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.

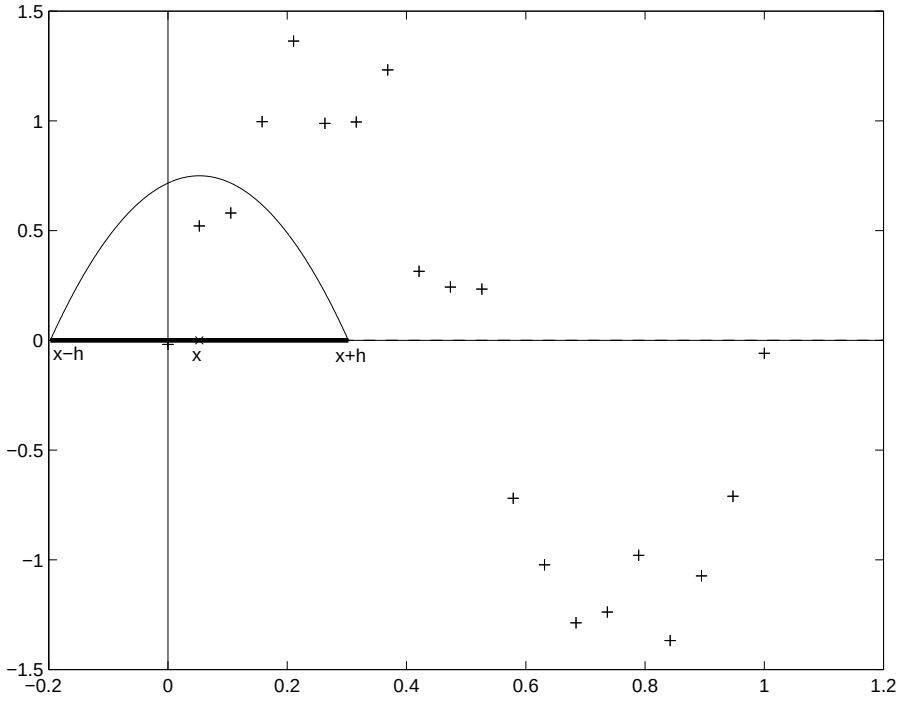

Figure 1: Demonstration of boundary effects.

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

$$x_{Li} = 2a - x_i\,,$$
$$Y_{Li} = 2\widehat{m}(a) - Y_i$$

are used. For obtaining right mirrors point $(b, \widehat{m}(b))$ and following relations

$$x_{Ri} = 2b - x_{n-i+1}\,,$$
$$Y_{Ri} = 2\widehat{m}(b) - Y_{n-i+1}$$

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.

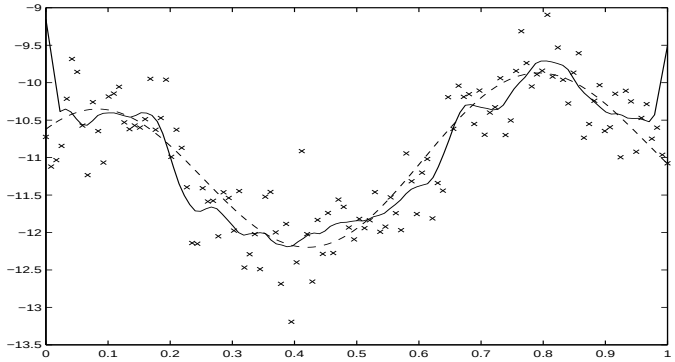

Figure 2: Graph of smoothness function with bandwidth $h = 0.0367$, the real regression function m, an estimate of m.

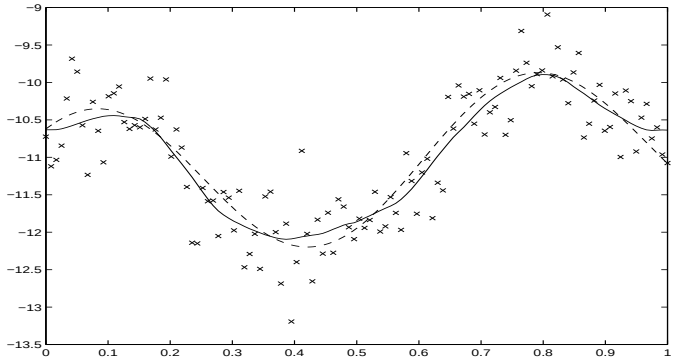

Figure 3: Graph of smoothness function with bandwidth $h = 0.0867$, the real regression function m, an estimate of m.

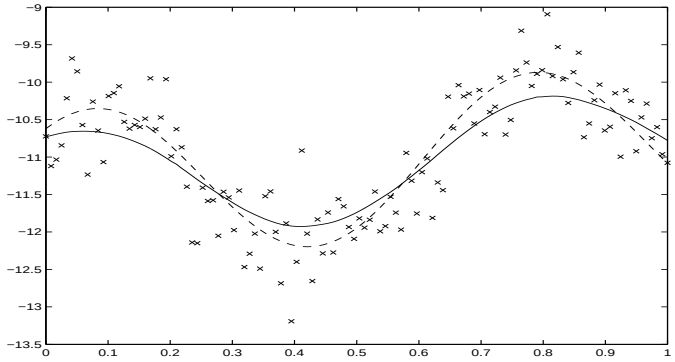

Figure 4: Graph of smoothness function with bandwidth $h = 0.2036$, the real regression function m an estimate of m.

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.

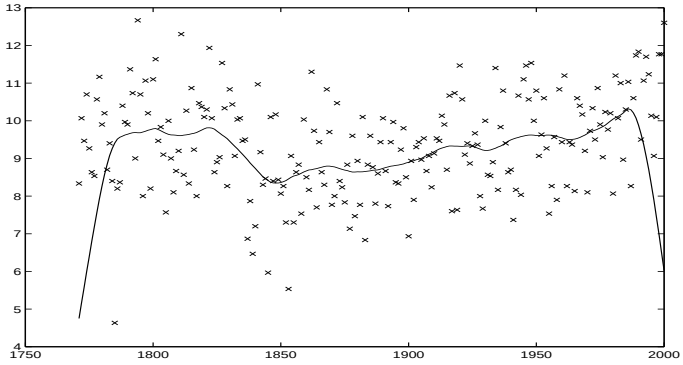

Figure 5: Graph of smoothness function with bandwidth $h = 0.0671$, an estimate of m.

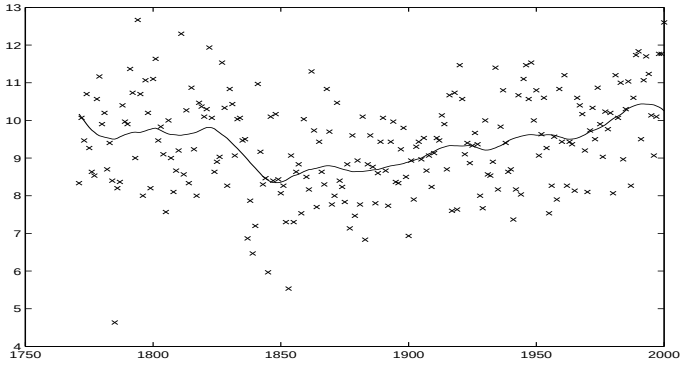

Figure 6: Graph of smoothness function with bandwidth $h = 0.0671$, an estimate of m.

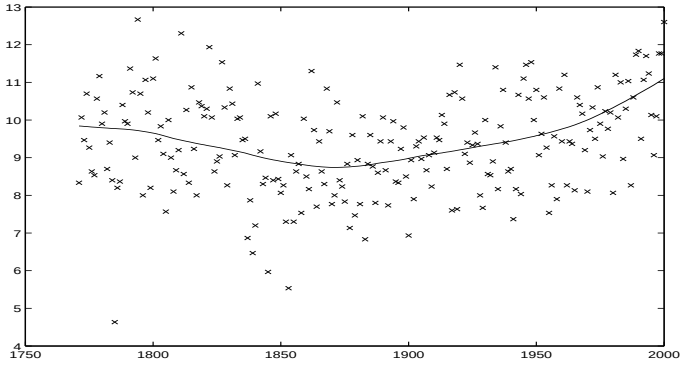

Figure 7: Graph of smoothness function with bandwidth $h = 0.2211$, an estimate of m.

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