)\mathbf{J}_{d^r}\right) f(\mathbf{x})d\mathbf{z}d\mathbf{x}\\
&= \operatorname{tr}\iint \Lambda(\mathbf{z})\left(\frac{1}{4!}\{(\mathbf{z}^T\mathbf{H}^{1/2}D)^4 D^{2r}\}f(\mathbf{x}) + o(\|\mathbf{H}^{1/2}\mathbf{z}\|^5)\mathbf{J}_{d^r}\right) f(\mathbf{x})d\mathbf{z}d\mathbf{x}\\
&= \frac{1}{4!}\operatorname{tr}\iint \Lambda(\mathbf{z})\{\operatorname{tr}^2(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})f(\mathbf{x})d\mathbf{z}d\mathbf{x} + o(\|\operatorname{vec}\mathbf{H}\|^{5/2}),
\end{aligned}
$$

using properties 5° and 6° we arrive at

$$
\begin{aligned}
&= \frac{1}{4}\operatorname{tr}\int \{\operatorname{tr}^2(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})f(\mathbf{x})d\mathbf{x} + o(\|\operatorname{vec}\mathbf{H}\|^{5/2})\\
&= \frac{(-1)^r}{4}\int \|\{\operatorname{tr}(\mathbf{H}D^2)D^r\}f(\mathbf{x})\|^2 d\mathbf{x} + o(\|\operatorname{vec}\mathbf{H}\|^{5/2}).
\end{aligned}
$$

To prove the second part of the Theorem it is sufficient to derive $\operatorname{Var}(\widehat{\mathrm{AISB}}_r(\mathbf{H}))$

$$
\operatorname{Var}(\widehat{\mathrm{AISB}}_r(\mathbf{H})) = \operatorname{Var}\left\{\frac{4}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_i-\mathbf{X}_j)\right\}.
$$

Since $\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}$ is symmetric about zero, we can use U-statistics, e.g., Wand and Jones (1995). In our case

$$
\begin{aligned}
\operatorname{Var}\frac{4}{n^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_i-\mathbf{X}_j) &= 32n^{-3}(n-1)\operatorname{Var}\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_2) + 64n^{-3}(n-1)(n-2)\\
&\quad\times \operatorname{Cov}\{\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_2), \operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_3)\}.
\end{aligned}
$$

Most of terms are asymptotically negligible, therefore the formula written above reduces to

$$
\begin{aligned}
&32n^{-2}\underbrace{E(\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_2))^2}_{\xi_2} - 64n^{-1}\underbrace{E^2\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_2)}_{\xi_0}\\
&\quad + 64n^{-1}\underbrace{E(\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_2)\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{X}_1-\mathbf{X}_3))}_{\xi_1}. \qquad (16)
\end{aligned}
$$

Let us express ξ_0, ξ_1 and ξ_2. From previous computations of the expected value one can see that ξ_0 is of order $o(\|\operatorname{vec}\mathbf{H}\|^3)$.

$$
\begin{aligned}
\xi_1 &= \iiint \operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{x}-\mathbf{y})\operatorname{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{x}-\mathbf{z})f(\mathbf{x})f(\mathbf{y})f(\mathbf{z})d\mathbf{x}d\mathbf{y}d\mathbf{z}\\
&= \iiint \Lambda(\mathbf{u})\Lambda(\mathbf{v})f(\mathbf{x})\operatorname{tr}D^{2r}f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{u})\operatorname{tr}D^{2r}f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{v})d\mathbf{x}d\mathbf{u}d\mathbf{v}\\
&= \iiint \Lambda(\mathbf{u})\Lambda(\mathbf{v})f(\mathbf{x})\operatorname{tr}\left[\sum_{i=0}^{5}\frac{(-1)^i}{i!}\{D^{2r}a^i\}f(\mathbf{x}) + o(\|\mathbf{H}^{1/2}\mathbf{u}\|^5)\mathbf{J}_{d^r}\right]\\
&\quad\times \operatorname{tr}\left[\sum_{i=0}^{5}\frac{(-1)^i}{i!}\{D^{2r}b^i\}f(\mathbf{x}) + o(\|\mathbf{H}^{1/2}\mathbf{v}\|^5)\mathbf{J}_{d^r}\right]d\mathbf{x}d\mathbf{u}d\mathbf{v}, \quad \text{where } a=\mathbf{u}^T\mathbf{H}^{1/2}D,\ b=\mathbf{v}^T\mathbf{H}^{1/2}D\\
&= \iiint \Lambda(\mathbf{u})\Lambda(\mathbf{v})f(\mathbf{x})\frac{1}{4!4!}\operatorname{tr}\{D^{2r}a^4\}f(\mathbf{x})\operatorname{tr}\{D^{2r}b^4\}f(\mathbf{x})d\mathbf{x}d\mathbf{u}d\mathbf{v} + o(\|\operatorname{vec}\mathbf{H}\|^4)\\
&= \frac{1}{4!4!}\int f(\mathbf{x})\left[\int \Lambda(\mathbf{z})\{\operatorname{tr}^2(\mathbf{H}^{1/2}D^2\mathbf{H}^{1/2}\mathbf{z}\mathbf{z}^T)D^{2r}\}f(\mathbf{x})d\mathbf{z}\right]^2 d\mathbf{x} + o(\|\operatorname{vec}\mathbf{H}\|^4)\\
&= \frac{1}{16}\int \{\operatorname{tr}^2(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})\{\operatorname{tr}^2(\mathbf{H}D^2)D^{2r}\}f(\mathbf{x})f(\mathbf{x})d\mathbf{x} + o(\|\operatorname{vec}\mathbf{H}\|^4).
\end{aligned}
$$

Thus ξ_1 is of order $o(\|\operatorname{vec}\mathbf{H}\|^3)$ and is negligible.

Finally

$$\begin{aligned}\xi_2 &= \iint \mathrm{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{x}-\mathbf{y})\mathrm{tr}D^{2r}\Lambda_{\mathbf{H}}(\mathbf{x}-\mathbf{y})f(\mathbf{x})f(\mathbf{y})d\mathbf{x}d\mathbf{y}\\ &= |\mathbf{H}|^{-1/2}\iint \mathrm{tr}^2(\mathbf{H}^{-r}D^{2r}\Lambda(\mathbf{z}))f(\mathbf{x})f(\mathbf{x}-\mathbf{H}^{1/2}\mathbf{z})d\mathbf{x}d\mathbf{z}\\ &= |\mathbf{H}|^{-1/2}\|\mathrm{vec}\mathbf{H}\|^{-2r}V(\mathrm{vec}D^{2r}\Lambda)V(f)+o(|\mathbf{H}|^{-1/2}\|\mathrm{vec}\mathbf{H}\|^{-2r}),\end{aligned}$$

which completes the proof of Theorem 3. □

Sketch of the proof of Theorem 4:

Proof. Since $\widehat{\Gamma}_r(\mathbf{H})\xrightarrow{P}\Gamma_r(\mathbf{H})$ then $\widehat{\mathbf{H}}_{\mathrm{IT}_r}\xrightarrow{P}\mathbf{H}_{\mathrm{AMISE},r}$ as $n\to\infty$ and we can adopt ideas of AMSE lemma (Duong, 2004). We expand

$$\begin{aligned}\widehat{\Gamma}_r(\widehat{\mathbf{H}}_{\mathrm{IT}_r}) &= (\widehat{\Gamma}_r-\Gamma_r)(\widehat{\mathbf{H}}_{\mathrm{IT}_r})+\Gamma_r(\widehat{\mathbf{H}}_{\mathrm{IT}_r})\\ &= (1+o(1))(\widehat{\Gamma}_r-\Gamma_r)(\mathbf{H}_{\mathrm{AMISE},r})+\Gamma_r(\mathbf{H}_{\mathrm{AMISE},r})\\ &\quad+(1+o(1))D_{\mathbf{H}}^T\Gamma_r(\mathbf{H}_{\mathrm{AMISE},r})\mathrm{vech}(\widehat{\mathbf{H}}_{\mathrm{IT}_r}-\mathbf{H}_{\mathrm{AMISE},r}).\end{aligned}$$

We multiply the equation by $\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}$ from the left side and remove all negligible terms. Then we obtain

$$\mathbf{0}=\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}(\widehat{\Gamma}_r-\Gamma_r)(\mathbf{H}_{\mathrm{AMISE},r})+\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}D_{\mathbf{H}}^T\Gamma_r(\mathbf{H}_{\mathrm{AMISE},r})\mathrm{vech}(\widehat{\mathbf{H}}_{\mathrm{IT}_r}-\mathbf{H}_{\mathrm{AMISE},r}).$$

It is easy to see that $D_{\mathbf{H}}^T\Gamma_r(\mathbf{H}_{\mathrm{AMISE},r})=\mathbf{a}^T n^{-2/(d+2r+4)}$ and $\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}=\mathbf{b}n^{-2/(d+2r+4)}$ for constant vectors $\mathbf{a}$ and $\mathbf{b}$, which implies

$$\mathrm{vech}(\widehat{\mathbf{H}}_{\mathrm{IT}_r}-\mathbf{H}_{\mathrm{AMISE},r})=\underbrace{-(\mathbf{b}\mathbf{a}^T)^{-1}}_{\mathbf{C}}n^{4/(d+2r+4)}\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}(\widehat{\Gamma}_r-\Gamma_r)(\mathbf{H}_{\mathrm{AMISE},r}).$$

Let us note that the matrix $\mathbf{b}\mathbf{a}^T$ can be singular in some cases (e.g., for a diagonal bandwidth matrix) and thus the matrix $\mathbf{C}=-(\mathbf{b}\mathbf{a}^T)^{-1}$ does not exist. But this fact does not take any effect for the rate of convergence.

Using results of Theorem 3 we express the convergence rate of MSE $\left\{(\widehat{\Gamma}_r-\Gamma_r)(\mathbf{H}_{\mathrm{AMISE},r})\right\}$

$$\begin{aligned}&= Bias^2(\widehat{\Gamma}_r(\mathbf{H}_{\mathrm{AMISE},r}))+\mathrm{Var}(\widehat{\Gamma}_r(\mathbf{H}_{\mathrm{AMISE},r}))\\ &= (o(\|\mathrm{vec}\mathbf{H}_{\mathrm{AMISE},r}\|^{5/2}))^2+O(n^{-2}|\mathbf{H}_{\mathrm{AMISE},r}|^{-1/2}\|\mathrm{vec}\mathbf{H}_{\mathrm{AMISE},r}\|^{-2r})\\ &= (O(\|\mathrm{vec}\mathbf{H}_{\mathrm{AMISE},r}\|^{3}))^2+O(n^{-2}|\mathbf{H}_{\mathrm{AMISE},r}|^{-1/2}\|\mathrm{vec}\mathbf{H}_{\mathrm{AMISE},r}\|^{-2r})\\ &= O(n^{-12/(d+2r+4)})+O(n^{-(d+8)/(d+2r+4)})\\ &= O(n^{-\min\{d+8,12\}/(d+2r+4)}).\end{aligned}$$

Then

$$\begin{aligned}\mathrm{MSE}\{\mathrm{vech}\widehat{\mathbf{H}}_{\mathrm{IT}_r}\} &= \mathrm{MSE}\left\{(\widehat{\Gamma}_r-\Gamma_r)(\mathbf{H}_{\mathrm{AMISE},r})\right\}\mathbf{C}\,\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE},r}\,\mathbf{C}^T n^{8/(d+2r+4)}\\ &= O\left(n^{-\min\{d+8,12\}/(d+2r+4)}\right)O\left(n^{8/(d+2r+4)}\mathbf{J}_{d^*}\right)\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE},r}\\ &= O\left(n^{-\min\{d,4\}/(d+2r+4)}\mathbf{J}_{d^*}\right)\mathrm{vech}\mathbf{H}_{\mathrm{AMISE},r}\mathrm{vech}^T\mathbf{H}_{\mathrm{AMISE},r}.\quad\square\end{aligned}$$

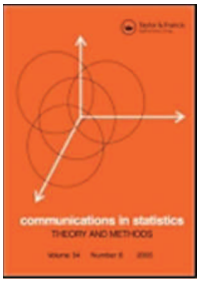
Acknowledgments

The research was supported by The Jaroslav Hájek Center for Theoretical and Applied Statistics (MŠMT LC 06024). K. Vopatová has been supported by the University of Defence through the Institutional development project UO FEM "Economics Laboratory". The authors thank the anonymous referees for their helpful comments and are also grateful to J.E. Chacón for a valuable discussion which contributed to improvement of this paper.

References

Aldershof, B., Marron, J., Park, B., Wand, M., 1995. Facts about the Gaussian probability density function. Applicable Analysis 59, 289–306.
Cao, R., Cuevas, A., González Manteiga, W., 1994. A comparative study of several smoothing methods in density estimation. Computational Statistics and Data Analysis 17, 153–176.
Chacón, J.E., Duong, T., 2012. Bandwidth selection for multivariate density derivative estimation, with applications to clustering and bump hunting. e-prints. http://arxiv.org/abs/1204.6160.
Chacón, J.E., 2010. Multivariate kernel estimation, lecture. Masaryk University, Brno.
Chacón, J.E., Duong, T., 2010. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. Test 19, 375–398.
Chacón, J.E., Duong, T., Wand, M.P., 2011. Asymptotics for general multivariate kernel density derivative estimators. Statistica Sinica 21, 807–840.

Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. Journal of the American Statistical Association 94, 807–823.
Duong, T., 2004. Bandwidth selectors for multivariate kernel density estimation. Ph.D. Thesis. School of Mathematics and Statistics. University of Western Australia.
Duong, T., Cowling, A., Koch, I., Wand, M.P., 2008. Feature significance for multivariate kernel density estimation. Computational Statistics and Data Analysis 52, 4225–4242.
Duong, T., Hazelton, M., 2005b. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics 32, 485–506.
Duong, T., Hazelton, M., 2005a. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. Journal of Multivariate Analysis 93, 417–433.
Godtliebsen, F., Marron, J.S., Chaudhuri, P., 2002. Significance in scale space for bivariate density estimation. Journal of Computational and Graphical Statistics 11, 1–21.
Horová, I., Koláček, J., Vopatová, K., 2012. Visualization and bandwidth matrix choice. Communications in Statistics—Theory and Methods 759–777.
Horová, I., Koláček, J., Zelinka, J., Vopatová, K., 2008. Bandwidth choice for kernel density estimates. In: Proceedings IASC. IASC, Yokohama, pp. 542–551.
Horová, I., Vieu, P., Zelinka, J., 2002. Optimal choice of nonparametric estimates of a density and of its derivatives. Statistics and Decisions 20, 355–378.
Horová, I., Vopatová, K., 2011. Kernel gradient estimate. In: Ferraty, F. (Ed.), Recent Advances in Functional Data Analysis and Related Topics. Springer-Verlag, Berlin, Heidelberg, pp. 177–182.
Horová, I., Zelinka, J., 2007a. Contribution to the bandwidth choice for kernel density estimates. Computational Statistics 22, 31–47.
Horová, I., Zelinka, J., 2007b. Kernel estimation of hazard functions for biomedical data sets. In: Härdle, W., Mori, Y., Vieu, P. (Eds.), Statistical Methods for Biostatistics and Related Fields. In: Mathematics and Statistics, Springer-Verlag, Berlin, Heidelberg, pp. 64–86.
Horová, I., Zelinka, J., Budíková, M., 2006. Kernel estimates of hazard functions for carcinoma data sets. Environmetrics 17, 239–255.
Härdle, W., Marron, J.S., Wand, M.P., 1990. Bandwidth choice for density derivatives. Journal of the Royal Statistical Society, Series B (Methodological) 52, 223–232.
Jones, M.C., Kappenman, R.F., 1991. On a class of kernel density estimate bandwidth selectors. Scandinavian Journal of Statistics 19, 337–349.
Jones, M.C., Marron, J.S., Park, B.U., 1991. A simple root n bandwidth selector. Annals of Statistics 19, 1919–1932.
Magnus, J.R., Neudecker, H., 1979. Commutation matrix—some properties and application. Annals of Statistics 7, 381–394.
Magnus, J.R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics, second ed.. Wiley.
Marron, J.S., Ruppert, D., 1994. Transformations to reduce boundary bias in kernel density estimation. Journal of the Royal Statistical Society, Series B (Methodological) 56, 653–671.
Park, B., Marron, J., 1990. Comparison of data-driven bandwidth selectors. Journal of the American Statistical Association 85, 66–72.
Sain, S., Baggerly, K., Scott, D., 1994. Cross-validation of multivariate densities. Journal of the American Statistical Association 89, 807–817.
Scott, D.W., 1992. Multivariate density estimation: theory, practice, and visualization. In: Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley.
Scott, D.W., Terrell, G.R., 1987. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association 82, 1131–1146.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer-Verlag, New York.
Taylor, C.C., 1989. Bootstrap choice of the smoothing parameter in kernel density estimation. Biometrika 76, 705–712.
Terrell, G.R., 1990. The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85, 470–477.
UNICEF, 2003. The state of the world's children 2003. http://www.unicef.org/sowc03/index.html.
Vopatová, K., Horová, I., Koláček, J., 2010. Bandwidth choice for kernel density derivative. In: Proceedings of the 25th International Workshop on Statistical Modelling. Glasgow, Scotland, pp. 561–564.
Wand, M., Jones, M., 1995. Kernel Smoothing. Chapman and Hall, London.
Wand, M.P., Jones, M.C., 1994. Multivariate plug-in bandwidth selection. Computational Statistics 9, 97–116.



Selection of bandwidth for kernel regression

Journal:	*Communications in Statistics – Theory and Methods*
Manuscript ID:	Draft
Manuscript Type:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Kolacek, Jan; Masaryk University, Dept. of Mathematics and Statistics Horova, Ivana; Masaryk University, Dept. of Mathematics and Statistics
Keywords:	kernel regression, bandwidth selection, iterative method
Abstract:	The most important factor in kernel regression is a choice of a bandwidth. Considerable attention has been paid to extension the idea of an iterative method known for a kernel density estimate. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated square bias. This approach leads to an iterative quadratically convergent process. The analysis of statistical properties shows the rationale of the proposed method. In order to see statistical properties of this method the consistency is determined. The utility of the method is illustrated through a simulation study and real data applications.

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

article.tex
references.bib

SELECTION OF BANDWIDTH FOR KERNEL REGRESSION

JAN KOLÁČEK, IVANA HOROVÁ

Abstract. The most important factor in kernel regression is a choice of a bandwidth. Considerable attention has been paid to extension the idea of an iterative method known for a kernel density estimate to kernel regression. Data-driven selectors of the bandwidth for kernel regression are considered. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated square bias. This approach leads to an iterative quadratically convergent process. The analysis of statistical properties shows the rationale of the proposed method. In order to see statistical properties of this method the consistency is determined. The utility of the method is illustrated through a simulation study and real data applications.

Keywords and Phrases: kernel regression, bandwidth selection, iterative method.

Mathematics Subject Classification: 62G08

1. Introduction

Kernel regression estimates are one of the most popular nonparametric estimates. In a univariate case, these estimates depend on a bandwidth, which is a smoothing parameter controlling smoothness of an estimated curve and a kernel which is considered as a weight function. The choice of the smoothing parameter is a crucial problem in the kernel regression. The literature on bandwidth selection is quite extensive, e.g., monographs [20, 17, 18], papers [7, 2, 3, 15, 19, 4, 5, 12, 13].

Although in practice one can try several bandwidths and choose a bandwidth subjectively, automatic (data-driven) selection procedures could be useful for many situations; see [16] for more examples. Most of these procedures are based on estimating of Average Mean Square Error. They are asymptotically equivalent and asymptotically unbiased (see [7, 2, 3]). However, in simulation studies ([12]), it is often observed that most selectors are biased toward undersmoothing and yield smaller bandwidths more frequently than predicted by asymptotic results.

Successful approaches to the bandwidth selection in kernel density estimation can be transferred to the case of kernel regression. The iterative method for the kernel density has been developed and widely discussed in [9]. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated square bias.

The paper is organized as follows: In Section 2 we describe kernel estimates of a regression function and give a form of the Mean Integrated Square Error and its asymptotic alternative. The next section is devoted to a data-driven bandwidth selection method. This method is based on an optimally balanced relation between the integrated variance and the integrated squared bias, see [9]. Similar ideas were

Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic.

applied to kernel estimates of hazard functions (see [11] or [10]). It seems that the basic idea can be also extended to a kernel regression and we are going to investigate this possibility. We discuss the statistical properties of the proposed method as well. Section 4 brings a simulation study and in the last section the developed theory is applied to real data sets.

2. Univariate kernel regression

Consider a standard regression model of the form

$$Y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{2.1}$$

where m is an unknown regression function, $Y_1, \dots, Y_n$ are observable data variables with respect to the design points $x_1, \dots, x_n$. The residuals $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed random variables for which

$$\mathrm{E}(\varepsilon_i) = 0, \qquad \mathrm{var}(\varepsilon_i) = \sigma^2 > 0, \qquad i = 1, \dots, n.$$

We suppose the *fixed equally spaced design*, *i.e.*, design variables are not random and $x_i = i/n,\ i = 1, \dots, n$. In the case of *random design*, where the design points $X_1, \dots, X_n$ are random variables with the same density f, all considerations are similar as for the fixed design. More detailed description of the random design can be found, *e.g.*, in [20].

The aim of kernel smoothing is to find a suitable approximation $\widehat{m}$ of the unknown function m.

We consider the estimator proposed by Pristley and Chao [14] which is defined as

$$\widehat{m}(x, h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) Y_i, \quad \text{for } x \in (0, 1). \tag{2.2}$$

The function K is called the kernel which is assumed to be symmetric about zero and be supported on $[-1, 1]$, be such that $V(K) = \int K(u)^2 du < \infty$ and have a finite second moment (*i.e.*, $\int u^2 K(u) du = \beta_2 < \infty$). Set $K_h(.) = \frac{1}{h} K(\frac{\cdot}{h})$, $h > 0$. A parameter h is called a *bandwidth*.

The quality of a kernel regression estimator can be locally described by the Mean Square Error (MSE) or by a global criterion the Mean Integrated Square Error (MISE), which can be written as a sum of the Integrated Variance (IV) and the Integrated Square Bias (ISB)

$$\begin{aligned} \mathrm{MISE}\{\widehat{m}(\cdot, h)\} &= E \int_0^1 [\widehat{m}(x, h) - m(x)]^2 dx \\ &= \underbrace{\int_0^1 \mathrm{Var}\, \widehat{m}(x, h) dx}_{\mathrm{IV}} + \underbrace{\int_0^1 [(K_h * m)(x) - m(x)]^2 dx}_{\mathrm{ISB}} + O\left(n^{-1}\right), \end{aligned} \tag{2.3}$$

where $*$ denotes a convolution.

Since the MISE is not mathematically tractable we employ the Asymptotic Mean Integrated Square Error (AMISE)

$$\text{AMISE}\{\widehat{m}(\cdot,h)\} = \underbrace{\frac{V(K)\sigma^2}{nh}}_{\text{AIV}} + \underbrace{\left(\frac{\beta_2}{2}\right)^2 V(m'')h^4}_{\text{AISB}}, \tag{2.4}$$

where $V(m'') = \int_0^1 (m''(x))^2\,dx$. The optimal bandwidth considered here is h_{opt}, the minimizer of (2.4), *i.e.*,

$$h_{opt} = \arg\min_{h\in H_n} \text{AMISE}\{\widehat{m}(\cdot,h)\},$$

where $H_n = [an^{-1/5}, bn^{-1/5}]$ for some $0 < a < b < \infty$.
The calculation gives

$$h_{opt} = \left(\frac{\sigma^2 V(K)}{n\beta_2^2 V(m'')}\right)^{\frac{1}{5}}. \tag{2.5}$$

In nonparametric regression estimation a critical and inevitable step is to choose the smoothing parameter (bandwidth) to control the smoothness of the curve estimate. The smoothing parameter considerably affects the features of the estimated curve.

One of the most widespread procedures for bandwidth selection is the cross-validation method, also known as "leave-one-out" method.

The method is based on modified regression smoother (2.2) in which one, say the j-th, observation is left out:

$$\widehat{m}_{-j}(x_j,h) = \frac{1}{n}\sum_{\substack{i=1\\ i\neq j}}^{n} K_h(x_i - x_j)Y_i, \quad j = 1,\ldots,n.$$

With using these modified smoothers, the error function which should be minimized takes the form

$$\text{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}_{-i}(x_i) - Y_i\}^2. \tag{2.6}$$

The function CV(h) is commonly called a "cross-validation" function. Let $\hat{h}_{\text{CV}}$ stand for minimization of CV(h), *i.e.*,

$$\hat{h}_{\text{CV}} = \arg\min_{h\in H_n} \text{CV}(h).$$

The literature on this criterion is quite extensive, *e.g.*, [19, 4, 7, 5].

3. Iterative method for kernel regression

The proposed method is based on the following relation. It is easy to show that the equation holds

$$\text{AIV}\,\{\widehat{m}(\cdot,h_{opt})\} - 4\,\text{AISB}\{\,\widehat{m}(\cdot,h_{opt})\} = 0, \tag{3.1}$$

where AIV and AISB are terms used in (2.4). For estimating of AIV and AISB in (3.1) we use

$$\widehat{\mathrm{AIV}}\left\{\widehat{m}(\cdot,h)\right\} = \frac{\hat{\sigma}^2 V(K)}{nh}, \text{ with } \hat{\sigma}^2 = \frac{1}{2n-2}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2$$

and

$$\begin{aligned}\widehat{\mathrm{AISB}}\left\{\widehat{m}(\cdot,h)\right\} &= \int_0^1 [(K_h * \widehat{m})(x,h) - \widehat{m}(x,h)]^2 dx \\ &= \frac{1}{4n^2 h}\sum_{\substack{i,j=1\\ i\neq j}}^{n} \Lambda\left(\frac{x_i - x_j}{h}\right) Y_i Y_j,\end{aligned}$$

where $\Lambda(z) = (K*K*K*K - 2K*K*K*K + K*K*K*K)(z)$ (see Complements for more details, for properties of $\Lambda(z)$ see [8]).
To find the bandwidth estimate $\hat{h}_{\mathrm{IT}}$ we solve the equation

$$\widehat{\mathrm{AIV}}\left\{\widehat{m}(\cdot,h)\right\} - 4\widehat{\mathrm{AISB}}\left\{\widehat{m}(\cdot,h)\right\} = 0, \tag{3.2}$$

which leads to finding a fixed point of the equation

$$h = \frac{\hat{\sigma}^2 V(K)}{4nh\widehat{\mathrm{AISB}}\left\{\widehat{m}(\cdot,h)\right\}}. \tag{3.3}$$

We use Steffensen's iterative method with the starting approximation $\hat{h}_0 = 2/n$. This approach leads to an iterative quadratically convergent process (see [9]).

The solution $\hat{h}_{\mathrm{IT}}$ of the equation (3.2) can be considered as a suitable approximation of h_{opt} as it is confirmed by the following theorem.

Theorem 3.1. *Let $m \in C^2[0,1]$, m'' be square integrable, $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh = \infty$. Let $\mathcal{P}(h)$ stand for the left side of (3.1) and $\widehat{\mathcal{P}}(h)$ for the left side of (3.2). Then*

$$\begin{aligned}E(\widehat{\mathcal{P}}(h)) &= \mathcal{P}(h) + O\left(n^{-1}\right), \\ \mathrm{var}(\widehat{\mathcal{P}}(h)) &= O\left(n^{-1}\right).\end{aligned} \tag{3.4}$$

Theorem 3.1 states that $\widehat{\mathcal{P}(h)}$ is a consistent estimate of $\mathcal{P}(h)$. This result confirms that the solution of (3.3) may be expected to be reasonably close to h_{opt}. Proof of Theorem 3.1 can be found in Complements.

4. Simulation study

We carry out two simulation studies to compare the performance of the bandwidth estimates. The comparison is done in the following way. The observations, Y_i, for $i = 1, \ldots, n = 100$, are obtained by adding independent Gaussian random variables with mean zero and variance σ^2 to some known regression function. Both regression functions used in our simulations are illustrated in Fig. 1.

One hundred series are generated. For each data set, we estimate the optimal bandwidth by both mentioned methods, *i.e.*, for each method we obtain 100 estimates. Since we know the optimal bandwidth, we compare it with the mean of estimates and look at their standard deviation, which describes the variability of methods. The Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}$ is used in all cases.

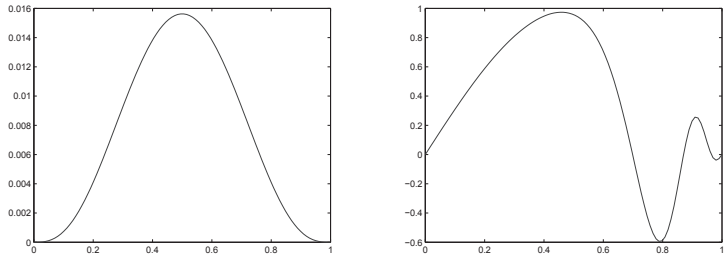


FIGURE 1. Regression functions.

Finally, we calculate the Integrated Square Error (ISE)

$$\mathrm{ISE}\{\widehat{m}(\cdot,h)\} = \int_0^1 \left(\widehat{m}(x,h) - m(x)\right)^2 dx$$

for each estimated regression function over all 100 replications. The logarithm of results are displayed in Tables 2, 4 and in Figures 3, 5. Here "IT" denotes the results for our proposed method, "CV" stands for the results of the cross-validation method.

4.1. **Simulation 1.** In this case, we use the regression function

$$m(x) = x^3(1-x)^3$$

with $\sigma^2 = 0.003^2$. Table 1 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 100 values and $std(\hat{h})$ is their standard deviation. Figure 2 illustrates the histogram of results of all 100 experiments.

	$h_{opt} = 0.1188$	
	$E(\hat{h})$	$std(\hat{h})$
CV	0.1057	0.0297
IT	0.1184	0.0200

TABLE 1. Means and standard deviations

Table 2 gives the mean and the standard deviations of log(ISE) for each method compared with log(ISE) for the regression estimate obtained with h_{opt}. Figure 3 illustrates the histogram of log(ISE) of all 100 experiments.

As we see, the standard deviation of all results obtained by the proposed method is less than the value for the case of cross-validation method and also the mean of these results is slightly closer to the theoretical optimal bandwidth. The comparison of results with respect to log(ISE) leads to the similar result. The reason is that the regression function is smooth and satisfies all the conditions supposed in the previous section. Thus the proposed method works very well in this case.

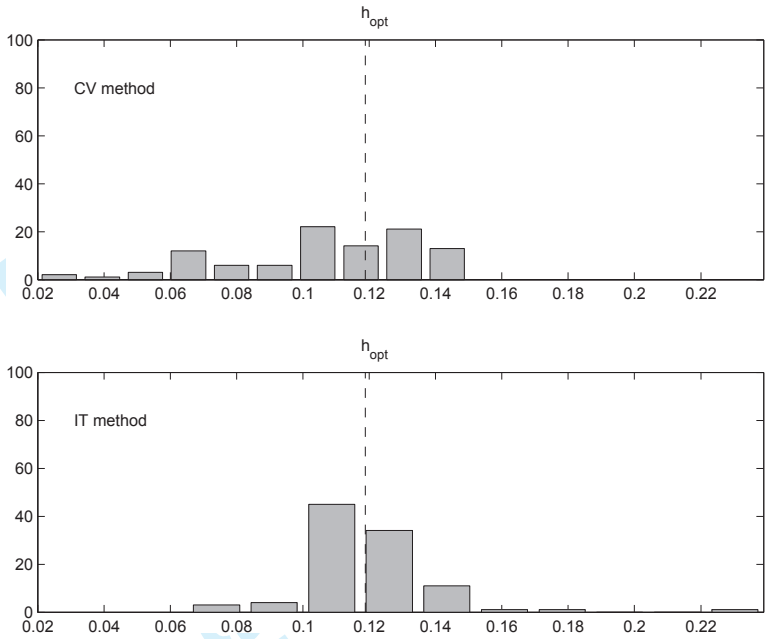


FIGURE 2. Distribution of $\hat{h}$ for both methods.

	$E(\log(\text{ISE}))$	$std(\log(\text{ISE}))$
h_{opt}	-14.4452	0.5421
IT	-14.3481	0.5193
CV	-14.2160	0.6276

TABLE 2. Means and standard deviations of log(ISE)

4.2. **Simulation 2.** In the second example, we use the regression function

$$m(x) = \sin(\pi\, x)\, \cos\left(3\,\pi\, x^5\right)$$

with $\sigma^2 = 0.05$. Table 3 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 100 values and $std(\hat{h})$ is their standard deviation. Figure 4 illustrates the histogram of results of all 100 experiments.

Table 4 brings the mean and the standard deviations of log(ISE) for each method compared with log(ISE) for the regression estimate obtained with h_{opt}. Figure 5 illustrates the histogram of log(ISE) of all 100 experiments.

Although the mean of $\hat{h}_{IT}$ is not so close to h_{opt} as the mean of $\hat{h}_{CV}$, the values of ISE are better. Also the variability of the proposed method seems to be smaller in this case. Thus we make a conclusion that the proposed method can provide better results for this regression model.

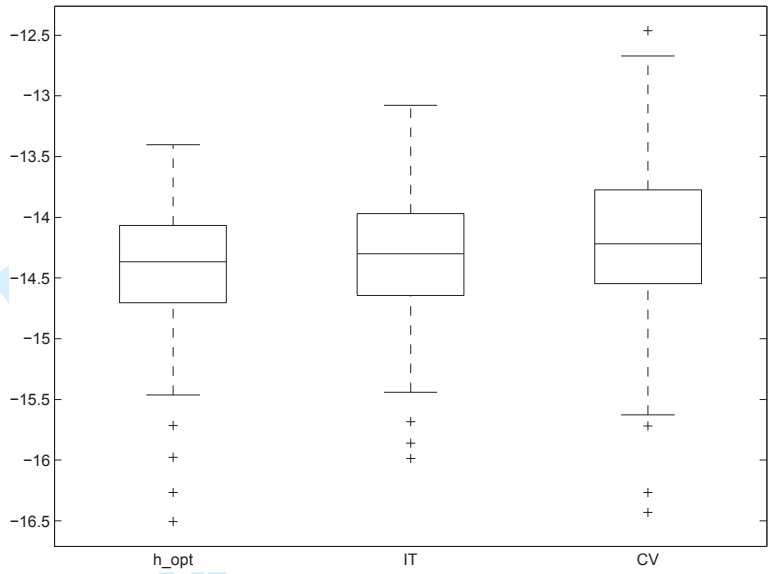


FIGURE 3. Logarithm of ISE.

	$h_{opt} = 0.0585$	
	$E(\hat{h})$	$std(\hat{h})$
CV	0.0633	0.0168
IT	0.0708	0.0072

TABLE 3. Means and standard deviations

	$E(\log(\text{ISE}))$	$std(\log(\text{ISE}))$
h_{opt}	−5.0932	0.3908
IT	−5.0560	0.3741
CV	−4.9525	0.3966

TABLE 4. Means and standard deviations of log(ISE)

5. APPLICATION TO REAL DATA

The main goal of this section is to make a comparison of mentioned bandwidth estimators on a real data set. We use data from [1] and follow annual measurements of the level, in feet, of Lake Huron 1875 – 1972, *i.e.*, the sample size is $n = 98$. We transform data to the interval $[0, 1]$ and use both selectors considered in the previous section to get the optimal bandwidth. We use the Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}$. All estimates of optimal bandwidth are listed in Table 5.

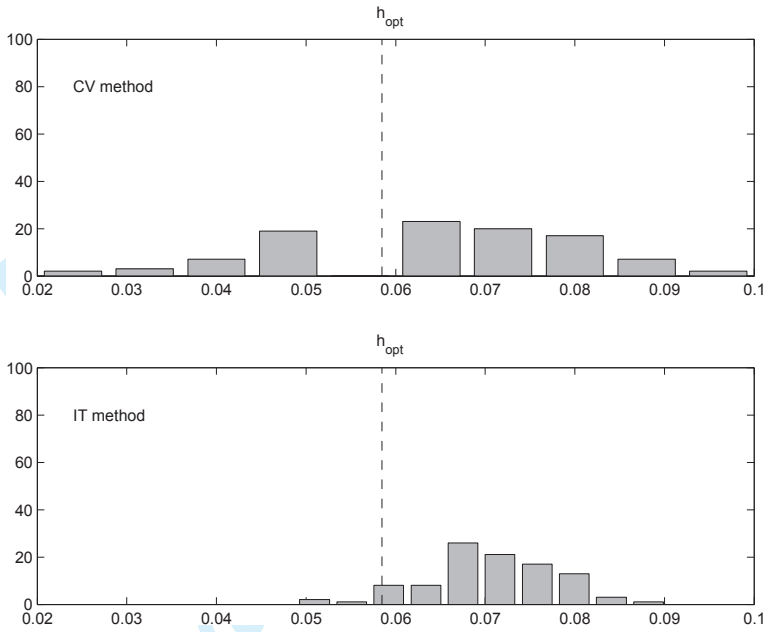


FIGURE 4. Distribution of $\hat{h}$ for both methods.

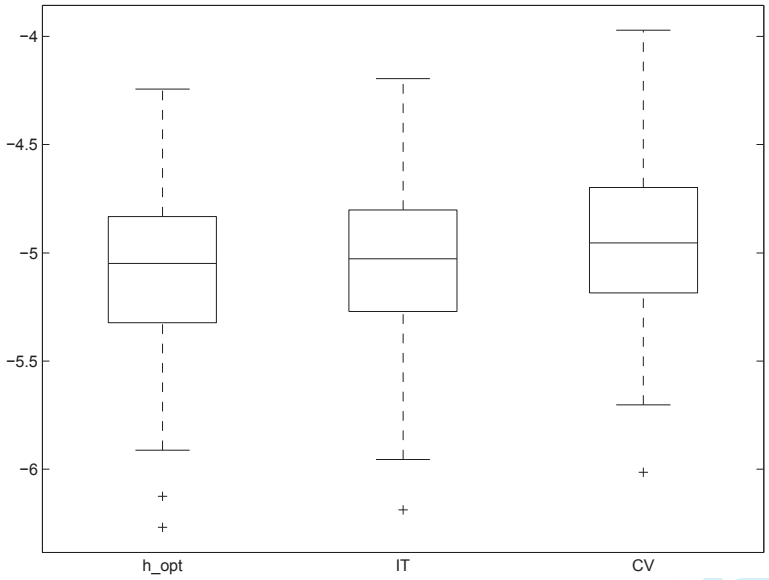


FIGURE 5. Logarithm of ISE.

Figure 6 illustrates the kernel regression estimate with the smoothing parameter $\hat{h}_{CV} = 0.0204$ which was obtained by cross-validation method.

Figure 7 shows the kernel regression estimate with the smoothing parameter $\hat{h}_{IT} = 0.0501$. This value was found by our proposed method

TABLE 5. Optimal bandwidth estimates for Lake Huron data.

iterative method	$\hat{h}_{IT} = 0.0501$
cross-validation	$\hat{h}_{CV} = 0.0204$

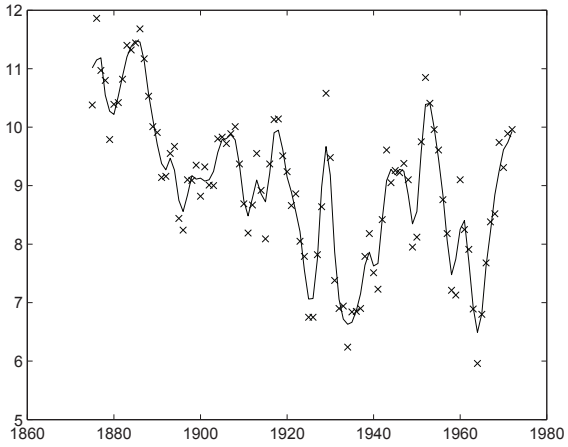


FIGURE 6. Kernel regression estimate with $\hat{h}_{CV} = 0.0204$.

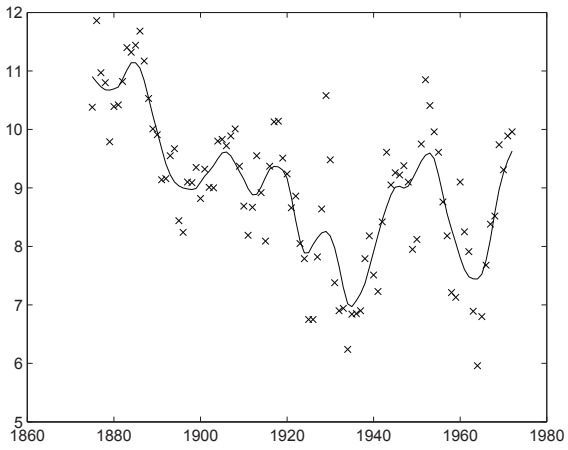


FIGURE 7. Kernel regression estimate with $\hat{h}_{IT} = 0.0501$.

Since we do not know the true regression function $m(x)$ it is hard to assess objectively which one of kernel estimates is better. It is very important to realize the fact that the final decision about the estimate is partially subjective because the estimates of the bandwidth are only asymptotically optimal. The values summarized in the table and figures show that the estimate with the smoothing parameter

obtained by cross-validation criterion is undersmoothed. In the context of these considerations, the estimate with parameter obtained by the iterative method appears to be sufficient.

6. Conclusion

A new bandwidth selector for kernel regression was proposed. The analysis of statistical properties shows the rationale of the proposed method. The advantage of the method is in computational aspects, since it makes possible to avoid the minimization process and only solves one nonlinear equation.

7. Acknowledgments

This research was supported by Masaryk University, project MUNI/A/1001/2009.

8. Complements

Proof of Theorem 3.1.
Let us denote

$$\mathcal{P}(h) = \frac{V(K)\sigma^2}{nh} - h^4\beta_2^2 V(m'') \tag{8.1}$$

and let

$$\widehat{\mathcal{P}}(h) = \frac{V(K)\hat{\sigma}^2}{nh} - \frac{1}{n^2 h}\sum_{\substack{i,j=1\\ i\neq j}}^{n} \Lambda\left(\frac{x_i - x_j}{h}\right) Y_i Y_j \tag{8.2}$$

stand for an estimate of $\mathcal{P}$. The proposed method aims to solve the equation

$$\widehat{\mathcal{P}}(h) = 0.$$

For a better clarity we use the notation $\int$ for $\int_0^1$ in next. As the first step, we prove the following lemma.

Lemma 8.1. *For $i, j = 1, \ldots, n,\ i \neq j$ the formula holds*

$$h\Lambda\left(\frac{x_i - x_j}{h}\right) Y_i Y_j = \int \left[(K*K)\left(\frac{x-x_i}{h}\right) - K\left(\frac{x-x_i}{h}\right)\right] \times \left[(K*K)\left(\frac{x-x_j}{h}\right) - K\left(\frac{x-x_j}{h}\right)\right] Y_i Y_j dx.$$

Proof.

$$\int \left[(K*K)\left(\frac{x-x_i}{h}\right) - K\left(\frac{x-x_i}{h}\right)\right]\left[(K*K)\left(\frac{x-x_j}{h}\right) - K\left(\frac{x-x_j}{h}\right)\right] dx$$
$$= \int (K*K)\left(\frac{x-x_i}{h}\right)(K*K)\left(\frac{x-x_j}{h}\right) dx - 2\int (K*K)\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right) dx + \int K\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right) dx.$$

Set the three integrals in the sum as η_1, η_2, η_3. We modify η_3 by substitution $t = \frac{x-x_j}{h}$. Using the parity of K we get

$$\eta_3 = h\int_{\frac{-x_j}{h}}^{\frac{1-x_j}{h}} K(t) K\left(t - \frac{x_i - x_j}{h}\right) dt.$$

Provided $x_j \in [0,1]$ then, as $h \to \infty$, $-x_j/h \to -\infty$ and $(1-x_j)/h \to \infty$. Therefore

$$\eta_3 = h(K * K)\left(\frac{x_i - x_j}{h}\right).$$

Similarly we can obtain

$$\eta_2 = h(K * K * K)\left(\frac{x_i - x_j}{h}\right),$$

$$\eta_1 = h(K * K * K * K)\left(\frac{x_i - x_j}{h}\right).$$

Thus $\eta_1 - 2\eta_2 + \eta_3 = h\Lambda\left(\frac{x_i - x_j}{h}\right)$. □

We start with an evaluation of $\frac{1}{n^2 h} E \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \Lambda\left(\frac{x_i - x_j}{h}\right) Y_i Y_j$:

$$\begin{aligned}
&\frac{1}{n^2 h} E \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \Lambda\left(\frac{x_i - x_j}{h}\right) Y_i Y_j \\
&\overset{L.8.1}{=} \frac{1}{n^2 h^2} E \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \int \left[(K*K)\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i}{h}\right)\right] \\
&\qquad \times \left[(K*K)\left(\frac{x - x_j}{h}\right) - K\left(\frac{x - x_j}{h}\right)\right] Y_i Y_j dx \\
&= \frac{1}{n^2 h^2} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \int \left[(K*K)\left(\frac{x - x_i}{h}\right) - K\left(\frac{x - x_i}{h}\right)\right] \\
&\qquad \times \left[(K*K)\left(\frac{x - x_j}{h}\right) - K\left(\frac{x - x_j}{h}\right)\right] m(x_i) m(x_j) dx \\
&= \int \left\{ \underbrace{\int_{-\infty}^{\infty} [(K*K)(t) - K(t)]\, m(x - ht) dt}_{I_1} \right\}^2 dx + O\left(n^{-1}\right).
\end{aligned}$$

Now, we approximate the integral I_1 by the Taylor's expansion of $m(x - th)$

$$I_1 = \int_{-\infty}^{\infty} [(K*K)(t) - K(t)] \left[m(x) - thm'(x) + \frac{t^2 h^2}{2} m''(x) + O(t^3 h^3)\right] dt.$$

It is an easy exercise to see the moment conditions for $(K*K)(t) - K(t)$: $\int_{-\infty}^{\infty} (K*K)(t) - K(t) dt = \int_{-\infty}^{\infty} t^{2k+1}[(K*K)(t) - K(t)]dt = 0,\ k \in \mathbb{N}$, $\int_{-\infty}^{\infty} t^2[(K*K)(t) - K(t)]dt = 2\beta_2$.
Thus

$$I_1 = h^2 \beta_2 m''(x) + O\left(h^4\right)$$

and

$$\frac{1}{n^2h}E\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)Y_iY_j = h^4\beta_2^2V(m'')+O\left(h^6\right)+O\left(n^{-1}\right).$$

Finally

$$E\widehat{\mathcal{P}}(h) = \frac{V(K)\sigma^2}{nh}-\beta_2^2V(m'')h^4+O\left(n^{-1}\right)$$

and

$$E\widehat{\mathcal{P}}(h) = \mathcal{P}(h)+O\left(n^{-1}\right). \tag{8.3}$$

Since it is assumed $\lim_{n\to\infty} nh=\infty$ then $E\widehat{\mathcal{P}}(h)\to\mathcal{P}(h)$.

Now, we derive the formula for $\mathrm{var}\widehat{\mathcal{P}}(h)$. As the first we express $\mathrm{var}\widehat{\mathrm{AISB}} = E(\widehat{\mathrm{AISB}})^2 - E^2\widehat{\mathrm{AISB}}$.

$$E(\widehat{\mathrm{AISB}})^2 = \frac{1}{16n^4h^2}E\left\{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)Y_iY_j\right\}^2$$

$$= \frac{1}{16n^4h^2}E\left\{\underbrace{\sum_{\substack{i,j,k,l=1\\ i\neq j\neq k\neq l}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)\Lambda\left(\frac{x_k-x_l}{h}\right)Y_iY_jY_kY_l}_{\zeta_1}\right.$$

$$\left.+\underbrace{\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)\Lambda\left(\frac{x_i-x_k}{h}\right)Y_i^2Y_jY_k}_{\zeta_2}+\underbrace{\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda^2\left(\frac{x_i-x_j}{h}\right)Y_i^2Y_j^2}_{\zeta_3}\right\}.$$

Then we compute

$$\frac{1}{16n^4h^2}E\zeta_1 = \frac{1}{16n^4h^2}\sum_{\substack{i,j,k,l=1\\ i\neq j\neq k\neq l}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)\Lambda\left(\frac{x_k-x_l}{h}\right)m(x_i)m(x_j)m(x_k)m(x_l)$$

$$= \frac{1}{16h^2}\iiint\Lambda\left(\frac{x-y}{h}\right)\Lambda\left(\frac{u-v}{h}\right)m(x)m(y)m(u)m(v)dxdydudv+O\left(n^{-1}\right)$$

$$= \frac{1}{16h^2}\left\{\iint\Lambda\left(\frac{x-y}{h}\right)m(x)m(y)dxdy\right\}^2+O\left(n^{-1}\right)$$

$$= \frac{1}{16}\left\{\int_{-\infty}^{\infty}\int\Lambda(t)m(x-th)m(x)dtdx\right\}^2+O\left(n^{-1}\right)$$

It is easy to see the moment conditions for $\Lambda(z)$: $\int_{-\infty}^{\infty}\Lambda(z)dz = \int_{-\infty}^{\infty}z^2\Lambda(z)dz = 0$, $\int_{-\infty}^{\infty}z^{2k-1}\Lambda(z)dz = 0$, $k\in\mathbb{N}$, $\int z^4\Lambda(z)dz = 6\beta_2^2$ ([8]). By using the second order

Taylor's expansion of $m(x-th)$ we obtain the result

$$\frac{1}{16n^4h^2}E\zeta_1 = \frac{1}{32}\left\{\int m''(x)m(x)dx \int_{-\infty}^{\infty}\left[\Lambda(t)t^2h^2 + O\left(h^3\right)\right]dt\right\}^2 + O\left(n^{-1}\right) = O\left(n^{-1}\right).$$

Similarly,

$$\begin{aligned}
\frac{1}{16n^4h^2}E\zeta_2 &= \frac{1}{16n^4h^2}\sum_{\substack{i,j,k=1\\ i\neq j\neq k}}^{n}\Lambda\left(\frac{x_i-x_j}{h}\right)\Lambda\left(\frac{x_i-x_k}{h}\right)\left[m^2(x_i)+\sigma^2\right]m(x_j)m(x_k)\\
&= \frac{1}{16nh^2}\iiint\Lambda\left(\frac{x-y}{h}\right)\Lambda\left(\frac{x-z}{h}\right)\left[m^2(x)+\sigma^2\right]m(y)m(z)dxdydz + O\left(n^{-1}\right)\\
&= \frac{1}{16n}\int\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Lambda(t)\Lambda(u)\left[m^2(x)+\sigma^2\right]m(x-th)m(x-uh)dtdudx + O\left(n^{-1}\right)\\
&= \frac{1}{64n}\int m''^2(x)\left[m^2(x)+\sigma^2\right]dx\left\{\int_{-\infty}^{\infty}\Lambda(t)t^2h^2dt\right\}^2 + O\left(h^6n^{-1}\right) + O\left(n^{-1}\right).
\end{aligned}$$

$$\begin{aligned}
\frac{1}{16n^4h^2}E\zeta_3 &= \frac{1}{16n^4h^2}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Lambda^2\left(\frac{x_i-x_j}{h}\right)\left[m^2(x_i)+\sigma^2\right]\left[m^2(x_j)+\sigma^2\right]\\
&= \frac{1}{16n^2h^2}\iint\Lambda^2\left(\frac{x-y}{h}\right)\left[m^2(x)+\sigma^2\right]\left[m^2(y)+\sigma^2\right]dxdy + O\left(n^{-1}\right)\\
&= \frac{1}{16n^2h}\int\int_{-\infty}^{\infty}\Lambda^2(t)\left[m^2(x)+\sigma^2\right]\left[m^2(x-th)+\sigma^2\right]dtdx + O\left(n^{-1}\right)\\
&= \frac{V(\Lambda)V(m^2+\sigma^2)}{16n^2h} + O\left(n^{-1}\right).
\end{aligned}$$

By combining results for $E(\widehat{\text{AISB}})^2$ and $E^2\widehat{\text{AISB}}$ we arrive at the expression

$$\text{var}\widehat{\text{AISB}} = O\left(n^{-1}\right).$$

Since $\hat{\sigma}^2$ is a consistent estimator of σ^2 (see [6]) and $\text{var}\widehat{\text{AISB}}$ is of order $O\left(n^{-1}\right)$, $\text{var}\widehat{\mathcal{P}}$ is a consistent estimator of $\text{var}\mathcal{P}$.

References

[1] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, 2009.

[2] S.T. Chiu. Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. *Biometrika*, 77(1):222–226, 1990.

[3] S.T. Chiu. Some stabilized bandwidth selectors for nonparametric regression. *Annals of Statistics*, 19(3):1528–1546, 1991.

[4] P Craven and G Wahba. Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403, 1979.

[5] Bernd Droge. Some comments on cross-validation. Technical Report 1994-7, Humboldt Universitaet Berlin, 1996.

[6] Peter Hall, J. W. Kay, and D. M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

[7] W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1st edition, 1990.
[8] I. Horová, J. Koláček, and J. Zelinka. *Kernel Smoothing in MATLAB*. World Scientific, Singapore, 2012.
[9] I. Horová and J. Zelinka. Contribution to the bandwidth choice for kernel density estimates. *Computational Statistics*, 22(1):31–47, 2007.
[10] I. Horová and J. Zelinka. Kernel estimation of hazard functions for biomedical data sets. In Wolfgang. Härdle, Yuichi. Mori, and Philippe Vieu, editors, *Statistical Methods for Biostatistics and Related Fields*, Mathematics and Statistics, pages 64–86. Springer-Verlag Berlin Heidelberg, 2007.
[11] I. Horová, J. Zelinka, and M. Budíková. Kernel estimates of hazard funcions for carcinoma data sets. *Environmetrics*, 17(3):239–255, 2006.
[12] Jan Koláček. *Kernel Estimation of the Regression Function (in Czech)*. PhD thesis, Masaryk University, Brno, feb 2005.
[13] Jan Koláček. Plug-in method for nonparametric regression. *Computational Statistics*, 23(1):63–78, 2008.
[14] M. B. Priestley and M. T. Chao. Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):385–392, 1972.
[15] J. Rice. Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12(4):1215–1230, 1984.
[16] Bernard W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47:1–52, 1985.
[17] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.
[18] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, New York, 1996.
[19] M Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 36(2):111–147, 1974.
[20] M.P. Wand and M.C. Jones. *Kernel smoothing*. Chapman and Hall, London, 1995.

Journal of Environmental Statistics

February 2013, Volume 4, Issue 2. http://www.jenvstat.org

Kernel Regression Model for Total Ozone Data

Horová I., Koláček J., Lajdová D.
Department of Mathematics and Statistics
Masaryk University Brno

Abstract

The present paper is focused on a fully nonparametric regression model for autocorrelation structure of errors in time series over total ozone data. We propose kernel methods which represent one of the most effective nonparametric methods.

But there is a serious difficulty connected with them – the choice of a smoothing parameter called a bandwidth. In the case of independent observations the literature on bandwidth selection methods is quite extensive. Nevertheless, if the observations are dependent, then classical bandwidth selectors have not always provided applicable results. There exist several possibilities for overcoming the effect of dependence on the bandwidth selection. In the present paper we use the results of Chu and Marron (1991) and Koláček (2008) and develop two methods for the bandwidth choice. We apply the above mentioned methods to the time series of ozone data obtained from the Vernadsky station in Antarctica. All discussed methods are implemented in Matlab.

Keywords: total ozone, kernel, bandwidth selection.

1. Introduction

Antarctica is significantly related to many environmental aspects and processes of the Earth. And thus its impact on the global climate system and water circulation in the world ocean is essential.

The stratosphere ozone depletion over Antarctica was discovered at the beginning of the 1990s. The lowest total ozone contents (TOC) in Antarctica are usually observed in the first week of October. The formation of ozone depletion begins approximately in the second half of August, culminates in the first half of October, and dissolves in November. During the ozone depletion, the average ozone concentration varied at the time of its culmination in October from the original value over 300 Dobson Units (DU) in 1950s and 1960s to a level between 100 and 150 DU in 1990-2000 (see Láska *et al.* (2009)). One DU is set as a 0.001 mm strong

layer of ozone under the pressure 1013 hPa and temperature 273 K.

One of the issues resolved within the Czech–Ukrainian scientific cooperation implemented on the Vernadsky Station in Antarctica is the measurement of total ozone content (TOC) in the stratosphere. The Vernadsky station is located on the west coast of Antarctic peninsula (65°S, 64°W). These data were obtained from ground measurements predominantly taken with the Dobson No 031 spectrophotometer. Data can be found at UAC (2012).

The data sets were processed as time points measuring the average daily amount of ozone. In order to analyze these data we have to take into account the autocorrelation structure of errors on such time series. We focus on kernel regression estimators of series of ozone data. These estimators depend on a smoothing parameter and it is well-known that selecting the correct smoothing parameter is difficult in the presence of correlated errors. There exist methods which are modifications of a classical cross-validation method for independent errors (the modified cross-validation method or the partitioned cross-validation method - see Chu and Marron (1991), Härdle and Vieu (1992)).

In the present paper we develop a new flexible plug-in approach for estimating the optimal smoothing parameter. The utility of this method is illustrated through a simulation study and application to TOC data measured in periods August to April 2004-2005, 2005-2006, 2006-2007.

2. Procedure Development

2.1. Kernel regression model

In nonparametric regression problems we are interested in estimating the mean function $\mathrm{E}(Y|x) = m(x)$ from a set of observations (x_i, Y_i), $i = 1, \ldots, n$. Many methods such as kernel methods, regression splines and wavelet methods are currently available. The papers in this filed have been mostly focused on case where an unknown function m is hidden by a certain amount of a white noise. The aim of a regression analysis is to remove the white noise and produce a reasonable approximation to the unknown function m.

Consider now the case when the noise is no longer white and instead contains a certain amount of a structure in the form of correlation. In particular, if data sets have been recorded over time from one object under a study, it is very likely that another response of the object will depend on its previous response. In this context we will be dealing with a time series case, where design points are fixed and equally spaced and thus our model takes the form

$$Y_i = m(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

and ε_i is an unknown ARMA process, i.e.,

$$\begin{aligned} \mathrm{E}(\varepsilon_i) &= 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n, \\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) &= \gamma_{|i-j|} = \sigma^2 \rho_{|i-j|}, \quad \mathrm{corr}(\varepsilon_i, \varepsilon_j) = \rho_{|i-j|} \end{aligned} \tag{2}$$

and the stationary process

$$\gamma_0 = \sigma^2, \quad \rho_t = \frac{\gamma_t}{\gamma_0},$$

where ρ_t is an autocorrelation function and γ_t is an autocovariance function. We consider the simplest situation (Opsomer *et al.* (2001), Chu and Marron (1991))

$$\rho_{t/n} = \rho_t.$$

Simple and the most widely used regression smoothers are based on kernel methods (see e.g. monographs Müller (1987), Härdle (1990), Wand and Jones (1995)). These methods are local weighted averages of the response Y. They depend on a kernel which plays the role of a weighted function, and a smoothing parameter called a bandwidth which controls the smoothness of the estimate.

Appropriate kernel regression estimators were proposed by Priestley and Chao (1972), Nadaraya (1964) and Watson (1964), Stone (1977), Cleveland (1979) and Gasser and Müller (1979).

These estimators were shown to be asymptotically equivalent (Lejeune (1985), Müller (1987), Wand and Jones (1995)) and without the lost of generality we consider the Nadaraya–Watson (NW) estimators $\widehat{m}$ of m. The NW estimator of m at the point $x \in (0,1)$ is defined as

$$\widehat{m}(x,h) = \frac{\sum_{i=1}^{n} K_h(x_i - x) Y_i}{\sum_{i=1}^{n} K_h(x_i - x)}, \tag{3}$$

for a kernel function K, where $K_h(.) = \frac{1}{h}K(\frac{\cdot}{h})$, and h is a nonrandom positive number $h = h(n)$ called the bandwidth.

Before studying the statistical properties of $\widehat{m}$ several additional assumptions on the statistical model and the parameters of the estimator are needed:

I. Let $m \in \mathrm{C}^2[0,1]$.

II. Let K be a real valued function continuous on $\mathbb{R}$ and satisfying the conditions:

(i) $|K(x) - K(y)| \le L|x-y|$ for a constant $L > 0$, $\forall x, y \in [-1,1]$,

(ii) $\mathrm{support}(K) = [-1,1], \quad K(-1) = K(1) = 0,$

(iii) $\int_{-1}^{1} x^j K(x)\mathrm{dx} = \begin{cases} 1 & j = 0, \\ 0 & j = 1, \\ \beta_2 \neq 0 & j = 2. \end{cases}$

Such a function is called a kernel of order 2 and a class of these kernels is denoted as S_{02}.

III. Let $h = h(n)$ be a sequence of nonrandom positive numbers, such that $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

IV. $\lim_{n\to\infty} \sum_{k=1}^{\infty} |\rho_k| < \infty, \quad$ i.e., $R = \sum_{k=1}^{\infty} \rho_k$ exists,

V. $\frac{1}{n} \sum_{k=1}^{\infty} k|\rho_k| = 0.$

Remark. The well-known kernels are, e.g.,

Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}$,

quartic kernel $K(x) = \frac{3}{4}(1-x^2)^2 I_{[-1,1]}$,

triweight kernel $K(x) = \frac{35}{32}(1-x^2)^2 I_{[-1,1]}$,

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

where $I_{[-1,1]}$ is an indicator function.

Though the Gaussian kernel does not satisfy the assumption II.(ii), it is very popular in many applications.

There is no problem with a choice of a suitable kernel. Symmetric probability density functions are commonly used (see Remark above). But choosing the smoothing parameter is a crucial problem in all kernel estimates. The literature on bandwidth selections is quite extensive in case of independent errors.

It is well known that when the kernel method is used to recover m, that correlated errors trouble bandwidth selection severely (see Altman (1990), Opsomer *et al.* (2001)). De Brabanter *et al.* (2010) developed a bandwidth selection procedure based on bimodal kernels which successfully removes the error correlation without requiring any prior knowledge about its structure.

The global quality of the estimate $\widehat{m}$ can be expressed by means of the Mean Integrated Squared Error (Altman (1990), Opsomer *et al.* (2001)). However more mathematically tractable is the Asymptotic Mean Integrated Squared Error (AMISE):

$$\mathrm{AMISE}(\widehat{m},h) = \underbrace{\frac{V(K)}{nh}S}_{\mathrm{AIV}(\widehat{m},h)} + \underbrace{\frac{\beta_2^2}{4}h^4 A_2}_{\mathrm{AISB}(\widehat{m},h)},$$

where

$$V(K) = \int K^2(x)\mathrm{dx}, \quad S = \sigma^2(1+2\sum_{k=1}^{\infty}\rho_k) = \sigma^2(1+2R), \quad A_2 = \int_0^1 m''(x)^2\mathrm{dx}.$$

The first term is called the asymptotic integrated variance (AIV) and the second one the asymptotic integrated squared bias (AISB). This decomposition provides an easier analysis and interpretation of the performance of the kernel regression estimator.

Using a standard procedure of mathematical analysis one can easily find that the bandwidth h_{opt} minimizing the AMISE is given by the formula

$$h_{opt} = \Big(\frac{V(K)S}{n\beta_2^2 A_2}\Big)^{1/5} = O(n^{-1/5}). \tag{4}$$

This formula provides a good insight into an optimal bandwidth, but unfortunately it depends on the unknown S and A_2.

Let us explain the impact of assuming an uncorrelated model.

If $R > 0$ (error correlation is positive), then $\mathrm{AIV}(\widehat{m}, h)$ is larger than in the corresponding uncorrelated case and $\mathrm{AMISE}(\widehat{m}, h)$ is minimized by a value h that is larger than in the uncorrelated case. It means that assuming wrongly uncorrelated errors causes that the bandwidth becomes too small.

If $R < 0$ (error correlation is negative), then $\mathrm{AIV}(\widehat{m}, h)$ is smaller and $\mathrm{AMISE}(\widehat{m}, h)$ optimal bandwidth is smaller than in the uncorrelated case.

In the next section the choosing of parameters S and A_2 will be treated.

2.2. Choosing the parameters

There are a number of data-driven bandwidth selection methods, but it can be shown that they fail in the case of correlated errors.

Among the earliest fully automatic and consistent bandwidth selectors are those based on cross-validation ideas. The cross-validation method employs an objective function

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}\Big(\widehat{m}_{-j}(x_j, h) - Y_j\Big)^2, \tag{5}$$

where $\widehat{m}_{-j}(x_j, h)$ is the estimate of $\widehat{m}(x_j, h)$ with x_j deleted, i.e., the leave-one-out estimator. The estimate of h_{opt} is then

$$\widehat{h}_{opt} = \arg\min_{h \in H_n} CV(h),$$

where $H_n = [an^{-1/5}, bn^{-1/5}], \quad 0 < a < b < \infty$.

Remark. If the design points are equally spaced then a recommended interval is $[\frac{1}{n}, 1)$.

However, this ordinary method is not suitable in the case of correlated observations. As it was shown in the papers Altman (1990) and Opsomer *et al.* (2001), if the observations are positively correlated, then the CV method produces too small a bandwidth, and if the observations are negatively correlated, then the CV method produces a large bandwidth.

We demonstrate this fact by the following example.

Consider the regression model (1), where

$$m(x) = \cos(3.15\pi x), \quad \varepsilon_i = \phi\varepsilon_{i-1} + e_i,$$
$$e_i \text{ – i.i.d. normal random variables } \mathrm{N}(0, \sigma^2),$$
$$\varepsilon_1 \text{ – } \mathrm{N}(0, \sigma^2/(1-\phi^2)),$$
$$\phi = 0.6, \quad \sigma = 0.5,$$

i.e, the regression errors are AR(1) process.

Figure 1 shows the result obtained by the CV method. It is evident, that the estimate is undersmoothed.

In order to overcome this problem, modified and partitioned CV methods were proposed by Härdle and Vieu (1992) and Chu and Marron (1991), respectively.

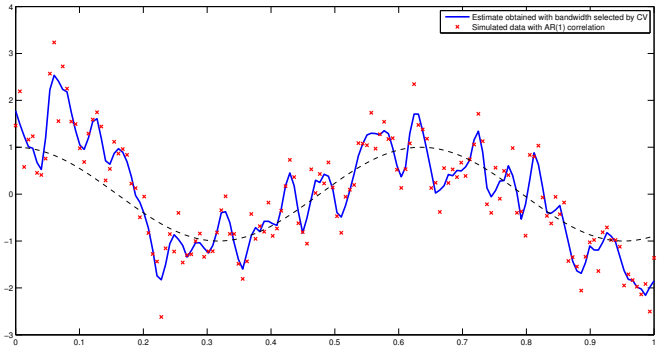


Figure 1: The estimate of simulated data with AR(1) errors

The modified cross-validation (MCV) method is a "leave-$(2l+1)$-out" version of CV ($l \geq 0$). The idea consists in minimizing of the modified cross-validation score:

$$CV_l(h) = \frac{1}{n}\sum_{j=1}^{n}\Big(\widehat{m}_{-j}(x_j,h) - Y_j\Big)^2, \tag{6}$$

where $\widehat{m}_{-j}(x_j,h)$ is the "leave-$(2l+1)$-out" estimate of $\widehat{m}(x_j,h)$, i.e., the observations (x_{j+i},Y_{j+i}), $-l \leq i \leq l$ are left out in constructing $\widehat{m}(x_j,h)$.

Then

$$\widehat{h}_{MCV} = \arg\min_{h\in H_n} CV_l(h).$$

The principle of the partitioned cross-validation method (PCV) can be described as follows.

For any natural number $g \geq 1$, the PCV involves splitting the observations into g groups by taking every g-th observation, calculating the ordinary cross-validation score $CV_{0,k}(h)$ of the k-th group of observations separately, for $k = 1, 2, \ldots, g$, and minimizing the average of these ordinary cross-validation scores

$$CV^*(h) = \frac{1}{g}\sum_{k=1}^{g} CV_{0,k}(h). \tag{7}$$

Let $\widehat{h}^*_{CV}$ stand for the minimizer of $CV^*(h)$:

$$\widehat{h}^*_{CV} = \arg\min_{h\in H_n} CV^*(h).$$

Since $\widehat{h}^*_{CV}$ is appropriate for the sample size n/g, the partitioned cross-validated bandwidth $\widehat{h}_{PCV(g)}$ is defined to be rescaled $\widehat{h}^*_{CV}$:

$$\widehat{h}_{PCV(g)} = g^{-1/5}\widehat{h}^*_{CV}.$$

When $g = 1$, the PCV is an ordinary cross-validation.

Remark. The number of subgroups is g and the number of observations in each group is $\eta = n/g$. If n is not a multiplier of g, then the values Y_j, $1 \le j \le g[n/g]$ are applied and the rest of the observations are dropped out ($[n/g]$ is the highest integer less or equal to n/g).

The asymptotic behavior of $\widehat{h}_{MCV(l)}$ and $\widehat{h}_{PCV(g)}$ was studied in the paper by Chu and Marron (1991). Furthemore we focus on the PCV method.

The PCV method needs to determine the factor g. A possible approach for the practical choice of g is based on an analogue of the mean squared error. Using the asymptotic variance and the asymptotic mean of $\widehat{h}_{PCV(g)}/h_{opt}$, the asymptotic mean squared error (AMSE) of this ratio is defined by

$$AMSE\big(\widehat{h}_{PCV(g)}/h_{opt}\big) = n^{-1/5}\mathrm{VAR}_{PCV(g)} + \big[C_{PCV(g)}/C - 1\big]^2, \tag{8}$$

where $\mathrm{VAR}_{PCV(g)}$, $C_{PCV(g)}$, C depend on γ_k, K, A_2 (see Chu and Marron (1991)).

Theoretically, if there exists a value $\widehat{g}$ which minimizes AMSE over $g \ge 1$, then this value is taken as the optimal value of g in the sense of AMSE:

$$\widehat{g}_{opt} = \arg\min_{g \ge 1} \mathrm{AMSE}\big(\widehat{h}_{PCV(g)}/h_{opt}\big).$$

Unfortunately the minimization of AMSE also depends on the unknown γ_k and A_2.

As far as the estimation of the variance component S is concerned, a common approach is the following (see e.g. Herrmann *et al.* (1992), Hart (1991), Opsomer *et al.* (2001), Chu and Marron (1991)):

$$\begin{aligned} \widehat{S} =& \hat{\gamma}_0\Big(1 + 2\sum_{k=1}^{n-1}\hat{\rho}_k\Big), \quad \hat{\gamma}_0 = \hat{\sigma}^2, \quad \hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}, \\ \hat{\gamma}_k =& \frac{1}{n-k}\sum_{t=1}^{n-k}\Big(Y_t - \overline{Y}\Big)\Big(Y_{t+k} - \overline{Y}\Big), \quad k = 0, \ldots, n-1. \end{aligned} \tag{9}$$

Nevertheless there is still a problem of how to estimate A_2. In paper Chu and Marron (1991) a simulation study was only conducted and no idea of estimating A_2 was given there.

We complete this method by adding a suitable estimate of A_2 and recommend to use an estimate of A_2 proposed by Koláček (2008). By means of the Fourier transformation he derived a suitable estimate $\widehat{A_2}$ of A_2. Therefore, A_2 in the AMSE formula is replaced by $\widehat{A_2}$. This approach is commonly known as a plug-in method.

Plug-in methods are also commonly used for selecting the bandwidth in the kernel regression. But these methods perform badly when the errors are correlated. In the paper Herrmann *et al.* (1992) a modified version of an existing plug-in bandwidth selectors is proposed. This method is based on the Gasser–Müller estimator of the second derivative and an iterative process is constructed. It is shown that under some additional assumptions this iterative process converges to a suitable estimate of the optimal bandwidth.

However we do not use this iterative method and propose to directly plug-in A_2 in the formula (4). This new version of a plug-in method is denoted as PI and the bandwidth estimate takes the form:

$$\widehat{h}_{PI} = \Big(\frac{V(K)\widehat{S}}{n\beta_2^2\widehat{A_2}}\Big)^{1/5}.$$

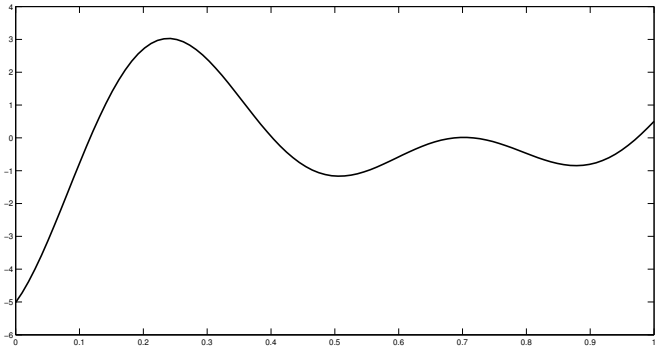


Figure 2: The regression function $m(x)$

$h_{opt} = 0.759$		
	$\mathrm{E}(\widehat{h})$	$\mathrm{std}(\widehat{h})$
PCV	0.1927	0.0649
PI	0.1513	0.0083

Table 1: The estimates $\widehat{h}$

We would like to point out the computational aspect of the plug-in method. It has preferable properties to classical methods, because it does not need any additional calculations such as the PCV method (see Koláček (2008) for details).

3. Case study

We conduct a simulation study to compare the PCV method and the PI method. The Epanechnikov kernel is used both in simulations and in applications.

Consider the regression model (1), where

$$m(x) = \frac{-6\sin 11x + 5}{\mathrm{cotg}(x-7)}, \quad \varepsilon_i = \phi\varepsilon_{i-1} + e_i$$

e_i – i.i.d. normal random variables $\mathrm{N}(0, \sigma^2)$

ε_1 – $\mathrm{N}(0, \sigma^2/(1-\phi^2))$

$\phi = 0.6, \quad \sigma = 0.5,$

for $i = 1, \ldots, n = 100$.

The graph of the regression function m is presented in Figure 2.

One hundred series are generated. For each data set, the optimal bandwidth is estimated by the PCV and PI method. Table 1 shows the comparison of means and standard deviations for these two methods.

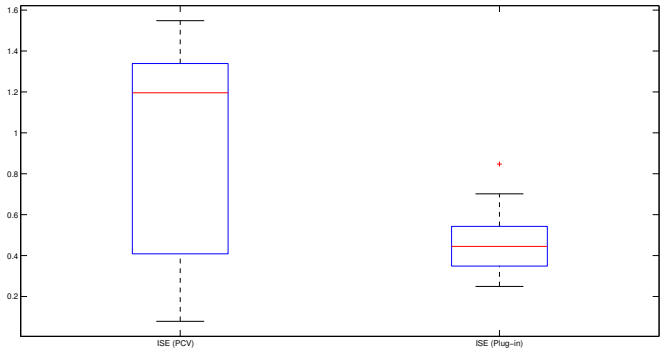


Figure 3: $ISE(\widehat{m}(.,h)) = \int_0^1 \Big(\widehat{m}(x,h) - m(x)\Big)^2 \mathrm{dx}$.

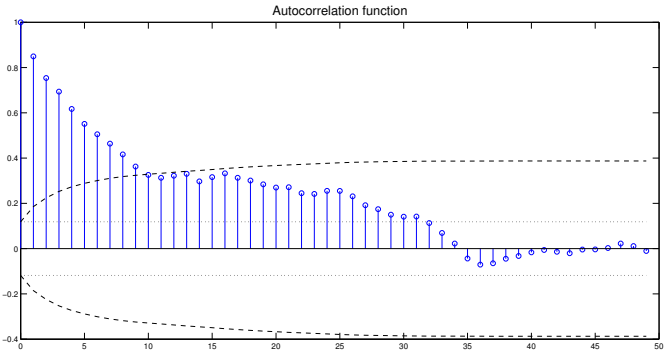


Figure 4: The autocorrelation function of the data set August 2004 – April 2005

The Integrated Square Error (ISE) is calculated for each estimate $\widehat{m}(.,h)$:

$$ISE(\widehat{m}(.,h)) = \int_0^1 \Big(\widehat{m}(x,h) - m(x)\Big)^2 \mathrm{dx}$$

for both PCV and PI methods and the results are displayed by means of the boxplots in Figure 3.

4. Results and discussion

In this section we apply the methods described above to ozone data. We analyze data which were measured in the period August to April in years 2004–2005, 2005–2006, 2006–2007. The sample size is $n = 273$ days. The observations are correlated as it can be seen in Figure 4. We transform data to the interval [0,1] and use the PCV method and the PI method to get the optimal bandwidth. Then we re-transform the bandwidth to the original sample and obtain the final kernel estimate.

Kernel estimates based on the PCV and PI methods are presented in Figure 6, Figure 7, or in Figure 8, respectively.

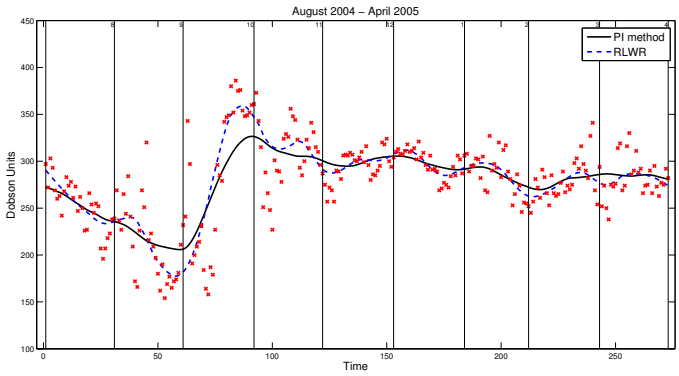


Figure 5: RLWR estimate with span = 40 (dashed line) and PI estimate with the bandwidth = 17.8 (solid line).

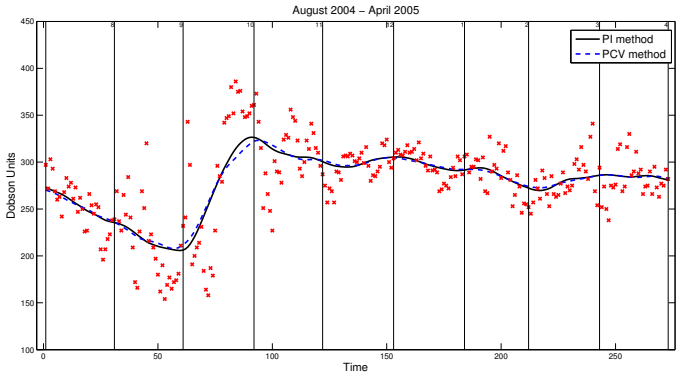


Figure 6: PCV estimate with the bandwidth = 20.9 (dashed line) and PI estimate with the bandwidth = 17.8 (solid line).

In paper Kalvová and Dubrovský (1995) the robust locally wighted regression (RLWR) is employed for data processing of TOC. They recommended to optimize h subjectively. This approach needs an experience and a special knowledge of the given data sets. The advantage of our methods consists in more complex approach. These methods are general and they allow to choose the value of h automatically. We used their methodology for data April 2004 - August 2005 and the comparison of the estimate obtained by the PI method and by the robust locally weighted regression can be seen in Figure 5. The PI method yields a rather oversmoothed estimate.

Our experience shows that both methods could be considered as a suitable tool for the choice of the bandwidth. But it seems that the PI method is sufficiently reliable and less time consuming than the PCV method.

Presented methods can be applied to other time series not only in environmetrics but also in economics or other fields.

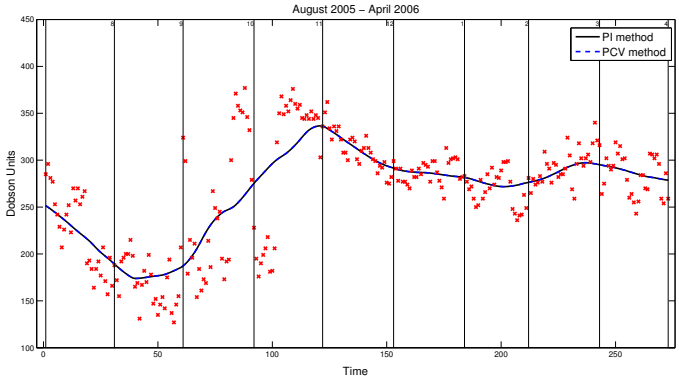


Figure 7: PCV estimate with the bandwidth = 20.4 (dashed line) and PI estimate with the bandwidth = 21.9 (solid line).

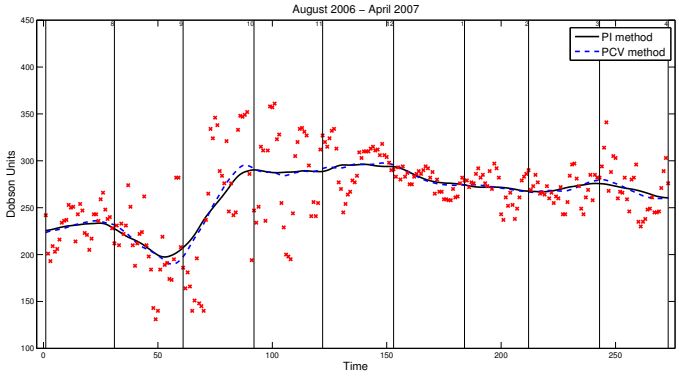


Figure 8: PCV estimate with the bandwidth = 17.2 (dashed line) and PI estimate with the bandwidth = 22.3 (solid line).

Acknowledgments

The research was supported by The Jaroslav Hájek Center for Theoretical and Applied Statistics (MŠMT LC 06024). The work was supported by the Student Project Grant at Masaryk university, rector's programme no. MUNI/A/1001/2009.

References

Altman N (1990). "Kernel Smoothing of Data With Correlated Errors." *Journal of the American Statistical Association*, **85**, 749–759.

Chu CK, Marron JS (1991). "Choosing a Kernel Regression Estimator." *Statistical Science*, **6**(4), 404–419. ISSN 08834237.

Cleveland WS (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association*, **74**(368), 829–836. ISSN 01621459.

De Brabanter K, De Brabanter J, Suykens J, De Moor B (2010). "Kernel Regression with Correlated Errors." *Computer Applications in Biotechnology*, pp. 13–18.

Gasser T, Müller HG (1979). "Kernel estimation of regression functions." In T Gasser, M Rosenblatt (eds.), *Smoothing Techniques for Curve Estimation*, volume 757 of *Lecture Notes in Mathematics*, pp. 23–68. Springer Berlin / Heidelberg.

Härdle W (1990). *Applied Nonparametric Regression*. 1st edition. Cambridge University Press, Cambridge.

Härdle W, Vieu P (1992). "Kernel Regression Smoothing of Time Series." *Journal of Time Series Analysis*, **13**(3), 209–232.

Hart JD (1991). "Kernel Regression Estimation with Time Series Errors." *Journal of the Royal Statistical Society*, **53**, 173–187.

Herrmann E, Gasser T, Kneip A (1992). "Choice of Bandwidth for Kernel Regression when Residuals are Correlated." *Biometrika*, **79**, 783–795.

Kalvová J, Dubrovský M (1995). "Assessment of the Limits Between Which Daily Average Values of Total Ozone Can Normally Vary." *Meteorol. Bulletin*, **48**, 9–17.

Koláček J (2008). "Plug-in Method for Nonparametric Regression." *Computational Statistics*, **23**(1), 63–78. ISSN 0943-4062.

Láska K, Prošek P, Budík L, Budíková M, Milinevsky G (2009). "Prediction of Erythemally Effective UVB Radiation by Means of Nonlinear Regression Model." *Environmetrics*, **20**(6), 633–646.

Lejeune M (1985). "Estimation Non-paramétrique par Noyaux: Régression Polynomiale Mobile." *Revue de Statistique Appliquée*, **33**(3), 43–67.

Müller HG (1987). "Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting." *Journal of the American Statistical Association*, **82**(397), 231–238. ISSN 01621459.

Nadaraya EA (1964). "On Estimating Regression." *Theory of Probability and its Applications*, **9**(1), 141–142.

Opsomer J, Wang Y, Yang Y (2001). "Nonparametric Regression with Correlated Errors." *Statistical Science*, **16**(2), 134–153.

Priestley MB, Chao MT (1972). "Non-Parametric Function Fitting." *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(3), 385–392. ISSN 00359246.

Stone CJ (1977). "Consistent Nonparametric Regression." *The Annals of Statistics*, **5**(4), 595–620. ISSN 00905364.

UAC (2012). "World Ozone and Ultraviolet Radiation Data Centre (WOUDC) [data]." URL http://www.woudc.org.

Wand M, Jones M (1995). *Kernel smoothing*. Chapman and Hall, London.

Watson GS (1964). "Smooth Regression Analysis." *Sankhya - The Indian Journal of Statistics, Series A*, **26**(4), 359–372. ISSN 0581572X.

Affiliation:

Ivana Horová
Masaryk University
Department of Mathematics and Statistics
Brno, Czech Republic
E-mail: horova@math.muni.cz
URL: https://www.math.muni.cz/~horova/

Journal of Statistics: Advances in Theory and Applications
Volume 7, Number 1, 2012, Pages 1-23

LIFT-BASED QUALITY INDEXES FOR CREDIT SCORING MODELS AS AN ALTERNATIVE TO GINI AND KS

MARTIN ŘEZÁČ and JAN KOLÁČEK

Department of Mathematics and Statistics
Masaryk University
Kotlářská 2, 61137 Brno
Czech Republic
e-mail: mrezac@math.muni.cz

Abstract

Assessment of risk associated with the granting of credits is very successfully supported by techniques of credit scoring. To measure the quality, in the sense of the predictive power, of the scoring models, it is possible to use quantitative indexes such as the Gini index (Gini), the K-S statistic (KS), the *c*-statistic, and lift. They are used for comparing several developed models at the moment of development as well as for monitoring the quality of the model after deployment into real business. The paper deals with the aforementioned quality indexes, their properties and relationships. The main contribution of the paper is the proposal and discussion of indexes and curves based on lift. The curve of ideal lift is defined; lift ratio (LR) is defined as analogous to Gini index. Integrated relative lift (IRL) is defined and discussed. Finally, the presented case study shows a case when LR and IRL are much more appropriate to use than Gini and KS.

2010 Mathematics Subject Classification: 62P05, 90B50.

Keywords and phrases: credit scoring, quality indexes, Gini index, lift, lift ratio, integrated relative lift.

Received February 14, 2012

1. Introduction

Banks and other financial institutions receive thousands of credit applications every day (in the case of consumer credits, it can be tens or hundreds of thousands every day). Since it is impossible to process them manually, automatic systems are widely used by these institutions for evaluating the credit reliability of individuals, who ask for credit. The assessment of the risk associated with the granting of credits has been underpinned by one of the most successful applications of statistics and operations research: credit scoring.

Credit scoring is the set of predictive models and their underlying techniques that aid financial institutions in the granting of credits. These techniques decide who will get credit, how much credit they should get, and what further strategies will enhance the profitability of the borrowers to the lenders. Credit scoring techniques assess the risk in lending to a particular client. They do not identify "good" or "bad" (negative behaviour is expected, e.g., default) applications on an individual basis, but forecast the probability that an applicant with any given score will be "good" or "bad". These probabilities or scores, along with other business considerations such as expected approval rates, profit, churn, and losses, are then used as a basis for decision making.

Several methods connected to credit scoring have been introduced during last six decades. The most well-known and widely used are logistic regression, classification trees, the linear programming approach, and neural networks.

The methodology of credit scoring models and some measures of their quality have been discussed in surveys including Hand and Henley [7], Thomas [14] or Crook et al. [4]. Even if ten years ago the list of books devoted to the issue of credit scoring was not extensive, the situation has improved in the last decade. In particular, this list now includes Anderson [1], Crook et al. [4], Siddiqi [11], Thomas et al. [15], and Thomas [16].

The aim of this paper is to give an overview of widely used techniques used to assess the quality of credit scoring models, to discuss the properties of these techniques, and to extend some known results. We review widely used quality indexes, their properties and relationships. The main part of the paper is devoted to lift. The curve of ideal lift is defined; lift ratio is defined as analogous to Gini index. Integrated relative lift is defined and discussed.

2. Measuring the Quality

We can consider two basic types of quality indexes: first, indexes based on a cumulative distribution function like the Kolmogorov-Smirnov statistic, Gini index or lift; second, indexes based on a likelihood density function like the mean difference (Mahalanobis distance) or informational statistic. For further available measures and appropriate remarks, see Wilkie [17], Giudici [6] or Siddiqi [11].

Assume that the realization $s \in R$ of a random variable S (score) is available for each client and put the following markings:

$$D = \begin{cases} 1, & \text{client is good,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Distribution functions, respectively, their empirical forms, of the scores of good (bad) clients are given by

$$F_{n.GOOD}(a) = \frac{1}{n}\sum_{i=1}^{N} I(s_i \le a \wedge D = 1),$$

$$F_{m.BAD}(a) = \frac{1}{m}\sum_{i=1}^{N} I(s_i \le a \wedge D = 0), \quad a \in [L, H], \tag{2}$$

where s_i is the score of i-th client, n is the number of good clients, m is the number of bad clients, and I is the indicator function, where $I(\text{true}) = 1$ and $I(\text{false}) = 0$. L is the minimum value of a given score, H is the maximum value. The empirical distribution function of the scores of all clients is given by

$$F_{N.ALL}(a) = \frac{1}{N}\sum_{i=1}^{N} I(s_i \le a), \quad a \in [L, H], \tag{3}$$

where $N = n + m$ is the number of all clients. We denote the proportion of bad (good) clients by

$$p_B = \frac{m}{n+m}, \qquad p_G = \frac{n}{n+m}. \tag{4}$$

An often-used characteristic in describing the quality of the model (scoring function) is the Kolmogorov-Smirnov statistic (K-S or KS). It is defined as

$$KS = \max_{a \in [L, H]} |F_{m.BAD}(a) - F_{n.GOOD}(a)|. \tag{5}$$

It takes values from 0 to 1. Value 0 corresponds to a random model, value 1 corresponds to the ideal model. The higher the KS, the better the scoring model.

The Lorenz curve (LC), sometimes called the ROC curve (receiver operating characteristic curve), can also be successfully used to show the discriminatory power of a scoring function, i.e., the ability to identify good and bad clients. The curve is given parametrically by

$$\begin{aligned} x &= F_{m.BAD}(a), \\ y &= F_{n.GOOD}(a), \qquad a \in [L, H]. \end{aligned} \tag{6}$$

Each point of the curve represents some value of a given score. If we consider this value as a cut-off value, we can read the proportion of rejected bad and good clients. An example of a Lorenz curve is given in Figure 1. We can see that by rejecting 20% of good clients, we also reject 50% of bad clients at the same time.

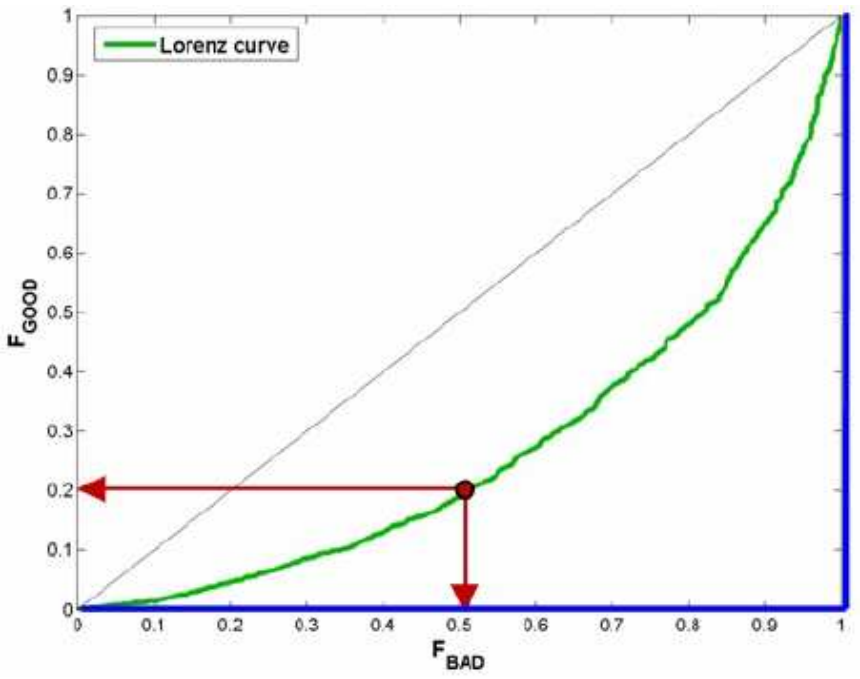


Figure 1. Lorenz curve (ROC).

The LC for a random scoring model is represented by the diagonal line from $[0, 0]$ to $[1, 1]$. It is the polyline from $[0, 0]$ through $[1, 0]$ to $[1, 1]$ in the case of an ideal model. It is obvious that the closer the curve is to the bottom right corner, the better is the model.

The definition and name (LC) is consistent with Müller and Rönz [8]. One can find the same definition of the curve, but called ROC, in Thomas et al. [15]. Siddiqi [11] used the name ROC for a curve with reversed axes and LC for a curve with the CDF of bad clients on the vertical axis and the CDF of all clients on the horizontal axis. This curve is also called the CAP (*cumulative accuracy profile*) or lift curve, see Sobehart et al. [12] or Thomas [16]. Furthermore, it is called a *gains chart* in the field of marketing; see Berry and Linoff [2]. An example of CAP is displayed in Figure 2. The ideal model is now represented by a polyline from $[0, 0]$

through $[p_B, 1]$ to $[1, 1]$. The advantage of this figure is that, one can easily read the proportion of rejected bads against the proportion of all rejected. For example, in the case of Figure 2, we can see that if we want to reject 70% of bads, we have to reject about 40% of all applicants.

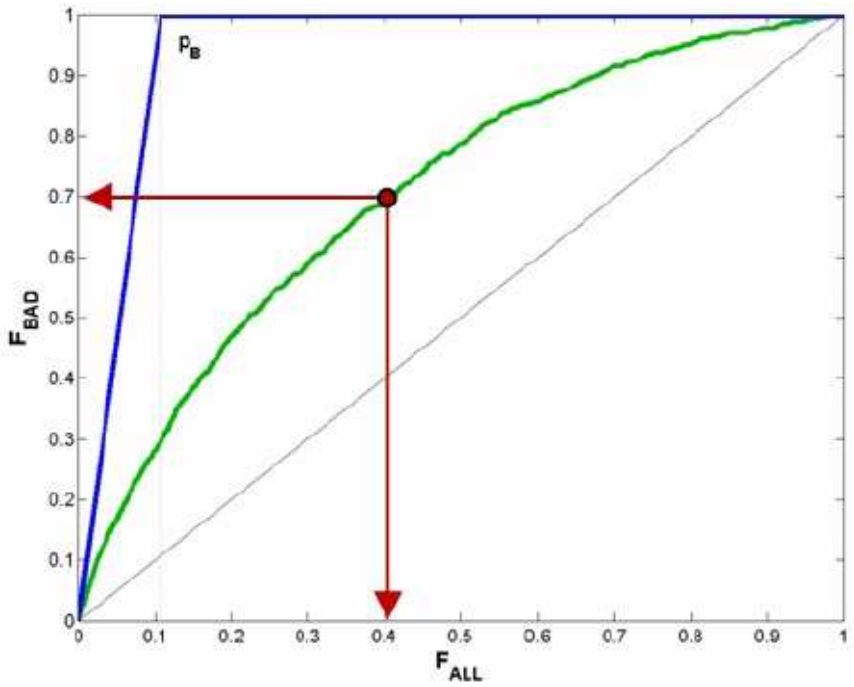


Figure 2. CAP.

In connection to LC, we consider the next quality measure, the Gini index. This index describes a global quality of the scoring model. It takes values from 0 to 1 (it can take negative values for contrariwise models). The ideal model, i.e., the scoring function that perfectly separates good and bad clients, has a Gini index equal to 1. On the other hand, a model that assigns a random score to the client, has a Gini index equal to 0. It can be shown that the Gini index is greater than or equal to KS for any scoring model. Using Figure 3, it can be defined as follows:

$$Gini = \frac{A}{A + B} = 2A. \tag{7}$$

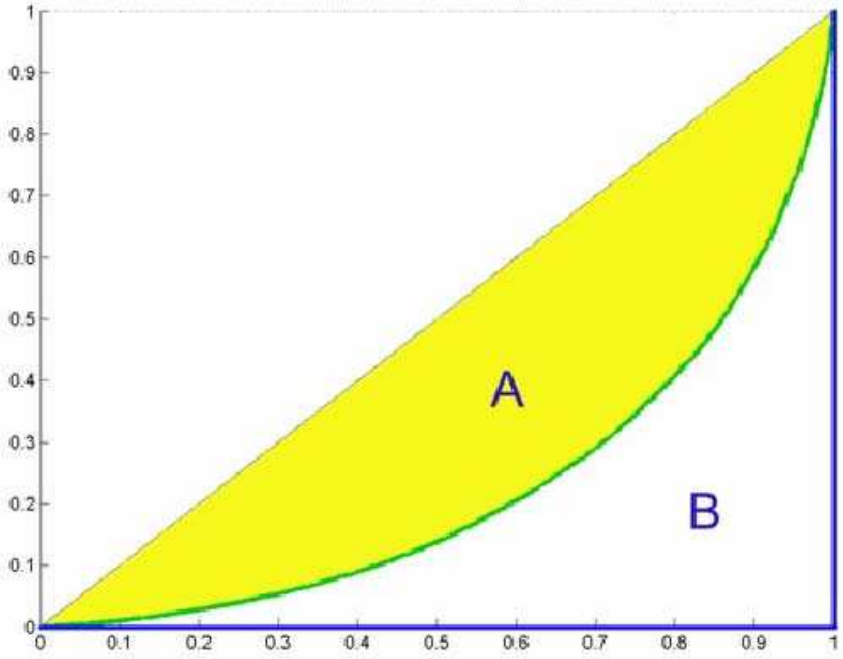


Figure 3. Lorenz curve, Gini index.

This means that, we compute the ratio of the area between the curve and the diagonal (which represents a random model) to the area between the ideal model's curve and the diagonal. Since the axes describe a unit square, the area $A + B$ is always equal to 0.5. Therefore, we can compute the Gini as two times the area A. Using previous markings, the computational formula of the Gini index is given by

$$Gini = 1 - \sum_{k=2}^{N} [(F_{m.BAD_k} - F_{m.BAD_k-1}) \times (F_{n.GOOD_k} + F_{n.GOOD_k-1})], \quad (8)$$

where $F_{m.BAD_k}(F_{n.GOOD_k})$ is the k-th vector value of the empirical distribution function of bad (good) clients. For further details, see Anderson [1] or Xu [18]. The Gini index is a special case of Somers' D (Somers [13]), which is an ordinal association measure. According to Thomas [16], one can calculate the Somers' D as

$$D_S = \frac{\sum_i g_i \sum_{j<i} b_j - \sum_i g_i \sum_{j>i} b_j}{n \cdot m}, \tag{9}$$

where $g_i(b_j)$ is the number of goods (bads) in the i-th interval of scores. Furthermore, it holds that D_S can be expressed by the Mann-Whitney U-statistic; see Nelsen [9] for further details.

When we use CAP instead of LC, we can define the accuracy rate (AR); see Thomas [16] or Sobehart et al. [12], where it is called the accuracy ratio. Again, it is defined by the ratio of some areas. We have

$$AR = \frac{\text{Area between CAP curve and diagonal}}{\text{Area between ideal model's CAP and diagonal}}$$

$$= \frac{\text{Area between CAP curve and diagonal}}{0.5(1 - p_B)}. \tag{10}$$

Although the ROC and CAP are not equivalent, it is true that Gini and AR are equal for any scoring model. Proof for discrete scores is given in Engelmann et al. [5]; for continuous scores, one can find it in Thomas [16].

In connection to the Gini index, the c-statistic (Siddiqi [11]) is defined as

$$c\_stat = \frac{1 + Gini}{2}. \tag{11}$$

It represents the likelihood that a randomly selected good client has a higher score than a randomly selected bad client, i.e.,

$$c\_stat = P(s_1 \geq s_2 \,|\, D_1 = 1 \wedge D_2 = 0). \tag{12}$$

It takes values from 0.5, for the random model, to 1, for the ideal model. An alternative name for the c-statistic can be found in the literature. It is known also as Harrell's c, which is a reparameterization of Somers' D (Newson [10]). Furthermore, it is called AUROC, e.g., in Thomas [16] or AUC, e.g., in Engelmann et al. [5].

3. Lift

Another possible indicator of the quality of scoring model is lift, which determines the number of times that, at a given level of rejection, the scoring model is better than random selection (the random model). More precisely, the ratio is the proportion of bad clients with a score less than a (where $a \in [L, H]$) to the proportion of bad clients in the general population. Formally, it can be expressed by

$$Lift(a) = \frac{CumBadRate(a)}{BadRate} = \frac{\dfrac{\sum_{i=1}^{N} I(s_i \le a \wedge D = 0)}{\sum_{i=1}^{N} I(s_i \le a)}}{\dfrac{\sum_{i=1}^{N} I(D = 0)}{\sum_{i=1}^{N} I(D = 0 \vee D = 1)}}$$

$$= \frac{\dfrac{\sum_{i=1}^{N} I(s_i \le a \wedge D = 0)}{\sum_{i=1}^{N} I(s_i \le a)}}{\dfrac{m}{N}}. \tag{13}$$

It can be easily verified that the lift can be equivalently expressed as

$$Lift(a) = \frac{F_{n.BAD}(a)}{F_{N.ALL}(a)}, \quad a \in [L, H]. \tag{14}$$

Now, we would like to discuss the form of the lift function for the case of the ideal model. This is the model for which sets of output scores of bad and good clients are disjoint. So there exists a cut-off point, for which

$$P(S \le a) = \begin{cases} P(S \le a \wedge D = 0), & a \le c, \\ P(D = 0) + P(S \le a \wedge D = 1), & a > c. \end{cases} \tag{15}$$

Thus, we can derive the form of the lift function

$$Lift_{ideal}(a) = \begin{cases} \dfrac{1}{p_B}, & a \le c, \\ \dfrac{1}{F_{N.ALL}(a)}, & a > c. \end{cases} \tag{16}$$

In practice, lift is computed corresponding to 10%, 20%, …, 100% of clients with the worst score (see Coppock [3]). Usually, it is computed by using a table with the numbers of both all and bad clients in given score bands (deciles). An example of such a table is given by Table 1.

Table 1. Lift (absolute and cumulative form) computational scheme

	Absolutely				Cumulatively		
Decile	**#Clients**	**# Bad clients**	**Bad rate**	**Abs. Lift**	**#Bad clients**	**Bad rate**	**Cum. Lift**
1	100	35	35.0%	3.50	35	35.0%	3.50
2	100	16	16.0%	1.60	51	25.5%	2.55
3	100	8	8.0%	0.80	59	19.7%	1.97
4	100	8	8.0%	0.80	67	16.8%	1.68
5	100	7	7.0%	0.70	74	14.8%	1.48
6	100	6	6.0%	0.60	80	13.3%	1.33
7	100	6	6.0%	0.60	86	12.3%	1.23
8	100	5	5.0%	0.50	91	11.4%	1.14
9	100	5	5.0%	0.50	96	10.7%	1.07
10	100	4	4.0%	0.40	100	10.0%	1.00
All	1000	100	10.0%				

It is possible to compute the lift value in each decile (absolute lift in the fifth column in Table 1), but usually, and in accordance with the definition of $Lift(a)$, the cumulative form is used. It holds that the value of lift has an upper limit of $1 / p_B$ and tends to a value of 1 when the score tends to infinity (or to its upper limit). In our case, we can see that the

best possible value of lift is equal to 10. We obtained the value 3.5 in the first decile, which is nothing excellent, but high enough for the model to be considered applicable in practice. Results are further illustrated in Figure 4.

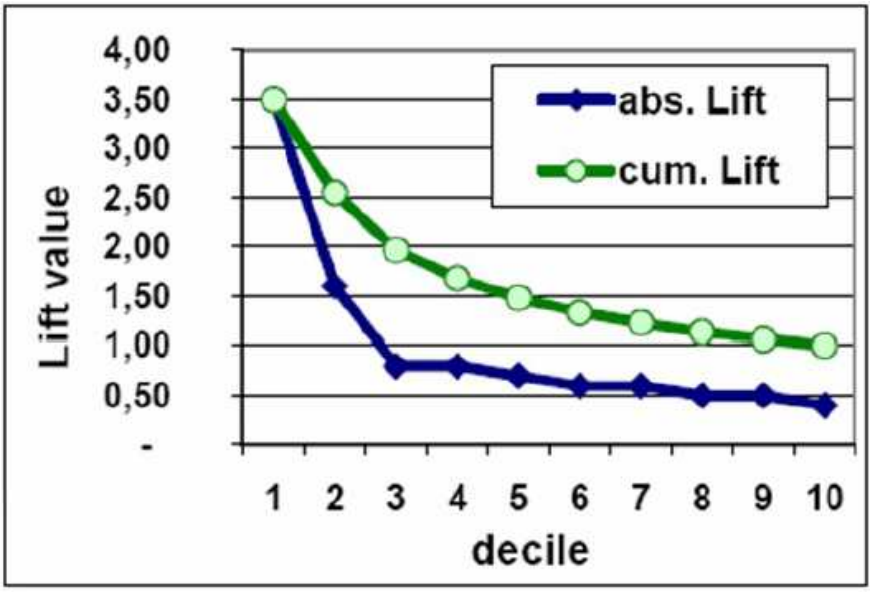


Figure 4. Lift value (absolute and cumulative).

In the context of this approach, we define

$$\begin{aligned} QLift(q) &= \frac{F_{m.BAD}(F^{-1}_{N.ALL}(q))}{F_{N.ALL}(F^{-1}_{N.ALL}(q))} \\ &= \frac{1}{q} F_{m.BAD}(F^{-1}_{N.ALL}(q)), \qquad q \in (0, 1], \end{aligned} \tag{17}$$

where q represents the score level of $100q\%$ of the worst scores and $F^{-1}_{N.ALL}(q)$ can be computed as

$$F^{-1}_{N.ALL}(q) = \min\{a \in [L, H],\ F_{N.ALL}(a) \geq q\}. \tag{18}$$

It can be easily shown that the lift function for the ideal model is now

$$\mathcal{Q}Lift_{ideal}(q) = \begin{cases} \frac{1}{p_B}, & q \in (0,\ p_B], \\ \frac{1}{q}, & q \in (p_{B,}\ 1]. \end{cases} \tag{19}$$

Figure 5, below, gives an example of the lift function for ideal, random, and actual models.

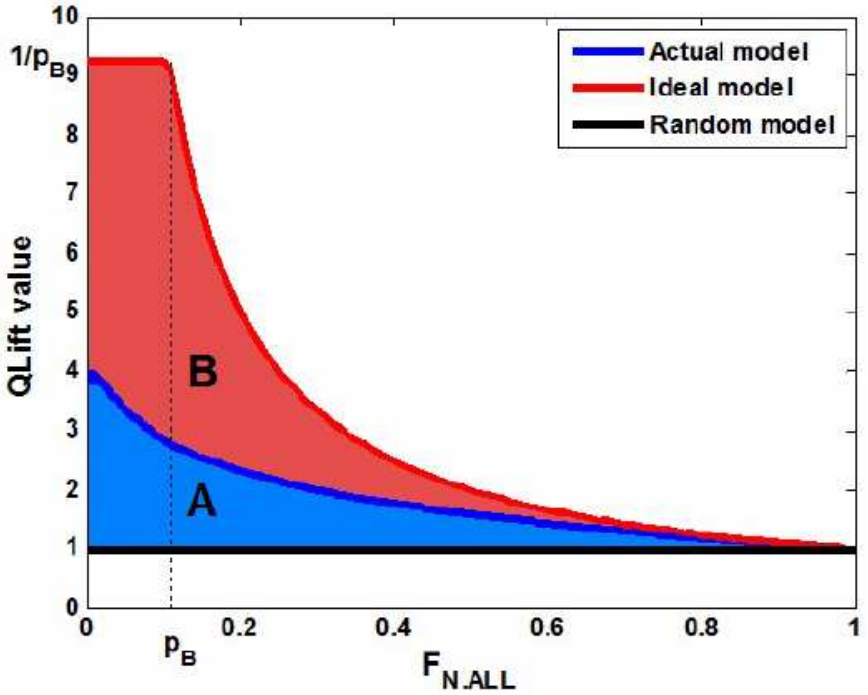


Figure 5. QLift function, lift ratio.

Using the previous Figure 5, we define lift ratio as analogous to Gini index

$$LR = \frac{A}{A+B} = \frac{\int_0^1 \mathcal{Q}Lift(q)dq - 1}{\int_0^1 \mathcal{Q}Lift_{ideal}(q)dq - 1}. \tag{20}$$

It is obvious that, it is a global measure of a model's quality and that it takes values from 0 to 1. Value 0 corresponds to the random model, value 1 matches the ideal model. The meaning of this index is quite simple: the higher, the better. An important feature is that lift ratio allows us to fairly compare two models developed on different data samples, which is not possible with lift.

Since lift ratio compares areas under the lift function corresponding to actual and ideal models, the next concept is focused on the comparison of lift functions themselves. We define the relative lift function by

$$RLift(q) = \frac{QLift(q)}{QLift_{ideal}(q)}, \qquad q \in (0, 1]. \tag{21}$$

An example of this function is presented in Figure 6. The definition domain of the function is $[0, 1]$; the range is a subinterval of $[0, 1]$. The graph starts at point $[q_{\min}, p_B \cdot QLift(q_{\min})]$, where $q_{\min}$ is a positive number near to zero. Then, it falls to a local minimum in point $[p_B, p_B \cdot QLift(p_B)]$ and then rises up to point $[1, 1]$. It is obvious that the graph of relative lift function for a better model is closer to the top line, which represents the function for the ideal model.

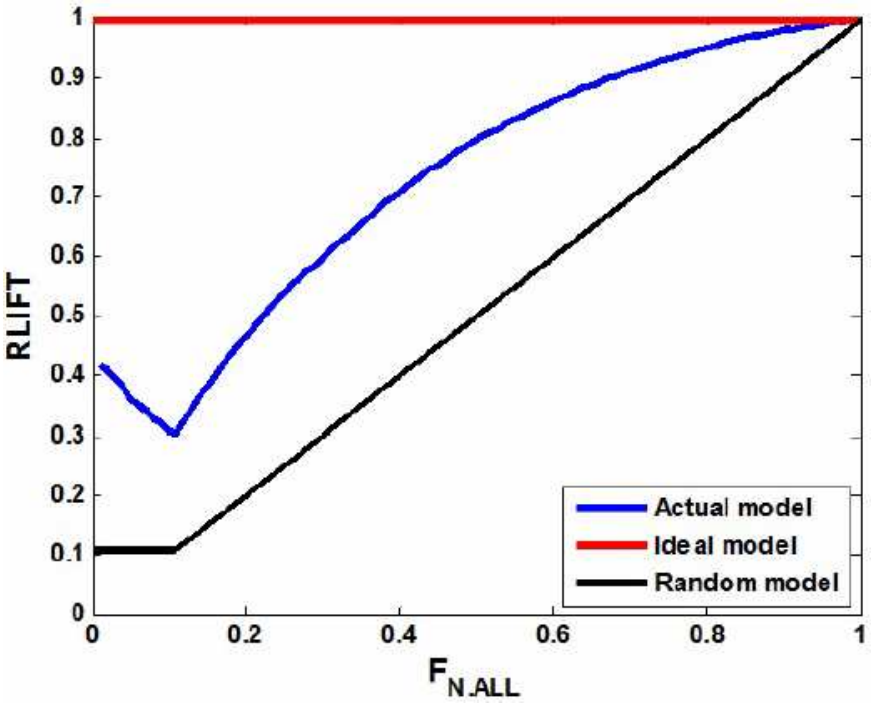


Figure 6. Relative lift function.

Now, it is natural to ask what we obtain when we integrate the relative lift function. We define the integrated relative lift (IRL) by

$$IRL = \int_0^1 RLift(q)dq. \tag{22}$$

It takes values from $0.5 + \frac{p_B^2}{2}$, for the random model, to 1, for the ideal model. Again the following holds: the higher, the better. This global measure of scoring a model's quality has an interesting connection to the *c*-statistic.

We made a simulation with scores generated from a normal distribution. The scores of bad clients had a mean equal to 0 and a variance equal to 1. The scores of good clients had a mean and variance

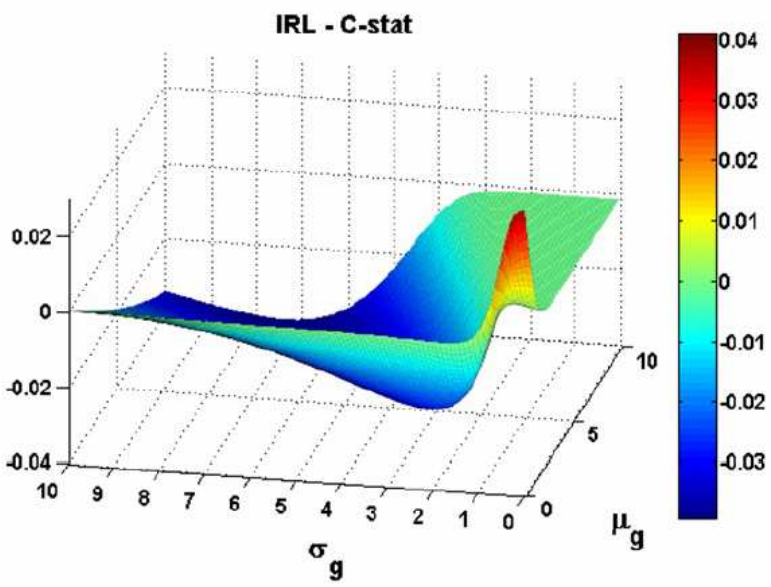
from 0.1 to 10 with a step equal 0.1. The number of samples and sample size were 1000, p_B was equal to 0.1. IRL and the *c*-statistic were computed for each sample and each value of the mean and variance of a good clients' scores. Finally, means of IRL and the *c*-statistic were computed. The results are presented in Figure 7. Part (b) represents the contour plot of the figure in part (a).

The simulation shows that IRL and the *c*-statistic are approximately equal when the variances of good and bad clients are equal. Furthermore, it shows that they significantly differ when the variances are different and the ratio of the mean and variance of good clients is near to 1.

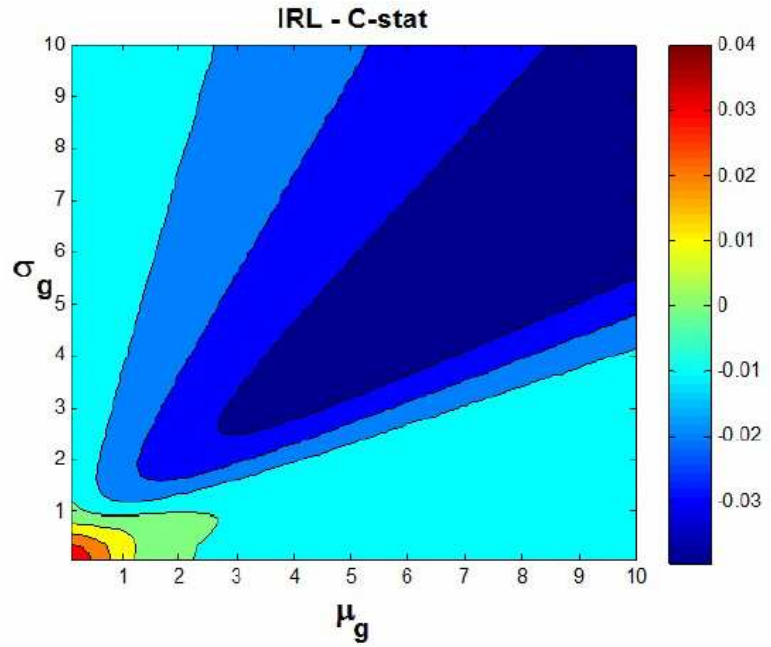
4. Case Study

To illustrate the advantage of the proposed indexes, we introduce a simple case study. We consider two scoring models with a score distribution given in Table 2.

Furthermore, we consider the standard meaning of scores, i.e., a higher score band means better clients (clients with the lowest scores, i.e., clients in score band 1, have the highest probability of default).



(a)



(b)

Figure 7. Difference of IRL and *c*-stat (a) and its contour plot (b).

Table 2. Score distribution and QLift of given scoring models

			Scoring model 1			Scoring model 2		
Score band	**#Clients**	***q***	**# Bad clients**	**Cumul. bad rate**	**QLift**	**#Bad clients**	**Cumul. bad rate**	**QLift**
1	100	0.1	20	20.0%	2.00	35	35.0%	3.50
2	100	0.2	18	19.0%	1.90	16	25.5%	2.55
3	100	0.3	17	18.3%	1.83	8	19.7%	1.97
4	100	0.4	15	17.5%	1.75	8	16.8%	1.68
5	100	0.5	12	16.4%	1.64	7	14.8%	1.48
6	100	0.6	6	14.7%	1.47	6	13.3%	1.33
7	100	0.7	4	13.1%	1.31	6	12.3%	1.23
8	100	0.8	3	11.9%	1.19	5	11.4%	1.14
9	100	0.9	3	10.9%	1.09	5	10.7%	1.07
10	100	1.0	2	10.0%	1.00	4	10.0%	1.00
All	1000		100			100		

The Gini index for each model is equal to 0.420. KS is equal to 0.356 for model 1 and to 0.344 for model 2. According to these numbers, one can say that both models are almost the same, maybe the first one is slightly better. However, if we look at the models in more detail, we find that they differ significantly. We get the first insight from their Lorenz curves in Figure 8.

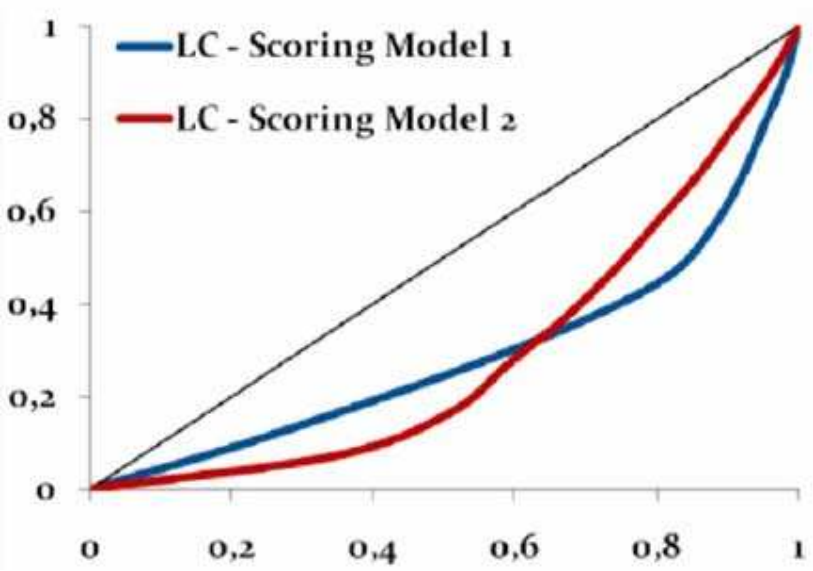


Figure 8. Lorenz curves for model 1 and model 2.

We can see that model 1 is stronger for higher score bands. This means that this model better separates the good from the best clients. On the other hand, model 2 is stronger for lower score bands, which means that it better separates the bad from the worst clients. We can read the same result from the figures of QLift and RLift in Figure 9.

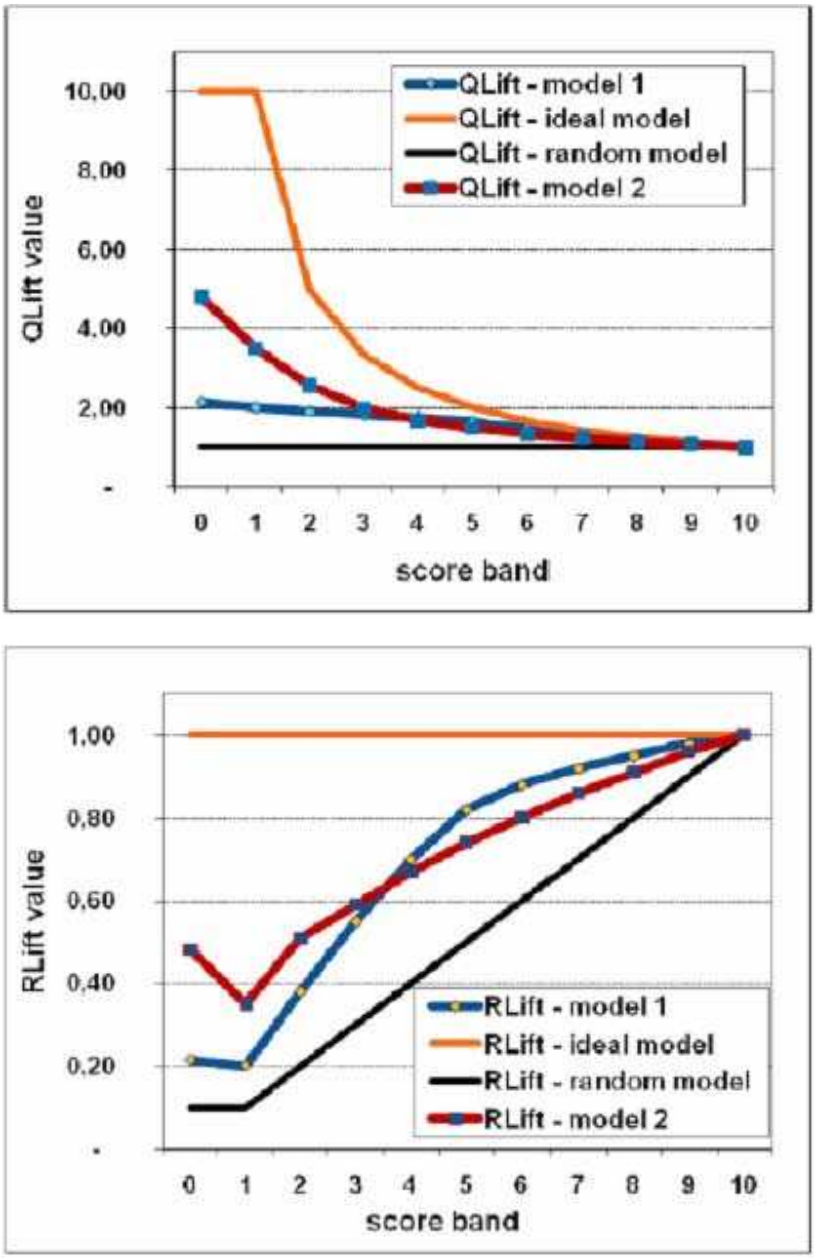


Figure 9. QLift and RLift for model 1 and model 2.

It is necessary to mention one computational problem at this point. In the discrete case, as in the case of Table 2, we do not know the value of QLift for q less than 0.1. Since QLift is not defined for $q = 0$, we need to extrapolate it somehow. According to the shape of the QLift curve, we propose using quadratic extrapolation, which yields

$$\mathcal{Q}Lift(0) = 3 \cdot \mathcal{Q}Lift(0.1) - 3 \cdot \mathcal{Q}Lift(0.2) + \mathcal{Q}Lift(0.3). \tag{23}$$

When we have a full data set, we can use formula (17). In this case, the extrapolation is not needed. Of course, we still do not have the value QLift (0). However, if we start the computation of QLift in some positive value of q, which is sufficiently near to zero, the final result is precise enough.

Overall, we can compare our two scoring models. Table 3, below, contains values of Gini indexes, K-S statistics, values of QLift(0.1), LR indexes, and IRL indexes. QLift(0.1) is a local measure of a model's quality; model 2 was designed to be better in the first score bands, hence it is natural that the value of QLift(0.1) is significantly higher for model 2, concretely 3.5 versus 2.0. On the other hand, all remaining indexes are global measures of a model's quality. Models were designed to have the same Gini index and similar KS. However, we can see that LR and IRL significantly differ for our models, 0.242 versus 0.372 and 0.699 versus 0.713, respectively.

Table 3. Quality indexes of two assessed scoring models

	Scoring model 1	**Scoring model 2**
Gini	0.420	0.420
KS	0.356	0.344
QLift(0.1)	2.000	3.500
LR	0.242	0.372
IRL	0.699	0.713

Finally, if the expected reject rate is up to 40%, which is a very natural assumption, using LR and IRL, we can state that model 2 is better than model 1 although their Gini indexes are equal and even their KS are in reverse order.

5. Conclusion

In Section 2, we presented widely used indexes for the assessment of credit scoring models. We focused mainly on the definitions of Lorenz curve, CAP, Gini index, AR, and lift. The Lorenz curve is sometimes confused with ROC. The discussion of their definitions is given within the paper. We suggest using the definition of the Lorenz curve given in Müller and Rönz [8], the definition of ROC given in Siddiqi [11], and the definition of CAP given in Sobehart et al. [12].

The main part of the paper, Section 3, was devoted to lift. Formulas for lift in basic and quantile form were presented as well as their forms for ideal models. These formulas allow the calculation of the value of lift for any given score and any given quantile level and comparison with the best obtainable results.

Lift ratio was presented as analogous to Gini index. An important feature is that LR allows the fair comparison of two models developed on different data samples, which is not possible with lift or QLift. Furthermore, a relative lift function was proposed, which shows the ratio of the QLifts of the actual and ideal models. Finally, integrated relative lift was defined. The connection to the *c*-statistic was presented by means of a simulation by using normally distributed scores. This simulation showed that IRL and the *c*-statistic are approximately equal in the case when the variances of good and bad clients are equal.

Despite the high popularity of the Gini index and KS, we conclude that the proposed lift based indexes are more appropriate for assessing the quality of credit scoring models. In particular, it is better to use them in the case of an asymmetric Lorenz curve. In such cases, using the Gini index or KS during the development process could lead to the selection of a weaker model.

Acknowledgement

This research was supported by our department and by The Jaroslav Hájek Center for Theoretical and Applied Statistics (grant No. LC 06024).

References

[1] R. Anderson, The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation, Oxford University Press, Oxford, 2007.

[2] M. J. A. Berry and G. S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd Edition, Wiley, Indianapolis, 2004.

[3] D. S. Coppock, Why Lift? DM Review Online, (2002). [Accessed on 1 December 2009].

www.dmreview.com/news/5329-1.html

[4] J. N. Crook, D. B. Edelman and L. C. Thomas, Recent developments in consumer credit risk assessment, European Journal of Operational Research 183(3) (2007), 1447-1465.

[5] B. Engelmann, E. Hayden and D. Tasche, Measuring the Discriminatory Power of Rating System, (2003). [Accessed on 4 October 2010].

http://www.bundesbank.de/download/bankenaufsicht/dkp/200301dkp_b.pdf

[6] P. Giudici, Applied Data Mining: Statistical Methods for Business and Industry, Wiley, Chichester, 2003.

[7] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: A review, Journal of the Royal Statistical Society, Series A 160(3) (1997), 523-541.

[8] M. Müller and B. Rönz, Credit Scoring using Semiparametric Methods, In: J. Franke, W. Härdle and G. Stahl (Eds.), Measuring Risk in Complex Stochastic Systems, Springer-Verlag, New York, 2000.

[9] R. B. Nelsen, Concordance and Gini's measure of association, Journal of Nonparametric Statistics 9(3) (1998), 227-238.

[10] R. Newson, Confidence intervals for rank statistics: Somers' D and extensions, The Stata Journal 6(3) (2006), 309-334.

[11] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, Wiley, New Jersey, 2006.

[12] J. Sobehart, S. Keenan and R. Stein, Benchmarking Quantitative Default Risk Models: A Validation Methodology, Moody's Investors Service, (2000). [Accessed on 4 October 2010].

http://www.algorithmics.com/EN/media/pdfs/Algo-RA0301-ARQ-DefaultRiskModels.pdf

[13] R. H. Somers, A new asymmetric measure of association for ordinal variables, American Sociological Review 27 (1962), 799-811.

[14] L. C. Thomas, A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers, International Journal of Forecasting 16(2) (2000), 149-172.

[15] L. C. Thomas, D. B. Edelman and J. N. Crook, Credit Scoring and its Applications, SIAM Monographs on Mathematical Modelling and Computation, Philadelphia, 2002.

[16] L. C. Thomas, Consumer Credit Models: Pricing, Profit, and Portfolio, Oxford University Press, Oxford, 2009.

[17] A. D. Wilkie, Measures for Comparing Scoring Systems, In: L. C. Thomas, D. B. Edelman and J. N. Crook (Eds.): Readings in Credit Scoring, Oxford University Press, Oxford, (2004), 51-62.

[18] K. Xu, How has the literature on Gini's index evolved in past 80 years? (2003). [Accessed on 1 December 2009].

economics.dal.ca/RePEc/dal/wparch/howgini.pdf

■

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Visualization and Bandwidth Matrix Choice

Ivana Horová [a], Jan Koláček [a] & Kamila Vopatová [a]
[a] Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic

Available online: 10 Jan 2012

To cite this article: Ivana Horová, Jan Koláček & Kamila Vopatová (2012): Visualization and Bandwidth Matrix Choice, Communications in Statistics - Theory and Methods, 41:4, 759-777

To link to this article: http://dx.doi.org/10.1080/03610926.2010.529539

Communications in Statistics—Theory and Methods, 41: 759–777, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2010.529539

Visualization and Bandwidth Matrix Choice

IVANA HOROVÁ, JAN KOLÁČEK,
AND KAMILA VOPATOVÁ

Department of Mathematics and Statistics, Masaryk University,
Brno, Czech Republic

Kernel smoothers are among the most popular nonparametric functional estimates. These estimates depend on a bandwidth that controls the smoothness of the estimate. While the literature for a bandwidth choice in a univariate density estimate is quite extensive, the progress in the multivariate case is slower. The authors focus on a bandwidth matrix selection for a bivariate kernel density estimate provided that the bandwidth matrix is diagonal. A common task is to find entries of the bandwidth matrix which minimizes the Mean Integrated Square Error (MISE). It is known that in this case there exists explicit solution of an asymptotic approximation of MISE (Wand and Jones, 1995). In the present paper we pay attention to the visualization and optimizers are presented as intersection of bivariate functional surfaces derived from this explicit solution and we develop the method based on this visualization. A simulation study compares the least square cross-validation method and the proposed method. Theoretical results are applied to real data.

Keywords Asymptotic mean integrated square error; Bandwidth matrix; Mean integrated square error; Product kernel.

Mathematics Subject Classification 62G07; 62H12.

1. Introduction

Methods for a bandwidth choice in a univariate density estimate have been developed in many papers and monographs (e.g., Cao et al., 1994; Chaudhuri and Marron, 1999; Härdle et al., 2004; Horová et al., 2002; Horová and Zelinka, 2007; Silverman, 1989; Taylor, 1989; Wand and Jones, 1995).

In this paper we focus on a problem of a data-driven choice of a bandwidth matrix in bivariate kernel density estimates. Bivariate kernel density estimation problem is an excellent setting for understanding aspects of multivariate kernel smoothing.

This problem, despite being the simplest multivariate density estimation problem, presents many challenges when it comes to selecting the correct amount of smoothing (i.e., choosing of a bandwidth matrix H). Most of popular bandwidth

Received July 19, 2010; Accepted September 28, 2010
Address correspondence to Ivana Horová, Department of Mathematics and Statistics, Masaryk University, Kotlarska 2, Brno 61137, Czech Republic; E-mail: horova@math.muni.cz

selection methods in a univariate case (e.g., Cao et al., 1994; Härdle et al., 2004) can be transferred into multivariate settings. The least squares cross-validation, the biased cross-validation, the smoothed cross-validation, and plug-in methods in multivariate case have been developed and widely discussed (Chacón and Duong, 2009; Duong and Hazelton, 2003, 2005a,b; Sain et al., 1994; Scott, 1992; Wand and Jones, 1994). The problem of the bandwidth matrix selection can be simplified by imposing constraints on H (Wand and Jones, 1995).

A common approach to the multivariate smoothing is to first rescale the data so the sample variances are equal in each dimension—this approach is called scaling or sphering the data so the sample covariance matrix is the identity (e.g., Duong, 2007; Wand and Jones, 1993). The aim of the present paper is to propose methods for the bandwidth matrix choice in bivariate case without using any pretransformations of the data.

It is well known that a visualization is an important component of a nonparametric data analysis (e.g., Chaudhuri and Marron, 1999; Godtliebsen et al., 2002). We use this effective strategy to clarify the process of the bandwidth matrix choice by using bivariate functional surfaces. The proposed method uses an optimally balanced relation between bias squared and variance and a suitable estimate of the asymptotic approximation of Mean Integrated Square Error (MISE).

The paper is organized as follows: In Section 2 we describe the basic properties of the multivariate density estimates. Section 3 is devoted to the mean integrated square error and its minimization. In Section 4 we deal with asymptotic MISE (AMISE) and its minimization. In Section 5 we describe the idea of our method and the theoretical results are explain by means of bivariate functional surfaces. In Section 6 we conduct a simulation study comparing the least squares cross-validation (LSCV) method and the proposed method. In Section 7 the theoretical results are applied to real data.

2. Kernel Density Estimation

Consider a d-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ coming from an unknown density f. We denote $X_{i1}, \ldots, X_{id}$ the components of $\mathbf{X}_i$ and a generic vector $\mathbf{x} \in \mathbb{R}^d$ has the representation $\mathbf{x} = (x_1, \ldots, x_d)^T$.

For a d-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ drawn from the density f the kernel density estimator is defined

$$\hat{f}(\mathbf{x}, H) = \frac{1}{n} \sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{X}_i), \tag{1}$$

where H is a symmetric positive definite $d \times d$ matrix called the bandwidth matrix, and $K_H(\mathbf{x}) = |H|^{-1/2} K(H^{-1/2}\mathbf{x})$, where $|H|$ stands for the determinant of H, and K is a d-variate kernel function. The kernel function K is often taken to be a d-variate probability density function.

There are two types of multivariate kernels created from a symmetric univariate kernel k—a product kernel K^P and a spherically symmetric kernel K^S:

$$K^P(\mathbf{x}) = \prod_{i=1}^{d} k(x_i), \quad K^S(\mathbf{x}) = c_k k\left(\sqrt{\mathbf{x}^T\mathbf{x}}\right)$$

where $c_k^{-1} = \int k\left(\sqrt{\mathbf{x}^T\mathbf{x}}\right)d\mathbf{x}$. The choice of a kernel does not influence the estimate as significantly as the bandwidth matrix.

The choice of the smoothing matrix H is of a crucial importance. This matrix controls the amount and the direction of the multivariate smoothing.

Let $\mathscr{H}_{\mathscr{F}}$ denote the class of symmetric, positive definite $d \times d$ matrices. The matrix $H \in \mathscr{H}_{\mathscr{F}}$ has $\frac{1}{2}d(d+1)$ independent entries which have to be chosen. A simplification can be obtained by imposing the restriction $H \in \mathscr{H}_{\mathscr{D}}$, where $\mathscr{H}_{\mathscr{D}} \subset \mathscr{H}_{\mathscr{F}}$ is the subclass of diagonal positive definite matrices: $H = \mathrm{diag}(h_1^2, \ldots, h_d^2)$. A further simplification follows from the restriction $H \in \mathscr{H}_{\mathscr{S}}$ where $\mathscr{H}_{\mathscr{S}} = \{h^2 I_d, h > 0\}$, I_d is $d \times d$ identity matrix and leads to the single bandwidth estimator (Wand and Jones, 1995). Using the single bandwidth matrix parametrization class $\mathscr{H}_{\mathscr{S}}$ is not advised for data which have different dispersions in the coordinate directions (Wand and Jones, 1993). On the other hand, the bandwidth selectors in the general $\mathscr{H}_{\mathscr{F}}$ class are able to handle differently dispersed data but are computationally intensive. So the $\mathscr{H}_{\mathscr{D}}$ diagonal matrix class is a compromise between computational speed with sufficient flexibility.

For this reason we turn our attention to the bivariate kernel density estimate provided that the bandwidth matrix is diagonal (i.e., $H = \mathrm{diag}(h_1^2, h_2^2)$). First, let us make some notation:

- $\int$ will be shorthand for $\iint$ and $d\mathbf{x}$ will be shorthand for $dx_1 dx_2$, $V(K) = \int K^2(\mathbf{x})d\mathbf{x}$, and
- $\mathscr{D}_f$ stands for the gradient and $\mathscr{D}_f^2$ for the Hessian matrix.

$$\mathscr{D}_f = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{pmatrix} \quad \mathscr{D}_f^2 = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} \end{pmatrix}.$$

For the next steps we need a few assumptions about the kernel function K, the bandwidth matrix H, and the density f:

(A1) K is a product bivariate kernel function satisfying

$$\int K(\mathbf{x})d\mathbf{x} = 1, \quad \int \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0, \quad \int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \beta_2(K)I_2.$$

(A2) $H = H_n$ is a sequence of diagonal bandwidth matrices such that $n^{-1}(h_1 h_2)^{-1}$ and h_1^2 and h_2^2 approach zero as $n \to \infty$.

(A3) Each entry of the Hessian matrix $\mathscr{D}_f^2$ is piecewise continuous and square integrable.

3. MISE and Its Minimization

The quality of the estimate (1) can be expressed in terms of MISE (Wand and Jones, 1995)

$$\mathrm{MISE}(H) = \int E\left(\hat{f}(\mathbf{x}, H) - f(\mathbf{x})\right)^2 d\mathbf{x} = \int \mathrm{var}(\hat{f}(\mathbf{x}, H))d\mathbf{x} + \int \mathrm{bias}^2(\hat{f}(\mathbf{x}, H))d\mathbf{x},$$

that is,

$$\mathrm{MISE}(H) = \frac{1}{nh_1h_2}V(K) + o\left((nh_1h_2)^{-1}\right) + \frac{1}{4}\beta_2^2(K)\left(h_1^4\psi_{4,0} + 2h_1^2h_2^2\psi_{2,2} + h_2^4\psi_{0,4}\right) + o\left((h_1^2+h_2^2)^2\right)$$

where

$$\psi_{k,\ell} = \int \left(\frac{\partial^2 f}{\partial x_1^2}\right)^{k/2}\left(\frac{\partial^2 f}{\partial x_2^2}\right)^{\ell/2} d\mathbf{x}, \quad k,\ell = 0,2,4, \quad k+\ell = 4.$$

Let H_{MISE} be a minimizer of MISE with respect to H, that is,

$$H_{\mathrm{MISE}} = \arg\min_{H\in\mathcal{H}_{\mathcal{D}}} \mathrm{MISE}.$$

The well known method of estimating H_{MISE} is the LSCV method (Duong and Hazelton, 2005b; Wand and Jones, 1995). The LSCV objective function is

$$\mathrm{LSCV}(H) = \int \left(\hat{f}(\mathbf{x},H)\right)^2 d\mathbf{x} - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(\mathbf{X}_i,H),$$

$$\hat{f}_{-i}(\mathbf{X}_i,H) = \frac{1}{(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n} K_H(\mathbf{X}_i - \mathbf{X}_j).$$

This function can be written in terms of convolutions $(f*g)(x) = \int_{\mathbb{R}} f(u)g(x-u)du$ (Duong and Hazelton, 2005b):

$$\mathrm{LSCV}(H) = n^{-2}\sum_{i,j=1}^{n}(K_H * K_H - 2K_H)(\mathbf{X}_i - \mathbf{X}_j) + 2n^{-1}K_H(0).$$

Moreover, $H_{\mathrm{LSCV}} = \arg\min_{H\in\mathcal{H}_{\mathcal{D}}} \mathrm{LSCV}$ is an unbiased estimate of H in the sense

$$E\left(\mathrm{LSCV}(H)\right) = \mathrm{MISE}\left(\hat{f}(\cdot,H)\right) - \int f^2(\mathbf{x})d\mathbf{x}.$$

4. AMISE and Its Minimization

Since MISE is not mathematically tractable, we employ an AMISE, which can be written as a sum of an asymptotic integrated variance and an asymptotic integrated square bias:

$$\mathrm{AMISE}(H) = \underbrace{\frac{V(K)}{nh_1h_2}}_{\mathrm{AIVar}} + \underbrace{\frac{1}{4}\beta_2(K)^2\left(h_1^4\psi_{4,0} + 2h_1^2h_2^2\psi_{2,2} + h_2^4\psi_{0,4}\right)}_{\mathrm{AIBias}^2} \tag{2}$$

and H_{AMISE} stands for minimum of AMISE

$$H_{\mathrm{AMISE}} = \arg\min_{H\in\mathcal{H}_{\mathcal{D}}} \mathrm{AMISE}.$$

First, we summarize properties of AMISE and H_{AMISE}. As a multivariate analogue of the functional, which minimization yields optimal kernels, we consider the functional

$$W(K) = V(K)^{2/3}\beta_2(K)^{2/3}.$$

Moreover, we define as a canonical factor

$$\gamma^3 = \frac{V(K)}{\beta_2(K)^2}.$$

Making some calculations we arrive at the following lemma.

Lemma 4.1. AMISE(H) *can be expressed in the form*

$$\mathrm{AMISE}(H) = W(K)\left\{\frac{\gamma}{nh_1h_2} + \frac{1}{4\gamma^2}(h_1^4\psi_{4,0} + 2h_1^2h_2^2\psi_{2,2} + h_2^4\psi_{0,4})\right\}. \tag{3}$$

It can be shown (Wand and Jones, 1995) that the entries of H_{AMISE} are equal to

$$\begin{aligned} h_{1,\mathrm{AMISE}}^2 &= \left[\frac{\psi_{0,4}^{3/4}V(K)}{n\beta_2(K)^2\psi_{4,0}^{3/4}\left(\psi_{2,2} + \psi_{0,4}^{1/2}\psi_{4,0}^{1/2}\right)}\right]^{1/3}, \\ h_{2,\mathrm{AMISE}}^2 &= \left[\frac{\psi_{4,0}^{3/4}V(K)}{n\beta_2(K)^2\psi_{0,4}^{3/4}\left(\psi_{2,2} + \psi_{0,4}^{1/2}\psi_{4,0}^{1/2}\right)}\right]^{1/3}. \end{aligned} \tag{4}$$

Thus $h_{i,\mathrm{AMISE}}^2 = O(n^{-1/3})$, $i = 1, 2$.

Inserting these quantities into the formula (2), we arrive at the following lemma.

Lemma 4.2. *Let* $H_{\mathrm{AMISE}} \in \mathscr{H}_{\mathscr{D}}$ *be a minimizator of AMISE with entries given by formula* (4). *Then*

$$\underbrace{\int \mathrm{var}\hat{f}(\mathbf{x}, H_{\mathrm{AMISE}})d\mathbf{x}}_{\mathrm{AIVar}} = 2\underbrace{\int \left(\mathrm{bias}\hat{f}(\mathbf{x}, H_{\mathrm{AMISE}}\right)^2 d\mathbf{x}}_{\mathrm{AIBias}^2}. \tag{5}$$

This relation is of great importance because it serves as a basis for a method we are going to present. It means that minimization of AMISE is equivalent to seeking for H_{AMISE} such that (5) is satisfied.

Further, the use of formulas (4) in the relation (3) yields

$$\mathrm{AMISE}(H_{\mathrm{AMISE}}) = \frac{3}{2}n^{-2/3}W(K)\left(\psi_{2,2} + \psi_{0,4}^{1/2}\psi_{4,0}^{1/2}\right)^{1/3}, \tag{6}$$

that is, $\mathrm{AMISE}(H_{\mathrm{AMISE}}) = O(n^{-2/3})$.

It is easy to show that

$$\frac{h_{2,\mathrm{AMISE}}}{h_{1,\mathrm{AMISE}}} = \left(\frac{\psi_{4,0}}{\psi_{0,4}}\right)^{1/4} \tag{7}$$

and

$$\left(\psi_{2,2}+\psi_{0,4}^{1/2}\psi_{4,0}^{1/2}\right)^{1/3}=\frac{\gamma}{n^{1/3}h_{1,\mathrm{AMISE}}\,h_{2,\mathrm{AMISE}}}$$

Then substituting $\left(\psi_{2,2}+\psi_{0,4}^{1/2}\psi_{4,0}^{1/2}\right)^{1/3}$ into (6) we obtain

$$\mathrm{AMISE}(H_{\mathrm{AMISE}})=\frac{3W(K)}{2nh_{1,\mathrm{AMISE}}h_{2,\mathrm{AMISE}}}.$$

This formula allows to separate kernel effects from bandwidth matrix effects in AMISE and thus offers a possibility to choose the kernel and the bandwidth matrix in some automatic and optimal way. For a univariate case an automatic procedure for simultaneous choice of a bandwidth, a kernel, and an order of the kernel was proposed previously (Horová et al., 2002).

Remark. The biased cross-validation methods and smoothed cross-validation method for estimating H_{AMISE} have been widely discussed previously (Duong and Hazelton, 2005b; Sain et al., 1994; Wand and Jones, 1994).

5. Proposed Methods

Our method is based on formula (5) and on a suitable estimate of AMISE.

In Horová et al. (2008) a suitable estimate of AMISE was used and the extension of the method for a univariate case was presented in Horová and Zelinka (2007). Here, we briefly describe this method and provide theoretical results.

Let

$$\widehat{\mathrm{AMISE}}(H)=\int \widehat{\mathrm{var}}\hat{f}(\mathbf{x},H)d\mathbf{x}+\int\left(\widehat{\mathrm{bias}}\hat{f}(\mathbf{x},H)\right)^2 d\mathbf{x},$$

where

$$\begin{aligned}\int \widehat{\mathrm{var}}\hat{f}(\mathbf{x},H)d\mathbf{x}&=\frac{1}{n}\int\left[\int K_H^2(\mathbf{x}-\mathbf{y})\hat{f}(\mathbf{y},H)d\mathbf{y}\right]d\mathbf{x}\\&=\frac{1}{n}|H|^{-1/2}\iint K^2(\mathbf{z})\hat{f}(\mathbf{x}-H^{1/2}\mathbf{z},H)d\mathbf{z}d\mathbf{x}\\&=\frac{1}{n}|H|^{-1/2}V(K)\int\hat{f}(\mathbf{x},H)d\mathbf{x}\\&=\frac{1}{n}|H|^{-1/2}V(K)\end{aligned}$$

and

$$\begin{aligned}\int\left(\widehat{\mathrm{bias}}\hat{f}(\mathbf{x},H)\right)^2 d\mathbf{x}&=\int\left[\int K_H(\mathbf{x}-\mathbf{y})\hat{f}(\mathbf{y},H)d\mathbf{y}-\hat{f}(\mathbf{x},H)\right]^2 d\mathbf{x}\\&=\int\left[\int K(\mathbf{z})\hat{f}(\mathbf{x}-H^{1/2}\mathbf{z},H)d\mathbf{z}-\hat{f}(\mathbf{x},H)\right]^2 d\mathbf{x}\end{aligned}$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} (K_H * K_H * K_H * K_H - 2K_H * K_H * K_H + K_H * K_H)(\mathbf{X}_i - \mathbf{X}_j).$$

Here a connection of the estimated squared bias term with the bootstrap method of Taylor (1989) can be seen.

Hereinafter, $\widehat{H}_{\mathrm{AMISE}} = \mathrm{diag}(\hat{h}^2_{1,\mathrm{AMISE}}, \hat{h}^2_{2,\mathrm{AMISE}})$ is the minimizer of $\widehat{\mathrm{AMISE}}$ over the class of diagonal bandwidth matrices $\mathscr{H}_{\mathscr{D}}$ (i.e., $\widehat{H}_{\mathrm{AMISE}} = \arg\min_{H \in \mathscr{H}_{\mathscr{D}}} \widehat{\mathrm{AMISE}}$).

Let $g(h_1, h_2)$ stand for the sum of convolutions in the form $\int \left(\widehat{\mathrm{bias}}\hat{f}(\mathbf{x}, H)\right)^2 d\mathbf{x}$, that is,

$$g(h_1, h_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} (K_H * K_H * K_H * K_H - 2K_H * K_H * K_H + K_H * K_H)(\mathbf{X}_i - \mathbf{X}_j).$$

The idea of our method is based on Lemma 4.2. Thus, we are seeking for $\hat{h}_1$, $\hat{h}_2$ such that

$$\frac{1}{n} \frac{1}{\hat{h}_1 \hat{h}_2} V(K) = 2 \frac{1}{n^2} g(\hat{h}_1, \hat{h}_2)$$

that is,

$$nV(K) = 2\hat{h}_1 \hat{h}_2 g(\hat{h}_1, \hat{h}_2) \tag{8}$$

It means that minimization of $\widehat{\mathrm{AMISE}}$ could be achieved through the solving Eq. (8).

But (8) is the nonlinear equation for two variables and thus we need another relation between h_1 and h_2. This problem will be dealt with in the next section. Now we explain the rationale of the proposed method.

Theorem 5.1. *Let assumptions* (A1), (A2), (A3) *be satisfied and let the density f have continuous partial derivatives of the fourth order. Then*

$$E \int K_H(\mathbf{x} - \mathbf{y}) \hat{f}(\mathbf{y}, H) d\mathbf{y} = f(\mathbf{x}) + \beta_2(K) tr(H \mathscr{D}_f^2(\mathbf{x})) + \frac{1}{4} \beta_2(K)^2 tr(H \mathscr{D}_f^2 H \mathscr{D}_f^2(\mathbf{x})) + o(trH).$$

The proof is given in the Appendix.

Corollary 5.1. *Under assumptions of Theorem* 5.1, *the relation*

$$E\left(\widehat{\mathrm{bias}}\hat{f}(\mathbf{x}, H)\right) = \mathrm{bias}\hat{f}(\mathbf{x}, H) + o(trH)$$

is valid.

The last relation confirms that the solution of Eq. (8) may be expected to be reasonably close to $\widehat{H}_{\mathrm{AMISE}}$.

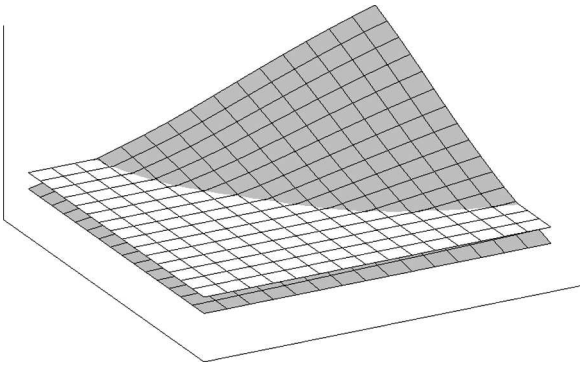


Figure 1. Optimal values of h_1 and h_2 lie on the curve $\Phi(h_1, h_2) = 2h_1h_2g(h_1, h_2) - nV(K) = 0$, which is an intersection of the surface $\Phi(h_1, h_2)$ (light gray) and the coordinate plane $z = 0$ (white).

Remark. Jones et al. (1991) was treated of the properties of the estimated square bias for a univariate case.

Remark. Wand and Jones (1995) reminded of solve-the-equation (STE) univariate selectors, which require solving nonlinear equation with respect to h. But their idea is different from that which we present.

Figure 1 shows the shape of the functional $\Phi(h_1, h_2) = 2h_1h_2g(h_1, h_2) - nV(K)$ and the point we are seeking lies on curve $\Phi(h_1, h_2) = 0$. Obviously, it has not a unique solution, and thus we need another relationship between h_1 and h_2 to get the unique solution. We propose two possibilities how to find this relationship.

5.1. *M1 Method*

Using Scott's rule (Scott, 1992) $\hat{h}_i = \hat{\sigma}_i n^{-1/6}$ for $i = 1, 2$ gives the other relationship between h_1 and h_2. It is easy to see that

$$h_2 = \hat{c}h_1, \quad \hat{c} = \frac{\hat{\sigma}_2}{\hat{\sigma}_1},$$

and $\hat{\sigma}$ can be estimated by a sample standard deviation, or by some robust method (e.g., a median deviation).

Now, the system of two equations for two unknowns h_1, h_2 has to be solved:

$$\text{M1} \begin{cases} 2h_1h_2g(h_1, h_2) = nV(K) \\ \\ h_2 = \hat{c}h_1 \end{cases} \tag{9}$$

Figure 2 demonstrates the solution of the system (9) as an intersection of the functional and planes.

As it will be shown in a simulation study, the method is rather inappropriate because the entries of covariance matrix are often not able to take into account the curvature of f and its orientation.

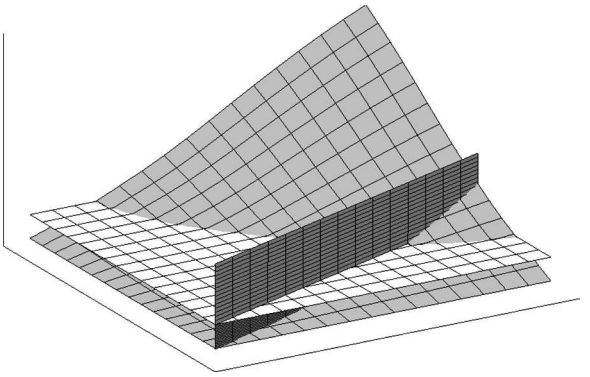


Figure 2. M1 method: The point $[\hat{h}_1, \hat{h}_2]$ we are looking for is an intersection of the plane $h_2 - \hat{c}h_1 = 0$ (dark gray) and the surface $\Phi(h_1, h_2)$ (light gray) and the coordinate plane $z = 0$ (white).

5.2. *M2 Method*

The second method can be considered as a hybrid of the biased cross-validation method (Duong and Hazelton, 2005b; Sain et al., 1994) and the plug-in method (Wand and Jones, 1994). We are concerned with fact (7), that is,

$$h_{2,\mathrm{AMISE}}^4 \cdot \psi_{0,4} = h_{1,\mathrm{AMISE}}^4 \cdot \psi_{4,0}, \tag{10}$$

where

$$\psi_{0,4} = \int \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 d\mathbf{x}, \quad \psi_{4,0} = \int \left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 d\mathbf{x}.$$

For the sake of simplicity in the next considerations the notation $h_1 = h_{1,\mathrm{AMISE}}$, $h_2 = h_{2,\mathrm{AMISE}}$ is used.

Relation (10) means that h_1, h_2 should be such that this equation is satisfied. At this step the estimates of $\psi_{0,4}$ and $\psi_{4,0}$ are needed. Since we assume that K is a product kernel we can express the estimates of $\psi_{0,4}$ and $\psi_{4,0}$ as the following

$$\hat{\psi}_{0,4} = n^{-2} \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_2^2} * \frac{\partial^2 K_H}{\partial x_2^2}\right)(\mathbf{X}_i - \mathbf{X}_j),$$

$$\hat{\psi}_{4,0} = n^{-2} \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_1^2} * \frac{\partial^2 K_H}{\partial x_1^2}\right)(\mathbf{X}_i - \mathbf{X}_j),$$

where instead of a pilot bandwidth matrix G in the plug-in method the bandwidth matrix H is used (i.e., $\hat{\psi}_{0,4}$, $\hat{\psi}_{4,0}$ estimate the density curvature in both directions).

Now, relation (10) yields

$$h_2^4 n^{-2} \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_2^2} * \frac{\partial^2 K_H}{\partial x_2^2}\right)(\mathbf{X}_i - \mathbf{X}_j) = h_1^4 n^{-2} \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_1^2} * \frac{\partial^2 K_H}{\partial x_1^2}\right)(\mathbf{X}_i - \mathbf{X}_j). \tag{11}$$

Hence, we have the second equation for h_1, h_2.

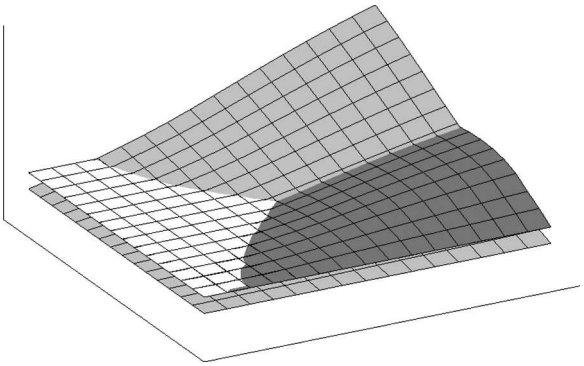


Figure 3. The searched point $[\hat{h}_1, \hat{h}_2]$ is an intersection of the surface $\Phi(h_1, h_2)$ (light gray), the coordinate plane $z = 0$ (white) and the surface $\Psi(h_1, h_2)$ (dark gray).

The proposed method is described by the system

$$\text{M2}\begin{cases} 2h_1h_2g(h_1, h_2) = nV(K) \\ h_2^4 \sum_{i,j=1}^{n} \left(\dfrac{\partial^2 K_H}{\partial x_2^2} * \dfrac{\partial^2 K_H}{\partial x_2^2} \right)(\mathbf{X}_i - \mathbf{X}_j) = h_1^4 \displaystyle\sum_{i,j=1}^{n} \left(\dfrac{\partial^2 K_H}{\partial x_1^2} * \dfrac{\partial^2 K_H}{\partial x_1^2} \right)(\mathbf{X}_i - \mathbf{X}_j) \end{cases} \tag{12}$$

The solution $[\hat{h}_1, \hat{h}_2]$ of this nonlinear system is an estimate of $[h_{1,\text{AMISE}}, h_{2,\text{AMISE}}]$. This system can be solved by Newton's method.

Table 1
Target densities

Normal I	$\mathcal{N}_2(0, 0; 1/4, 1, 0)$
Normal II	$\frac{1}{2}\mathcal{N}_2(-3/2, 0; 1/16, 1; 0) + \frac{1}{2}\mathcal{N}_2(3/2, 0; 1/16, 1; 0)$
Normal III	$\frac{1}{2}\mathcal{N}_2(0, 0; 1, 1, 0) + \frac{1}{2}\mathcal{N}_2(3, 0; 1, 1/2, 0)$
Normal IV	$\frac{1}{3}\mathcal{N}_2(0, 0; 1, 1, 0) + \frac{1}{3}\mathcal{N}_2(0, 4; 1, 4, 0) + \frac{1}{3}\mathcal{N}_2(4, 0; 4, 1, 0)$
Normal V	$\frac{1}{4}\mathcal{N}_2(0, 0; 1, 1, 0) + \frac{3}{4}\mathcal{N}_2(4, 3; 4, 3, 0)$
Normal VI	$\frac{1}{5}\mathcal{N}_2(0, 0; 1, 1, 0) + \frac{1}{5}\mathcal{N}_2(1/2, 1/2; 4/9, 4/9, 0)$ $+\frac{3}{5}\mathcal{N}_2(13/12, 13/12; 25/81, 25/81, 0)$
Normal VII	$\frac{1}{3}\mathcal{N}_2(0, -3; 1, 1/16, 0) + \frac{1}{3}\mathcal{N}_2(0, 0; 1, 1/16, 0) + \frac{1}{3}\mathcal{N}_2(0, 3; 1, 1/16, 0)$
Normal VIII	$\frac{1}{3}\mathcal{N}_2(0, -3; 1, 1/16, 0) + \frac{1}{3}\mathcal{N}_2(0, 0; 1/2, 1/16, 0) + \frac{1}{3}\mathcal{N}_2(0, 3; 1/8, 1/16, 0)$
Normal IX	$\frac{1}{3}\mathcal{N}_2(-6/5, 0; 9/16, 9/16, 7/10) + \frac{1}{3}\mathcal{N}_2(0, 0; 9/16, 9/16, -7/10)$ $+\frac{1}{3}\mathcal{N}_2(6/5, 0; 9/16, 9/16, 7/10)$
Beta Beta	$\mathcal{B}(2, 4) \cdot \mathcal{B}(2, 6)$
Beta Weibull	$\mathcal{B}(2, 4) \cdot \mathcal{W}(2, 3)$
Gamma Beta	$\mathcal{G}(2, 1) \cdot \mathcal{B}(2, 6)$
LogNormal	$\mathcal{LN}_2(0, 0; 1, 1, 0)$

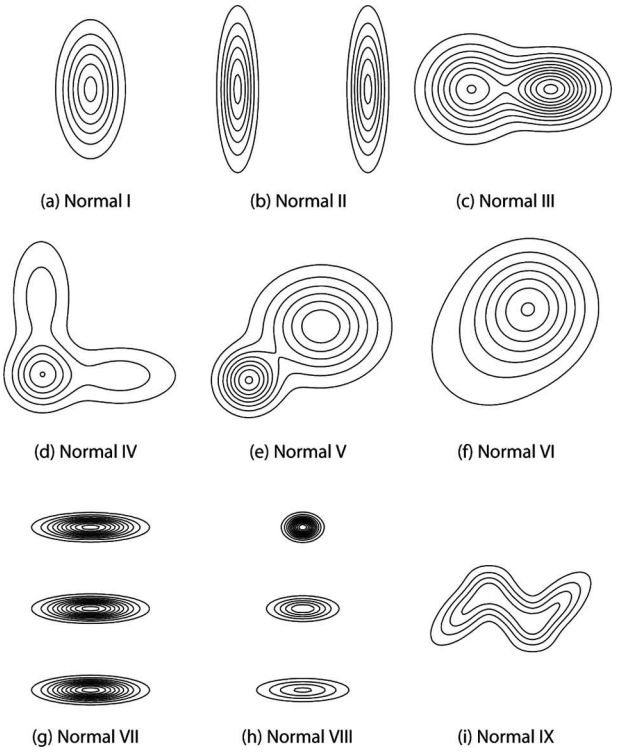


Figure 4. Contour plots of normal target densities.

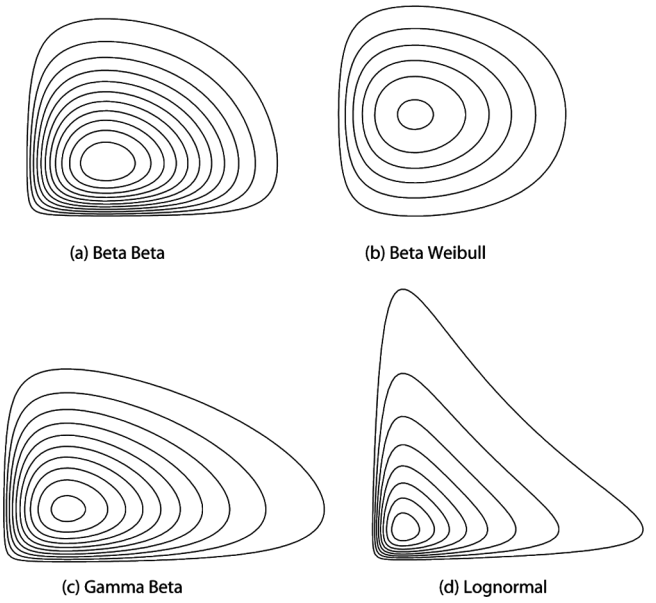


Figure 5. Contour plots of nonnormal target densities.

In Fig. 3 the graphs of the surfaces

$$\Phi(h_1, h_2) = 2h_1h_2g(h_1, h_2) - nV(K)$$

and

$$\Psi(h_1, h_2) = h_2^4 \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_2^2} * \frac{\partial^2 K_H}{\partial x_2^2} \right) (\mathbf{X}_i - \mathbf{X}_j) - h_1^4 \sum_{i,j=1}^{n} \left(\frac{\partial^2 K_H}{\partial x_1^2} * \frac{\partial^2 K_H}{\partial x_1^2} \right) (\mathbf{X}_i - \mathbf{X}_j)$$

are presented. The solution of this system yields the estimates $\hat{h}_1$ and $\hat{h}_2$.

Remark. It is clear that $h_{i,\mathrm{AMISE}}^2 = O(n^{-1/3})$. Asymptotic properties and a rate of convergence of $\widehat{H}_{\mathrm{AMISE}}$ to H_{AMISE} can be treated of in a similar way as in (Duong and Hazelton, 2005a,b) and (Duong and Hazelton, 2005a) showed that the discrepancy between H_{AMISE} and H_{MISE} is asymptotically negligible.

6. Simulation Study

In this section we conduct a simulation study comparing the LSCV method with the M1 and M2 methods. Samples of the size $n = 100$ were drawn from densities listed in Table 1. Bandwidth matrices were selected for 100 random samples generated from each density. Contour plots of target densities are displayed in Figures 4 and 5.

As a criterion for comparison of data driven bandwidth matrix selectors the average of integrated square errors, that is,

$$\overline{\mathrm{ISE}} = \mathrm{avg}_H \int (\hat{f}(\mathbf{x}, H) - f(\mathbf{x}))^2 d\mathbf{x}, \tag{13}$$

Table 2
$\overline{\mathrm{ISE}}$: The average of ISE with a standard error in parentheses

Density	LSCV	M1	M2
Normal I	$1.58 \cdot 10^{-2}(0.150 \cdot 10^{-2})$	$0.91 \cdot 10^{-2}(0.041 \cdot 10^{-2})$	$0.92 \cdot 10^{-2}(0.042 \cdot 10^{-2})$
Normal II	$1.82 \cdot 10^{-2}(0.068 \cdot 10^{-2})$	$3.59 \cdot 10^{-2}(0.045 \cdot 10^{-2})$	$1.39 \cdot 10^{-2}(0.043 \cdot 10^{-2})$
Normal III	$0.62 \cdot 10^{-2}(0.040 \cdot 10^{-2})$	$0.47 \cdot 10^{-2}(0.016 \cdot 10^{-2})$	$0.49 \cdot 10^{-2}(0.017 \cdot 10^{-2})$
Normal IV	$0.28 \cdot 10^{-2}(0.024 \cdot 10^{-2})$	$0.20 \cdot 10^{-2}(0.007 \cdot 10^{-2})$	$0.23 \cdot 10^{-2}(0.008 \cdot 10^{-2})$
Normal V	$0.23 \cdot 10^{-2}(0.013 \cdot 10^{-2})$	$0.18 \cdot 10^{-2}(0.005 \cdot 10^{-2})$	$0.18 \cdot 10^{-2}(0.005 \cdot 10^{-2})$
Normal VI	$1.55 \cdot 10^{-2}(0.110 \cdot 10^{-2})$	$1.00 \cdot 10^{-2}(0.045 \cdot 10^{-2})$	$1.01 \cdot 10^{-2}(0.045 \cdot 10^{-2})$
Normal VII	$1.23 \cdot 10^{-2}(0.063 \cdot 10^{-2})$	$5.51 \cdot 10^{-2}(0.146 \cdot 10^{-2})$	$1.11 \cdot 10^{-2}(0.075 \cdot 10^{-2})$
Normal VIII	$2.92 \cdot 10^{-2}(0.126 \cdot 10^{-2})$	$5.52 \cdot 10^{-2}(0.144 \cdot 10^{-2})$	$2.76 \cdot 10^{-2}(0.124 \cdot 10^{-2})$
Normal IX	$1.98 \cdot 10^{-2}(0.084 \cdot 10^{-2})$	$1.91 \cdot 10^{-2}(0.044 \cdot 10^{-2})$	$1.81 \cdot 10^{-2}(0.048 \cdot 10^{-2})$
Beta Beta	$3.07 \cdot 10^{-1}(0.194 \cdot 10^{-1})$	$1.93 \cdot 10^{-1}(0.071 \cdot 10^{-1})$	$1.99 \cdot 10^{-1}(0.104 \cdot 10^{-1})$
Beta Weibull	$5.92 \cdot 10^{-2}(0.420 \cdot 10^{-2})$	$3.72 \cdot 10^{-2}(0.151 \cdot 10^{-2})$	$4.12 \cdot 10^{-2}(0.248 \cdot 10^{-2})$
Gamma Beta	$5.93 \cdot 10^{-2}(0.324 \cdot 10^{-2})$	$4.05 \cdot 10^{-2}(0.128 \cdot 10^{-2})$	$4.27 \cdot 10^{-2}(0.221 \cdot 10^{-2})$
LogNormal	$2.49 \cdot 10^{-2}(0.060 \cdot 10^{-2})$	$2.51 \cdot 10^{-2}(0.065 \cdot 10^{-2})$	$2.81 \cdot 10^{-2}(1.601 \cdot 10^{-2})$

is used, where the average is taken over simulated realizations. Table 2 brings the results of this comparison. It can be also considered the criterion $\overline{\mathrm{IAE}} = \mathrm{avg}_H \int |\hat{f}(\mathbf{x}, H) - f(\mathbf{x})| d\mathbf{x}$.

Figures 6–8 show distributions of the entries $\hat{h}_1$ and $\hat{h}_2$ of bandwidth matrices $\widehat{H}_{\mathrm{AMISE}}$ in the (h_1, h_2) coordinate plane. We observe that LSCV estimates of H_{AMISE} suffer from large variability. This fact could be explained by the fact that MISE(H) surface is rather flatter near H_{MISE}. The M1 and M2 methods perform very similarly; however, the M1 estimator fails for densities Normal II, Normal VII and Normal VIII. It is due to the fact that the use of the Scott's (1992) rule does not quite account for the curvature of f. The same problem occurs in application to real data, shown in the next section. The advantage of the M1 method is contained in its simplicity.

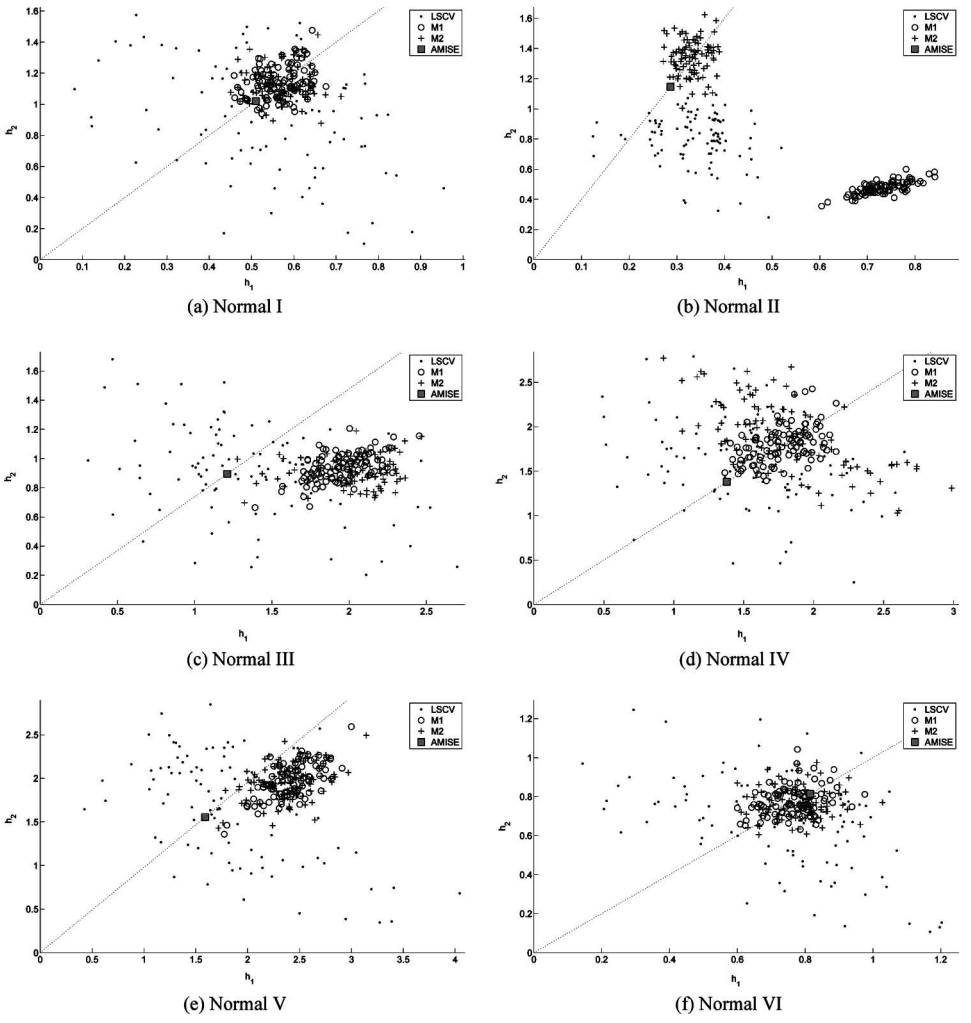


Figure 6. Distribution of $\hat{h}_1$ and $\hat{h}_2$—normal densities.

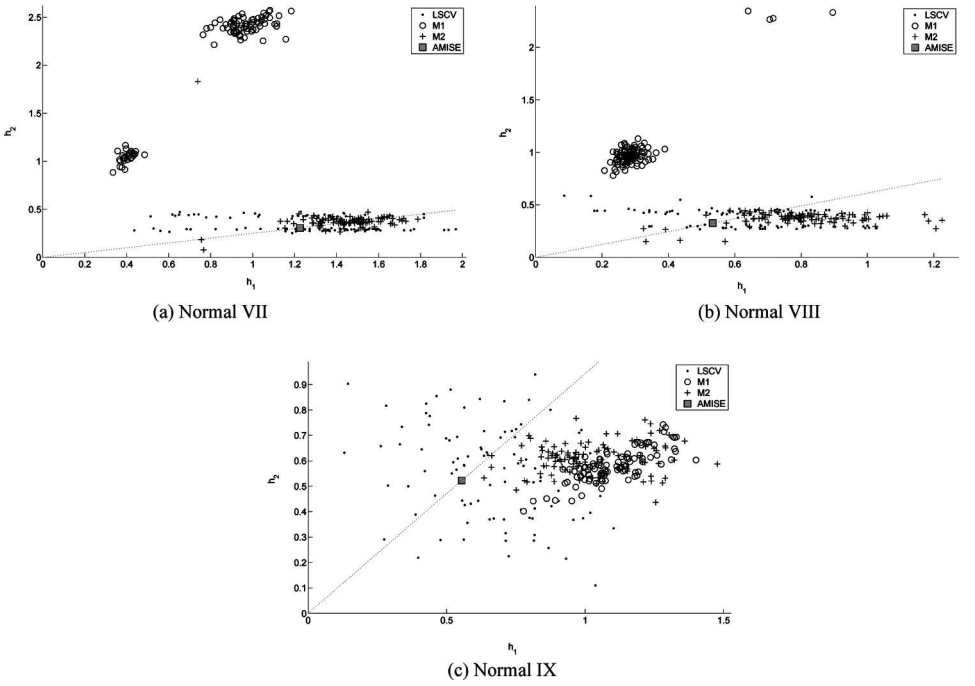


Figure 7. Distribution of $\hat{h}_1$ and $\hat{h}_2$—normal densities.

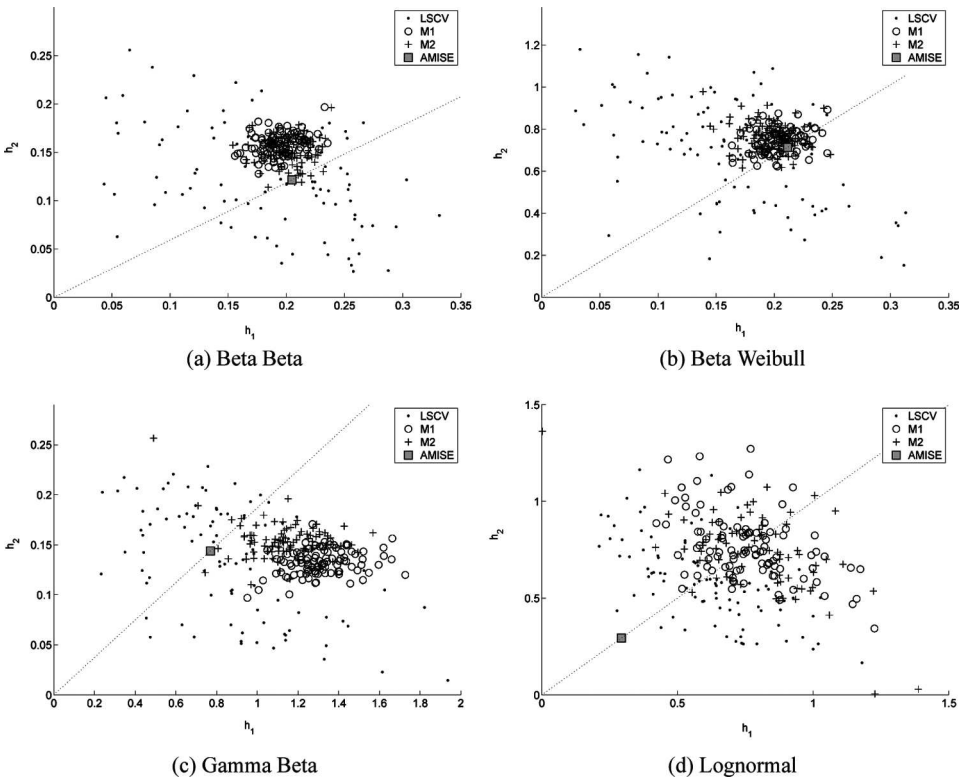


Figure 8. Distribution of $\hat{h}_1$ and $\hat{h}_2$—nonnormal densities.

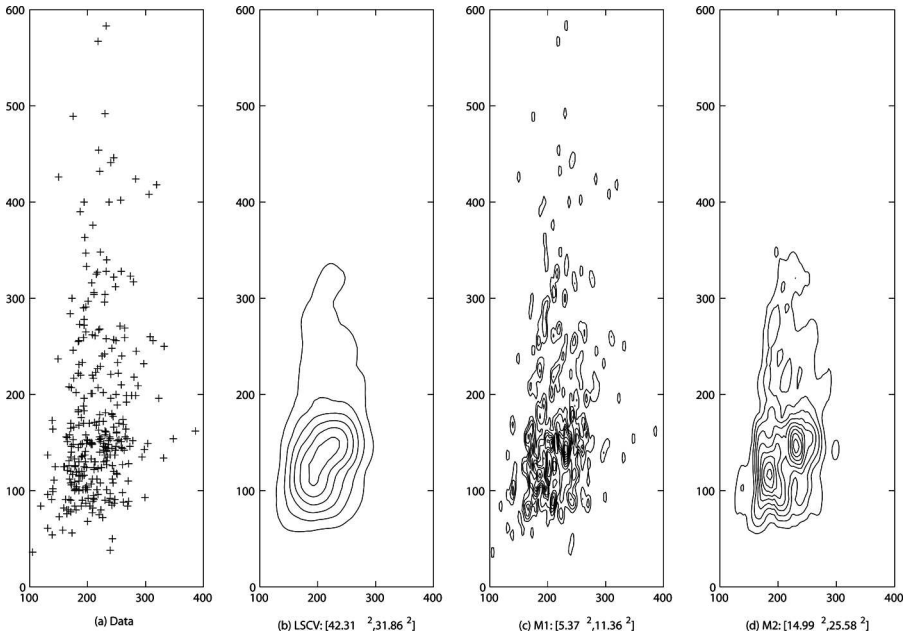


Figure 9. Kernel estimate of plasma lipid data.

On the other hand, it is obvious that the LSCV method performs rather well in mixtures of normal densities (Normal VII, Normal VIII) with respect to M1. The M2 method seems to be sufficiently reliable and easy to implement (using the product kernel). This fact is also confirmed by examing these methods on real data sets in the next section.

7. Application to Real Data

We applied the proposed methods to the plasma lipid data—a bivariate data set consisting of concentration of plasma cholesterol and plasma triglycerides taken on 320 patients with chest pain in a heart disease study (Scott, 1992). A scatterplot of the data is shown in Fig. 9a. Figures 9c and 9d represent reconstructed probability density functions using the bandwidth matrix $\widehat{H}_{\mathrm{M1}} = \mathrm{diag}(5.37^2, 11.63^2)$ and $\widehat{H}_{\mathrm{M2}} = \mathrm{diag}(14.99^2, 25.58^2)$, respectively. It can be compared with the reconstructed probability density function using $\widehat{H}_{\mathrm{LSCV}} = \mathrm{diag}(42.31^2, 31.86^2)$ shown in Fig. 9b. The authors of the original case study (Scott et al., 1978) found two primary clusters in these data set as well as the method M2 has found. See also papers by Ćwik and Koronacki (1997), Sain et al. (1994), Silverman (1989), and Wand and Jones (1995). Interestingly, while the LSCV and M1 methods fail to recognize the density bimodality, the M2 estimate is clearly bimodal.

8. Conclusion

The advantage of these methods is in their flexibility and in the fact that they are very easy to implement, especially for product kernels. Due to the fast computations of convolutions these methods seem to be less time consuming. Simulations show

that M2 estimates provides a sufficiently reliable way of estimating arbitrary densities.

We would like to emphasize that we restrict ourselves on the use of the Epanechnikov product kernel, because it has an optimality property (Wand and Jones, 1995) and corresponding integrals can be easily evaluated by means of convolutions. On the other hand, this kernel does not satisfy smoothness conditions for bias cross-validation methods and the plug-in method. Thus the simulation study compares the proposed methods with the LSCV method. Moreover, the proposed methods essentially minimize the MISE as the LSCV does.

Further assessment of their practical performance and comparison with other matrix bandwidth selectors through a large-scale simulation study would be very important further research.

Appendix

Proof of Theorem 5.1. The proof requires some notations: for a $m \times n$ matrix A vec is the vector operation (i.e., vec A is a $mn \times 1$ vector of stacked columns of the matrix A), and $A \otimes B$ denotes the Kronecker product of matrices A and B.

Let us denote

$$\begin{aligned} I(\mathbf{x}) &= E \int K_H(\mathbf{x}-\mathbf{y})\hat{f}(\mathbf{y}, H)d\mathbf{y} \\ &= E \int K(\mathbf{z})\hat{f}(\mathbf{x}-H^{1/2}\mathbf{z}, H)d\mathbf{z} \\ &= \int K(\mathbf{z})E\hat{f}(\mathbf{x}-H^{1/2}\mathbf{z}, H)d\mathbf{z}. \end{aligned}$$

And now compute

$$I_1(\mathbf{z}) = E\hat{f}(\mathbf{x}-H^{1/2}\mathbf{z}, H) = \int K_H(\mathbf{x}-H^{1/2}\mathbf{z}-\mathbf{y})f(\mathbf{y})d\mathbf{y}.$$

Substitutions yield

$$I_1(\mathbf{z}) = \int K(\mathbf{w}-\mathbf{z})f(\mathbf{x}-H^{1/2}\mathbf{w})d\mathbf{w} = \int K(\mathbf{u})f(\mathbf{x}-H^{1/2}\mathbf{u}-H^{1/2}\mathbf{z})d\mathbf{u}.$$

We use Taylor expansion in the form

$$\begin{aligned} f(\mathbf{x}-H^{1/2}\mathbf{u}-H^{1/2}\mathbf{z}) = f(\mathbf{x}-H^{1/2}\mathbf{z}) &- \left(H^{1/2}\mathbf{u}\right)^T \mathscr{D}_f(\mathbf{x}-H^{1/2}\mathbf{z}) \\ &+ \frac{1}{2}\left(H^{1/2}\mathbf{u}\right)^T \mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z})H^{1/2}\mathbf{u} + o(\operatorname{tr} H), \end{aligned}$$

Hence, using properties (A1) of the kernel

$$\begin{aligned} I_1(\mathbf{z}) &= f(\mathbf{x}-H^{1/2}\mathbf{z}) + \frac{1}{2}\int \left(H^{1/2}\mathbf{u}\right)^T \mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z})H^{1/2}\mathbf{u}K(\mathbf{u})d\mathbf{u} + o(\operatorname{tr} H) \\ &= f(\mathbf{x}-H^{1/2}\mathbf{z}) + \frac{1}{2}\beta_2(K)\operatorname{tr} H\mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z}) + o(\operatorname{tr} H). \end{aligned}$$

Further

$$\operatorname{tr} H\mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z}) = (\operatorname{vec}H)^T \operatorname{vec}\mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z})$$

(Magnus and Neudecker, 2007).

Now, we need Taylor expansion of $\operatorname{vec}\mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z})$:

$$\operatorname{vec}\mathscr{D}_f^2(\mathbf{x}-H^{1/2}\mathbf{z}) = \operatorname{vec}\mathscr{D}_f^2(\mathbf{x}) - (\mathscr{D}_f \otimes \mathscr{D}_f^2)(\mathbf{x})H^{1/2}\mathbf{z} + \frac{1}{2}(\mathscr{D}_f^2 \otimes \mathscr{D}_f^2)(\mathbf{x})\operatorname{vec}(\mathbf{z}\mathbf{z}^T H) + \mathbf{O}(\|\operatorname{vec}H^2\|).$$

Thus

$$I_1(\mathbf{z}) = f(\mathbf{x}-H^{1/2}\mathbf{z}) + \frac{1}{2}\beta_2(K)\,(\operatorname{vec}H)^T \left\{\operatorname{vec}\mathscr{D}_f^2(\mathbf{x}) - (\mathscr{D}_f \otimes \mathscr{D}_f^2)(\mathbf{x})H^{1/2}\mathbf{z} + \frac{1}{2}(\mathscr{D}_f^2 \otimes \mathscr{D}_f^2)(\mathbf{x})\operatorname{vec}(\mathbf{z}\mathbf{z}^T H) + \mathbf{O}(\|\operatorname{vec}H^2\|)\right\} + o(\operatorname{tr} H).$$

Hence

$$\begin{aligned} I(\mathbf{x}) &= \int K(\mathbf{z})I_1(\mathbf{z})d\mathbf{z} \\ &= \int K(\mathbf{z})f(\mathbf{x}-H^{1/2}\mathbf{z})d\mathbf{z} + \frac{1}{2}\beta_2(K)\,(\operatorname{vec}H)^T \operatorname{vec}\mathscr{D}_f^2(\mathbf{x}) \\ &\quad + \frac{1}{4}\beta_2(K)^2\,(\operatorname{vec}H)^T\,\mathscr{D}_f^2 \otimes \mathscr{D}_f^2(\mathbf{x})\operatorname{vec}H + o(\operatorname{tr} H) \\ &= E\hat{f}(\mathbf{x}, H) + \frac{1}{2}\beta_2(K)\operatorname{tr}(H\mathscr{D}_f^2(\mathbf{x})) \\ &\quad + \frac{1}{4}\beta_2(K)^2\operatorname{tr} H\mathscr{D}_f^2(\mathbf{x})H\mathscr{D}_f^2(\mathbf{x}) + o(\operatorname{tr} H), \end{aligned}$$

where we use again the results from (Magnus and Neudecker, 2007): A, B, C, D square matrices $\Rightarrow \operatorname{tr}ABCD = (\operatorname{vec}D)^T(A \otimes C)^T\operatorname{vec}B^T$. In our case $D = B = H$, $A = C = \mathscr{D}_f^2(\mathbf{x})$. All matrices are symmetrical and from this statement the last expression follows immediately. Since

$$E\hat{f}(\mathbf{x}, H) = f(\mathbf{x}) + \frac{1}{2}\beta_2(K)\operatorname{tr} H\mathscr{D}_f^2(\mathbf{x}) + o(\operatorname{tr}H)$$

the statement of Theorem 5.1 is valid. □

Proof of Corollary 5.1.

$$\begin{aligned} E(\widehat{\operatorname{bias}}\hat{f}(\mathbf{x}, H)) &= E\left(\int K_H(\mathbf{x}-\mathbf{y})\hat{f}(\mathbf{y}, H)d\mathbf{y} - \hat{f}(\mathbf{x}, H)\right) \\ &= f(\mathbf{x}) + \beta_2(K)\operatorname{tr}(H\mathscr{D}_f^2(\mathbf{x})) + \frac{1}{4}\beta_2^2(K)\operatorname{tr}(H\mathscr{D}_f^2(\mathbf{x})H\mathscr{D}_f^2(\mathbf{x})) \\ &\quad + o(\operatorname{tr} H) - E(\hat{f}(\mathbf{x}, H)) \end{aligned}$$

$$= f(\mathbf{x}) + \beta_2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})) + \frac{1}{4}\beta_2^2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})H\mathscr{D}_f^2(\mathbf{x}))$$
$$- f(\mathbf{x}) - \frac{1}{2}\beta_2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})) + o(\mathrm{tr}\, H)$$
$$= \frac{1}{2}\beta_2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})) + \frac{1}{4}\beta_2^2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})H\mathscr{D}_f^2(\mathbf{x})) + o(\mathrm{tr}\, H).$$

Further, $E(\hat{f}(\mathbf{x}, H) - f(\mathbf{x})) = \frac{1}{2}\beta_2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})) + o(\mathrm{tr}\, H)$, then

$$E(\widehat{\mathrm{bias}}\hat{f}(\mathbf{x}, H)) = \mathrm{bias}\hat{f}(\mathbf{x}, H) + \frac{1}{4}\beta_2^2(K)\mathrm{tr}(H\mathscr{D}_f^2(\mathbf{x})H\mathscr{D}_f^2(\mathbf{x})) + o(\mathrm{tr}\, H)$$
$$= \mathrm{bias}\hat{f}(\mathbf{x}, H) + o(\mathrm{tr}\, H).$$

□

Acknowledgment

This research was supported by Ministry of Education, Youth and Sports of the Czech Republic under the project LC06024 and by Masaryk University under the Student Project Grant MUNI/A/1001/2009. The authors would like to thank José E. Chacón for his very helpful and constructive comments and suggestions.

References

Cao, R., Cuevas, A., González Manteiga, W. (1994). A comparative study of several smoothing methods on density estimation. *Comput. Statist. Data Anal.* 17:153–176.

Chacón, J. E., Duong, T. (2009). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test* 19:375–398.

Chaudhuri, P., Marron, J. S. (1999). SiZer for exploration of structure in curves. *J. Amer. Statist. Assoc.* 94:807–823.

Ćwik, J., Koronacki, J. (1997). A combined adaptive-mixtures/plug-in estimator of multivariate probability densities. *Comput. Statist. Data Anal.* 26:199–218.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Soft.* 21:1–16.

Duong, T., Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.* 15:17–30.

Duong, T., Hazelton, M. L. (2005a). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *J. Multivariate Anal.* 93:417–433.

Duong, T., Hazelton, M. L. (2005b). Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.* 32:485–506.

Godtliebsen, F., Marron, J. S., Chaudhuri, P. (2002). Significance in scale space for density estimation. *J. Comput. Graph. Statist.* 11:1–21.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. [On-line]. Retrieved from http://fedc.wiwi.hu-berlin.de/xplore/ebooks/html/-spm/

Horová, I., Vieu, P., Zelinka, J. (2002). Optimal choice of nonparametric estimates of a density and of its derivatives. *Statistics & Decisions* 20:355–378.

Horová, I., Koláček, J., Zelinka, J., Vopatová, K. (2008). Bandwidth choice for kernel density estimates. *Proc. IASC*, 542–551.

Horová, I., Zelinka, J. (2007). Contribution to the bandwidth choice for kernel density estimates. *Comput. Statist.* 22:31–47.

Jones, M. C., Marron, J. S., Park, B. U. (1991). A simple root *n* bandwidth selector. *Ann. Statist.* 19:1919–1932.

Magnus, J. R., Neudecker, H. (2007). *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Chichester, England: Wiley.

Sain, S. R., Baggerly, K. A., Scott, D. W. (1994). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc.* 89:807–817.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Scott, D. W., Gorry, G. A., Hoffman, R. G., Barboriak, J. J., Gotto, A. M. (1978). A new approach for evaluating risk factors in coronary artery disease: a study of lipid concentrations and severity of disease in 1847 males. *Circulation* 62:477–484.

Silverman, B. W. (1989). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* 76:705–712.

Wand, M. P., Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* 88:520–528.

Wand, M. P., Jones, M. C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* 9:97–116.

Wand, M. P., Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman & Hall.

Journal of Statistics: Advances in Theory and Applications
Volume 8, Number 2, 2012, Pages 91-103

ITERATIVE BANDWIDTH METHOD FOR KERNEL REGRESSION

JAN KOLÁČEK and IVANKA HOROVÁ

Department of Mathematics and Statistics
Masaryk University
Brno
Czech Republic
e-mail: kolacek@math.muni.cz

Abstract

The aim of the contribution is to extend the idea of an iterative method known for a kernel density estimate to kernel regression. The method is based on a suitable estimate of the mean integrated square error. This approach leads to an iterative quadratically convergent process. We conduct a simulation study comparing the proposed method with the well-known cross-validation method. Results are implemented in Matlab.

1. Univariate Kernel Density Estimator

Let $X_1, \ldots, X_n$ be independent real random variables having the same continuous density f. The symbol $\hat{f}$ will be used to denote whatever density estimation is currently being considered.

Definition 1.1. Let k be an even nonnegative integer and K be a real valued function continuous on $\mathbb{R}$ and satisfying the conditions:

(i) $|K(x) - K(y)| \le L|x - y|$ for a constant $L > 0$, $\forall x, y \in [-1, 1]$,

2010 Mathematics Subject Classification: 62G08.

Keywords and phrases: kernel regression, bandwidth selection, iterative method.

Received November 1, 2012

(ii) support $(K) = [-1, 1]$, $K(-1) = K(1) = 0$,

$$\text{(iii)} \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 < j < k, \\ 1, & j = \nu, \\ \beta_k \neq 0, & j = k. \end{cases}$$

Such a function is called a *kernel* of order k and a class of these kernels is denoted as S_{0k}.

Remark 1.2. The well-known kernels are, e.g.,

(a) Epanechnikov kernel: $K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}$,

(b) quartic kernel: $K(x) = \frac{3}{4}(1 - x^2)^2 I_{[-1,1]}$,

(c) triweight kernel: $K(x) = \frac{35}{32}(1 - x^2)^2 I_{[-1,1]}$,

(d) Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$,

where $I_{[-1,1]}$ is an indicator function. Though the Gaussian kernel does not satisfy the assumption (ii), it is very popular in many applications.

Let $K \in S_{0k}$, set $K_h(.) = \frac{1}{h} K(\frac{\cdot}{h})$, $h > 0$. A parameter h is called a *bandwidth*. The kernel estimator of f at the point $x \in \mathbb{R}$ is defined as

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

The problem of choosing the smoothing parameter is of a crucial importance and will be treated in the next sections. Our analysis requires the specification of an appropriate error criterion for measuring the error when estimating the density at a single point as well as the error when

estimating the density over the whole real line. A useful criterion when estimating at a single point is the mean square error (MSE) defined by

$$MSE\{\hat{f}(x, h)\} = E\{\hat{f}(x, h) - f(x)\}^2.$$

As concerns a global criterion, we consider the mean integrated square error

$$MISE\{\hat{f}(\cdot, h)\} = E\int \{\hat{f}(x, h) - f(x)\}^2 dx.$$

Since MISE is not mathematically tractable, we employ the asymptotic mean integrated square error (AMISE), which can be written as a sum of the asymptotic integrated variance and the asymptotic integrated square bias

$$AMISE\{\hat{f}(\cdot, h)\} = \underbrace{\frac{V(K)}{nh}}_{AIV\hat{f}} + \underbrace{h^{2k}\,\frac{\beta_k^2}{k!^2}\,V(f^{(k)})}_{AISB\hat{f}}, \tag{1.1}$$

where $V(g) = \int g^2(x)dx$. Now, by minimizing (1.1) with respect to h, we obtain the AMISE-optimal bandwidth $h_{opt,k} = \arg\min AMISE\{\hat{f}(\cdot, h)\}$, which takes the form

$$h_{opt,k}^{2k+1} = \frac{V(K)}{2knV(f^{(k)})}\,\frac{k!^2}{\beta_k^2}.$$

For more details, see, e.g., [9], [14].

2. Iterative Method for Kernel Density Estimation

The problem of choosing how much to smooth, i.e., how to choose the bandwidth is a crucial common problem in kernel smoothing. Methods for a bandwidth choice have been developed in many papers and monographs, see, e.g., [1, 2, 5, 7, 8, 11, 12, 14], and many others. However, there does not exist any universally accepted approach to this serious problem yet.

The iterative method is based on the relation

$$AIV\ \hat{f}(\cdot,\ h_{opt,k}) = 2kAISB\ \hat{f}(\cdot,\ h_{opt,k}), \tag{2.1}$$

with estimates of AIV and $AISB$

$$\widehat{AIV}\ \hat{f}\ (\cdot,\ h) = \frac{V\ (K)}{nh},$$

and

$$\widehat{AISB}\ \hat{f}\ (\cdot,\ h) = \int \left(\int K\ (x)\ \hat{f}\ (x - hy,\ h)\, dy - \hat{f}\ (x,\ h) \right)^2 dx$$

$$= \frac{1}{n^2 h} \sum_{i,\,j=1}^{n} \Lambda\left(\frac{X_i - X_j}{h} \right),$$

where

$$\Lambda(z) = (K * K * K * K - 2K * K * K + K * K)(z),$$

and $*$ denotes the convolution, i.e., $K * K(u) = \int K(t)K(u - t)dt.$ The bandwidth estimate $\hat{h}_{IT,k}$ is a solution of the equation

$$\frac{V(K)}{nh} - \frac{2k}{n^2 h} \sum_{i,\,j=1}^{n} \Lambda\left(\frac{X_i - X_j}{h} \right) = 0. \tag{2.2}$$

In the paper [8], this nonlinear equation was solved by Steffensen's method. But this equation can be rewritten as

$$\frac{2k}{n} \sum_{\substack{i,\,j=1 \\ i \neq j}}^{n} \Lambda\left(\frac{X_i - X_j}{h} \right) - V(K) = 0. \tag{2.3}$$

Since the first derivative of the function standing on the left hand side of this equation is easy to compute by using convolutions, Newton's method can be used. For more details, see [9].

3. Univariate Kernel Regression

Consider a standard regression model of the form

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where m is an unknown regression function, $Y_1, \ldots, Y_n$ are observable data variables with respect to the design points $x_1, \ldots, x_n$. The residuals $\varepsilon_1, \ldots, \varepsilon_n$ are independent identically distributed random variables for which

$$E(\varepsilon_i) = 0, \quad \operatorname{var}(\varepsilon_i) = \sigma^2 > 0, \quad i = 1, \ldots, n.$$

The aim of kernel smoothing is to find a suitable approximation $\widehat{m}$ of the unknown function m.

To avoid boundary effects, the estimate is obtained by applying the kernel on the extended series $\widetilde{Y}_i$, $i = -n+1, -n+2, \ldots, 2n$, where $\widetilde{Y}_{j\pm n} = Y_j$ for $j = 1, \ldots, n$. Similarly, $x_i = i/n$, $i = -n+1, -n+2, \ldots, 2n$.

The assumption of the cyclic model leads to the kernel regression estimator

$$\widehat{m}(x_j, h) = \frac{1}{C_n} \sum_{i=-n+1}^{2n} K_h(x_i - x_j)\, \widetilde{Y}_i, \quad j = 1, \ldots, n, \tag{3.2}$$

where $C_n = \sum_{i=-n+1}^{n-1} K_h(x_i)$. For more details about this estimator, see [9] and [10].

The quality of a kernel regression estimator can be locally described by the mean square error (MSE) or by a global criterion the mean integrated square error (MISE). According to same reasons as in kernel density estimation, we employ the asymptotic mean integrated square error (AMISE), which can be written as a sum of the asymptotic integrated variance and asymptotic integrated square bias

$$AMISE\{\widehat{m}(\cdot, h)\} = \underbrace{\frac{V(K)\sigma^2}{nh}}_{AIV} + \underbrace{\left(\frac{\beta k}{k!}\right)^2 A_k h^{2k}}_{AISB}, \tag{3.3}$$

where $A_k = \int (m^{(k)}(x))^2 dx$. The optimal bandwidth considered here is $h_{opt,k}$, the minimizer of (3.3), i.e.,

$$h_{opt,k} = \arg\min_{h \in H_n} AMISE\{\widehat{m}(\cdot, h)\},$$

where $H_n = [an^{-1/(2k+1)}, bn^{-1/(2k+1)}]$ for some $0 < a < b < \infty$.

The calculation gives

$$h_{opt,k} = \left(\frac{\sigma^2 V(K)(k!)^2}{2kn\beta_k^2 A_k}\right)^{\frac{1}{2k+1}}. \tag{3.4}$$

In nonparametric regression estimation, like in density estimation, a critical and inevitable step is to choose the smoothing parameter (bandwidth) to control the smoothness of the curve estimate. The smoothing parameter considerably affects the features of the estimated curve.

One of the most widespread procedures for bandwidth selection is the cross-validation method, also known as "leave-one-out" method.

The method is based on modified regression smoother (3.2) in which one, say the j-th, observation is left out

$$\widehat{m}_{-j}(x_j, h) = \frac{1}{C_n} \sum_{\substack{i=-n+1 \\ i \neq j}}^{2n} K_h(x_i - x_j)\widetilde{Y}_i, \quad j = 1, \ldots, n.$$

With using these modified smoothers, the error function which should be minimized takes the form

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{m}_{-i}(x_i) - Y_i\}^2. \tag{3.5}$$

The function $CV(h)$ is commonly called a "*cross-validation*" function. Let $\hat{h}_{CV}$ stand for minimization of $CV(h)$, i.e.,

$$\hat{h}_{CV} = \arg\min_{h \in H_n} CV(h).$$

The literature on this criterion is quite extensive, e.g., [3, 4, 6, 13].

4. Iterative Method for Kernel Regression

The proposed method is based on the similar relation as in the kernel density estimation. It is easy to show that the following equation holds:

$$AIV\,\widehat{m}\,(\cdot,\, h_{opt,k}) = 2kASB\,\widehat{m}\,(\cdot,\, h_{opt,k}), \tag{4.1}$$

where

$$AIV\,\widehat{m}\,(\cdot,\, h) = \frac{\sigma^2 V(K)}{nh},$$

and

$$ASB\,\widehat{m}\,(\cdot,\, h) = \frac{1}{n}\sum_{i=1}^{n}\left\{E\,\widehat{m}\,(x_i,\, h) - m(x_i)\right\}^2.$$

For estimating of AIV and ASB in (4.1), we use

$$\widehat{AIV}\,\widehat{m}\,(\cdot,\, h) = \frac{\hat{\sigma}^2 V(K)}{nh}, \text{ with } \hat{\sigma}^2 = \frac{1}{2n-2}\sum_{i=2}^{n}(Y_i - Y_{i-1})^2,$$

and

$$\widehat{ASB}\,\widehat{m}\,(\cdot,\, h) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{C_n}\sum_{i=-n+1}^{2n} K_h(x_i - x_j)\left[\widetilde{Y}_i - \frac{1}{C_n}\sum_{l=-n+1}^{2n} K_h(x_l - x_i)\widetilde{Y}_l\right]\right)^2.$$

To find the bandwidth estimate $\hat{h}_{IT,k}$, we solve the following equation:

$$h = \frac{\hat{\sigma}^2 V(K)}{2kn\widehat{ASB}\,\widehat{m}(\cdot, h)}. \tag{4.2}$$

We use Steffensen's iterative method with the starting approximation $\hat{h}_0 = k/n$. This approach leads to an iterative quadratically convergent process.

5. Simulation Study

We carry out two simulation studies to compare the performance of the bandwidth estimates. The comparison is done by the following way. The observations, Y_i, for $i = 1, \ldots, n = 100$, are obtained by adding independent Gaussian random variables with mean zero and variance σ^2 to some known regression function. Both regression functions used in our simulations are illustrated in Figure 1. They are not chosen randomly for our comparison. The first one is suitable for the extension to the cyclic model, on the other side, the second function does not satisfy the assumption for the cyclic model.

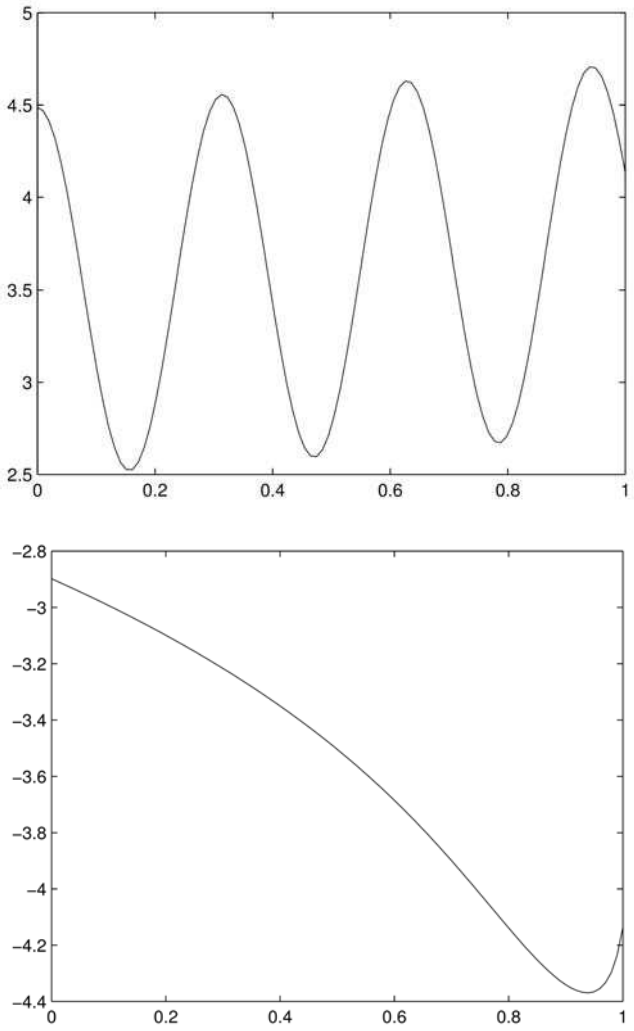


Figure 1. Regression functions.

One hundred series are generated. For each data set, we estimate the optimal bandwidth by both mentioned methods, i.e., for each method, we obtain 100 estimates. Since we know the optimal bandwidth, we compare it with the mean of estimates and look at their standard deviation, which describes the variability of all methods. The Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]}$ is used in all cases.

5.1. Simulation 1. In this case, we use the regression function

$$m(x) = \cos(20x) + 2\sin\left(4 - \frac{x}{6}\right) + 5,$$

with $\sigma^2 = 0.3$. Table 1 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 100 values and $std(\hat{h})$ is their standard deviation. Figure 2 illustrates the histogram of results of all 100 experiments.

Table 1. Means and standard deviations

	$h_{opt,2} = 0.0560$	
	$E(\hat{h})$	$std(\hat{h})$
CV	0.0550	0.0120
IT	0.0556	0.0048

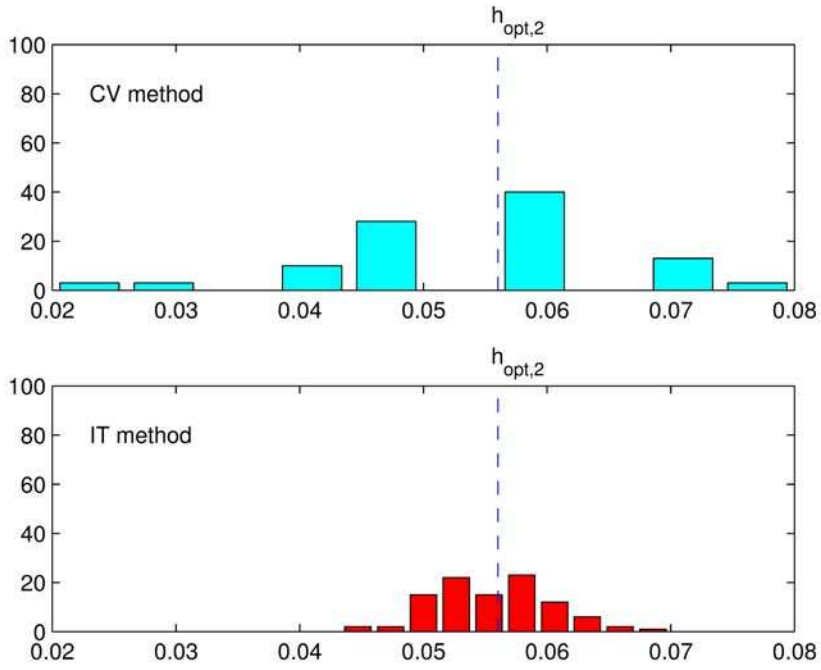


Figure 2. Distribution of $\hat{h}$ for both methods.

As we see, the standard deviation of all results obtained by the proposed method is less than the value for the case of cross-validation method and also the mean of these results is a little bit closer to the theoretical optimal bandwidth. The reason is that the regression function is smooth and satisfies the conditions for the extension to the cyclic design. Thus, the proposed method works very well in this case.

5.2. Simulation 2. In the second example, we use the regression function

$$m(x) = \ln(11 - 10x) + \cot(5 - x^{13}) - 5,$$

with $\sigma^2 = 0.05$. Table 2 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 100 values and $std(\hat{h})$ is their standard deviation. Figure 3 illustrates the histogram of results of all 100 experiments.

Table 2. Means and standard deviations

	$h_{opt,2} = 0.0707$	
	$E(\hat{h})$	$std(\hat{h})$
CV	0.1466	0.0443
IT	0.0592	0.0112

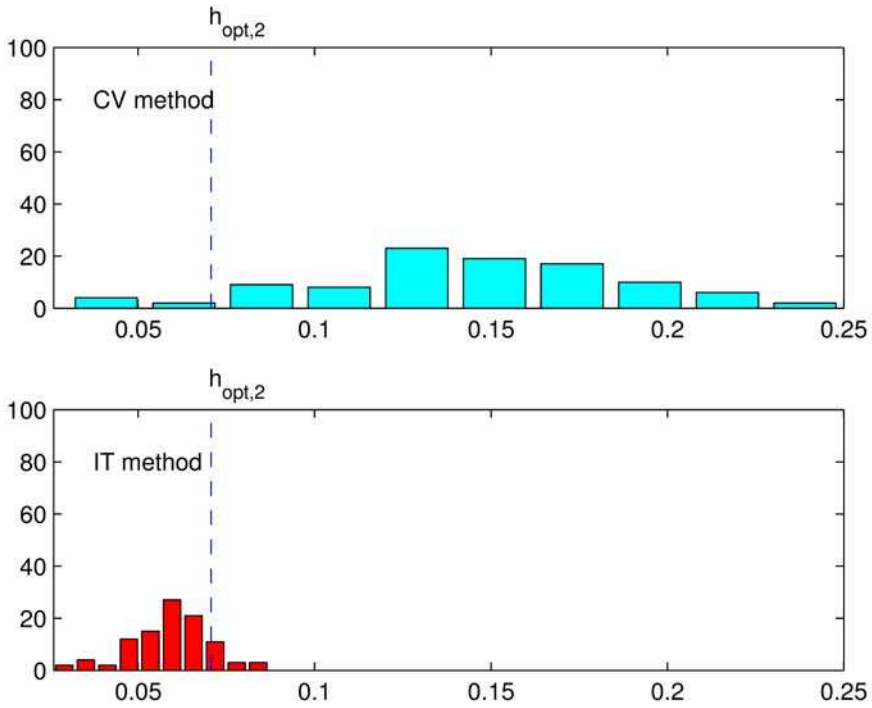


Figure 3. Distribution of $\hat{h}$ for both methods.

It is evident that better results are obtained by the proposed method. This method is successful despite the fact that the regression function does not meet assumptions for the extension to the cyclic model. The cross-validation method often results in smaller bandwidths. The variance of this criterion is also significant.

Acknowledgement

This research was supported by Masaryk University under the project MUNI/A/1001/2009.

References

[1] R. Cao, A. Cuevas and W. González Manteiga, A comparative study of several smoothing methods in density estimation, Computational Statistics and Data Analysis 17(2) (1994), 153-176.

[2] P. Chaudhuri and J. S. Marron, Sizer for exploration of structures in curves, Journal of the American Statistical Association 94(447) (1999), 807-823.

[3] P. Craven and G. Wahba, Smoothing noisy data with spline functions estimating the correct degree of smoothing by the method of generalized cross-validation, Numerische Mathematik 31(4) (1979), 377-403.

[4] Bernd Droge, Some Comments on Cross-Validation, Technical Report 1994-7, Humboldt Universitaet Berlin, 1996.

[5] Jianqing Fan and Irene Gijbels, Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation, Journal of the Royal Statistical Society, Series B (Methodological) 57(2) (1995), 371-394.

[6] W. Härdle, Applied Nonparametric Regression, 1st Edition, Cambridge University Press, Cambridge, 1990.

[7] W. Härdle, M. Müller, S. Sperlich and A. Wewatz, Nonparametric and Semiparametric Models, 1st Edition, Springer, Heidelberg, 2004.

[8] I. Horová and J. Zelinka, Contribution to the bandwidth choice for kernel density estimates, Computational Statistics 22(1) (2007), 31-47.

[9] I. Horová, J. Koláček and J. Zelinka, Kernel Smoothing in MATLAB, World Scientific, Singapore, 2012.

[10] Jan Koláček, Plug-in method for nonparametric regression, Computational Statistics 23(1) (2008), 63-78.

[11] David W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, 1992.

[12] Bernard W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[13] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society Series B-Statistical Methodology 36(2) (1974), 111-147.

[14] M. P. Wand and M. C. Jones, Kernel Smoothing, Chapman and Hall, London, 1995.

■

Journal of Applied Probability and Statistics
Vol. 6, No. 1&2, pp. 73-85
ISOSS Publications 2012

A GENERALIZED REFLECTION METHOD FOR KERNEL DISTRIBUTION AND HAZARD FUNCTIONS ESTIMATION

Jan Koláček

Jan Koláček, Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic
Email: kolacek@math.muni.cz

Rohana J. Karunamuni

Rohana J. Karunamuni, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada
Email: R.J.Karunamuni@ualberta.ca

SUMMARY

In this paper we focus on kernel estimates of cumulative distribution and hazard functions (rates) when the observed random variables are nonnegative. It is well known that kernel distribution estimators are not consistent when estimating a distribution function near the point $x = 0$. This fact is rather visible in many applications, for example in kernel ROC curve estimation [10]. In order to avoid this problem we propose a bias reducing technique that is a kind of generalized reflection method. Our method is based on ideas of [8] and [19] developed for boundary correction in kernel density estimation. The proposed estimators are compared with the traditional kernel estimator and with the estimator based on "classical" reflection method using simulation studies.

Keywords and phrases: kernel estimation, reflection, distribution function, hazard function.

AMS Classification: 30C40, 62G30.

1 Introduction

The most commonly used nonparametric estimate of a cumulative distribution function F is the empirical distribution function F_n, where $F_n(x) = n^{-1} \sum_{i=1}^{n} I[X_i \leq x]$ with X_1, ..., X_n being the observations. But F_n is a step function even in the case that F is a continuous function. Another type of nonparametric estimator for F is derived from kernel smoothing methods. Kernel smoothing is most widely used because it is easy to apply and produce estimators which have good small and asymptotic properties. Kernel smoothing has received a lot of attention in density estimation. Good references in this area are [3], [16] and [17].

However, results in kernel distribution function estimation are relatively few. Theoretical properties of kernel distribution function estimator have been investigated by [12], [14] and [1]. Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied. The same can be said about estimation of hazard function (rates) estimation.

In this paper, we develop a new kernel type estimator of the cumulative distribution and hazard rates that removes boundary effects near the end points of the support. Our estimator is based on a new boundary corrected kernel estimator of distribution function and it is based on ideas of [6], [7], [8] and [19] developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We derive expressions for the bias and variance of the proposed estimators. Furthermore, the proposed estimators are compared with the traditional estimator and with the estimator based on "classical" reflection method using simulation studies. We observe that the proposed estimators successfully remove boundary effects and performs considerably better than the others two.

Kernel smoothing in distribution function estimation and boundary effects are discussed in the next section. The proposed estimator of distribution functions is given in Section 3. Section 4 discusses estimation of hazard functions (rates). Simulation results are given in Section 5 and our results are applied on real data in Section 6. Finally, some concluding remarks are given in Section 7.

2 Kernel distribution estimator and boundary effects

Let f denote a continuous density function with support $[0, a]$, $0 < a \leq \infty$, and consider nonparametric estimation of the cumulative distribution function F of f based on a random sample X_1, ..., X_n from f. Suppose that $F^{(j)}$, the j-th derivative of F, exists and is continuous on $[0, a]$, $j = 0, 1, 2$, with $F^{(0)} = F$ and $F^{(1)} = f$. Then the traditional kernel estimator of F is given by

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt \tag{2.1}$$

where K is a unimodal symmetric density function with support $[-1, 1]$ and h is the bandwidth ($h \to 0$ as $n \to \infty$). Set $\beta_2 = \int_{-1}^{1} t^2 K(t)dt$. The basic properties of $\widehat{F}_{h,K}(x)$ at interior points are well-known (e.g., [11]), and under some smoothness assumptions these include, for $h \leq x \leq a - h$,

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = \frac{1}{2}\beta_2 f^{(1)}(x)h^2 + o(h^2)$$

$$n\mathrm{Var}(\widehat{F}_{h,K}(x)) = F(x)(1-F(x)) + hf(x)\int_{-1}^{1} W(t)(W(t)-1)dt + o(h).$$

The performance of $\widehat{F}_{h,K}(x)$ at boundary points, i.e., for $x \in [0,h)\cup(a-h,a]$, however, differs from the interior points due to so-called "boundary effects" that occur in nonparametric curve estimation problems. More specifically, the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$ at boundary points while the variance of $\widehat{F}_{h,K}(x)$ is of the same order. This fact can be clearly seen by examining the behavior of $\widehat{F}_{h,K}$ inside the left boundary region $[0,h]$. Let x be a point in the left boundary, i.e., $x \in [0,h]$. Then we can write $x = ch,\ 0 \le c \le 1$. It can be shown that the bias and variance of $\widehat{F}_{h,K}(x)$ at $x = ch$ are of the form

$$\begin{aligned}\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = {} & hf(0)\int_{-1}^{-c} W(t)dt \\ & + h^2 f^{(1)}(0)\left\{\frac{c^2}{2} + c\int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt\right\} + o(h^2)\end{aligned} \tag{2.2}$$

$$n\mathrm{Var}(\widehat{F}_{h,K}(x)) = F(x)(1-F(x)) + hf(0)\left\{\int_{-1}^{c} W^2(t)dt - c\right\} + o(h). \tag{2.3}$$

From the expression (2.2) it is now clear that the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$. To remove this boundary effect in kernel distribution estimation we investigate a new class of estimators in the next section.

3 The proposed estimator

In this section we propose a class of estimators of the distribution function F of the form

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}, \tag{3.1}$$

where h is the bandwidth, W is a cumulative distribution function defined by (2.1) and g_1 and g_2 are two transformations that need to be determined. We assume that $g_i,\ i = 1,2$, are nonnegative, continuous and monotonically increasing functions defined on $[0,\infty)$. Further assume that g_i^{-1} exists, $g_i(0) = 0,\ g_i^{(1)}(0) = 1$, and that $g_i^{(2)}$ exists and is continuous on $[0,\infty)$, where $g_i^{(j)}$ denotes the j-th derivative of g_i, with $g_i^{(0)} = g_i$ and g_i^{-1} denoting the inverse function of $g_i,\ i = 1,2$. We will choose g_1 and g_2 so that $\widetilde{F}_{h,K}(x) \ge 0$ everywhere. Note that the i-th term of the sum in (3.1) can be expressed as

$$W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right) = \int_{\frac{-x+g_1(X_i)}{h}}^{\frac{x+g_2(X_i)}{h}} K(t)dt.$$

The preceding integral is non-negative provided the inequality $\frac{-x+g_1(X_i)}{h} \leq \frac{x+g_2(X_i)}{h}$ holds. Since $x \geq 0$, the preceding inequality will be satisfied if g_1 and g_2 are such that $g_1(X_i) \leq g_2(X_i)$ for $i = 1, \ldots, n$. Thus we will assume that g_1 and g_2 are chosen such that $g_1(x) \leq g_2(x)$ for $x \in [0, \infty)$ for the proposed estimator. Now, we can obtain the bias and variance of (3.1) at $x = ch, 0 \leq c \leq 1$, as

$$\begin{aligned} \mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = h^2 \Bigg\{ & f^{(1)}(0) \left(\frac{c^2}{2} + 2c \int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt \right) \\ & - f(0) g_1^{(2)}(0) \int_{-1}^{c} (c-t) W(t)dt \\ & - f(0) g_2^{(2)}(0) \int_{-1}^{-c} (c+t) W(t)dt \Bigg\} + o(h^2). \end{aligned} \tag{3.2}$$

$$\begin{aligned} n\mathrm{Var}(\widetilde{F}_{h,K}(x)) = F(x)(1-F(x)) + hf(0) \Bigg\{ & \int_{-1}^{c} W^2(t)dt \\ & - 2 \int_{-1}^{c} W(t)W(t-2c)dt + \int_{-1}^{-c} W^2(t)dt \Bigg\} + o(h). \end{aligned} \tag{3.3}$$

The proofs of (3.2) and (3.3) are given in [10]. Similarly we can express the bias and variance of (3.1) at "interior" points $x = c > 1$. Note that the contribution of g_2 on the bias vanishes as $c \to 1$. By comparing expressions (2.2), (3.2), (2.3) and (3.3) at boundary points we can see that the variances are of the same order and the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ while the bias of $\widetilde{F}_{h,K}(x)$ is of order $O(h^2)$. So our estimator removes boundary effects in kernel distribution estimation since the bias at boundary points is of the same order as the bias at interior points.

It is clear that there are various possible choices available for the pair (g_1, g_2). However, we will choose g_1 and g_2 so that the condition $\widetilde{F}_{h,K}(0) = 0$ will be satisfied because of the fact that $F(0) = 0$. A sufficient (but not necessary) condition for the preceding to be satisfied is that g_1 and g_2 must be equal. Thus we need to construct a single transformation function g such that $g = g_1 = g_2$. Other important properties that are desirable in the estimator $\widetilde{F}_{h,K}$ are the local adaptivity, that is the transformation function g depends on c.

Some discussion on the choice of g_c and other various improvements that can be made would be appropriate here. The trivial choice is $g_c(y) = y$, which represents the "classical" reflection method estimator. However, it is possible to construct functions g_c's that improve the bias further under some additional conditions. For instance, if one examines the right hand side of the bias expansion (3.2) then it is not difficult to see that the terms inside bracket (i.e., the coefficient of h^2) can be made equal to zero if g_c is appropriately chosen.

Set

$$A_c = \begin{cases} d_1 \frac{\frac{c^2}{2}+2cI_1-I_2}{c^2+2cI_1-I_2}, & \text{for } 0 \le c < 1 \\ d_1 \frac{\beta_2}{c^2+\beta_2}, & \text{for } c > 1 \end{cases}$$

where $d_1 = \frac{f^{(1)}(0)}{f(0)}$, $I_1 = \int_{-1}^{-c} W(t)dt$, $I_2 = \int_{-c}^{c} tW(t)dt$.

If g_c is chosen such that $g_c^{(2)}(0) = A_c$ then the bias of $\widetilde{F}_{h,K}(x)$ would be theoretically of order $O(h^3)$. For such a function g_c, the second derivative at zero $g_c^{(2)}(0)$ will be dependent on the ratio $d_1 = \frac{f^{(1)}(0)}{f(0)}$. Then the problem of estimation of d_1 naturally arises as in the papers of [6], [7], [8] and [9]. For example, the ratio $d_1 = \frac{f^{(1)}(0)}{f(0)}$ is estimated in [9] as the first derivative of natural logarithm of f at zero. For more details, especially for the exact formula for $\widehat{d}_1$ and for some statistical properties, especially for the asymptotic convergence rate, see the preceding paper.

Summarizing all the assumptions, it is clear now that g_c should satisfy the following conditions:

(i) $g_c : [0, \infty) \to [0, \infty)$, g_c is continuous, monotonically increasing and $g_c^{(i)}$ exists, $i = 1, 2$,

(ii) $g_c^{-1}(0) = 0$, $g_c^{(1)}(0) = 1$

(iii) $g_c^{(2)}(0) = A_c$.

Functions satisfying conditions (i) – (iii) are easy to construct. We will consider the following transformation. For $y \ge 0$, let us define

$$g_c(y) = y + \frac{1}{2}\widehat{A}_c y^2 + \lambda \widehat{A}_c^2 y^3, \tag{3.4}$$

where $\widehat{A}_c$ is an estimator of A_c based on an estimator $\widehat{d}_1$ of d_1, and λ is a positive constant such that $\lambda > \frac{1}{12}$. This condition on λ is necessary for $g_c(y)$ to be an increasing function of y. Based on extensive simulations, we find that this transformation adapts well to various shapes of distribution functions with setting $\lambda = 0.1$.

4 Estimation of hazard rates

Given a distribution F with probability density function f, the hazard rate is defined by

$$z(t) = \frac{f(t)}{1 - F(t)}. \tag{4.1}$$

The hazard rate is also called the age-specific or conditional failure rate. It is useful particularly in the context of reliability theory and survival analysis and hence in fields as diverse as engineering and medical statistics. See [2] for a discussion of the role of the hazard rate in understanding and modeling survival data. [16] provides a survey of some methods for nonparametric estimation of hazard estimation.

Given a sample X_1, ..., X_n from the density f, a natural nonparametric estimator of the hazard rate is $\widehat{z}(t) = \widehat{f}(t)/(1-\widehat{F}(t))$, where $\widehat{f}$ is a suitable density estimator of f based on X_1, ..., X_n and $\widehat{F}(t) = \int_{-\infty}^{t} \widehat{f}(x)dx$ estimates $F(t)$. If $\widehat{f}$ is the traditional kernel estimator with a kernel K and bandwidth h, then $\widehat{F}(t)$ can be obtained by $\widehat{F}(t) = n^{-1}\sum_{i=1}^{n} K_1((t-X_i)/h)$, where $K_1(u) = \int_{-\infty}^{u} K(t)dt$. [18] introduced and discussed $\widehat{z}(t)$ and various alternative nonparametric estimators of $z(t)$. For further properties of $\widehat{z}(t)$ with kernel and other related estimators see, e.g., [15], [13] (Section 4.3) and [16] (Section 6.5).

It has been observed that consideration of errors involved in the construction of $\widehat{z}$ show that, to a first approximation, the main contribution to the error will be due to the numerator of $\widehat{z}$, i.e., due to the estimator $\widehat{f}$; see, e.g., [16] (Section 6.5). Thus, to obtain the best possible estimate of the hazard rate, one should aim to minimize the error in the estimation of density f. If the support of f is the interval $[0,\ a]$, $0 < a \leq \infty$, which is usually the case in survival and reliability data, then the traditional kernel estimators of density f suffer from boundary effects. Therefore, it is advisable to use boundary adjusted estimators of density f and the distribution F in this context. For this purpose here we implement a boundary adjusted kernel density estimator similar to the one proposed in [6] and the boundary adjusted distribution function estimator $\widetilde{F}_{h,K}$ given above. Thus, the proposed estimator of the hazard rate $z(t)$ is given by, for $t = ch$, $c \geq 0$,

$$\widetilde{z}(t) = \frac{\widetilde{f}(t)}{1 - \widetilde{F}_{h,K}(t)}, \tag{4.2}$$

where $\widetilde{F}_{h,K}$ is defined by (3.1) and $\widetilde{f}$ is defined by

$$\widetilde{f}(t) = \frac{1}{nh}\sum_{i=1}^{n}\left\{K\left(\frac{t - \widehat{g}_{1,c}(X_i)}{h}\right) + K\left(\frac{t + \widehat{g}_{1,c}(X_i)}{h}\right)\right\}, \tag{4.3}$$

where

$$\widehat{g}_{1,c}(x) = x + \frac{1}{2}\widehat{d}_1 k_c x^2 + \lambda_0(\widehat{d}_1 k_c)^2 y^3, \tag{4.4}$$

with $\widehat{d}_1$ as defined in [9], λ_0 is a positive constant such that $12\lambda_0 > 1$, and k_c given by, for $c \geq 0$,

$$k_c = \frac{2\int_c^1 (u-c)K(u)du}{c + 2\int_c^1 (u-c)K(u)du}. \tag{4.5}$$

Theorem 1. *The mean squared error (MSE) of $\widetilde{z}(t)$ is given by, for $t = ch$, $c \geq 0$,*

$$E(\widetilde{z}(t) - z(t))^2 = \left(\frac{1-F(t)}{w_1 w_2}\right)^2 \frac{f(0)}{nh}\left[2\int_c^1 K(t)K(2c-t)dt + V(K)\right] + o\left(\frac{1}{nh}\right), \tag{4.6}$$

where $w_i, i = 1, 2$ are finite constants satisfying $1 - \widetilde{F}_{h,K}(t) \geq w_1 > 0$ and $1 - F(t) \geq w_2 > 0$ and $V(K) = \int_{-1}^{1} K^2(x)dx$.

Proof. For a detailed proof see Appendix. □

5 A simulation study

To test the effectiveness of our estimator, we simulated its performance against the classical reflection method. The simulation is based on 1 000 replications. In each replication, the random variables $X \sim Exp(1)$ were generated and the estimate of the hazard function was computed. Let us note that the real hazard function in this case is constant equal to one.

In all replications the sample size of $n = 100$ was used. In this case, the actual global optimal bandwidth (see [1]) for F is $h_F = 0.8479$ and for f is $h_f = 0.7860$ (see [16]). For kernel estimation of both needed functions (distribution and density) we have used the Epanechnik kernel $K(x) = \frac{3}{4}(1-x^2)I_{[-1,1]}(x)$, where I_A is the indicator function on the set A.

For each estimated hazard function we have calculated the mean integrated squared Error (MISE) on the interval $[0, h_F]$ over all 1 000 replications and have displayed the results in a boxplot in Figure 1. The variance of each estimator can be accurately gauged by the whiskers of the plot. The values of means and standard deviations for MISE of each method are given in Table 1. As we can see the reflection method gives the smaller values of MISE than the classical estimator, but the variance is not so small. From this point of view the proposed estimator seems to be better.

To get more detailed information about estimators we have calculated the Mean Squared error (MSE) at four points in the boundary region $x = ch_F$, $c = 0, 0.25, 0.5, 0.75$. The boxplot of MSE for each estimator over all 1 000 replications is illustrated in Figure 2. The values of means and standard deviations for MSE at each point for each method are given in Table 2. These values describe the performance of our proposed method with respect to MSE when compared with the classical and reflection method estimators. The values of mean and also of the variance were smallest in the case of our proposed estimator. This is

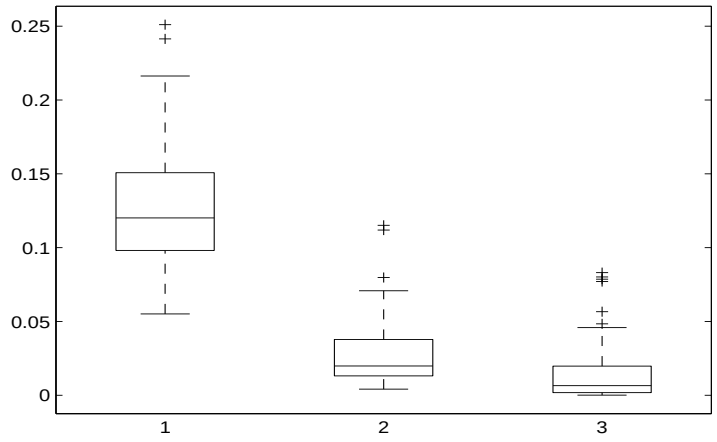


Figure 1: MISE for estimates of $z(t)$ for the classical estimator with boundary effects (1), the reflection method (2) and for our proposed method (3).

Table 1: Means and STD's for MISE

Method	*Mean*	*STD*
Classical	0.1265	0.0376
Reflection	0.0273	0.0209
Proposed	0.0142	0.0185

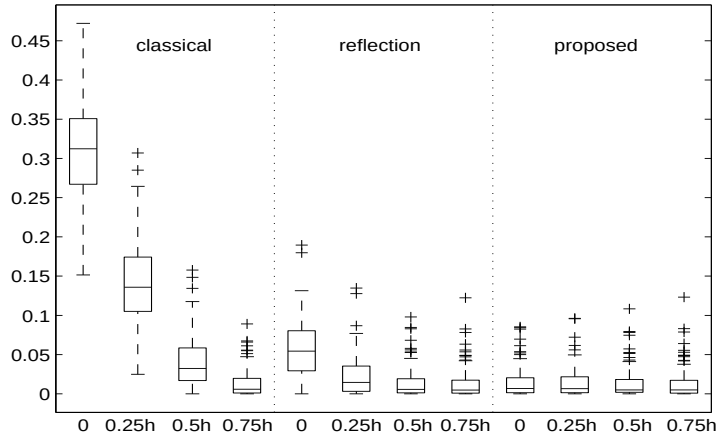


Figure 2: MSE at points $x = ch_F$, $c = 0, 0.25, 0.5, 0.75$ for the classical estimator with boundary effects, the reflection method and for our proposed method.

caused by a local adaptivity of our estimator. On other hand, the classical and reflection method estimators are not locally adaptive. From the figures and tables it is clear that the proposed estimator performed the best among the three compared. It captures the features of the distribution and hazard functions correctly with minimum bias while holding onto a low variance.

Table 2: Means and STD's for MSE at $x = ch_F$.

	Classical		*Reflection*		*Proposed*	
c	*Mean*	*STD*	*Mean*	*STD*	*Mean*	*STD*
0.00	0.3103	0.0591	0.0582	0.0369	0.0149	0.0195
0.25	0.1398	0.0528	0.0229	0.0261	0.0144	0.0194
0.50	0.0421	0.0346	0.0140	0.0198	0.0137	0.0198
0.75	0.0140	0.0183	0.0139	0.0210	0.0139	0.0210

6 Real data

In this section we apply our results to a real data set. For our analysis, we have used the suicide data from [16]. The proposed hazard rate estimate is given in Figure 3. The solid line represents our proposed estimator (4.2) and the dashed line is for the traditional kernel estimator (with boundary effects). When choosing the optimal bandwidths for the density and distribution function estimation, we used iterative methods described in [5] and [4]. The optimal bandwidths for the density and the distribution function were estimated as $\hat{h}_f = 132.01$ and $\hat{h}_F = 144.83$, respectively. The proposed estimator of hazard rate again captures proper features of the actual hazard rate, while the traditional estimator dip near the left end point due to boundary effects.

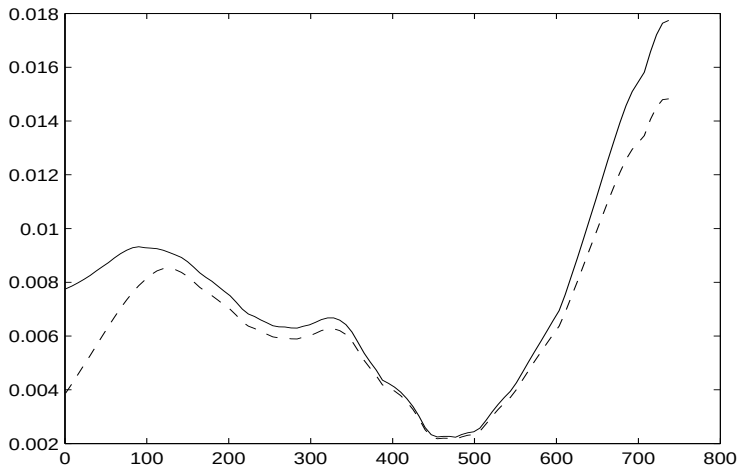


Figure 3: Hazard rate estimates constructed from the suicide data.

7 Conclusion

In this paper we proposed new kernel-type estimators to estimate the distribution and hazard functions without boundary effects near the endpoints of the support. The technique implemented is a kind of generalized reflection method involving reflecting a transformation of the data. The proposed method generates a class of boundary corrected estimators and it is based on ideas of boundary corrections for kernel density estimators presented in [6], [7] and [8]. We showed some good properties of our proposed method (e.g., local adaptivity). Furthermore, it is shown that bias of the proposed estimator is better than that of the "classical" one. The proposed estimators performed quite well in the numerical studies compared to the classical and reflection method estimators.

8 Acknowledgements

The research was supported by The Jaroslav Hájek center for theoretical and applied statistics (MŠMT LC 06024).

Appendix: Proof of Theorem 1

Theorem 1 *The mean squared error (MSE) of* $\widetilde{z}(t)$ *is given by, for* $t = ch$, $c \geq 0$,

$$E(\widetilde{z}(t) - z(t))^2 = \left(\frac{1-F(t)}{w_1 w_2}\right)^2 \frac{f(0)}{nh}\left[2\int_c^1 K(t)K(2c-t)dt + V(K)\right] + o\left(\frac{1}{nh}\right),$$

where $w_i, i = 1, 2$ *are finite constants satisfying* $1 - \widetilde{F}_{h,K}(t) \geq w_1 > 0$ *and* $1 - F(t) \geq w_2 > 0$ *and* $V(K) = \int_{-1}^{1} K^2(x)dx$.

Proof. The difference $\widetilde{z}(t) - z(t)$ is equal to, for $t = ch$, $c \geq 0$,

$$\begin{aligned}\widetilde{z}(t) - z(t) &= \frac{\widetilde{f}(t)}{1-\widetilde{F}_{h,K}(t)} - \frac{f(t)}{1-F(t)} \\ &= \frac{\widetilde{f}(t)\,(1-F(t)) - f(t)(1-\widetilde{F}_{h,K}(t)}{(1-\widetilde{F}_{h,K}(t))(1-F(t))}.\end{aligned}$$

Since we are only concerned about the behavior of $\widetilde{z}(t)$ near the left boundary, i.e., $t = ch$, $c \geq 0$, we only need to study the difference near the left endpoint 0. For $t = ch$, $c \geq 0$ we can assume that $1 - \widetilde{F}_{h,K}(t) \geq w_1 > 0$ and $1 - F(t) \geq w_2 > 0$, where $w_i, i = 1, 2$ are finite constants. The preceding conditions are reasonable, since $\widetilde{F}_{h,K}(0) = 0$, $F(0) = 0$ and $\widetilde{F}_{h,K}$ and F are continuous functions. Therefore we obtain

$$(\widetilde{z}(t) - z(t))^2 \leq (w_1 w_2)^{-2}(\widetilde{f}(t)\,(1-F(t)) - f(t)(1-\widetilde{F}_{h,K}(t)))^2.$$

To get the formula for MSE of $\widetilde{z}(t)$ we need to express $E(\widetilde{f}(t)\,(1-F(t)) - f(t)(1-\widetilde{F}_{h,K}(t))^2$.

$$
\begin{aligned}
&E(\widetilde{f}(t)\,(1-F(t)) - f(t)(1-\widetilde{F}_{h,K}(t))^2 \\
&\quad = \quad (1-F(t))^2 E\widetilde{f}^2(t) + f^2(t)E(1-\widetilde{F}_{h,K}(t))^2 - 2f(t)(1-F(t))E\widetilde{f}(t)(1-\widetilde{F}_{h,K}(t)) \\
&\quad = \quad (1-F(t))^2\left[var\widetilde{f}(t) + (E\widetilde{f}(t))^2\right] + f^2(t)\left[var\widetilde{F}_{h,K}(t) + (1-E\widetilde{F}_{h,K}(t))^2\right] \\
&\qquad -2f(t)(1-F(t))\left[E\widetilde{f}(t)(1-E\widetilde{F}_{h,K}(t)) + o\left(\frac{1}{nh}\right)\right] \\
&\quad = \quad (1-F(t))^2\left\{\frac{f(0)}{nh}\left[2\int_c^1 K(u)K(2c-u)du + V(K)\right] + o\left(\frac{1}{nh}\right) + f^2(t) + o(h)\right\} \\
&\qquad +f^2(t)\left\{\frac{1}{n}F(t)(1-F(t)) + \frac{hf(0)}{n}\left[\int_{-1}^{c} W^2(u)du - 2\int_{-1}^{c} W(u)W(u-2c)du\right.\right. \\
&\qquad\qquad \left.\left. + \int_{-1}^{-c} W^2(u)du\right] + o(h) + (1-F(t))^2 + o(h^2)\right\} \\
&\qquad -2f(t)(1-F(t))\left[f(t)+o(h)\right]\left[1-F(t)+o(h^2)\right] + o\left(\frac{1}{nh}\right) \\
&\quad = \quad (1-F(t))^2\frac{f(0)}{nh}\left[2\int_c^1 K(u)K(2c-u)du + V(K)\right] + o\left(\frac{1}{nh}\right).
\end{aligned}
$$

□

References

[1] A. Azzalini, (1981). A note on the estimation of a distribution function and quantiles by a kernel method, *Biometrika*, 68, 326–328.

[2] D. Cox and D. Oakes, (1984). *Analysis of survival data*, London, New York: Chapman and Hall.

[3] T. Gasser, H. Müller and V. Mammitzsch, (1985). Kernels for nonparametric curve estimation, *Journal of the Royal Statistical Society. Series B*, 47, 238–252.

[4] I. Horová, J. Koláček, J. Zelinka and A.H. El-Shaarawi, (2008). Smooth Estimates of Distribution Functions with Application in Environmental Studies, *Advanced topics on mathematical biology and ecology*, pp. 122–127.

[5] I. Horová and J. Zelinka, (2007). Contribution to the bandwidth choice for kernel density estimates, *Computational Statistics*, 22, 31–47.

[6] R. Karunamuni and T. Alberts, (2005a). A generalized reflection method of boundary correction in kernel density estimation, *Canad. J. Statist.*, 33, 497–509.

[7] R. Karunamuni and T. Alberts, (2005b). On boundary correction in kernel density estimation, *Statistical Methodology*, 2, 191–212.

[8] R. Karunamuni and T. Alberts, (2006). A locally adaptive transformation method of boundary correction in kernel density estimation, *J. Statist. Plann. Inference*, 136, 2936–2960.

[9] R. Karunamuni and S. Zhang, (2007). Some improvements on a boundary corrected kernel density estimator, *Statistics & Probability Letters*, 78, 497–507.

[10] J. Koláček and R. Karunamuni, (2009). On boundary correction in kernel estimation of ROC curves, *Austrian Journal of Statistics*, 38, 17–32.

[11] M. Lejeune and P. Sarda, (1992). Smooth estimators of distribution and density functions, *Computational Statistics & Data Analysis*, 14, 457–471.

[12] E. Nadaraya, (1964). Some new estimates for distribution functions, *Theory Probab. Appl.*, 15, 497–500.

[13] B. Prakasa Rao, (1983). *Nonparametric functional estimation*, Academic Press.

[14] R. Reiss, (1981). Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, 8, 116–119.

[15] J. Rice and M. Rosenblatt, (1976). Estimation of the log survivor function and hazard function, *Sankhya*, 38, 60–78.

[16] W. Silverman, (1986). *Density estimation for statistics and data analysis*, London: Chapman and Hall.

[17] M. Wand and M. Jones, (1995). *Kernel smoothing*, London: Chapman and Hall.

[18] G. Watson and M. Leadbetter, (1964). Hazard Analysis I, *Biometrika*, 51, 175–184.

[19] S. Zhang, R. Karunamuni and M. Jones, (1999). An improved estimator of the density function at the boundary, *J. Amer. Statist. Assoc.*, 94, 1231–1241.

AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 1, 17–32

On Boundary Correction in Kernel Estimation of ROC Curves

Jan Koláček[1] and Rohana J. Karunamuni[2]

[1]Dept. of Mathematics and Statistics, Brno
[2]Dept. of Mathematical and Statistical Sciences, University of Alberta

Abstract: The Receiver Operating Characteristic (ROC) curve is a statistical tool for evaluating the accuracy of diagnostics tests. The empirical ROC curve (which is a step function) is the most commonly used non-parametric estimator for the ROC curve. On the other hand, kernel smoothing methods have been used to obtain smooth ROC curves. The preceding process is based on kernel estimates of the distribution functions. It has been observed that kernel distribution estimators are not consistent when estimating a distribution function near the boundary of its support. This problem is due to "boundary effects" that occur in nonparametric functional estimation. To avoid these difficulties, we propose a generalized reflection method of boundary correction in the estimation problem of ROC curves. The proposed method generates a class of boundary corrected estimators.

Zusammenfassung: Die Receiver Operating Characteristic (ROC) Kurve ist ein statistisches Werkzeug zur Bewertung der Präzision diagnostischer Tests. Die empirische ROC Kurve (sie ist eine Treppenfunktion) ist der am weitesten verbreitete nicht-parametrische Schätzer der ROC Kurve. Andererseits wurden Kerngättungsmethoden verwendet, um glatte ROC Kurven zu erhalten. Der vorangehende Prozess basiert dabei auf Kernschätzungen der Verteilungsfunktionen. Es wurde beobachtet, dass Kernschätzer der Verteilung nicht konsistent sind falls die Verteilungsfunktion in der Nähe des Randes ihres Trägers geschätzt wird. Dieses Problem beruht auf dem "Randeffekt" der in der nicht-parametrischen funktionalen Schätzung auftritt. Um derartige Schwierigkeiten zu vermeiden, empfehlen wir eine verallgemeinerte Reflexionsmethode der Randkorrektur im Schätzproblem von ROC Kurven. Die vorgeschlagene Methode generiert eine Klasse von randkorrigierten Schätzern.

Keywords: Reflection, Distribution Estimation.

1 Introduction

The Receiver Operating Characteristic (ROC) describes the performance of a diagnostic test which classifies subjects into either group without condition $\mathcal{G}_0$ or group with condition $\mathcal{G}_1$ by means of a continuous discriminant score X, i.e., a subject is classified as $\mathcal{G}_1$ if $X \geq d$ and $\mathcal{G}_0$ otherwise for a given cutoff point $d \in \mathbb{R}$. The ROC is defined as a plot of probability of false classification of subjects from $\mathcal{G}_1$ versus the probability of true classification of subjects from $\mathcal{G}_0$ across all possible cutoff point values of X. Specifically, let

F_0 and F_1 denote the distribution functions of X in the groups $\mathcal{G}_0$ and $\mathcal{G}_1$, respectively. Then, the ROC curve can be written as

$$R(p) = 1 - F_1\left(F_0^{-1}(1-p)\right), \qquad 0 < p < 1,$$

where p is the false positive rate in $(0, 1)$ as the corresponding cut-off point ranges from $-\infty$ to $+\infty$ and F_0^{-1} denotes the inverse function of F_0.

A simple non-parametric estimator for $R(p)$ is to use the empirical distribution functions for F_0 and F_1. The resulting ROC curve is a step function and it is called the empirical ROC curve. Another type of non-parametric estimator for $R(p)$ is derived from kernel smoothing methods. Kernel smoothing is most widely used mainly because it is easy to derive and has good asymptotic and small sample properties. Kernel smoothing has received a considerable attention in density estimation context; see, for example the monographs of Silverman (1986) and Wand and Jones (1995). However, applications of kernel smoothing in distribution function estimation are relatively few. Some theoretical properties of a kernel distribution function estimator have been investigated by Nadaraya (1964), Reiss (1981), and Azzalini (1981). Lloyd (1998) proposed a nonparametric estimator of ROC by using kernel estimators for the distribution functions F_0 and F_1.

Lloyd and Yong (1999) showed that Lloyd's estimator has better mean squared error properties than the empirical ROC curve estimator. However, his estimator has some drawbacks. For example, Lloyd's estimator is unreliable near the end points of the support of the ROC curve due to so-called "boundary effects" that occur in nonparametric functional estimation. Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied.

In this paper, we develop a new kernel type estimator of the ROC curve that removes boundary effects near the end points of the support. Our estimator is based on a new boundary corrected kernel estimator of distribution functions and it is based on ideas of Karunamuni and Alberts (2005a, 2005b, 2006), Zhang and Karunamuni (1998, 2000), (Karunamuni and Zhang, 2008), and Zhang, Karunamuni, and Jones (1999) developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We derive expressions for the bias and variance of the proposed estimator. Furthermore, the proposed estimator is compared with the "classical estimator" using simulation studies. We observe that the proposed estimator successfully remove boundary effects and performs considerably better than the "classical estimator".

Kernel smoothing in distribution function and ROC curve estimation is discussed in the next section. The proposed estimator is given in Section 3. Simulation results are given in Section 4. A real data example is analyzed in Section 5. Finally, some concluding remarks are given in Section 6.

2 Kernel Smoothing

2.1 Kernel ROC Estimator

Suppose that independent samples $X_{01}, \ldots, X_{0n_0}$ and $X_{11}, \ldots, X_{1n_1}$ are available from some two unknown distributions F_0 and F_1, respectively, where $F_0 \in \mathcal{G}_0$ and $F_1 \in \mathcal{G}_1$ and $\mathcal{G}_0$ and $\mathcal{G}_1$ denote two groups of continuous distribution functions. Then a simple nonparametric estimator of the ROC curve $R(p) = 1 - F_1\left(F_0^{-1}(1-p)\right)$, $0 < p < 1$, is known as the *empirical ROC* curve given by

$$\widetilde{R}_E(p) = 1 - \widetilde{F}_1\left(\widetilde{F}_0^{-1}(1-p)\right), \qquad 0 \le p \le 1,$$

where $\widetilde{F}_0$ and $\widetilde{F}_1$ denote the empirical distribution functions of F_0 and F_1 based on the data $X_{01}, \ldots, X_{0n_0}$ and $X_{11}, \ldots, X_{1n_1}$, respectively; that is

$$\widetilde{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} I(X_{0i} \le x), \qquad \widetilde{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} I(X_{1i} \le x).$$

Note that $\widetilde{R}$ is not a continuous function. In fact, it is a step function on the interval $[0,1]$. This is a notable weakness of the empirical ROC curve $\widetilde{R}(p)$. Since the ROC curve is a smooth function of p, we would expect to have an estimator that is smooth as well. Lloyd (1998) proposed a smooth estimator using kernel smoothing techniques. His idea is to replace unknown distribution F_0 and F_1 by two smooth kernel estimators. Specifically, he employed following kernel estimators of F_0 and F_1:

$$\widehat{F}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} W\left(\frac{x - X_{0i}}{h_0}\right), \qquad \widehat{F}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} W\left(\frac{x - X_{1i}}{h_1}\right),$$

where $W(x) = \int_{-1}^{x} K(t)dt$, h_0 and h_1 denote bandwidths ($h_0 \to 0$ and $h_1 \to 0$ as $n_0 \to \infty$ and $n_1 \to \infty$, respectively), and K is a unimodal symmetric density function with support $[-1,1]$. The corresponding estimator of the ROC curve $R(p)$ is then given by

$$\widehat{R}(p) = 1 - \widehat{F}_1\left(\widehat{F}_0^{-1}(1-p)\right), \qquad 0 \le p \le 1.$$

An example of a smooth estimate of $R(p)$ using $\widehat{R}(p)$ is illustrated in Figure 1.

When $\mathcal{G}_0$ and $\mathcal{G}_1$ contain distributions with finite support then the estimator $\widehat{R}$ exhibits boundary effects near the endpoints of the support due to the same boundary effects that occur in the uncorrected kernel estimators $\widehat{F}_0$ and $\widehat{F}_1$. The main purpose of this article is to improve the kernel distribution estimators and thereby to avoid boundary effects of smooth kernel ROC estimators. Details of the boundary problem with $\widehat{F}_0$ and $\widehat{F}_1$ are described in the next section.

2.2 Kernel Distribution Estimator and Boundary Effects

Let f denote a continuous density function with support $[0,a]$, $0 < a \le \infty$, and consider nonparametric estimation of the cumulative distribution function F of f based on a random sample $X_1, \ldots, X_n$ from f. Suppose that $F^{(j)}$, the j-th derivative of F, exists and is

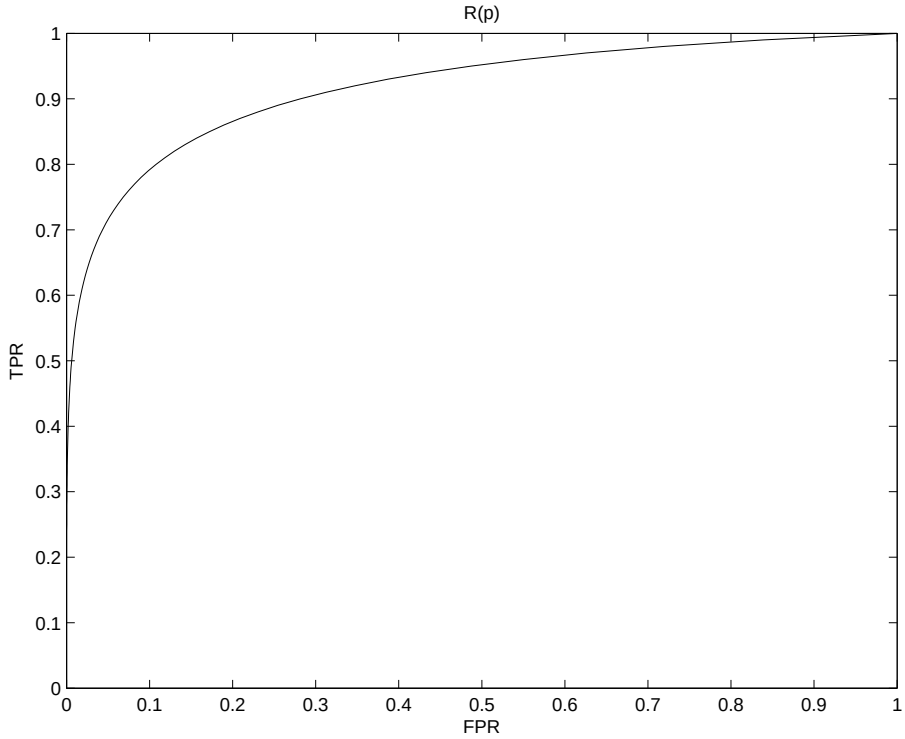


Figure 1: Smooth estimate of $R(p)$.

continuous on $[0, a]$, $j = 0, 1, 2$, with $F^{(0)} = F$ and $F^{(1)} = f$. Then the traditional kernel estimator of F is given by

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt\,,$$

where K is a symmetric density function with support $[-1, 1]$ and h is the bandwidth ($h \to 0$ as $n \to \infty$). The basic properties of $\widehat{F}_{h,K}(x)$ at interior points are well-known (e.g. Lejeune and Sarda, 1992), and under some smoothness assumptions these include, for $h \le x \le a - h$,

$$\begin{aligned}
\mathrm{E}\left(\widehat{F}_{h,K}(x)\right) - F(x) &= \frac{1}{2}\beta_2 f^{(1)}(x)h^2 + o(h^2) \\
n\mathrm{var}\left(\widehat{F}_{h,K}(x)\right) &= F(x)\left(1 - F(x)\right) + hf(x)\int_{-1}^{1} W(t)\left(W(t) - 1\right)dt + o(h)\,.
\end{aligned}$$

The performance of $\widehat{F}_{h,K}(x)$ at boundary points, i.e., for $x \in [0, h) \cup (a - h, a]$, however, differs from the interior points due to so-called "boundary effects" that occur in nonparametric curve estimation problems. More specifically, the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$ at boundary points, while the variance of $\widehat{F}_{h,K}(x)$ is of the same order. This fact can be clearly seen by examining the behavior of $\widehat{F}_{h,K}$ inside the left boundary region $[0, h]$. Let x be a point in the left boundary, i.e., $x \in [0, h]$. Then we can write

$x = ch$, $0 \le c \le 1$. The bias and variance of $\widehat{F}_{h,K}(x)$ at $x = ch$ are of the form

$$\mathrm{E}\left(\widehat{F}_{h,K}(x)\right) - F(x) = hf(0)\int_{-1}^{-c} W(t)dt + h^2 f^{(1)}(0)\left\{\frac{c^2}{2} + c\int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt\right\} + o(h^2) \quad (1)$$

$$n\mathrm{var}\left(\widehat{F}_{h,K}(x)\right) = F(x)(1-F(x)) + hf(0)\left\{\int_{-1}^{c} W^2(t)dt - c\right\} + o(h)\,. \quad (2)$$

From expression (1) it is now clear that the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$. To remove this boundary effect in kernel distribution estimation we investigate a new class of estimators in the next section.

3 The Proposed Estimator

In this section we propose a class of estimators of the distribution function F of the form

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right)\right\}, \quad (3)$$

where h is the bandwidth, K is a symmetric density function with support $[-1, 1]$, and g_1 and g_2 are two transformations that need to be determined. The same type of estimator in density estimation case has been discussed in Zhang et al. (1999). As in the preceding paper, we assume that g_i, $i = 1, 2$, are nonnegative, continuous and monotonically increasing functions defined on $[0, \infty)$. Further assume that g_i^{-1} exists, $g_i(0) = 0$, $g_i^{(1)}(0) = 1$, and that $g_i^{(2)}$ exists and is continuous on $[0, \infty)$, where $g_i^{(j)}$ denotes the j-th derivative of g_i, with $g_i^{(0)} = g_i$ and g_i^{-1} denoting the inverse function of g_i, $i = 1, 2$. We will choose g_1 and g_2 such that $\widetilde{F}_{h,K}(x) \ge 0$ everywhere. Note that the i-th term of the sum in (3) can be expressed as

$$W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right) = \int_{\frac{-x+g_1(X_i)}{h}}^{\frac{x+g_2(X_i)}{h}} K(t)dt\,.$$

The preceding integral is non-negative provided the inequality $-x + g_1(X_i) \le x + g_2(X_i)$ holds. Since $x \ge 0$, the preceding inequality will be satisfied if g_1 and g_2 are such that $g_1(X_i) \le g_2(X_i)$ for $i = 1, \ldots, n$. Thus we will assume that g_1 and g_2 are chosen such that $g_1(x) \le g_2(x)$ for $x \in [0, \infty)$ for our proposed estimator. Now, we can obtain the

bias and variance of (3) at $x = ch$, $0 \leq c \leq 1$, as

$$\begin{aligned} \mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x) &= h^2 \left\{ f^{(1)}(0) \left(\frac{c^2}{2} + 2c \int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt \right) \right. \\ &\quad - f(0) g_1^{(2)}(0) \int_{-1}^{c} (c-t) W(t)dt \\ &\quad \left. - f(0) g_2^{(2)}(0) \int_{-1}^{-c} (c+t) W(t)dt \right\} + o(h^2) \end{aligned} \tag{4}$$

$$\begin{aligned} n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) &= F(x)(1-F(x)) + hf(0) \left\{ \int_{-1}^{c} W^2(t)dt \right. \\ &\quad \left. - 2\int_{-1}^{c} W(t)W(t-2c)dt + \int_{-1}^{-c} W^2(t)dt \right\} + o(h)\,. \end{aligned} \tag{5}$$

The proofs of (4) and (5) are given in the Appendix. Note that the contribution of g_2 on the bias vanishes as $c \to 1$. By comparing expressions (1), (4), (2), and (5) at boundary points we can see that the variances are of the same order and the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ whereas the bias of $\widetilde{F}_{h,K}(x)$ is of order $O(h^2)$. So our proposed estimator removes boundary effects in kernel distribution estimation since the bias at boundary points is of the same order as the bias at interior points.

It is clear that there are various possible choices available for the pair (g_1, g_2). However, we will choose g_1 and g_2 so that the condition $\widetilde{F}_{h,K}(0) = 0$ will be satisfied because of the fact that $F(0) = 0$. A sufficient (but not necessary) condition for the preceding condition to be satisfied is that g_1 and g_2 must be equal. Thus we need to construct a single transformation function g such that $g = g_1 = g_2$. Other important properties that are desirable in the estimator $\widehat{F}_{h,K}$ are the local adaptivity (i.e., the transformation function g depends on c) and that $\widetilde{F}_{h,K}(x)$ being equal to the usual kernel estimator $\widehat{F}_{h,K}(x)$ at interior points. For the latter, g must satisfy that $g(y) \to y$ as $c \to 1$. In order to display the dependance of g on c, $0 \leq c \leq 1$, we shall denote g by g_c in what follows.

Summarizing all the assumptions, it is clear now that g_c should satisfy the conditions

(i) $g_c : [0, \infty) \to [0, \infty)$, g_c is continuous, monotonically increasing and $g_c^{(i)}$ exists, $i = 1, 2$.

(ii) $g_c^{-1}(0) = 0$ and $g_c^{(1)}(0) = 1$.

(iii) $g_c(y) \to y$ for $c \to 1$.

Functions satisfying conditions (i) to (iii) are easy to construct. The trivial choice is $g_c(y) = y$, which represents the "classical" reflection method estimator. Based on extensive simulations, we observed that the following transformation adapts well to various shapes of distributions:

$$g_c(y) = y + \frac{1}{2} I_c y^2\,, \tag{6}$$

for $y \geq 0$ and $0 \leq c \leq 1$, where $I_c = \int_{-1}^{-c} W(t)dt$.

Remark: Some discussion on the above choice of g_c and other various improvements that can be made would be appropriate here. It is possible to construct functions g_c that improve the bias further under some additional conditions. For instance, if one examines

the right hand side of bias expansion (4) then it is not difficult to see that the terms inside bracket (i.e., the coefficient of h^2) can be made equal to zero if g_c is appropriately chosen. Indeed, if g_c is chosen such that

$$f(0)g_c^{(2)}(0)\left\{\int_{-1}^{c}(c-t)W(t)dt+\int_{-1}^{-c}(c+t)W(t)dt\right\} = f^{(1)}(0)\left(\frac{c^2}{2}+2c\int_{-1}^{-c}W(t)dt-\int_{-c}^{c}tW(t)dt\right),$$

then the bias of $\widetilde{F}_{h,K}(x)$ would be theoretically of order $O(h^3)$. For such a function g_c, the second derivative at zero, $g_c^{(2)}(0)$, will depend on the ratio $d_1 = f^{(1)}(0)/f(0)$. In this case, the function g_c would probably be some cubic polynomial; see e.g. Karunamuni and Alberts (2005a, 2005b, 2006). Then the problem of estimation of d_1 naturally arises as in the preceding paper. Another problem that one would face is that the second derivative $g_c^{(2)}(0)$ may not go to 0 when $c \to 1$ as in the case of density estimation context. Thus one may not be able to find any function g_c which satisfies condition (iii) and hence the estimator $\widetilde{F}_{h,K}$ loses the property of "natural extension" to the classical estimator outside the boundary points. These are basically the main reasons why we decided to implement a quadratic function defined in (6) as our choice of transformation.

4 Simulation

To test the effectiveness of our estimator, we simulated its performance against the reflection method. The simulation is based on 1000 replications. In each replication, the random variables $X_0 \sim \text{Exp}(2)$ and $X_1 \sim \text{Gamma}(3,2)$ were generated and the estimate of the ROC curve was computed. The probability distributions of both groups $\mathcal{G}_0$ and $\mathcal{G}_1$ are illustrated in Figure 2.

In all replications sample sizes of $n_0 = n_1 = 50$ were used. In this case, the actual global optimal bandwidths (see Azzalini, 1981) for F_0 and F_1 are $h_{F_0} = 2.9149$ and $h_{F_1} = 5.8298$, respectively. For the kernel estimation of the cumulative distributions we used the quartic kernel $K(x) = \frac{15}{16}(1-x^2)^2 I_{[-1,1]}$, where I_A is the indicator function on the set A. In our experience, the quality of estimated curve by using this kernel is not too sensitive to an optimal bandwidth choice. Hence we used this kernel also in the next section.

For each ROC curve we have calculated the mean integrated squared error (MISE) on the interval $[0,1]$ over all 1000 replications and have displayed the results in a boxplot in Figure 3. The variance of each estimator can be accurately gauged by the whiskers of the plot. The values of means and standard deviations for MISE of each method are given in Table 1.

We also obtained 10 typical realizations of each estimator and displayed these in Figure 4 for comparison purposes with the theoretical ROC curve. The solid line represents the theoretical ROC curve and the dotted lines illustrate the 10 realizations.

The final estimate of the ROC curve depends on estimates of the cumulative distribution functions F_0 and F_1. While boundary effects cause problems by estimating F_0 and

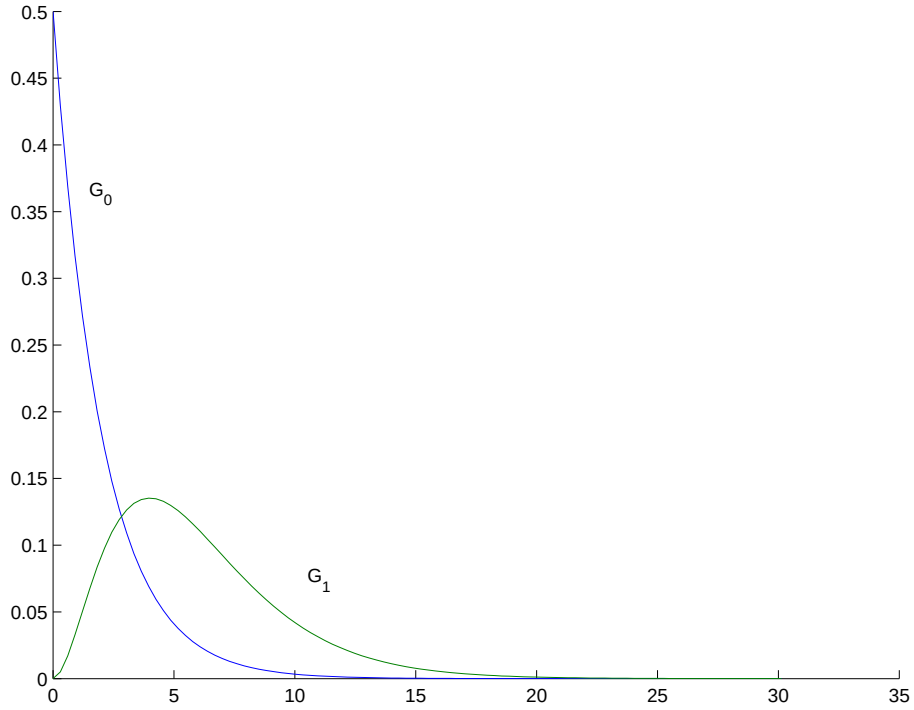


Figure 2: The probability distribution of groups $\mathcal{G}_0$ and $\mathcal{G}_1$.

Table 1: Means and standard deviations of the MISE.

Method	Mean	STD
Proposed	0.0053	0.0047
Reflection	0.0065	0.0050
Classical	0.0084	0.0054

F_1 inside the left boundary region, the quality of the final estimate of the ROC can also be influenced by these effects near the right boundary of the interval $[0, 1]$ as well. As we can see in Figure 4, the biggest difference between the above mentioned methods is in the second half part of the interval $[0, 1]$. Table 1 describes the performance of our proposed method with respect to the MISE. The values of the mean and the standard deviation for the MISE were smallest in case of our proposed estimator. Although the theoretical bias of our estimator is of the same order as in the case of the reflection method, the numerical results of estimators of the ROC curves were better for our estimator in the simulation. In our opinion, this is due to the fact that our estimator is locally adaptive.

5 Consumer Loans Data

In this example we used some (unspecified) scoring function to predict the solidity of a client. The goal here is to determine which clients are able to pay their loans. We considered a test set of 332 clients; 309 paid their loans (group $\mathcal{G}_0$) and 22 had problems with

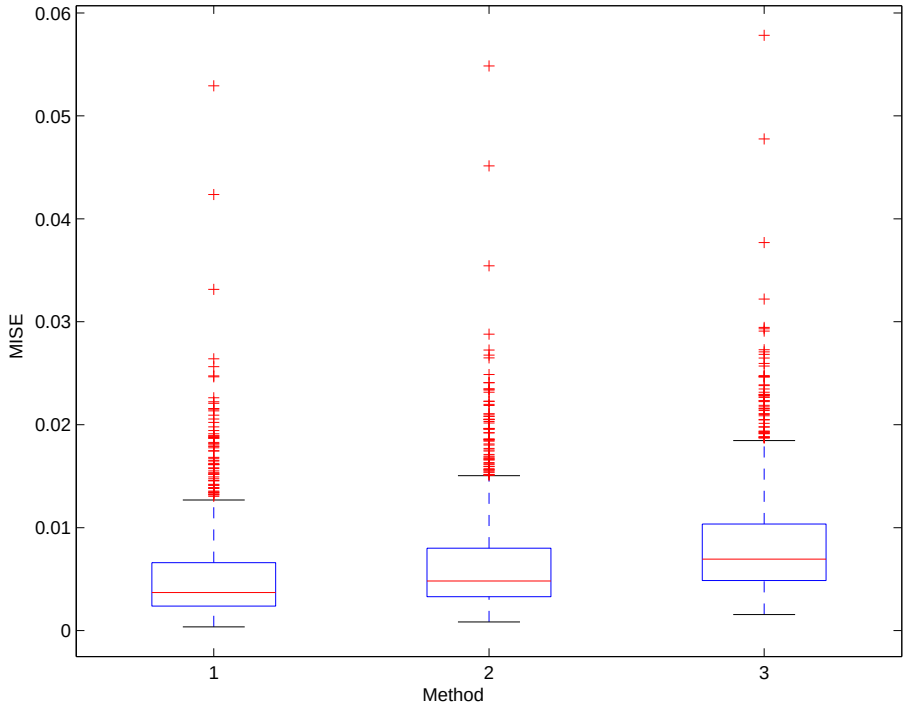
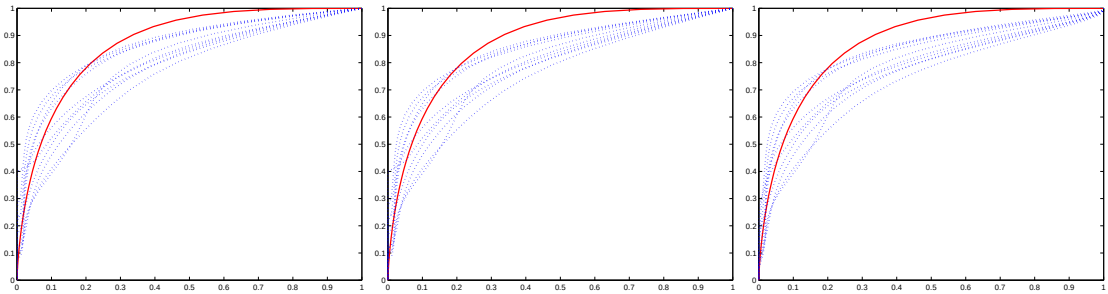


Figure 3: Boxplots of the MISE over $[0, 1]$ for our proposed method (1), the reflection method (2), and the classical estimator with boundary effects (3).



(1) (2) (3)

Figure 4: Estimates of the ROC for our proposed method (1), the reflection method (2), and the classical estimator with boundary effects (3).

payments or did not pay (group $\mathcal{G}_1$). We used the ROC curve to assess the discrimination between clients with and without a good solidity. It is of interest for us to know here if our scoring function is a good predictor of the solidity.

Estimates of ROC are illustrated in Figure 5. The dashed line represents the estimate obtained by our proposed method and the solid line is for the kernel ROC with boundary effects. When choosing the optimal bandwidths for distribution function estimation, we used the method described in Horová, Koláček, Zelinka, and El-Shaarawi (2008). A somewhat similar method for density estimation is given in Sheather and Jones (1991). The optimal bandwidths for distribution functions F_0 and F_1 were estimated as $\hat{h}_{F_0} =$

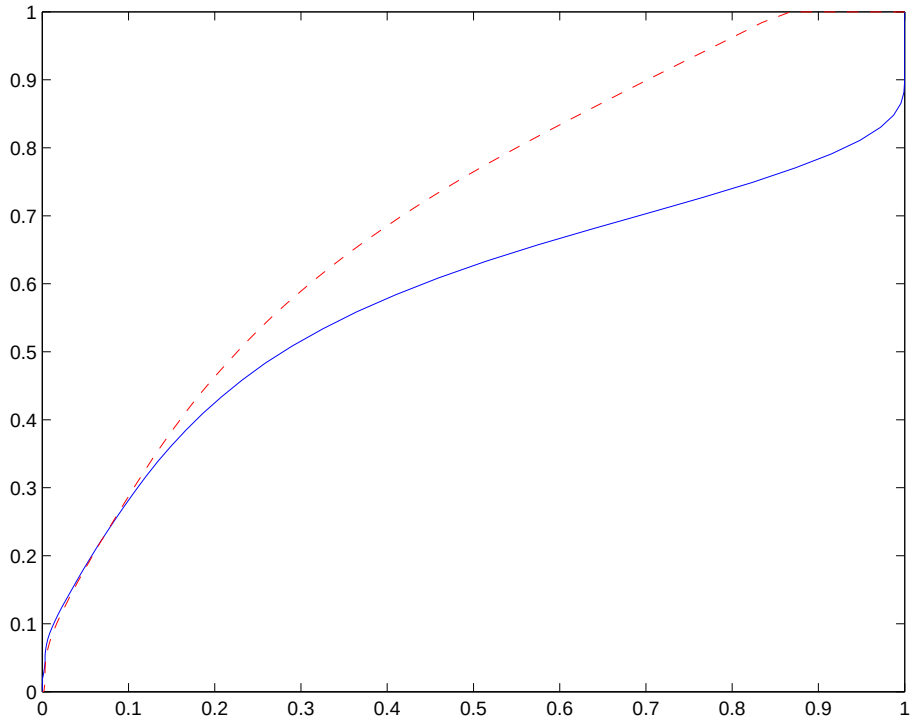


Figure 5: The estimate of the ROC for consumer the loans data.

0.0068 and $\hat{h}_{F_1} = 0.0286$, respectively.

From the estimates of the ROC one can see that the scoring function is not a good predictor of the solidity of a client. This fact could be also affected by the different sizes of both groups. When group $\mathcal{G}_1$ is too small it causes larger boundary effects. It is clearly visible that the estimate of the ROC obtained by the classical estimator (solid line) has some values under the diagonal of the unit square. However, this situation does not show up theoretically. Thus there is a larger influence of boundary effects to the quality of final estimates of the ROC.

6 Conclusion

In this paper we proposed a new kernel-type distribution estimator to avoid the difficulties near the boundary. The technique implemented is a kind of generalized reflection method involving reflecting a transformation of the data. The proposed method generates a class of boundary corrected estimators and it is based on ideas of boundary corrections for kernel density estimators presented in Karunamuni and Alberts (2005a, 2005b, 2006). We showed some good properties of our proposed method (e.g., local adaptivity). Furthermore, it is shown that bias of the proposed estimator is smaller than that of the "classical" case.

Acknowledgements

The research was supported by the Jaroslav Hájek center for theoretical and applied statistics (grant No. LC 06024). The second author's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Proof of (4). For $x = ch$, $0 \le c \le 1$, using the property $W(t) = 1 - W(-t)$ we obtain

$$\begin{aligned}
\mathrm{E}(\widetilde{F}_{h,K}(x)) &= \mathrm{E}\left(W\left(\frac{x-g_1(X_i)}{h}\right)\right) - \mathrm{E}\left(W\left(-\frac{x+g_2(X_i)}{h}\right)\right) \\
&= \int_0^\infty W\left(\frac{x-g_1(y)}{h}\right) f(y)dy - \int_0^\infty W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= h\int_{-1}^{c} W(t)\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt - h\int_{-1}^{-c} W(t)\frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}dt \\
&= h\int_{-1}^{-c} W(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} - \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt \\
&\quad + h\int_{-c}^{c}(1-W(-t))\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\
&= h\int_{-1}^{-c} W(t)\left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} - \frac{f\left(g_2^{-1}(-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\}dt \\
&\quad + F\left(g_1^{-1}(2ch)\right) - h\int_{-c}^{c} W(t)\frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)}dt\,.
\end{aligned}$$

Using a Taylor expansion of order 2 on the function $F\left(g_1^{-1}(\cdot)\right)$ we have

$$F\left(g_1^{-1}(2ch)\right) = F(0) + f(0)2ch + \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)2c^2h^2 + o(h^2)\,.$$

By the existence and continuity of $F^{(2)}(\cdot)$ near 0, we obtain for $x = ch$

$$\begin{aligned}
F(0) &= F(x) - f(x)ch + \frac{1}{2}f^{(1)}(x)c^2h^2 + o(h^2) \\
f(x) &= f(0) + f^{(1)}(0)ch + o(h) \\
f^{(1)}(x) &= f^{(1)}(0) + o(1)\,.
\end{aligned}$$

Therefore,

$$F\left(g_1^{-1}(2ch)\right) = F(x) + f(0)ch + \left(\frac{3}{2}f^{(1)}(0) - 2f(0)g_1^{(2)}(0)\right)c^2h^2 + o(h^2)\,. \qquad (7)$$

Now, (7) and a Taylor expansion of order 1 of the functions

$$\frac{f\left(g_1^{-1}(\cdot)\right)}{g_1^{(1)}\left(g_1^{-1}(\cdot)\right)} \qquad \text{and} \qquad \frac{f\left(g_2^{-1}(\cdot)\right)}{g_2^{(1)}\left(g_2^{-1}(\cdot)\right)}$$

give

$$
\begin{aligned}
&\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x) \\
&= h\int_{-1}^{-c} W(t)\left\{2f^{(1)}(0)ch - f(0)h\left((c-t)g_1^{(2)}(0) + (c+t)g_2^{(2)}(0)\right) + o(h)\right\}dt \\
&\quad + f(0)ch + \left\{\frac{3}{2}f^{(1)}(0) - 2f(0)g_1^{(2)}(0)\right\}c^2h^2 + o(h^2) \\
&\quad - h\int_{-c}^{c} W(t)\left\{f(0) + \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)(c+t)h + o(h)\right\}dt \\
&= h\left\{f(0)c - f(0)\int_{-c}^{c} W(t)dt\right\} + h^2\left\{\frac{3}{2}f^{(1)}(0)c^2 + 2f^{(1)}(0)c\int_{-1}^{-c} W(t)dt\right. \\
&\quad - 2f(0)g_1^{(2)}(0)c^2 - f(0)g_1^{(2)}(0)\int_{-1}^{-c}(c-t)W(t)dt - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c+t)W(t)dt \\
&\quad \left. - \left(f^{(1)}(0) - f(0)g_1^{(2)}(0)\right)\int_{-c}^{c}(c+t)W(t)dt\right\} + o(h^2)\,.
\end{aligned}
$$

From the symmetry of K and the definition $W(x)$, one can write $W(x) = \frac{1}{2} + b(x)$, where $b(x) = -b(-x)$ for all x such that $|x| \leq 1$. Thus $\int_{-c}^{c} W(t)dt = c$ and therefore the coefficient of h is zero. So after some algebra we obtain the bias expression as

$$
\begin{aligned}
\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right) - F(x) = h^2\left\{f^{(1)}(0)\left(\frac{c^2}{2} + 2c\int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt\right)\right. \\
\left. -f(0)g_1^{(2)}(0)\int_{-1}^{c}(c-t)W(t)dt - f(0)g_2^{(2)}(0)\int_{-1}^{-c}(c+t)W(t)dt\right\} + o(h^2)\,.
\end{aligned}
$$

Proof of (5). Observe that for $x = ch$, $0 \leq c \leq 1$, we have

$$
\begin{aligned}
n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) &= \frac{1}{n}\mathrm{var}\left\{\sum_{i=1}^{n}\left[W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right)\right]\right\} \\
&= \mathrm{E}\left\{W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right)\right\}^2 \\
&\quad - \left\{\mathrm{E}\left[W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right)\right]\right\}^2 \\
&= A_1 - A_2\,,
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= \mathrm{E}\left\{W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right\}^2 \\
&= \int_0^\infty \left\{W\left(\frac{x-g_1(y)}{h}\right) - W\left(-\frac{x+g_2(y)}{h}\right)\right\}^2 f(y)dy \\
&= \int_0^\infty \left\{W^2\left(\frac{x-g_1(y)}{h}\right) + W^2\left(-\frac{x+g_2(y)}{h}\right)\right\} f(y)dy \\
&\quad - \int_0^\infty 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= h\int_{-1}^{-c} W^2(t) \left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} + \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\} dt \\
&\quad + h\int_{-c}^{c} W^2(t) \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} dt \\
&\quad - \int_0^\infty 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\
&= A_{1,1} + A_{1,2} - A_{1,3}\,.
\end{aligned}
$$

Using a Taylor expansion as in the last proof, it can be shown that

$$
\begin{aligned}
A_{1,1} &= h\int_{-1}^{-c} W^2(t) \left\{\frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} + \frac{f\left(g_2^{-1}((-c-t)h)\right)}{g_2^{(1)}\left(g_2^{-1}((-c-t)h)\right)}\right\} dt \\
&= h\int_{-1}^{-c} W^2(t)\left(2f(0) + o(1)\right) dt\,.
\end{aligned}
$$

For $A_{1,2}$ we use the identity $W(t) = 1 - W(-t)$ and similarly as in the last proof we get

$$
\begin{aligned}
A_{1,2} &= h\int_{-c}^{c} W^2(t) \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} dt \\
&= h\int_{-c}^{c} \left(1 - 2W(-t) + W^2(-t)\right) \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} dt \\
&= h\int_{-c}^{c} \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)} dt - 2h\int_{-c}^{c} W(t) \frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)} dt \\
&\quad + h\int_{-c}^{c} W^2(t) \frac{f\left(g_1^{-1}((c+t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c+t)h)\right)} dt \\
&= F\left(g_1^{-1}(2ch)\right) - 2h\int_{-c}^{c} W(t)\left(f(0) + o(1)\right) dt + h\int_{-c}^{c} W^2(t)\left(f(0) + o(1)\right) dt \\
&= F(x) - f(0)ch + hf(0)\int_{-c}^{c} W^2(t)dt + o(h)\,.
\end{aligned}
$$

Using the continuity of $g_i^{(2)}$, $g_i(0) = 0$, and $g_i^{(1)}(0) = 1$, $i = 1, 2$, and by a Taylor

expansion of order 2 on $g_2\left(g_1^{-1}(\cdot)\right)$, we have

$$\begin{aligned} g_2\left(g_1^{-1}((c-t)h)\right) &= g_2\left(g_1^{-1}(0)\right) + \frac{g_2^{(1)}\left(g_1^{-1}(0)\right)}{g_1^{(1)}\left(g_1^{-1}(0)\right)}(c-t)h + o(h) \\ &= (c-t)h + o(h)\,. \end{aligned}$$

With the preceding expansion we obtain

$$\begin{aligned} A_{1,3} &= \int_0^\infty 2W\left(\frac{x-g_1(y)}{h}\right) W\left(-\frac{x+g_2(y)}{h}\right) f(y)dy \\ &= 2h\int_{-1}^{c} W(t)W\left(-\frac{x}{h} - \frac{g_2\left(g_1^{-1}((c-t)h)\right)}{h}\right) \frac{f\left(g_1^{-1}((c-t)h)\right)}{g_1^{(1)}\left(g_1^{-1}((c-t)h)\right)}dt \\ &= 2h\int_{-1}^{c} W(t)W\left(\frac{-ch-(c-t)h-o(h)}{h}\right)(f(0)+o(1))\,dt \\ &= 2hf(0)\int_{-1}^{c} W(t)W(t-2c)dt + o(h)\,. \end{aligned}$$

Now we can express A_1 as

$$\begin{aligned} A_1 &= A_{1,1} + A_{1,2} - A_{1,3} \\ &= 2hf(0)\int_{-1}^{-c} W^2(t)dt + F(x) - f(0)ch + hf(0)\int_{-c}^{c} W^2(t)dt \\ &\quad -2hf(0)\int_{-1}^{c} W(t)W(t-2c)dt + o(h) \\ &= F(x) + hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} \\ &\quad +o(h)\,. \end{aligned}$$

With the expression obtained for the bias we obtain the expression for A_2 as

$$\begin{aligned} A_2 &= \left\{\mathrm{E}\left[W\left(\frac{x-g_1(X_i)}{h}\right) - W\left(-\frac{x+g_2(X_i)}{h}\right)\right]\right\}^2 \\ &= \left\{\mathrm{E}\left(\widetilde{F}_{h,K}(x)\right)\right\}^2 \\ &= F^2(x) + o(h)\,. \end{aligned}$$

Finally, we obtain the variance of the estimator as

$$\begin{aligned} &n\mathrm{var}\left(\widetilde{F}_{h,K}(x)\right) = A_1 - A_2 \\ &= F(x) + hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} \\ &\quad -F^2(x) + o(h) \\ &= F(x)(1-F(x)) \\ &\quad +hf(0)\left\{2\int_{-1}^{-c} W^2(t)dt - c + \int_{-c}^{c} W^2(t)dt - 2\int_{-1}^{c} W(t)W(t-2c)dt\right\} + o(h)\,. \end{aligned}$$

References

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, *68*, 326-328.

Horová, I., Koláček, J., Zelinka, J., and El-Shaarawi, A. H. (2008). Smooth estimates of distribution functions with application in environmental studies. *Advanced topics on mathematical biology and ecology*, 122-127.

Karunamuni, R. J., and Alberts, T. (2005a). A generalized reflection method of boundary correction in kernel density estimation. *Canadian Journal of Statistics*, *33*, 497-509.

Karunamuni, R. J., and Alberts, T. (2005b). On boundary correction in kernel density estimation. *Statistical Methodology*, *2*, 191-212.

Karunamuni, R. J., and Alberts, T. (2006). A locally adaptive transformation method of boundary correction in kernel density estimation. *Journal of Statistical Planning and Inference*, *136*, 2936-2960.

Karunamuni, R. J., and Zhang, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters*, *78*, 497-507.

Lejeune, M., and Sarda, P. (1992). Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis*, *14*, 457-471.

Lloyd, C. J. (1998). The use of smoothed ROC curves to summarise and compare diagnostic systems. *Journal of the American Statistical Association*, *93*, 1356-1364.

Lloyd, C. J., and Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, *44*, 221-228.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability and its Application*, *15*, 497-500.

Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, *8*, 116-119.

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, *53*, 683-690.

Silverman, W. R. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Zhang, S., and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *J. Statist. Planning and Inference*, *70*, 301–316.

Zhang, S., and Karunamuni, R. J. (2000). On nonparametric density estimation at the boundary. *Nonparametric Statistics*, *12*, 197–221.

Zhang, S., Karunamuni, R. J., and Jones, M. C. (1999). An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, *94*, 1231–1241.

Authors' addresses:

Jan Koláček
Department of Mathematics and Statistics
Faculty of Science
Kotlářská 2
611 37 Brno
Czech Republic

E-Mail: kolacek@math.muni.cz

Rohana J. Karunamuni
Department of Mathematical and Statistical Sciences
University of Alberta
T6G 2G1 Edmonton
Canada

E-Mail: R.J.Karunamuni@ualberta.ca

Computational Statistics (2008) 23:63–78
DOI 10.1007/s00180-007-0068-6

ORIGINAL PAPER

Plug-in method for nonparametric regression

Jan Koláček

Accepted: 5 October 2006 / Published online: 25 September 2007

Abstract The problem of bandwidth selection for non-parametric kernel regression is considered. We will follow the Nadaraya–Watson and local linear estimator especially. The circular design is assumed in this work to avoid the difficulties caused by boundary effects. Most of bandwidth selectors are based on the residual sum of squares (RSS). It is often observed in simulation studies that these selectors are biased toward undersmoothing. This leads to consideration of a procedure which stabilizes the RSS by modifying the periodogram of the observations. As a result of this procedure, we obtain an estimation of unknown parameters of average mean square error function (AMSE). This process is known as a plug-in method. Simulation studies suggest that the plug-in method could have preferable properties to the classical one.

Keywords Bandwidth selection · Fourier transform · Kernel estimation · Nonparametric regression

1 Introduction

In nonparametric regression estimation, a critical and inevitable step is to choose the smoothing parameter (bandwidth) to control the smoothness of the curve estimate. The smoothing parameter considerably affects the features of the estimated curve. Although in practice one can try several bandwidths and choose a bandwidth subjectively, automatic (data-driven) selection procedures could be useful for many situations; see Silverman (1985) for more examples.

Supported by the MSMT: LC 06024.

J. Koláček (✉)
Faculty of Science, Masaryk University, Janackovo nam. 2a, Brno, Czech Republic
e-mail: kolacek@math.muni.cz

Several automatic bandwidth selectors have been proposed and studied in Craven and Wahba (1979), Härdle (1990), Härdle et al. (1988), Droge (1996), and references given therein. It is well recognized that these bandwidth estimates are subject to large sample variation. The kernel estimates based on the bandwidths selected by these procedures could have very different appearances. Due to the large sample variation, classical bandwidth selectors might not be very useful in practice.

In the simulation study of Chiu (1990), it was observed that Mallows' criterion gives smaller bandwidths more frequently than predicted by the asymptotic theorems. Chiu (1990) provided an explanation for the cause and suggested a procedure to overcome the difficulty. By applying the procedure, we introduce a new method for bandwidth selection which gives much more stable bandwidth estimates.

2 Kernel regression

Consider a standard regression model of the form

$$Y_t = m(x_t) + \varepsilon_t, \quad t = 0, \ldots, T-1, \quad T \in \mathbb{N},$$

where m is an unknown regression function, x_t are design points, Y_t are measurements and ε_t are independent random variables for which

$$E(\varepsilon_t) = 0, \quad \operatorname{var}(\varepsilon_t) = \sigma^2 > 0, \quad t = 0, \ldots, T-1.$$

The aim of kernel smoothing is to find suitable approximation $\widehat{m}$ of the unknown function m.

In next we will assume

1. The design points x_t are equidistantly distributed on the interval $[0, 1]$, that is $x_t = t/T,\ t = 0, \ldots, T-1$.
2. We use a "cyclic design", that is, suppose $m(x)$ is a smooth periodic function and the estimate is obtained by applying the kernel on the extended series $\widetilde{Y}_t$, where $\widetilde{Y}_{t+kT} = Y_t$ for $k \in \mathbb{Z}$. Similarly $x_t = t/T, t \in \mathbb{Z}$.

$Lip[a, b]$ denotes the class of continuous functions satisfying the inequality

$$|g(x) - g(y)| \le L|x - y| \quad \forall x, y \in [a, b],\ L > 0,\ L \text{ is a constant.}$$

Definition Let κ be a nonnegative even integer and assume $\kappa \ge 2$. The function $K \in Lip[-1, 1]$, support$(K) = [-1, 1]$, satisfying the following conditions

(i) $K(-1) = K(1) = 0$

(ii) $\int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 < j < \kappa \\ 1, & j = 0 \\ \beta_\kappa \neq 0, & j = \kappa, \end{cases}$

is called a *kernel* of order κ and a class of all these kernels is marked $S_{0\kappa}$.

These kernels are used for an estimation of the regression function (see Wand and Jones 1995).

Let $K \in S_{0\kappa}$, set $K_h(.) = \frac{1}{h}K(\frac{\cdot}{h})$, $h \in (0, 1)$. A parameter h is called a *bandwidth*.

Commonly used non-parametric methods for estimating $m(x)$ are the kernel estimators

1. **Nadaraya–Watson estimator** (Nadaraya 1964; Watson 1964)

$$\widehat{m}_{NW}(x; h) = \frac{\sum_{k=-T}^{2T-1} K_h(x_k - x)\widetilde{Y}_k}{\sum_{k=-T}^{2T-1} K_h(x_k - x)}$$

2. **Local linear estimator** (Stone 1977; Cleveland 1979)

$$\widehat{m}_{LL}(x; h) = \frac{1}{T}\sum_{k=-T}^{2T-1} \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(x_k - x)\}K_h(x_k - x)\widetilde{Y}_k}{\hat{s}_2(x; h)\hat{s}_0(x; h) - \hat{s}_1(x; h)^2}$$

where

$$\hat{s}_r(x; h) = \frac{1}{T}\sum_{k=-T}^{2T-1} (x_k - x)^r K_h(x_k - x).$$

In the cyclic design, the kernel estimators can be generally expressed as

$$\widehat{m}(x; h) = \sum_{k=-T}^{2T-1} W_k^{(j)}(x)\widetilde{Y}_k,$$

where the weights $W_k^{(j)}(x)$, $j \in \{NW, LL\}$ correspond to the weights of estimators $\widehat{m}_{NW}, \widehat{m}_{LL}$. The assumption of the circular model leads to the fact, that the weights of Nadaraya–Watson and local linear estimator are identical at design points, that is

$$W_k^{(LL)}(x_t) = W_k^{(NW)}(x_t),$$

for $k \in \{-T, -T-1, \ldots, 2T-1\}$, $t \in \{0, 1, \ldots, T-1\}$, so in next, we will write only $W_k(x_t)$ without upper index.

Let $K \in S_{0\kappa}$, $h \in (0, 1)$, $t \in \{0, \ldots, T-1\}$. Then the sum $\sum_{k=-T}^{2T-1} K_h(x_k - x_t) = \sum_{k=-T+1}^{T-1} K_h(x_k)$ is independent on t. Set $C_T := \sum_{k=-T+1}^{T-1} K_h(x_k)$. We can simply write the value of weight functions at design points x_t, $t = 0, \ldots, T-1$

$$W_k(x_t) = \frac{1}{C_T}K_h(x_k - x_t).$$

The optimal bandwidth considered here is h_{opt}, the minimizer of the average mean squared error

$$(AMSE) \qquad R_T(h) = \frac{1}{T}E\sum_{t=0}^{T-1}\{m(x_t) - \widehat{m}(x_t; h)\}^2.$$

Let $K \in S_{0\kappa}$. Under some mild conditions, $AMSE$ converges to

$$\overline{R_T}(h) = \frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa, \tag{1}$$

where

$$V(K) = \int_{-1}^{1} K^2(x)dx, \quad \beta_\kappa = \int_{-1}^{1} x^\kappa K(x)dx, \quad A_\kappa = \int_0^1 \left(m^{(\kappa)}(x)\right)^2 dx.$$

This function has an unique minimum h_{opt}

$$h_{\text{opt}} = \left(\frac{\sigma^2 V(K)(\kappa!)^2}{2\kappa T\beta_\kappa^2 A_\kappa}\right)^{\frac{1}{2\kappa+1}} \tag{2}$$

(for more details, see Wand and Jones 1995).

There exist many estimators of this error function, which are asymptotically equivalent and asymptotically unbiased (see Härdle 1990; Chiu 1990, 1991). However, in simulation studies, it is often observed that most selectors are biased toward undersmoothing and give smaller bandwidths more frequently than predicted by asymptotic results. Most bandwidth selectors are based on the residual sum of squares

$$(RSS) \qquad RSS_T(h) = \frac{1}{T}\sum_{t=0}^{T-1}\{Y_t - \widehat{m}(x_t; h)\}^2.$$

For example Rice (see Rice 1984) considered

$$\widehat{R}_T(h) = RSS_T(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0, \tag{3}$$

where $\hat{\sigma}^2$ is an estimate of σ^2

$$\hat{\sigma}^2 = \frac{1}{2T-2}\sum_{t=1}^{T-1}(Y_t - Y_{t-1})^2. \tag{4}$$

The estimate $\hat{h}_{\text{opt}}$ of optimal bandwidth is defined as

$$\hat{h}_{\text{opt}} = \arg\min \widehat{R}_T(h).$$

3 Use of Fourier transformation

Let $M_t = m(x_t),\ t = 0, \ldots, T-1$. The periodogram of the vector of observations $\boldsymbol{Y}$ is defined by I_{Y_λ}

$$I_{Y_\lambda} = |Y_\lambda^-|^2/2\pi T,$$

where

$$Y_\lambda^- = \sum_{k=0}^{T-1} Y_k e^{-\frac{i2\pi k\lambda}{T}}$$

is the finite Fourier transform of the vector $\boldsymbol{Y}$. This transformation is denoted by $\boldsymbol{Y}^- = DFT^-(\boldsymbol{Y})$.

The periodograms and Fourier transforms of the series $\boldsymbol{\varepsilon}$ and $\boldsymbol{M}$ are defined similarly. Under mild conditions, the periodogram ordinates I_{ε_t} on Fourier frequencies $\frac{2\pi t}{T}$, for $t = 1, \ldots, N = \left[\frac{T-1}{2}\right]$, are approximately independently and exponentially distributed with means $\frac{\sigma^2}{2\pi}$. Here $[x]$ means the greatest integer less or equal to x.

Definition Let $\boldsymbol{x} = (x_0, \ldots, x_{T-1}), \boldsymbol{y} = (y_0, \ldots, y_{T-1}) \in \mathbb{C}^T$;

$$z_t = \sum_{k=0}^{T-1} x_{\langle t-k\rangle_T} y_k,$$

where $\langle t-k\rangle_T$ marks $(t-k) \bmod T$. Then $\boldsymbol{z} = (z_0, \ldots, z_{T-1})$ is called *the discrete cyclic convolution* of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$; we write $\boldsymbol{z} = \boldsymbol{x} \circledast \boldsymbol{y}$.

Let us define a vector $\boldsymbol{w} := (w_0, w_1, \ldots, w_{T-1})$, where

$$w_t = W_0(x_t - 1) + W_0(x_t) + W_0(x_t + 1).$$

Let $h \in (0, 1)$, $K \in S_{0k}$, $t \in \{0, \ldots, T-1\}$. Then we can write $\widehat{m}(x_t; h)$ as a discrete cyclic convolution of vectors $\boldsymbol{w}$ and $\boldsymbol{Y}$.

$$\widehat{m}(x_t; h) = \sum_{k=0}^{T-1} w_{\langle t-k\rangle_T} Y_k. \tag{5}$$

Applying Parseval's formula yields

$$RSS_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} I_{Y_t} \left\{1 - w_t^-\right\}^2, \tag{6}$$

where $w_t^- = \sum_{k=-T+1}^{T-1} W_0(x_k) e^{-\frac{i2\pi kt}{T}}$ is the finite Fourier transform of $\boldsymbol{w}$ (see Chiu 1990, for details). From (3) and (6) we arrive at the equivalent expression for $\widehat{R}_T(h)$

$$\widehat{R}_T(h) = \frac{4\pi}{T} \sum_{t=1}^{N} I_{Y_t} \{1 - w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0. \tag{7}$$

Similarly,

$$R_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N}\left\{I_{M_t} + \frac{\sigma^2}{2\pi}\right\}\{1 - w_t^-\}^2 - \sigma^2 + 2\sigma^2 w_0. \tag{8}$$

4 The motivation and the plug-in method

Let $D(h) = \widehat{R}_T(h) - R_T(h)$. From previous expressions we obtain

$$D(h) = \frac{4\pi}{T}\sum_{t=1}^{N}\left\{I_{Y_t} - I_{M_t} - \frac{\sigma^2}{2\pi}\right\}\{1 - w_t^-\}^2. \tag{9}$$

The periodogram ordinates I_{M_t} decrease rapidly for smooth $m(x)$. So I_{Y_t} do not contain much information about I_{M_t} at high frequencies (for the rigorous proof see Rice 1984). This leads to the consideration of the procedure proposed by Chiu (1991). The main idea is to modify RSS to make it less variable. We find the first index J_1 such that $I_{Y_{J_1}} < c\hat{\sigma}^2/2\pi$ for some constant $c > 1$, where $\hat{\sigma}^2$ is an estimate of σ^2. The constant c sets a threshold. In our experience, setting $1 < c < 3$ yields good results.

The modified residual sum of squares is defined by

$$\mathrm{MRSS}_T(h) = \frac{2\pi}{T}\sum_{t=0}^{T-1}\tilde{I}_{Y_t}\{1 - w_t^-\}^2,$$

where

$$\tilde{I}_{Y_t} = \begin{cases} I_{Y_t}, & t < J_1 \\ \hat{\sigma}^2/2\pi, & t \geq J_1, \end{cases}$$

(see Figs. 1, 2).

Thus, the proposed selector is

$$\widetilde{R}_T(h) = \mathrm{MRSS}_T(h) - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \tag{10}$$

and the new estimate of optimal bandwidth

$$\hat{h}_{\mathrm{opt}} = \arg\min \widetilde{R}_T(h)$$

[for more details see Chiu (1990, 1991)].

To simplify the discussion below, set $c = 2$ and rewrite (10) to the formula in next lemma.

Lemma 1 *Let J_1 be the least index, that $I_{Y_{J_1}} < \hat{\sigma}^2/\pi T$. Then*

$$\widetilde{R}_T(h) = \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2 + \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\left\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\right\}\{1 - w_t^-\}^2.$$

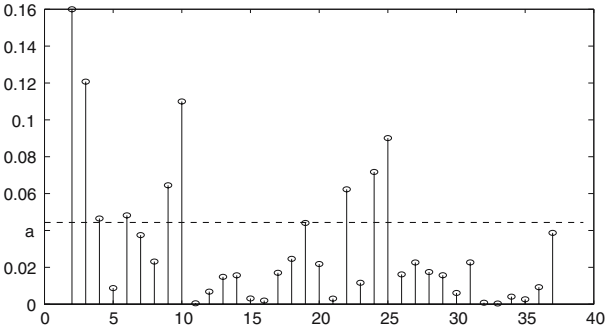


Fig. 1 *The periodogram ordinates* I_{Y_t} *as a function of* t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$

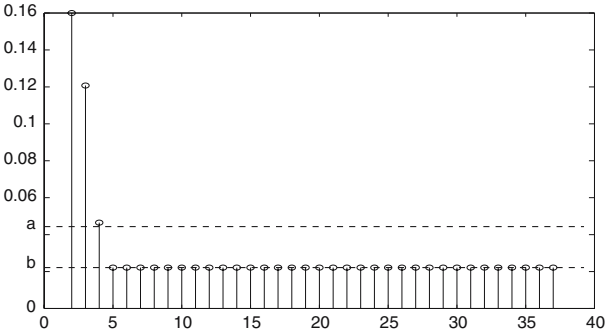


Fig. 2 *The modified periodogram ordinates* $\tilde{I}_{Y_t}$ *as a function of* t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$, $b = \frac{\hat{\sigma}^2}{2\pi}$

The main idea of plug-in method is to estimate unknown parameters σ^2 and A_κ in the expression (2) for the optimal bandwidth h_{opt}, which is the minimum of $\overline{R_T}(h)$

$$\overline{R_T}(h) = \frac{\sigma^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2 A_\kappa.$$

As an estimate of σ^2 we can use (4), but for A_κ the situation is more complicated. From the previous considerations we can replace the error function $\overline{R_T}(h)$ by the selector $\widetilde{R_T}(h)$ expressed in Lemma 1. If we compare these two error functions, we arrive at results described in next theorems.

Theorem 1 *Let* $\boldsymbol{w}^-$ *be the discrete Fourier transformation of vector* $\boldsymbol{w}$. *Then it holds*

$$\sum_{t=0}^{T-1}(w_t^-)^2 = \frac{1}{h}V(K) + O(T^{-1}). \tag{11}$$

The previous theorem implies that the first term of $\widetilde{R_T}(h)$ estimates the first term of $\overline{R_T}(h)$, that is

$$\frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2 = \frac{\sigma^2 V(K)}{Th} + O(T^{-2}).$$

In next, we will compare the second terms in these error functions to obtain an estimator for A_κ.

Let $\varepsilon > 0$, $h \in (0, 1)$, set J_2 the last index from $\{0, \ldots, T-1\}$ for which

$$J_2 \leq \frac{\sqrt[\kappa+1]{\varepsilon(\kappa+1)!}}{2\pi h}.$$

Let's remark that the parameter ε is an error of Taylor's approximation used in the proof of Theorem 2 and the parameter h is some "starting" approximation of h_{opt}. In our experience, setting $\varepsilon = 10^{-3}$ and $h = \frac{\kappa}{T}$ yields good results. In next we will request both conditions for indexes J_1 and J_2 hold at the same time, so we will define the index J

$$J = \min\{J_1, J_2 + 1\}. \tag{12}$$

Theorem 2 *Let J be the index defined by (12). Then for all $j \in \mathbb{N}$, $1 \leq j \leq J-1$, it holds*

$$\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) = (-1)^{\frac{\kappa}{2}+1}\frac{h^\kappa}{\kappa!}\beta_\kappa + c + O(T^{-1}), \tag{13}$$

where c is a constant satisfying $|c| < \varepsilon$.

By using the result of this theorem we can deduce the estimator of unknown parameter A_κ.

Definition Let J be the index defined by (12). Then the estimator of the parameter A_κ is of the form

$$\widehat{A}_\kappa = \frac{4\pi}{T}\sum_{j=1}^{J-1}(2\pi j)^{2\kappa}\left\{I_{Y_j} - \frac{\hat{\sigma}^2}{2\pi}\right\}.$$

So we can estimate the error function (1)

$$\widehat{\overline{R}}_T(h) = \frac{\hat{\sigma}^2 V(K)}{Th} + \frac{h^{2\kappa}}{(\kappa!)^2}\beta_\kappa^2\widehat{A}_\kappa, \tag{14}$$

and its minimum

$$\hat{h}_{\text{opt}} = \left(\frac{\hat{\sigma}^2 V(K)(\kappa!)^2}{2\kappa T\beta_\kappa^2\widehat{A}_\kappa}\right)^{\frac{1}{2\kappa+1}}. \tag{15}$$

Table 1 Kernels of class $S_{0\kappa}$

κ	$K(x)$
2	$-\frac{3}{4}(x^2-1)$
4	$\frac{15}{32}(x^2-1)(7x^2-3)$
6	$-\frac{105}{256}(x^2-1)(33x^4-30x^2+5)$

Table 2 Summary of sample means and standard deviations of bandwidth estimates

	$\kappa=2; h_{\text{opt}}=0.1374$		$\kappa=4; h_{\text{opt}}=0.3521$		$\kappa=6; h_{\text{opt}}=0.5783$	
	$E(\hat{h}_{\text{opt}})$	$\text{std}(\hat{h}_{\text{opt}})$	$E(\hat{h}_{\text{opt}})$	$\text{std}(\hat{h}_{\text{opt}})$	$E(\hat{h}_{\text{opt}})$	$\text{std}(\hat{h}_{\text{opt}})$
Rice	0.1269	0.0402	0.3354	0.0938	0.4432	0.1078
Plug-in	0.1383	0.0074	0.3422	0.0348	0.5604	0.0623

The parameter $\hat{h}_{\text{opt}}$ given by (15) is the estimator of the theoretical optimal bandwidth h_{opt} obtained by plug-in method. We would like to point out the computational aspect of the plug-in method. It has preferable properties to classical methods, because there is no problem of minimization of any error function. Also the sample size necessary to compute the estimation is far less than for classical methods. On the other side, a small disadvantage could be the fact, that we need some "starting" approximation of unknown parameter h.

5 A simulation study

We carried out a small simulation study to compare the performance of the bandwidth estimates. The observations, Y_t, for $t = 0, \ldots, T-1 = 74$, were obtained by adding independent Gaussian random variables with mean zero and variance $\sigma^2 = 0.2$ to the function

$$m(x) = \sin(2\pi x).$$

Table 1 describes kernels used in our simulation study. The theoretical optimal bandwidth (see Wand and Jones 1995; Koláček 2005) for these cases are given in Table 2.

Two hundred series were generated. Table 2 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 200 values and $std(\hat{h})$ is their standard deviation.

Figure 3 illustrates the histogram of results of all 200 experiments for $\kappa = 2$.

As we can see, the standard deviation of all results obtained by plug-in method is less than the value of case of Rice's selector and also the mean of these results is closer to theoretical optimal bandwidth.

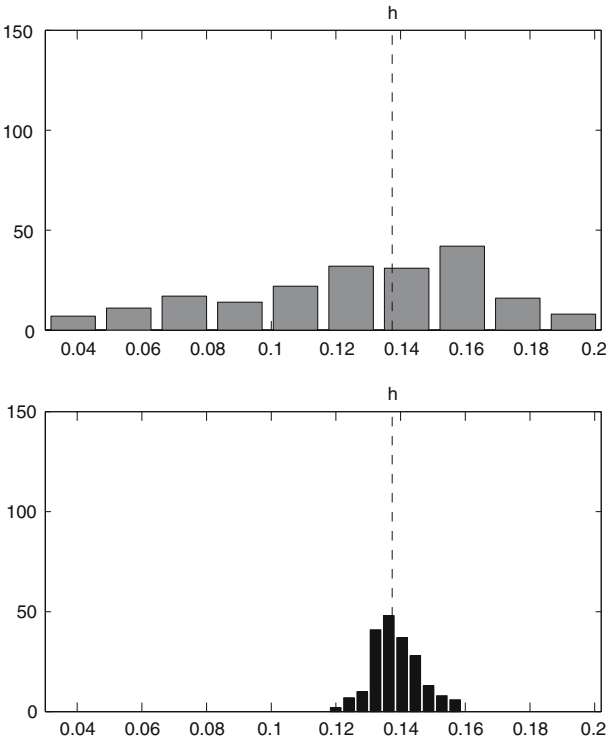


Fig. 3 The histogram of results of all 200 experiments obtained by Rice's selector (*grey*) and by plug-in method (*black*)

6 Examples

In this section, we will solve some practical examples. We used the data from Eurostat[1] and followed the count of marriages in Austria and Switzerland in May in 1950–2003. We transformed the data to the interval [0, 1] and used two selectors to get the optimal bandwidth. Firstly, we found the optimal bandwidth by the Rice's selector $\widehat{R}_T(h)$, which is the classical bandwidth selector. Then we used our proposed selector $\widehat{\widehat{R}}_T(h)$. We made estimations of the regression function with both bandwidths by using the kernel of order (0, 4)

$$K(x) = \begin{cases} \frac{15}{16}(\frac{7}{2}x^4 - 5x^2 + \frac{3}{2}), & |x| \leq 1 \\ 0, & |x| > 1. \end{cases}$$

We used Nadaraya–Watson estimator to obtain final result.

[1] see http://epp.eurostat.cec.eu.int.

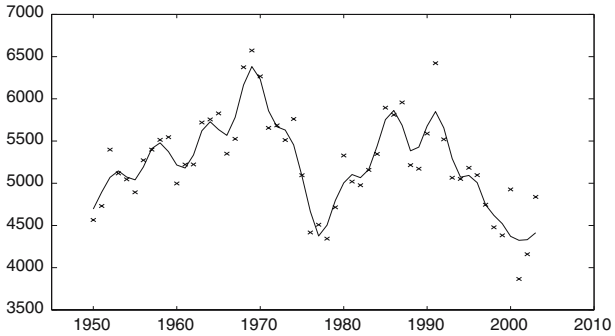


Fig. 4 Estimation of the regression function (*solid line*). The parameter $h = 0.0740$ was found by Rice's selector $\widehat{R}_T(h)$

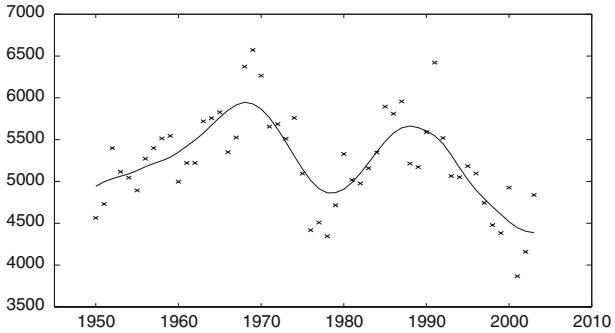


Fig. 5 Estimation of the regression function (*solid line*). The parameter $h = 0.2180$ was found by plug-in method $\widehat{\widehat{R}}_T(h)$

Marriages in Switzerland

In this example we followed the count of marriages in Switzerland in May in 1950–2003.

In this case, the bandwidth obtained by Rice's selector is too small and the final curve is undersmoothed (Figs. 4, 5).

Marriages in Austria

In this example we followed the count of marriages in Austria in May in 1950–2003.

In this case, we think that the value of the bandwidth obtained by Rice's selector is too large and the final curve is oversmoothed (Figs. 6, 7). If we compare results of both examples we can see, that the plug-in method is more stable then the classical one.

7 Conclusion

The problem of bandwidth selection for non-parametric kernel regression is considered. In many studies, there was often observed that classical methods give smaller

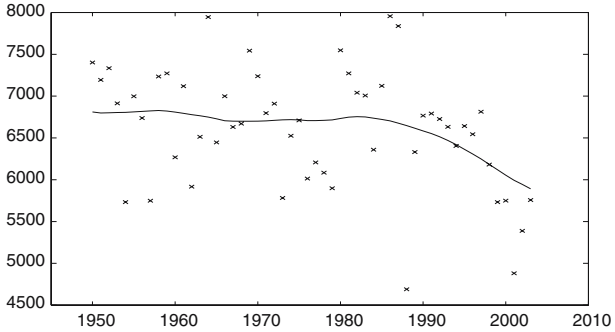


Fig. 6 Estimation of the regression function (*solid line*). The parameter $h = 0.4084$ was found by Rice's selector $\widehat{R}_T(h)$

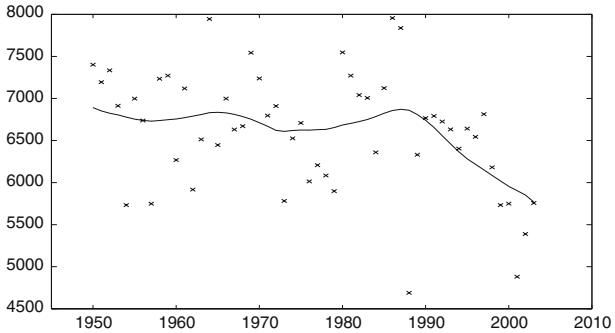


Fig. 7 Estimation of the regression function (*solid line*). The parameter $h = 0.2945$ was found by plug-in method $\widehat{\widehat{R}}_T(h)$

bandwidths more frequently than predicted by the asymptotic theorems. Chiu (1990) provided an explanation for the cause and suggested a procedure to overcome the difficulty. By applying the procedure, we introduced a new approach to estimate unknown parameters of average mean square error function (AMSE) (this process is known as a plug-in method). Let us remark that Chiu's procedure was proposed for Pristley–Chao estimator and for a special class of symmetric probability density functions from S_{02} as kernels. We followed the Nadaraya–Watson and local-linear estimator especially and extended the procedure to these estimators. It was shown they are identical in circular model (see Koláček 2005). In this paper, this approach has been generalized for kernels from the class $S_{0\kappa}$, κ even. The main result of this work is in Theorem 2 and in the resulting definition, where the unknown parameter A_κ is estimated. Simulation study and practical examples suggest that our proposed method could have preferable properties to the classical one.

We remark that the proposed method is developed for a rather limited case: circular design and equally spaced design points. Further research is required for more general situations.

8 Appendix

Lemma 1 *Let J_1 be the least index, that $I_{Y_{J_1}} < \hat{\sigma}^2/\pi T$. Then*

$$\widetilde{R}_T(h) = \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2 + \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\left\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\right\}\{1-w_t^-\}^2.$$

Proof

$$\begin{aligned}
\widetilde{R}_T(h) &= \frac{4\pi}{T}\sum_{t=1}^{N}\tilde{I}_{Y_t}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1} I_{Y_t}\{1-w_t^-\}^2 + \frac{4\pi}{T}\sum_{t=J_1}^{N}\frac{\hat{\sigma}^2}{2\pi}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\left\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\right\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}\{1-w_t^-\}^2 - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\left\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\right\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\left(T - 2Tw_0 + \sum_{t=0}^{T-1}(w_t^-)^2\right) \\
&\quad - \hat{\sigma}^2 + 2\hat{\sigma}^2 w_0 \\
&= \frac{4\pi}{T}\sum_{t=1}^{J_1-1}\left\{I_{Y_t} - \frac{\hat{\sigma}^2}{2\pi}\right\}\{1-w_t^-\}^2 + \frac{\hat{\sigma}^2}{T}\sum_{t=0}^{T-1}(w_t^-)^2.
\end{aligned}$$

Lemma 2 *Let $t \in \{0,\ldots,T-1\}$, then*

$$W_0(x_t) = \frac{1}{T}K_h(x_t) + O(T^{-2}).$$

Proof

$$W_0(x_t) = \frac{1}{TC_T}K_h(x_t),$$

where

$$C_T = \frac{1}{T}\sum_{k=-T+1}^{T-1}K_h(x_k).$$

We can express this constant in another way

$$C_T = \int_{-1}^{1}K(x)dx + O(T^{-1}) = 1 + O(T^{-1})$$

and after substitution we arrive at the result

$$W_0(x_t) = \frac{1}{T(1+O(T^{-1}))}K_h(x_t) = \frac{1}{T}K_h(x_t) + O(T^{-2}).$$

Theorem 1 *Let $\boldsymbol{w}^-$ be the discrete Fourier transformation of vector $\boldsymbol{w}$. Then it holds*

$$\sum_{t=0}^{T-1}(w_t^-)^2 = \frac{1}{h}V(K) + O(T^{-1}).$$

Proof

$$\begin{aligned}
\sum_{t=0}^{T-1}(w_t^-)^2 &= \sum_{t=0}^{T-1}|w_t^-|^2 = \sum_{t=0}^{T-1} w_t^-\overline{w_t^-} \\
&= \sum_{t=0}^{T-1}\sum_{j=-T+1}^{T-1}\sum_{k=-T+1}^{T-1} W_0(x_j)W_0(x_k)e^{\frac{i2\pi(k-j)t}{T}} \\
&= \sum_{j=-T+1}^{T-1}\sum_{k=-T+1}^{T-1} W_0(x_j)W_0(x_k)\sum_{t=0}^{T-1} e^{\frac{i2\pi(k-j)t}{T}} \\
&= T\sum_{k=-T+1}^{T-1} W_0^2(x_k) = \sum_{k=-T+1}^{T-1}\frac{1}{T}K_h^2(x_k) + O(T^{-1}) \\
&= \int_{-1}^{1} K_h^2(u)du + O(T^{-1}) = \frac{1}{h}\int_{-1}^{1} K^2(x)dx + O(T^{-1}).
\end{aligned}$$

Theorem 2 *Let J be the index defined by (12). Then for all $j \in \mathbb{N}$, $1 \le j \le J-1$, it holds*

$$\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) = (-1)^{\frac{\kappa}{2}+1}\frac{h^\kappa}{\kappa!}\beta_\kappa + c + O(T^{-1}), \tag{16}$$

where c is a constant satisfying $|c| < \varepsilon$.

Proof

$$\begin{aligned}
\frac{1}{(2\pi j)^\kappa}(1 - w_j^-) &= \frac{1}{(2\pi j)^\kappa}\left\{1 - 2\sum_{t=0}^{T-1} W_0(x_t)\cos\left(\frac{2\pi tj}{T}\right)\right\} \\
&= \frac{1}{(2\pi j)^\kappa}\left\{1 - 2\sum_{t=0}^{T-1}\frac{1}{T}K_h(x_t)\cos\left(\frac{2\pi tj}{T}\right)\right\} + O(T^{-1})
\end{aligned}$$

$$= \frac{1}{(2\pi j)^{\kappa}} \left\{ 1 - 2 \int_0^1 K_h(u) \cos(2\pi j u) du \right\} + O(T^{-1})$$

$$= \frac{1}{(2\pi j)^{\kappa}} \left\{ \int_{-1}^1 K_h(u) du - \int_{-1}^1 K_h(u) \cos(2\pi j u) du \right\} + O(T^{-1})$$

$$= \frac{1}{(2\pi j)^{\kappa}} \int_{-1}^1 \{1 - \cos(2\pi j u)\} K_h(u) du + O(T^{-1}).$$

We can replace the function $1 - \cos(2\pi j u)$ by Taylor's polynomial of degree κ. Let R_κ is an error of this approximation

$$\frac{1}{(2\pi j)^{\kappa}}(1 - w_j^-) = \frac{1}{(2\pi j)^{\kappa}} \int_{-1}^1 \left\{ \frac{(2\pi j u)^2}{2} - \frac{(2\pi j u)^4}{24} + \cdots + \frac{(-1)^{\frac{\kappa}{2}+1}(2\pi j u)^{\kappa}}{\kappa!} \right\}$$

$$\times K_h(u) du + \frac{R_\kappa}{(2\pi j)^{\kappa}} + O(T^{-1})$$

$$= \frac{(-1)^{\frac{\kappa}{2}+1}}{\kappa!} \int_{-1}^1 u^{\kappa} K_h(u) du + \frac{R_\kappa}{(2\pi j)^{\kappa}} + O(T^{-1})$$

$$= (-1)^{\frac{\kappa}{2}+1} \frac{h^{\kappa}}{\kappa!} \int_{-1}^1 x^{\kappa} K(x) dx + \frac{R_\kappa}{(2\pi j)^{\kappa}} + O(T^{-1}).$$

The last two terms are negligible, because $O(T^{-1})$ tends to zero with $T \to \infty$ and from the assumptions for index j holds $\left| \frac{R_\kappa}{(2\pi j)^{\kappa}} \right| \leq \frac{\varepsilon}{(2\pi)^{\kappa}}$ for any $\varepsilon > 0$.

References

Cleveland WS (1979) Robust locally weighted regression and smoothing scatter plots. J Am Stat Assoc 74:829–836

Craven P, Wahba G (1979) Smoothing noisy data with spline function. Numer Math 31:377–403

Chiu ST (1991) Some stabilized bandwidth selectors for nonparametric regression. Ann Stat 19:1528–1546

Chiu ST (1990) Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. Biometrika 77:222–226

Droge B (1996) Some comments on cross-validation. Stat Theory Comput Aspects Smooth 178–199

Härdle W (1990) Applied nonparametric regression. Cambridge University Press, Cambridge

Härdle W, Hall P, Marron JS (1988) How far are automatically chosen regression smoothing parameters from their optimum? J Am Stat Assoc 83:86–95

Koláček J (2005) Kernel estimation of the regression function. PhD-thesis, Brno

Nadaraya EA (1964) On estimating regression. Theory Probab Appl 10:186–190

Rice J (1984) Bandwidth choice for nonparametric regression. Ann Stat 12:1215–1230

Silverman BW (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J Roy Stat Soc Ser B 47:1–52
Stone CJ (1977) Consistent nonparametric regression. Ann Stat 5:595–645
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London
Watson GS (1964) Smooth regression analysis. Shankya Ser A 26:359–372

AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 281–288

A Comparative Study of Boundary Effects for Kernel Smoothing

Jan Koláček[1] and Jitka Poměnková
Masaryk University, Brno, Czech Republic

Abstract: The problem of boundary effects for nonparametric kernel regression is considered. We will follow the problem of bandwidth selection for Gasser-Müller estimator especially. There are two ways to avoid the difficulties caused by boundary effects in this work. The first one is to assume the circular design. This idea is effective for smooth periodic regression functions mainly. The second presented method is reflection method for kernel of the second order. The reflection method has an influence on the estimate outside edge points. The method of penalizing functions is used as a bandwidth selector. This work compares both techniques in a simulation study.

Keywords: Bandwidth Selection, Kernel Estimation, Nonparametric Regression.

1 Basic Terms and Definitions

Consider a standard regression model of the form

$$Y_i = m(x_i) + \varepsilon_i\,, \qquad i = 1, \ldots, n\,, \quad n \in \mathbb{N}\,,$$

where m is an unknown regression function, x_i are design points, Y_i are measurements and ε_i are independent random variables for which

$$\mathrm{E}(\varepsilon_i) = 0\,, \quad \mathrm{var}(\varepsilon_i) = \sigma^2 > 0\,, \qquad i = 0, \ldots, n\,.$$

The aim of kernel smoothing is to find suitable approximation $\widehat{m}$ of an unknown function m.

In next we will assume the design points x_i are equidistantly distributed on the interval $[0, 1]$, that is $x_i = (i-1)/n$, $i = 1, \ldots, n$.

$Lip[a, b]$ denotes the class of continuous functions satisfying the inequality

$$|g(x) - g(y)| \le L|x - y|\,, \quad \forall x, y \in [a, b]\,, \quad L > 0\,, \quad L \text{ is a constant.}$$

Definition. Let κ be a nonnegative even integer and assume $\kappa \ge 2$. The function $K \in Lip[-1, 1]$, support(K) = $[-1, 1]$, satisfying the following conditions

1. $K(-1) = K(1) = 0$
2. $\int_{-1}^{1} x^j K(x)\,dx = \begin{cases} 0, & 0 < j < \kappa \\ 1, & j = 0 \\ \beta_\kappa \neq 0, & j = \kappa, \end{cases}$

is called a *kernel* of order κ and a class of all these kernels is marked $S_{0\kappa}$. These kernels are used for an estimation of the regression function (see Wand and Jones, 1995). Let $K \in S_{0\kappa}$, set $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, $h \in (0, 1)$. A parameter h is called a *bandwidth*.

[1]Supported by the GACR: 402/04/1308

2 Kernel Estimation of the Regression Function

Commonly used non-parametric methods for estimating $m(x)$ are the kernel estimators **Gasser–Müller estimators** (1979)

$$\widehat{m}_{GM}(x;h) = \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_h(t-x)\,dt\,,$$

where

$$s_i = \frac{x_i + x_{i+1}}{2}\,, \qquad i = 1,\ldots,n-1\,, \qquad s_0 = 0\,, \quad s_n = 1\,.$$

The kernel estimators can be generally expressed as

$$\widehat{m}(x;h) = \sum_{i=1}^{n} W_i(x) Y_i\,,$$

where the weights $W_i(x)$ correspond to the weights of the estimators $\widehat{m}_{GM}$.

The quality of the estimated curve is affected by the smoothing parameter h, which is called a bandwidth. The optimal bandwidth considered here is h_{opt}, the minimizer of the average mean squared error

$$(AMSE) \qquad R_n(h) = \frac{1}{n}\mathrm{E}\sum_{i=1}^{n}(m(x_i) - \widehat{m}(x_i;h))^2\,.$$

Let $K \in S_{0\kappa}$. There exist many estimators of this error function, which are asymptotically equivalent and asymptotically unbiased (see Chiu, 1991, 1990; Härdle, 1990). Most of them are based on the residual sum of squares

$$(RSS) \qquad RSS_n(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \widehat{m}(x_i;h)]^2\,.$$

We will use the method of penalizing functions (see Koláček, 2005, 2002) for choosing the smoothing parameter. So the prediction error $RSS_n(h)$ is adjusted by some penalizing function $\Xi(n^{-1}W_i(x_i))$, that is, modified to

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n}[\widehat{m}(x_i;h) - Y_i]^2 \cdot \Xi(n^{-1}W_i(x_i))\,.$$

The reason for this adjustment is that the correction function $\Xi(n^{-1}W_i(x_i))$ penalizes values of h too low. For example Rice (see Rice, 1984) considered

$$\Xi_R(u) = \frac{1}{1-2u}\,.$$

This penalizing function will be used.

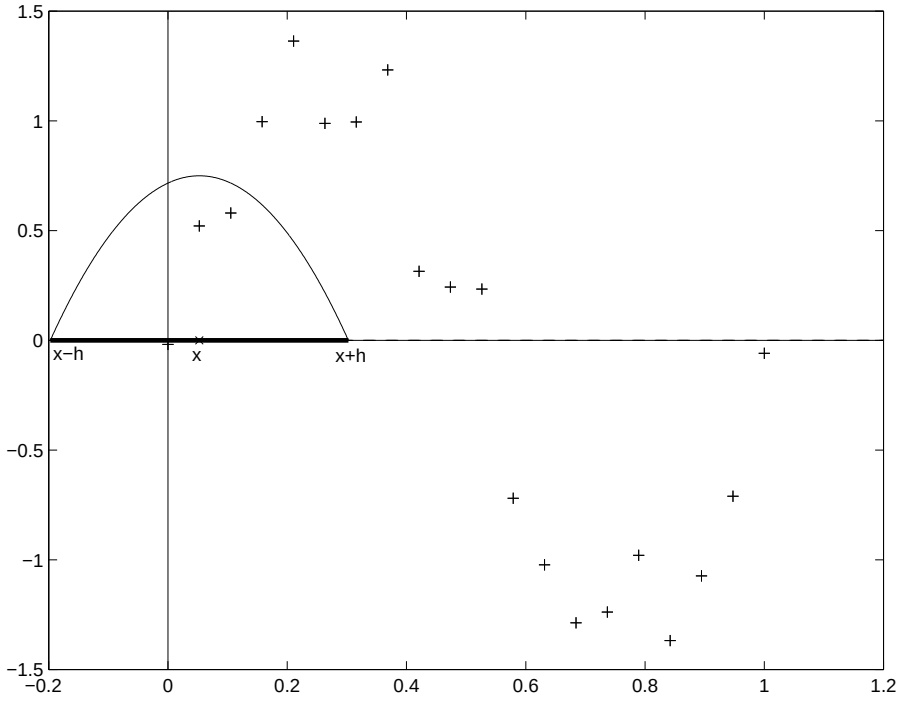


Figure 1: Demonstration of boundary effects.

3 Boundary Effects

In the finite sample situation, the quality of the estimate in the boundary region $[0, h] \cup [1-h, 1]$ is affected since the effective window is $[x-h, x+h] \not\subset [0, 1]$ so, that the finite equivalent of the moment conditions on the kernel function does not apply any more. There are several methods to avoid the difficulties caused by boundary effects.

3.1 Cyclic Model

One of possible ways to solve problem of boundary effects is to use a cyclic design. That is, suppose $m(x)$ is a smooth periodic function and the estimate is obtained by applying the kernel on the extended series $\widetilde{Y}_i$, where $\widetilde{Y}_{i+kn} = Y_i$ for $k \in \mathbb{Z}$. Similarly $x_i = (i-1)/n$, $i \in \mathbb{Z}$.

In the cyclic design, the kernel estimators can be generally expressed as

$$\widehat{m}(x; h) = \sum_{i=-n+1}^{2n} W_i(x) \widetilde{Y}_i \,,$$

where the weights $W_i(x)$ correspond to the weights of estimators $\widehat{m}_{GM}$

$$W_i(x) = \int_{s_{i-1}}^{s_i} K_h(t-x)dt,$$

where

$$s_i = \frac{x_i + x_{i+1}}{2} \,, \qquad i = -n+1, \ldots, 2n-1 \,, \quad s_{-n} = -1 \,, \quad s_{2n} = 2 \,.$$

Let us define a vector $\boldsymbol{w} := (w_1, \ldots, w_n)$, where

$$w_i = W_1(x_i - 1) + W_1(x_i) + W_1(x_i + 1)\,.$$

Let $h \in (0,1)$, $K \in S_{0\kappa}$, $i \in \{1, \ldots, n\}$. Then we can write $\widehat{m}(x_i; h)$ as a discrete cyclic convolution of vectors $\boldsymbol{w}$ and $\boldsymbol{Y}$.

$$\widehat{m}(x_i; h) = \sum_{k=1}^{n} w_{<i-k>_n} Y_k\,, \tag{1}$$

where $< i - k >_n$ marks $(i - k) \bmod n$. We write

$$\widehat{\boldsymbol{m}} = \boldsymbol{w} \circledast \boldsymbol{Y}\,,$$

where $\widehat{\boldsymbol{m}} = (\widehat{m}(x_1; h), \ldots, \widehat{m}(x_n; h))$.

As the bandwidth selector the method of Rice's penalizing function will be used. In the case of cyclic model, we can simplify the error function $\widehat{R}_n(h)$, because the weights $W_i(x_i)$ are independent on i. Set

$$I(h) := \int_{-1/2n}^{1/2n} K_h(x) dx\,.$$

Then we can express $\widehat{R}_n(h)$ as

$$\widehat{R}_n(h) = \frac{n}{n - 2\,I(h)}\, RSS_n(h) \tag{2}$$

and the estimate $\hat{h}_{opt}$ of optimal bandwidth is defined as

$$\hat{h}_{opt} = \operatorname*{arg\,min}_{h \in (0,1)} \widehat{R}_T(h)\,.$$

3.2 Reflection Technique

Let's have observations (x_i, Y_i), $i = 1, \ldots, n$, regression model described in Section 1 and design points $x_i \in [0, 1]$ such that

$$0 = a \le x_1 \le \cdots \le x_n \le b = 1\,.$$

Now, technique for design points reflection will be discussed. We may begin by estimating the function m at edge points a and b with corresponding bandwidth for these points, h_a and h_b, and edge kernels $K_L, K_R \in S_{02}$:

$$\widehat{m}(a) = \frac{1}{h_a} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_L\left(\frac{a - u}{h_a}\right) du\,,$$

$$\widehat{m}(b) = \frac{1}{h_b} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K_R\left(\frac{b - u}{h_b}\right) du\,.$$

For the bandwidth choice h_a, h_b and the edge kernels K_L, K_R for $\widehat{m}(a), \widehat{m}(b)$ see Poměnková (2005). Further data reflection will be made. We proceed from original data set (x_i, Y_i), $i = 1, \ldots, n$. For obtaining left mirrors point $(a, \widehat{m}(a))$ and following relations

$$\begin{aligned} x_{Li} &= 2a - x_i \,, \\ Y_{Li} &= 2\widehat{m}(a) - Y_i \end{aligned}$$

are used. For obtaining right mirrors point $(b, \widehat{m}(b))$ and following relations

$$\begin{aligned} x_{Ri} &= 2b - x_{n-i+1} \,, \\ Y_{Ri} &= 2\widehat{m}(b) - Y_{n-i+1} \end{aligned}$$

are used. Then original data set (x_i, Y_i) is connected with left mirrors (x_{Li}, Y_{Li}) and with right mirrors (x_{Ri}, Y_{Ri}). By this connection new data set which is called pseudodata and denoted as $(\overline{x}_j, \overline{Y}_j)$, $j = 1, \ldots, 3n$.

How to find the bandwidth for an estimate on pseudodata at the design points will be in next. Finally, the function m in design points including points a and b using the pseudodata is estimated.

Let $K \in S_{02}$ be a symmetric second-order kernel with support $[-1, 1]$. The final estimate of function $\widehat{m}$ at points of plan x_i, $i = 0, \ldots, n+1$, where $x_0 = a$, $x_{n+1} = b$ on pseudodata $\overline{x}_j$, $j = 1, \ldots, 3n$, with kernel K and bandwidth h is defined

$$\widehat{m}(x) = \frac{1}{h} \sum_{j=1}^{3n} \overline{Y}_j \int_{s_{j-1}}^{s_j} K\left(\frac{x-u}{h}\right) du \,,$$

where

$$s_j = \frac{\overline{x}_i + \overline{x}_{i+1}}{2} \,, \qquad j = 1, \ldots, 3n-1 \,, \quad s_0 = -1 \,, \quad s_{3n} = 2 \,.$$

Bandwidth selection for pseudodata

In this part an estimate of the bandwidth for pseudodata will be searched. Note that estimates at edge points $\widehat{m}(a), \widehat{m}(b)$ are functions of h. Therefore, for any chosen value $h \in H = [1/n, 2]$ values $\widehat{m}(a), \widehat{m}(b)$ have to be enumerated, then data reflection is made and pseudodata are obtained. Hereafter, on this pseudodata minimum of the function is searched.

To find value h using a Rice penalization function is proposed. Consider pseudodata $(\overline{x}_j, \overline{Y}_j)$, $j = 1, \ldots, 3n$, $\bar{x}_j \in [-1, 2]$, $\widehat{m}(x)$ defined as above. Then

$$\widehat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} [\widehat{m}(x_i; h) - Y_i]^2 \cdot \frac{1}{1 - 2x_i} \,.$$

The resulting bandwidth $h = \hat{h}_{opt}$ is the value h that corresponds to the minimum of the function $\widehat{R}_n(h)$, i.e.

$$\hat{h}_{opt} = \arg\min_{h \in H} \widehat{R}_n(h) \,. \tag{3}$$

4 A Simulation Study

We carried out a small simulation study to compare the performance of the bandwidth estimates. The observations Y_i, for $i = 1, \ldots, n = 75$, were obtained by adding independent Gaussian random variables with mean zero and variance $\sigma^2 = 0.2$ to the function

$$m(x) = \cos(9x - 7) - (3 + x^{12})/6 + 8^{x-1}\,.$$

We made estimations of the regression function by using the kernel of order 2

$$K(x) = \begin{cases} -\frac{3}{4}(x^2 - 1), & |x| \leq 1 \\ 0, & |x| > 1\,. \end{cases}$$

In this case, there was selected $\hat{h} = 0.0367$ by using an estimate without any elimination of boundary effects (Figure 2). At the second, there was selected $\hat{h} = 0.0867$ by using the method of cyclic model (Figure 3) and at the third, there was selected $\hat{h} = 0.2036$ by using the reflection method (Figure 4).

From the figures it can be seen that both, cyclic model and reflection method, are very useful for removing problems caused by boundary effects.

5 A Practical Example

We carried out a short real application to compare the performance of the bandwidth estimates. The observations Y_i, for $i = 1, \ldots, n = 230$, were average spring temperatures measured in Prague between 1771 – 2000. The data were obtained from Department of Geography, Masaryk University. We made estimations of the regression function by using the kernel of order 2

$$K(x) = \begin{cases} -\frac{3}{4}(x^2 - 1), & |x| \leq 1 \\ 0, & |x| > 1\,. \end{cases}$$

In this case, there was selected $\hat{h} = 0.0671$ by using an estimate without any elimination of boundary effects (Figure 5). At the second, there was selected $\hat{h} = 0.0671$ by using the method of cyclic model (Figure 6) and at the third, there was selected $\hat{h} = 0.2211$ by using the reflection method (Figure 7). These figures show that both, cyclic model and reflection method, are very useful for removing problems caused by boundary effects.

References

Chiu, S. (1990). Why bandwidth selectors tend to choose smaller bandwidths, and a remedy. *Biometrika*, *77*, 222-226.

Chiu, S. (1991). Some stabilized bandwidth selectors for nonparametric regression. *Annals of Statistics*, *19*, 1528-1546.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Koláček, J. (2002). Kernel estimation of the regression function – bandwidth selection. *Summer School DATASTAT'01 Proceedings FOLIA*, *1*, 129-138.

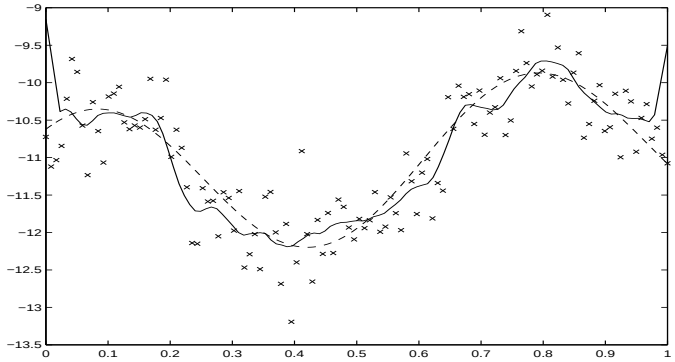


Figure 2: Graph of smoothness function with bandwidth $h = 0.0367$, the real regression function m, an estimate of m.

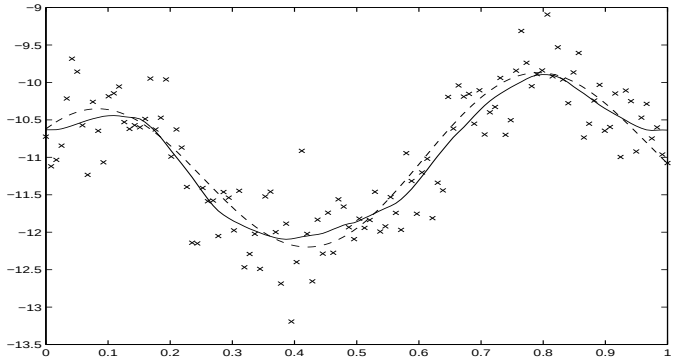


Figure 3: Graph of smoothness function with bandwidth $h = 0.0867$, the real regression function m, an estimate of m.

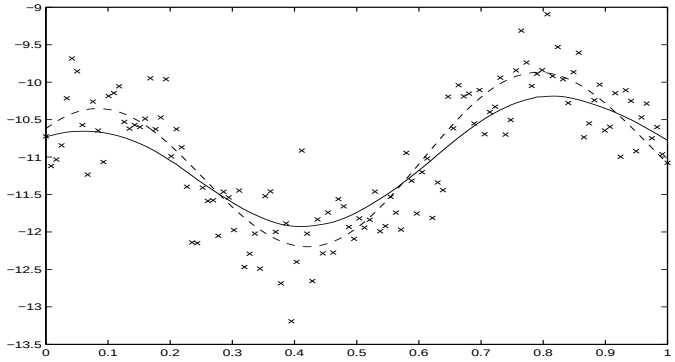


Figure 4: Graph of smoothness function with bandwidth $h = 0.2036$, the real regression function m an estimate of m.

Koláček, J. (2005). *Kernel Estimators of the Regression Function*. Brno: PhD-Thesis.

Poměnková, J. (2005). *Some Aspects of Regression Function Smoothing (in Czech)*. Ostrava: PhD-Thesis.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, *12*, 1215-1230.

Wand, M., and Jones, M. (1995). *Kernel Smoothing*. London: Chapman & Hall.

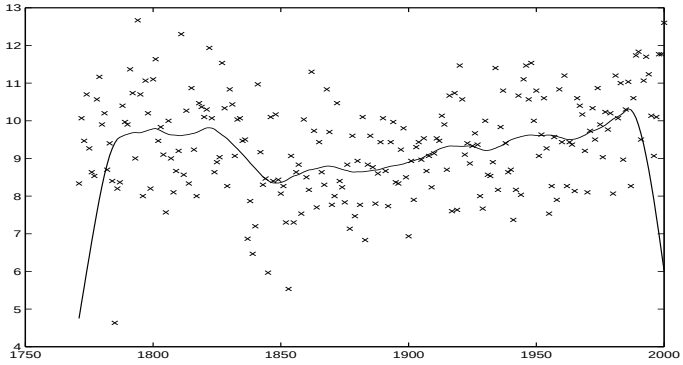


Figure 5: Graph of smoothness function with bandwidth $h = 0.0671$, an estimate of m.

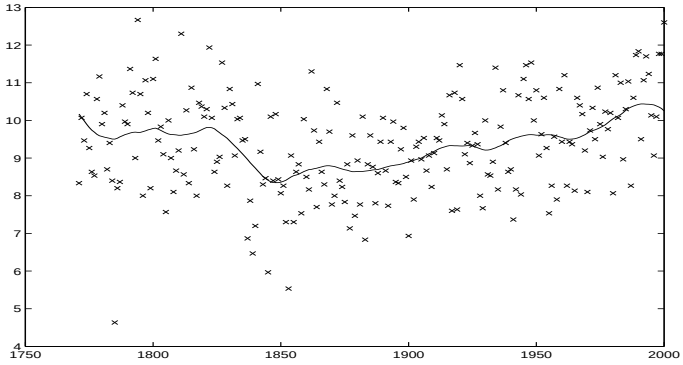


Figure 6: Graph of smoothness function with bandwidth $h = 0.0671$, an estimate of m.

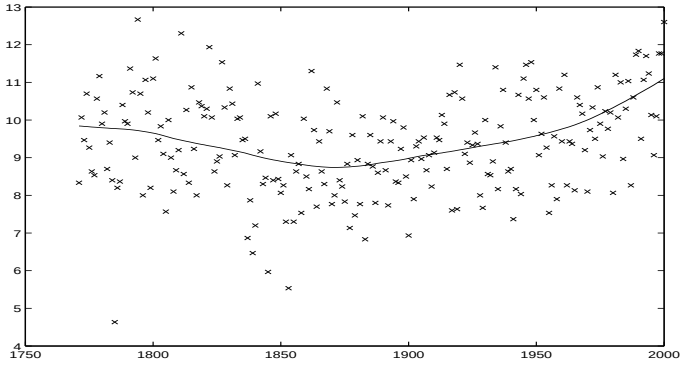


Figure 7: Graph of smoothness function with bandwidth $h = 0.2211$, an estimate of m.

Authors' address:

Jan Koláček, Jitka Pomněnková
Masaryk University in Brno
Department of Applied Mathematics
Janáčkovo náměstí 2a
CZ-602 00 Brno
Czech Republic
E-mail: kolacek@math.muni.cz