Tatra Mt. Math. Publ. **17** (1999), 241–250

TATRA MOUNTAINS
Mathematical Publications

ADDITIVE MODELS AND KERNEL SMOOTHING

JIŘÍ ZELINKA

ABSTRACT. Nonparametric regression methods are often used to estimate an unknown function $(x_1, \dots, x_p)$ in a regression model

$$Y = m(X_1, \dots, X_p) + \varepsilon$$

for random variables $X_1, \dots, X_p, Y$ and error ε. An additive model can be used for the function m in the special form

$$m(x_1, \dots, x_p) = m_1(x_1) + \dots m_p(x_p)\,.$$

The application of kernel smoothing to additive models is demonstrated in this contribution together with some practical results.

1. Kernel smoothing

Let X, Y be random variables and let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a bivariate random sample from the population (X, Y). We consider the model

$$Y = m(X) + \varepsilon\,, \tag{1}$$

where ε is an error with $E\varepsilon = 0$ and $\operatorname{var}\varepsilon = \sigma^2$. It is well-known, that the conditional expectation $E\bigl(Y\,|\,X\bigr)$ minimizes the squared error $E\bigl(Y - m(X)\bigr)^2$ so our goal is to estimate this function $m(x) = E\bigl(Y\,|\,X = x\bigr)$ using the random sample $(X_1, Y_1), \dots, (X_n, Y_n)$.

There are a lot of parametric and nonparametric methods for this estimation. In this paper we shall deal with the *kernel estimation*. The properties of the kernel estimation are described, for example, in [4], [7], [8] or [10] so we can use results from these papers.

2000 Mathematics Subject Classification: 62G05, 62G07, 46C99.
Keywords: linear smoother, smoother matrices, additive model, orthogonal projection.

The work was accomplished under the financial support of the Grant Agency of the Czech Republic, reg. no. 201/96/0665.

The kernel estimation approach uses so-called *kernel functions* or, briefly, *kernels* K with the properties

$$K(-x) = K(x)\,, \qquad \int\limits_{-\infty}^{\infty} K(x)\,dx = 1\,.$$

Other conditions are required to improve the properties of the kernel estimators, for example, the condition $K(x) > 0$ is very often used.

The *Epanechnikov kernel* is the most popular kernel function but some other kernels also give good results. The following table shows the formulas of several kernels.

Epanechnikov kernel	$K(x) = \frac{3}{4}(1-x^2)I_{\langle -1,1\rangle}\,,$
Normal density kernel	$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,,$
Quartic (biweight) kernel	$K(x) = \frac{15}{16}\left(1-x^2\right)^2 I_{\langle -1,1\rangle}\,,$
Triangular kernel	$K(x) = \left(1-\lvert x\rvert\right)I_{\langle -1,1\rangle}\,.$

The notation $K_h(x) = \frac{1}{h}K(\frac{x}{h})$ is often used in kernel estimation. The parameter h is a positive constant called the *bandwidth* and determines which data from a random sample will be used for the estimation. It can be calculated, for example by the application of the cross-validation method.

The kernel estimation $\hat{\boldsymbol{m}}$ of the function m can be written as

$$\hat{\boldsymbol{m}}(x) = \sum_{i=1}^{n} W_{i,h}(x)Y_i\,, \qquad x \in \mathbb{R}\,, \tag{2}$$

where the weights $W_{i,h}$ at the point x depend on the bandwidth h and are constructed with the aid of kernels. Throughout this paper two kinds of weights were used: the *Nadaraya–Watson estimator* and the *local linear estimator*. They can be defined in the following way (see [10]).

For a non-negative integer p the *local polynomial estimator* $\hat{\boldsymbol{m}}(x;p,h)$ is obtained by fitting the polynomial $a_0 + a_1(\cdot - x) + \dots + a_p(\cdot - x)^p$ to the random sample (X_i, Y_i) using least squares with the kernel weight $K_h(X_i - x)$. The value of $\hat{\boldsymbol{m}}(x;p,h)$ is the height of $\hat{a}_0$ where $\hat{\boldsymbol{a}} = (\hat{a}_0, \dots, \hat{a}_p)^T$ minimizes the sum

$$\sum_{i=1}^{n}\left[Y_i - a_0 - a_1(X_i - x) - \dots - a_p(X_i - x)^p\right]^2 \cdot K_h(X_i - x)\,.$$

The vector $\hat{\boldsymbol{a}}$ can be expressed in matrix notation. If M_K is the $n\times n$ diagonal matrix

$$M_K = \begin{bmatrix} K_h(X_1-x) & 0 & \dots & 0 \\ 0 & K_h(X_2-x) & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & K_h(X_n-x) \end{bmatrix},$$

and M_X is the $n\times(p+1)$ matrix

$$M_X = \begin{bmatrix} 1 & X_1-x & \dots & (X_1-x)^p \\ 1 & X_2-x & \dots & (X_2-x)^p \\ \vdots & & \ddots & \vdots \\ 1 & X_n-x & \dots & (X_n-x)^p \end{bmatrix},$$

then assuming the invertibility of $M_X^T M_K M_X$ we get

$$\hat{\boldsymbol{a}} = \left(M_X^T M_K M_X\right)^{-1} M_X^T M_K Y \tag{3}$$

for the response vector $Y = (Y_1,\dots,Y_n)^T$. Now we can obtain the weights $W_{i,h}(x)$ from (2) as the first row of the matrix $\left(M_X^T M_K M_X\right)^{-1} M_X^T M_K$.

For $p=0$ we attach the *Nadaraya–Watson estimator*. In this case, the order of the matrix $M_X^T M_K M_X$ is equal to one,

$$M_X^T M_K M_X = \sum_{i=1}^n K_h(X_i-x).$$

The explicit formula

$$\hat{\boldsymbol{m}}(x) = \frac{\sum\limits_{i=1}^n K_h(X_i-x)Y_i}{\sum\limits_{i=1}^n K_h(X_i-x)}. \tag{4}$$

can be easily derived. The weights in (2) take the form

$$W_{i,h}(x) = \frac{K_h(X_i-x)}{\sum\limits_{j=1}^n K_h(X_j-x)}, \qquad i=1,\dots,n.$$

If $p=1$, we get the *local linear estimator*. The matrix $M_X^T M_K M_X$ can be expressed as

$$M_X^T M_K M_X = \begin{bmatrix} \sum\limits_{i=1}^n K_h(X_i-x) & \sum\limits_{i=1}^n (X_i-x)K_h(X_i-x) \\ \sum\limits_{i=1}^n (X_i-x)K_h(X_i-x) & \sum\limits_{i=1}^n (X_i-x)^2 K_h(X_i-x) \end{bmatrix}.$$

For non-negative kernels it can be proved that this matrix is singular iff all X_i with property $K_h(X_i - x) > 0$ have the same value. The explicit formula for a local linear estimator can be written in the following way:

$$\hat{\boldsymbol{m}}(x) = n^{-1} \sum_{i=1}^{n} \frac{\big[s_2(x) - s_1(x)(X_i - x)\big] K_h(X_i - x) Y_i}{s_2(x) s_0(X) - s_1(x)^2} ,$$

where

$$s_r(x) = m^{-1} \sum_{i=1}^{n} (X_i - x)^r K_h(X_i - x) .$$

For a fixed plan design $X_1, \dots, X_n$ (often denoted by lower case characters $x_1, \dots, x_n$) the estimation $\hat{\boldsymbol{m}}$ is evaluated at these points. For a random plan it is useful to count the values of $\hat{\boldsymbol{m}}$ on any equidistant set of points $x_1, \dots, x_m$ (m and n need not be the same) and to construct the *smoother matrix*

$$S = \begin{bmatrix} W_{1,h}(x_1) & W_{2,h}(x_1) & \dots & W_{n,h}(x_1) \\ W_{1,h}(x_2) & W_{2,h}(x_2) & \dots & W_{n,h}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ W_{1,h}(x_m) & W_{2,h}(x_m) & \dots & W_{n,h}(x_m) \end{bmatrix} . \tag{5}$$

Then $\big(\hat{\boldsymbol{m}}(x_1), \dots, \hat{\boldsymbol{m}}(x_m)\big)^T = SY$.

2. Additive model

Now we shall suppose that the random variable Y depends on the random variables $X_1, \dots, X_p$, i.e.,

$$Y = m(X_1, \dots, X_p) + \varepsilon , \tag{6}$$

where the variables $X_1, \dots, X_p$ are independent and $E\varepsilon = 0$, $\operatorname{var} \varepsilon = \sigma^2$. Our aspiration is to estimate the function m as the conditional expectation $E(Y|X_1, \dots, X_p)$ in the same way as in the one-dimensional case. There are some difficulties with the estimation of the function m using methods such as kernel estimation because we need a huge random sample for relatively small values of p (about 5).

The situation becomes easier if we suppose the function m to be in the form

$$m(x_1, \dots, x_p) = m_1(x_1) + \dots + m_p(x_p) . \tag{7}$$

This case is called an *additive model* and the functions m_i have the form $m_i(x_i) = E\big(Y|X_i = x_i\big)$ for $i = 1, \dots, p$. We may derive the equations for the determination of the functions m_i, $i = 1, \dots, p$ as follows (see [3], [6]).

Let $\mathcal{C}$ be the space of random variables with finite second moment (assuming that the random variables X and X' are identified if the probability $P(X = X')$ is equal to one). Then $\mathcal{C}$ is a Hilbert space with inner product $\langle X, X'\rangle = EXX'$ and norm $\|X\| = \sqrt{EXX}$. We shall assume that $Y \in \mathcal{C}$ and $m(X_1, \dots, X_p) \in \mathcal{C}$.

Now let $\mathcal{H} = \mathcal{H}(X_1, \dots, X_p)$ be the subspace of $\mathcal{C}$ containing the random variables in the form $\Phi(X_1, \dots, X_p)$ where Φ is the Borel function for which the relation $\Phi(X_1, \dots, X_p) \in \mathcal{C}$ holds. We may say that the subspace $\mathcal{H}$ is generated by the random variables $X_1, \dots, X_p$. Similarly, we may construct the subspaces $\mathcal{H}_1 = \mathcal{H}_1(X_1), \dots, \mathcal{H}_p = \mathcal{H}_P(X_p)$ each generated by one random variable. Finally, let $\mathcal{H}_{ADD}$ be the subspace of $\mathcal{C}$ generated by functions having the form $\Phi(X_1, \dots, X_p) = \Phi_1(X_1) + \cdots + \Phi_p(X_p)$.

It is clear that $\mathcal{H}_i \subset \mathcal{H}_{ADD} \subset \mathcal{H}$, $m \in \mathcal{H}_{ADD}$ and $m_i \in \mathcal{H}_i$ for $i = 1, \dots, p$. Let us denote by $\mathcal{P}_1, \dots, \mathcal{P}_p$ and $\mathcal{P}_{ADD}$ the orthogonal projections from the space $\mathcal{C}$ to the subspaces $\mathcal{H}_1, \dots, \mathcal{H}_p$ and $\mathcal{H}_{ADD}$ respectively.

The random variable $m(X_1, \dots, X_p)$ minimizes the mean squared error

$$E\big(Y - m(X_1, \dots, X_p)\big)^2 = \big\|Y - m(X_1, \dots, X_p)\big\|^2$$

over $\mathcal{H}_{ADD}$ if $Y - m(X_1, \dots, X_p)$ is orthogonal to $\mathcal{H}_{ADD}$. But then it must be orthogonal to every subspace $\mathcal{H}_i$ or equivalently

$$P_i\Big(Y - m(X_1, \dots, X_p)\Big) = 0 \qquad \text{for} \quad i = 1, \dots, p.$$

As $m_i(X_i) \in \mathcal{H}_i$ we have $P_i m_i(X_i) = m_i(X_i)$ for $i = 1, \dots, p$ so we can write

$$m_i(X_i) = P_i\Big(Y - \sum_{j \neq i} m_j(X_j)\Big) \qquad \text{for} \quad i = 1, \dots, p.$$

This equation can be rewritten using matrix and operator forms:

$$\begin{bmatrix} I & \mathcal{P}_1 & \mathcal{P}_1 & \dots & \mathcal{P}_1 \\ \mathcal{P}_2 & I & \mathcal{P}_2 & \dots & \mathcal{P}_2 \\ & \vdots & & \ddots & \vdots \\ \mathcal{P}_p & \mathcal{P}_p & \mathcal{P}_p & \dots & I \end{bmatrix} \begin{bmatrix} m_1(X_1) \\ m_2(X_2) \\ \vdots \\ m_p(X_p) \end{bmatrix} = \begin{bmatrix} \mathcal{P}_1 Y \\ \mathcal{P}_2 Y \\ \vdots \\ \mathcal{P}_p Y \end{bmatrix}. \tag{8}$$

Now let us suppose that we have the random sample $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ from the population $(X_1, \dots, X_p, Y)$ where $\mathbf{X}_i = (X_{1i}, \dots, X_{pi})$. We shall use this random sample to find an estimation $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}_1 + \cdots + \hat{\boldsymbol{m}}_p$ of the function m.

Let us put $\hat{\boldsymbol{m}}_i = \big(\hat{\boldsymbol{m}}_i(X_{i1}), \dots, \hat{\boldsymbol{m}}_i(X_{in})\big)^T$ for $i = 1, \dots, p$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. If $S_1, \dots, S_p$ are the smoothing matrices for ith component of $\mathbf{X}$

at the points $X_{i1}, \dots, X_{in}$ and I is the identity matrix, we obtain the following data version of (8)

$$\begin{bmatrix} I & S_1 & S_1 & \dots & S_1 \\ S_2 & I & S_2 & \dots & S_2 \\ & \vdots & & \ddots & \vdots \\ S_p & S_p & S_p & \dots & I \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{m}}_1 \\ \hat{\boldsymbol{m}}_2 \\ \vdots \\ \hat{\boldsymbol{m}}_p \end{bmatrix} = \begin{bmatrix} S_1\mathbf{Y} \\ S_2\mathbf{Y} \\ \vdots \\ S_p\mathbf{Y} \end{bmatrix}. \tag{9}$$

Solving this system, we get the estimations $\hat{\boldsymbol{m}}_1, \dots, \hat{\boldsymbol{m}}_p$ and also the function $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}_1 + \dots + \hat{\boldsymbol{m}}_p$.

The solution of (9) is non-unique as the constant random variables belong to in all the subspaces $\mathcal{H}_i$, thus $\mathcal{H}_{ADD}$ is not the direct sum of the spaces $\mathcal{H}_1, \dots, \mathcal{H}_p$. This problem can be solved using model

$$m(x_1, \dots, x_p) = m_1(x_1) + \dots + m_p(x_p) + a\,, \tag{10}$$

where $Em_1(X_1) = \dots = Em_p(X_p) = 0$ and $a = EY$. Using a similar procedure to that used above we obtain the system of equations

$$\begin{bmatrix} I & S_1^* & S_1^* & \dots & S_1^* \\ S_2^* & I & S_2^* & \dots & S_2^* \\ & \vdots & & \ddots & \vdots \\ S_p^* & S_p^* & S_p^* & \dots & I \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{m}}_1 \\ \hat{\boldsymbol{m}}_2 \\ \vdots \\ \hat{\boldsymbol{m}}_p \end{bmatrix} = \begin{bmatrix} S_1^*\mathbf{Y} \\ S_2^*\mathbf{Y} \\ \vdots \\ S_p^*\mathbf{Y} \end{bmatrix}. \tag{11}$$

The matrices S_i^*, $i = 1, \dots, p$, were obtained by centering the matrices S_i, i.e.,

$$S_i^* = (I - \frac{1}{n}J)S_i\,, \qquad J = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix},$$

and the constant a is the mean of $Y_1, \dots, Y_n$.

3. Practical results

The theory described above was used for smoothing data in the two-dimensional case. We had two data sets. The first one is the set of real basketball data acquired from the internet address

`http://www.stern.nyu.edu/SOR/SmoothMeth.`

The variable X_1 is height of the basketball player, X_2 is the time period of playing and the response variable Y gives the average number of points gained by the player during one minute of the match.

The second data set was simulated. X_1 and X_2 were normally distributed random variables, the data version $\mathbf{X}_1$ and $\mathbf{X}_2$ was transformed to the interval $\langle 0,1\rangle$ for easier counting. The response variable Y was given as the sum of the exponential function in X_1, the polynomial function in X_2 and the normally distributed error.

There are several methods for solving the system (9). We have tried the following:

- M–P pseudoinverse matrix;
- semiiterative method; (see [2], [5]);
- back fitting algorithm (or blocked Gauss–Seidel iterative method). This can be used according the following scheme:

$$\begin{aligned}\hat{\boldsymbol{m}}_1^{(k+1)} &= S_1\big(\mathbf{Y} - \hat{\boldsymbol{m}}_2^{(k)}\big)\\ \hat{\boldsymbol{m}}_2^{(k+1)} &= S_2\big(\mathbf{Y} - \hat{\boldsymbol{m}}_1^{(k+1)}\big)\end{aligned}$$

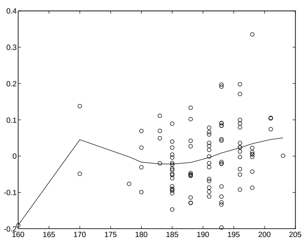
  derived from (9) for the appropriate starting iteration.

The solution of (11) can be expressed directly (see [9]) for $p = 2$ as.

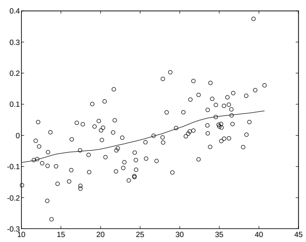
$$\begin{aligned}\hat{\boldsymbol{m}}_1 &= \big[I - (I - S_1^* S_2^*)^{-1}(I - S_1^*)\big]\mathbf{Y}\,,\\ \hat{\boldsymbol{m}}_2 &= [I - (I - S_2^* S_1^*)^{-1}(I - S_2^*)]\mathbf{Y}\,,\\ \hat{\boldsymbol{m}} &= \bar{\mathbf{Y}} + \hat{\boldsymbol{m}}_1 + \hat{\boldsymbol{m}}_2\,.\end{aligned} \tag{12}$$

For practical computation this form is better than the direct calculation of solution (11).

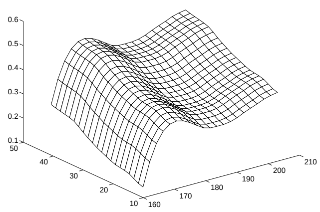
The following pictures demonstrate the shape of the estimator $\hat{\boldsymbol{m}}_1$, $\hat{\boldsymbol{m}}_2$ and $\hat{\boldsymbol{m}}$ for basketball data. The pictures in the first row were acquired using the Nadaraya–Watson estimator and in the second row by using the local linear estimator. The Epanechnikov kernel was used.



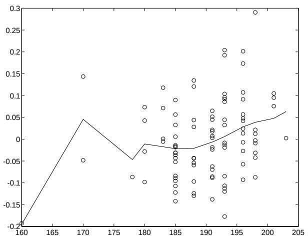
Estimation of m_1.



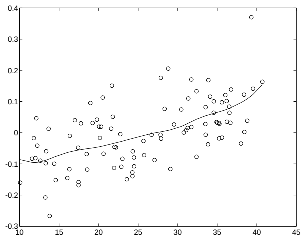
Estimation of m_2.



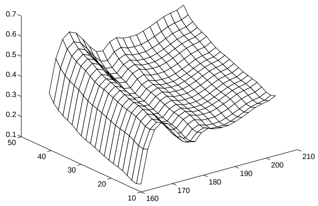
Estimation of m.



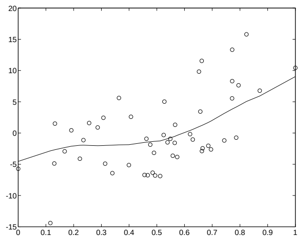
Estimation of m_1.



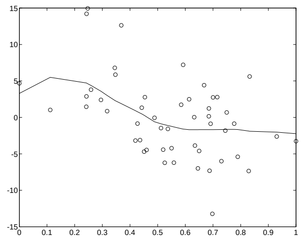
Estimation of m_2.



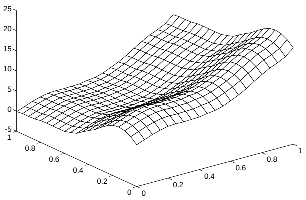
Estimation of m.



Estimation of m_1.



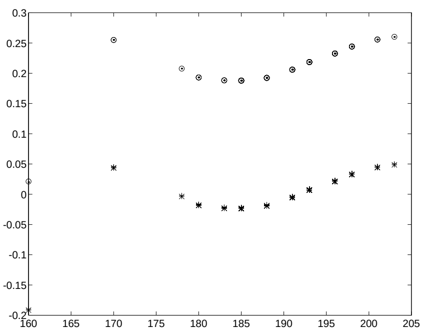
Estimation of m_2.



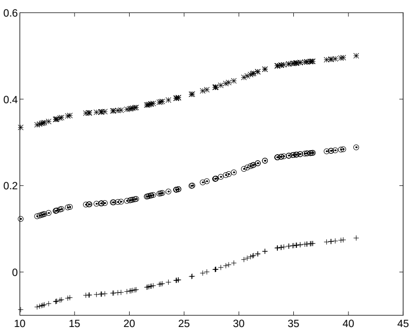
Estimation of m.

Other pictures show the results for the simulated data. The results presented were acquired using the Nadaraya–Watson estimator and the Epanechnikov kernel.

The last two pictures show the results of the various methods of computing the estimators $\hat{m}_1$ and $\hat{m}_2$ for the basketball data. It may be seen that the results may differ by a constant but their sum $\hat{\boldsymbol{m}}$ is the same.



Estimation of m_1.



Estimation of m_2.

$+$	– direct method	time : 3.080 s ,	flops : 14340735
$\circ$	– semi-iterative method	time : 9.620 s ,	flops : 32299256
$\cdot$	– using pseudoinverse matrix	time : 19.490 s ,	flops : 101728196
$*$	– back fitting algorithm	time : 0.050 s ,	flops : 149092

The above table gives the time period for evaluating and the number of float operations. For the computation the `MATLAB` system was used. The back fitting algorithm is very quick and insensitive to the starting iteration, but it does not give precise results. In three or four iteration it is near to the solution (relative error about 0.1 percent) but then alternates periodically. For this reason, the direct solution of (11) is the optimal method. It is rather fast and precise.

Conclusion

It is seen from the practical examples that the Nadaraya–Watson estimators are better than local linear estimators for the application of kernel smoothing to additive models. It is rather interesting because local linear estimators are better in one-dimensional kernel smoothing. Direct calculation using (12) seems to be the optimal method for computing in two dimensions. For higher dimensions it is possible to use (11).

REFERENCES

[1] BUJA, A.—HASTIE, T.—TIBSHIRANI, R.: *Linear smoothers and additive models*, Ann. Statist. **17** (1989), 453–509.

[2] EIERMANN, M.—MAREK, I.—NIETHAMMER, W.: *On the solution of singular linear systems of algebraic equations by semiiterative methods*, Numer. Math. **53** (1988), 265–283.

[3] HASTIE, T.—TIBSHIRANI, R.: *Generalized Additive Model*, Chapman & Hall, 1990.

[4] HOROVÁ, I.: *Boundary Kernels*, in: Proceedings of Summer School MATLAB '94, '95, Folia Fac. Sci. Natur. Univ. Masaryk Brun. Math. (Brno), 1997, pp. 17–24.

[5] MAREK, I.—SZYLD, D. B.: *Iterative and semiiterative methods for computing probability vectors of Markov operators*, Math. Comp. **61** (1993), 719–731.

[6] MICHÁLEK, J.: *Application of kernel smoothing to additive models*, Proceedings of Summer School MATLAB '94, '95, Fac. Sci. Natur. Univ. Masaryk Brun. Math. (Brno), 1997, pp. 103–111.

[7] MICHÁLEK, J.: *Jádrové odhady – základní vlastnosti a optimální výběr parametrů odhadu*, in: Proceedings ROBUST '94, JČMF, Prague, 1994, pp. 102–113.

[8] MÜLLER, H. G.: *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, Berlin, 1988.

[9] OPSOMER, J. D.—RUPPERT, D.: *Fitting a bivariate additive model by local polynomial regression*, Ann. Statist. **25** (1997), 186–211.

[10] WAND, M. P.—JONES, M. C.: *Kernel Smoothing*, Chapman & Hall, 1995.

Received May 4, 1998

Department of Applied Mathematics
Faculty of Science
Masaryk University Brno
Janáčkovo nám. 2a
CZ–662 95 Brno
CZECH REPUBLIC
E-mail: zelinka@math.muni.cz