Detection of Structural Changes in Regression

Marie Hušková and Jaromír Antoch

Charles University of Prague, Department of Statistics, Sokolovská 83,
CZ – 186 75 Praha 8 – Karlín, Czech Republic
marie.huskova@karlin.mff.cuni.cz, jaromir.antoch@karlin.mff.cuni.cz

Abstract: Some results on testing for changes in linear models are presented, approximations to the critical values based on permutation principle developed and their properties studied. Finally, selected conclusions of an application to both real and simulated data are presented.

2000 Mathematical Subject Class: 62G20, 62E20, 60F17

Keywords: Linear regression, structural changes, L_2-procedures, permutation principle, Monte Carlo, change-point problem.

Acknowledgement. The work was supported by grants GAČR 201/00/0769 and MSM 113200008.

1 Introduction

We consider the regression model with a change after an unknown time point m_n, i.e.

$$Y_{in} = \boldsymbol{x}_{in}^T\boldsymbol{\beta} + \boldsymbol{x}_{in}^T\boldsymbol{\delta}_n \cdot I\{i > m_n\} + e_i, \quad i = 1\ldots, n, \tag{1.1}$$

where $m_n\,(\leq n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\delta}_n = (\delta_{1n}, \ldots, \delta_{pn})^T \neq \mathbf{0}$ are unknown parameters, $\boldsymbol{x}_{in} = (x_{i1,n}, \ldots, x_{ip,n})^T$, $x_{i1,n} = 1$, $i = 1, \ldots, n$, are known design points and $e_1, \ldots, e_n$ are iid random errors fulfilling regularity conditions specified below. Function $I\{A\}$ denotes the indicator of the set A.

Model (1.1) describes the situation where the first m_n observations follow the linear model with the parameter $\boldsymbol{\beta}$ and the remaining $n - m_n$ observations follow the linear regression model with the parameter $\boldsymbol{\beta} + \boldsymbol{\delta}_n$. The parameter m_n is usually called the *change point*.

Such models are usually considered if one suspects that the regression parameters might change at an unknown time point, i.e., when one is not sure whether a change(s) has occurred or not. If the observations indicate a change, then an estimator of its location is of interest. Such problems occur in various situations, e.g., in hydrological, meteorological or econometric time series. For recent references see, e.g., Csörgő and Horváth (1997) and Antoch et al. (2002).

In the present paper we focus on the testing problem:

$$H_0 : m_n = n \qquad \text{against} \qquad H_1 : m_n < n. \tag{1.2}$$

We consider test procedures based the partial sums

$$\boldsymbol{S}_{kn} = \sum_{i=1}^{k} \boldsymbol{x}_{in}\big(Y_{in} - \boldsymbol{x}_{in}^T \widehat{\boldsymbol{\beta}}_n\big), \quad k = 1, \ldots, n, \tag{1.3}$$

$$S_{kn}^{*} = \sum_{i=1}^{k} \big(Y_{in} - \boldsymbol{x}_{in}^T \widehat{\boldsymbol{\beta}}_n\big), \qquad k = 1, \ldots, n, \tag{1.4}$$

where

$$\widehat{\boldsymbol{\beta}}_n = \boldsymbol{C}_{nn}^{-1} \sum_{i=1}^{n} \boldsymbol{x}_{in} Y_{in}, \quad \boldsymbol{C}_{kn} = \sum_{i=1}^{k} \boldsymbol{x}_{in}\boldsymbol{x}_{in}^T, \quad \boldsymbol{C}_{kn}^{o} = \boldsymbol{C}_{nn} - \boldsymbol{C}_{kn}, \quad k = 1, \ldots, n.$$

Both partial sums $\boldsymbol{S}_{kn}$ and S_{kn}^{*}, $k = 1, \ldots, n$, can be viewed as partial weighted sums of the L_2-type residuals

$$\widehat{e}_{in} = Y_{in} - \boldsymbol{x}_{in}^T \widehat{\boldsymbol{\beta}}_n, \quad i = 1, \ldots, n. \tag{1.5}$$

Procedures based on $\boldsymbol{S}_{kn}$, $k = 1, \ldots, n$, used for testing H_0 against H_1, are based on either of the following test statistics:

$$T_n = \max_{p<k<n-p} \Big\{ \widehat{\sigma}_n^{-2} \boldsymbol{S}_{kn}^T \boldsymbol{C}_{kn}^{-1} \boldsymbol{C}_{nn} \boldsymbol{C}_{kn}^{o-1} \boldsymbol{S}_{kn} \Big\}, \tag{1.6}$$

$$T_n(q) = \sup_{0<t<1} \Big\{ q^{-2}(t) \widehat{\sigma}_n^{-2} \boldsymbol{S}_{\lfloor (n+1)t \rfloor n}^T \boldsymbol{C}_{nn}^{-1} \boldsymbol{S}_{\lfloor (n+1)t \rfloor n} \Big\}, \tag{1.7}$$

where $\lfloor a \rfloor$ denotes the integer part of a, $q(\cdot)$ is a positive weight function and $\widehat{\sigma}_n^2$ is an estimator of σ^2 with the property

$$\widehat{\sigma}_n^2 - \sigma^2 = o_p\big((\log\log n)^{-1/2}\big) \quad \text{as} \quad n \to \infty. \tag{1.8}$$

While the test statistic T_n is related to the likelihood ratio test statistic when the errors have $N(0, \sigma^2)$ distribution, $T_n(q)$ is its natural modification. For more details see, e.g., Chapter 3 in Csörgő and Horváth (1997). Notice, moreover, that $\widehat{\boldsymbol{\beta}}_n$ is the least squares estimator of the vector parameter $\boldsymbol{\beta}$ in the model (1.1) with $m_n = n$, i.e. no change, and that

$$\widehat{\sigma}_n^2 = \frac{1}{n-p} \min_{p<k<n-p} \Big\{ \sum_{i=1}^{k} (Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k)^{\,2} + \sum_{i=k+1}^{n} (Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k^{0})^{\,2} \Big\}, \tag{1.9}$$

where $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\beta}}_k^{0}$ are the LSE based on $Y_1, \ldots, Y_k$ and $Y_1, \ldots, Y_k$, respectively, fulfills (1.8) in the case that there is at most one change (otherwise it should be properly modified).

It can be checked by direct, however tedious, calculation that

$$\widehat{\sigma}_n^{\,2} = \frac{1}{n-p}\Big\{ \sum_{i=1}^{n} \widehat{e}_{in}^{\,2} - \max_{p<k<n-p} \boldsymbol{S}_{kn}^T \boldsymbol{C}_{kn}^{-1} \boldsymbol{C}_{nn} \boldsymbol{C}_{kn}^{o-1} \boldsymbol{S}_{kn} \Big\} \tag{1.10}$$

and therefore T_n can be equivalently expressed as

$$T_n = \Big(Q_n^{-1} - \frac{1}{n-p} \Big)^{-1} \tag{1.11}$$

with

$$Q_n = \frac{\max_{p<k<n-p} \boldsymbol{S}_{kn}^T \boldsymbol{C}_{kn}^{-1} \boldsymbol{C}_{nn} \boldsymbol{C}_{kn}^{o-1} \boldsymbol{S}_{kn}}{\frac{1}{n-p} \sum_{i=1}^{n} \widehat{e}_{in}^{\,2}}. \tag{1.12}$$

The test procedures based on S_{kn}^*, $k = 1, \ldots, n$, are either of the form

$$T_n^* = \max_{1 \le k < n} \left\{ \sqrt{\frac{n}{k(n-k)}} \cdot \frac{|S_{kn}^*|}{\widehat{\sigma}_n} \right\} \quad \text{or} \quad T_n^*(q) = \sup_{0<t<1} \left\{ \frac{|S_{\lfloor (n+1)t \rfloor n}^*|}{\sqrt{n}\, q(t)\, \widehat{\sigma}_n} \right\}. \tag{1.13}$$

Notice, that the estimator (1.9) of σ^2 used in statistic (1.13) is quite complicated. However, those who prefer the computational simplicity to the efficiency may use instead

$$\widetilde{\sigma}_n^{\,2} = \frac{1}{n} \sum_{i=1}^{n} \widehat{e}_i^{\,2}.$$

Since large values of the above described test statistics indicate that the null hypothesis is violated, the corresponding critical regions have the form

$$T_n > c_n(\alpha), \quad T_n^* > c_n^*(\alpha), \quad T_n(q) > c_n(\alpha, q) \quad \text{and} \quad T_n^*(q) > c_n^*(\alpha, q), \tag{1.14}$$

where $c_n(\alpha)$, $c_n^*(\alpha)$, $c_n(\alpha, q)$ and $c_n^*(\alpha, q)$ are critical values corresponding to the level α. Approximations to these critical values can be obtained through the limit distribution of the respective test statistics under H_0, however, such approximations are usually not satisfactory. For details see, e.g., Csörgő and Horváth (1997) and Antoch et al. (2002). Therefore, in this paper we propose another possibility, namely the approximations based on the application of the permutational principle, of course, suitably modified for the situation of regression models.

Recall that the test statistic $T_n(q)$ with $q(t) = 1$, $t \in [0, 1]$ is quite often used in detection of changes in econometric models. More information about the recent development in the area of change point analysis in regression models can be found, e.g., in Horváth (1995), Csörgő and Horváth (1997), Bai and Perron (1999) and Hušková (1997, 2000).

In the following section we formulate the assumptions and remind the results on the limit distribution of the considered statistics under H_0. In Section 3 modified permutational tests are introduced and their limit properties investigated. Section 4 describes selected results of the application of our approach both to the real and simulated data. Finally, Section 5 summarizes crucial steps of the proofs.

2 Assumptions and limit behavior under H_0

In this section we formulate the assumptions and remind assertions on the limit distributions of the test statistics introduced in Section 1. Their proofs can be found, e.g., in Csörgő and Horváth (1997). Nevertheless, let us start with necessary assumptions.

• Concerning the design points $\boldsymbol{x}_{in} = \big(x_{i1,n}, \ldots, x_{ip,n}\big)^T,\ i = 1, \ldots, n$, we assume that they satisfy:

A.1. $x_{i1,n} = 1,\ i = 1, \ldots, n$, and $\sum_{i=1}^{n} x_{ij,n} = 0,\ j = 2, \ldots, p$.

A.2. There exists a positive definite $p \times p$ matrix $\boldsymbol{C}$ such that for any sequence $\big\{l_n\big\}$, $\lim_{n\to\infty} l_n = \infty$, $l_n \leq n$,

$$\left\| \frac{1}{l_n}\big(\boldsymbol{C}_{k+l_n n} - \boldsymbol{C}_{kn}\big) - \boldsymbol{C} \right\| = o\Big(\big(\log\, l_n\big)^{-1}\Big)$$

uniformly for $1 \leq k \leq n - l_n$; $\|\cdot\|$ denotes the Euclidean norm.

A.3. It holds, as $n \to \infty$, that

$$\max_{1\leq k\leq n} \left\{ \frac{1}{k} \sum_{i=1}^{k} \ \big\|\boldsymbol{x}_{in}\big\|^4 + \frac{1}{n-k} \sum_{i=k+1}^{n} \ \big\|\boldsymbol{x}_{in}\big\|^4 \right\} = O(1).$$

• Concerning the distribution of the error terms e_i's, following set of assumptions should be satisfied:

B.1. $e_1, e_2, \ldots$ are iid random variables with zero mean, nonzero variance σ^2 and finite moment $E|e_i|^{2+\Delta_1}$ with some $\Delta_1 > 0$.

• Finally, for the weight function $q(\cdot)$ it should hold:

C.1. $q(\cdot)$ is positive on $(0, 1)$, nondecreasing in a neighborhood of 0, nonincreasing in a neighborhood of 1, $\inf\{q(t); t \in (\eta, 1-\eta)\} > 0$ for all $\eta \in (0, 1/2)$ and for some $c > 0$

$$\int_0^1 \frac{1}{s(1-s)} \exp\Big\{ - \frac{cq^2(s)}{s(1-s)} \Big\}\, ds < \infty.$$

The assumptions imposed on the design points $\boldsymbol{x}_{in}$'s are slightly stronger then those considered in standard linear regression problems. However, the formulated assumptions are still fulfilled for a broad spectrum of situations.

Now, we state two theorems on limit behavior of the considered test statistics under the null hypothesis (corresponding to "no change").

Theorem 2.1. *Let assumptions A.1. – A.3. and B.1 be satisfied. Then*

$$\lim_{n\to\infty} P\Big(a\big(\log n\big)\sqrt{T_n} \le t + b_p\big(\log n\big)\Big) = \exp\big\{-2e^{-t}\big\}, \quad t \in \mathcal{R}_1, \tag{2.1}$$

$$\lim_{n\to\infty} P\Big(a\big(\log n\big)\, T_n^* \le t + b_1\big(\log n\big)\Big) = \exp\big\{-2e^{-t}\big\}, \quad t \in \mathcal{R}_1, \tag{2.2}$$

where

$$a(y) = \sqrt{2\log y},\ b_p(y) = 2\log y + \tfrac{p}{2}\log\log y - \log\big(\Gamma(\tfrac{p}{2})\big), \quad y > 1, \tag{2.3}$$

and $\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}\,dt$.

Proof: Can be found in Chapter 3 of Csörgő and Horváth (1997). □

Theorem 2.2. *Let assumptions A.1. – A.3. and B.1 be satisfied. Let the weight function* $q(\cdot)$ *satisfies assumption C.1. Then, as* $n \to \infty$,

$$\sqrt{T_n(q)} \xrightarrow{\mathcal{D}} \sup_{0<t<1}\left\{\frac{\sqrt{\sum_{i=1}^p B_i^2(t)}}{q(t)}\right\} \quad \text{and} \quad T_n^*(q) \xrightarrow{\mathcal{D}} \sup_{0<t<1}\left\{\frac{|B_1(t)|}{q(t)}\right\}, \tag{2.4}$$

where $\big\{B_j(t); t \in (0,1)\big\}_{j=1}^p$ *are independent Brownian bridges.*

Proof: Can be found in Chapter 3 of Csörgő and Horváth (1997). □

Remark 2.1. The assertions of both theorems remain true also for random design. Particularly, if $\boldsymbol{x}_{1n}, \dots, \boldsymbol{x}_{nn}$ are random vectors that do not depend on $e_1, \dots, e_n$ and that fulfill the assumptions (A.1) – (A.3) in probability.

3 Permutation test procedures

In the present section we apply a modified permutational principle in order to get permutational counterparts of the test statistics considered in the previous section. Then the limit behavior of these statistics is derived.

At first note that random errors $\big(e_1, \dots, e_n\big)$ have the same distribution as $\big(e_{R_1}, \dots, e_{R_n}\big)$, where $\boldsymbol{R} = \big(R_1, \dots, R_n\big)$ is a random permutation of $\big(1, \dots, n\big)$. The basic idea is that since we do not know $e_1, \dots, e_n$, we randomly permute their estimators $\widehat{e}_{1n}, \dots, \widehat{e}_{nn}$, where $\widehat{e}_{in}$ is the L_2-residual defined by (1.5), and apply them repeatedly when calculating statistics $\boldsymbol{S}_{kn}$ and S_{kn}^*.

More precisely, note that

$$\boldsymbol{S}_{kn} = \sum_{i=1}^k \boldsymbol{x}_{in}\widehat{e}_{in} - \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\sum_{j=1}^n \boldsymbol{x}_{jn}\widehat{e}_{jn}, \quad k = 1, \dots, n. \tag{3.1}$$

Applying the (modified) permutation principle, we have the permutational version of $\boldsymbol{S}_{kn}$ in the form

$$\boldsymbol{S}_{kn}(\boldsymbol{R}) = \sum_{i=1}^{k} \boldsymbol{x}_{in}\widehat{e}_{R_i n} - \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\sum_{j=1}^{n} \boldsymbol{x}_{jn}\widehat{e}_{R_j n}, \quad k = 1, \ldots, n. \tag{3.2}$$

Similarly, the permutational version of S_{kn}^{*} has the form

$$S_{kn}^{*}(\boldsymbol{R}) = \sum_{i=1}^{k} \widehat{e}_{R_i n} - \frac{k}{n}\sum_{j=1}^{n} \widehat{e}_{R_j n}, \quad k = 1, \ldots, n. \tag{3.3}$$

Permutational versions $T_n(\boldsymbol{R})$ and $T_n(q;\boldsymbol{R})$ of T_n and $T_n(q)$ are defined by (1.6) and (1.7) with $\boldsymbol{S}_{kn}$, $k = 1, \ldots, n$, and $\widehat{\sigma}_n$ replaced by $\boldsymbol{S}_{kn}(\boldsymbol{R})$, $k = 1, \ldots, n$, and $\widehat{\sigma}_n(\boldsymbol{R})$. Here $\widehat{\sigma}_n(\boldsymbol{R})$ is defined as

$$\widehat{\sigma}_n^{\,2}(\boldsymbol{R}) = \frac{1}{n-p}\Big\{ \sum_{i=1}^{n} \widehat{e}_{in}^{\,2} - \max_{p<k<n-p} \boldsymbol{S}_{kn}^{T}(\boldsymbol{R})\boldsymbol{C}_{kn}^{-1}\boldsymbol{C}_{nn}\boldsymbol{C}_{kn}^{o-1}\boldsymbol{S}_{kn}(\boldsymbol{R})\Big\}.$$

The permutational versions $T_n^{*}(\boldsymbol{R})$ and $T_n^{*}(q;\boldsymbol{R})$ are defined accordingly. In case of $T_n(\boldsymbol{R})$ one can use the relation (1.11) with $Q_n(\boldsymbol{R})$ defined accordingly.

Next, we study the conditional distribution of $T_n(\boldsymbol{R})$ given the original observations $\boldsymbol{Y}_n = \big(Y_{1n}, \ldots, Y_{nn}\big)^T$, i.e. we consider only the randomness generated by the random permutation $\boldsymbol{R} = (R_1, \ldots, R_n)^T$. Since the distribution of $\boldsymbol{R}$ is known, the conditional distribution of $T_n(\boldsymbol{R})$ given $\boldsymbol{Y}_n$ is known and can be calculated, which means to calculate $T_n(\boldsymbol{r})$ for all permutations $\boldsymbol{r}$ of $\{1, \ldots, n\}$. Since the number of possible permutations is $n!$, in reality one cannot calculate $T_n(\boldsymbol{r})$ for all permutations but only for very small part of them. However, with the current state of computers one can calculate a reasonable approximation even for moderate values of n. Particularly, one chooses independently and randomly permutations $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_B$, where B is large enough but still affordable with computers. Details are given in Section 4. With other permutational versions of statistics one can proceed quite analogously.

Finally, we derive the conditional limit behavior of $T_n(\boldsymbol{R})$ and $T_n^{*}(\boldsymbol{R})$ under quite general assumptions. Particularly, it is shown that their conditional limit distributions, given $\boldsymbol{Y}_n$ that follows model (1.1), is with no assumptions on m_n the same as that of T_n and T_n^{*} under H_0.

Theorem 3.1. *Let* $\big(Y_1, \boldsymbol{x}_{1n}^T\big), \ldots, \big(Y_n, \boldsymbol{x}_{nn}^T\big)$ *follow the model* (1.1) *and the assumptions A.1. – A.3. and B.1. be satisfied. Then*

$$\lim_{n\to\infty} P\Big(a\big(\log n\big)\sqrt{T_n(\boldsymbol{R})} \le t + b_p\big(\log n\big)\,\big|\,\boldsymbol{Y}_n\Big) = \exp\big\{-2e^{-t}\big\}, \quad t \in \mathcal{R}_1, \tag{3.4}$$

in probability, and

$$\lim_{n\to\infty} P\Big(a\big(\log n\big)\,T_n^{*}(\boldsymbol{R}) \le t + b_1\big(\log n\big)\,\big|\,\boldsymbol{Y}_n\Big) = \exp\big\{-2e^{-t}\big\}, \quad t \in \mathcal{R}_1, \tag{3.5}$$

in probability, where $a(y)$ *and* $b_p(y)$ *are given by* (2.3).

Concerning the permutational versions $T_n(q; \boldsymbol{R})$ and $T_n^*(q; \boldsymbol{R})$ related to $T_n(q)$ and $T_n^*(q)$, their limit distribution can be approximated as stated bellow.

Theorem 3.2. *Let assumptions of Theorem 3.1 be satisfied. Let the weight function $q(\cdot)$ satisfy assumption C.1. Then, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\sup_y \left| P\Big(\sqrt{T_n(q; \boldsymbol{R})} \le y \,\Big|\, \boldsymbol{Y}_n\Big) - P\Big(\sup_{0<t<1} \frac{\sqrt{\sum_{i=1}^p B_i^2(t)}}{q(t)} \le y\Big) \right| \ge \epsilon \right) = 0$$

in probability, and

$$\lim_{n\to\infty} P\left(\sup_y \left| P\Big(T_n^*(q; \boldsymbol{R}) \le y \,\Big|\, \boldsymbol{Y}_n\Big) - P\Big(\sup_{0<t<1} \frac{|B_1(t)|}{q(t)} \le y\Big) \right| \ge \epsilon \right) = 0$$

in probability, where $\big\{B_j(t); t \in (0,1)\big\}_{j=1}^p$ are independent Brownian bridges.

Remark 3.1. Proofs are postponed to the Section 5.

Remark 3.2. Notice that the assumptions of Theorems 3.1 and 3.2 cover both the null hypothesis (no change) and alternatives. Moreover, the limit conditional distributions of $T_n(\boldsymbol{R})$ does not depend on the original observations $\boldsymbol{Y} = \big(Y_{n1}, \ldots, Y_{nn}\big)^T$ and coincide with the limit distribution of T_n under the null hypothesis. This means that the $100(1-\alpha)\%$-quantile $d_n(1-\alpha, \boldsymbol{Y}_n)$ of the conditional distribution of $T_n(\boldsymbol{R})$ provides an approximation for the critical value for the test based on T_n. Therefore, the resulting test with approximate level α based on the permutational principle has the rejection region

$$T_n > d_n(1-\alpha, \boldsymbol{Y}_n). \tag{3.6}$$

The same also holds for other test statistics mentioned above.

4 Example

Large scale deforestation may cause the soil to lose its capability for water retention. The researchers of the Czech Research Institute for Forest Management studied the effect of controlled deforestation on the rainfalls-runoffs relationship, for details see Jarušková (1997). The objective of statistical inference was to decide whether this relationship changed during the study.

To simplify the model it is supposed that the relation between rainfalls and runoffs is linear. The problem is one of many change point problems in linear regression. One can either suppose that the change might occur in the intercept only or that it might occur in both parameters, i.e. in the intercept and/or the slope.

The test statistic T_n^*, cf (1.13), attains the value 8.22, and the test statistic T_n, cf (1.6), value 8.11. The asymptotic critical value, cf Theorem 2.1, is equal to 4.04, and therefore we can conclude that the null hypothesis claiming that the rainfalls-runoffs relationship is stationary rejected.

The least squares estimator $\widehat{m}_n$ of the change point m_n is equal to 26. Figure 1 shows the linear relationship between rainfalls and run-offs in the first as well in the second time period; i.e. $y = -0.194 + 0.800\,x$ and $y = -0.033 + 0.825\,x$. By $\star$ we denote the observations related to the first 26 years of observation, while by $\circ$ the observations from the last 10 years.

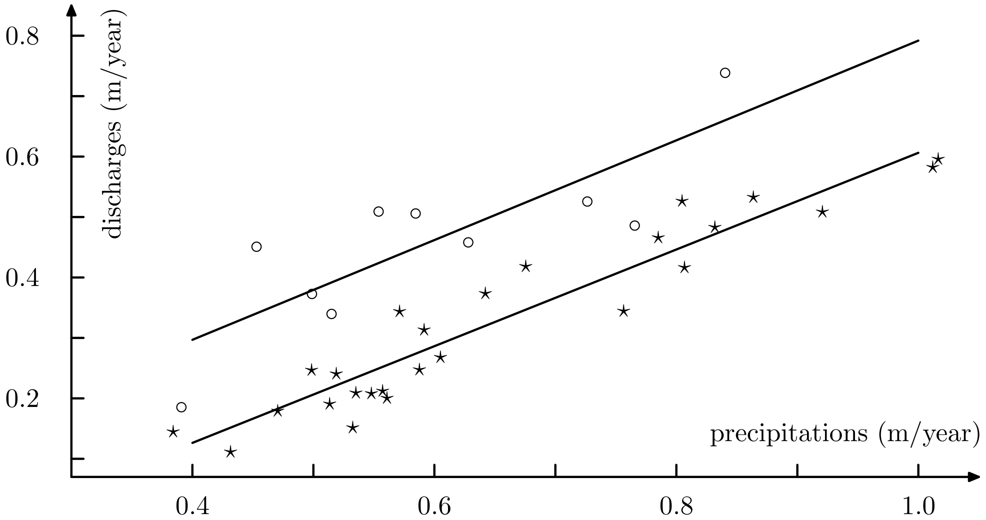

Figure 1. Malá Ráztoka: Data and model.

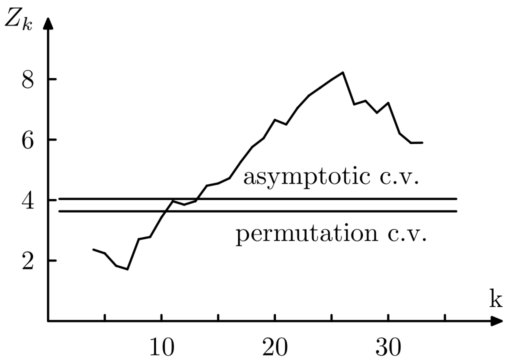

Figure 2a. Statistics Z_k.

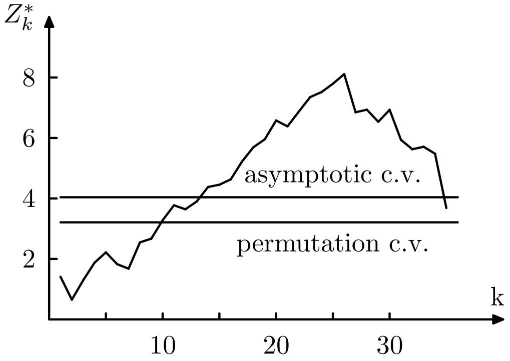

Figure 2b. Statistics Z_k^*.

Figure 2a shows the values of statistics

$$Z_k^* = \sqrt{\frac{n}{k(n-k)}} \cdot \frac{|S_{kn}^*|}{\widehat{\sigma}_n},$$

cf. (1.13) for details. Analogously, Figure 2b shows the values of statistics

$$Z_k = \widehat{\sigma}_n^{-2} \boldsymbol{S}_{kn}^T \boldsymbol{C}_{kn}^{-1} \boldsymbol{C}_{nn} \boldsymbol{C}_{kn}^{o-1} \boldsymbol{S}_{kn},$$

cf. (1.6) and (1.12) for details. In both Figures 2a and 2b the asymptotic and permutational critical values are plotted. According to the expectation, the permutational critical values are smaller than the asymptotic ones.

To understand better the sensitivity and usefulness of the permutational principle, we prepared following semi-simulated study. At first we decided to keep the same model as for the real Ráztoka data (RRD). More precisely:

C.1. As the base model we used two straight lines with the slopes equal to the estimated ones in the real data set, i.e.,

$$y_1 = a_1 + 0.800x + \epsilon_1 \quad \text{and} \quad y_2 = a_2 + 0.825x + \epsilon_2.$$

The values x were simulated from the uniform distribution on $[0.4, 1.0]$. The "error terms" were simulated from the normal distribution $\mathcal{N}(0, 0.0456^2)$ and $\mathcal{N}(0, 0.0799^2)$, where the values 0.0456^2 and 0.0799^2 correspond to the estimated variability of the two least squares fits from the RRD.

C.2. We "played" with the number of observations prior the "deforestification" and after it. Denote the lengths of these periods n_1 and n_2.

C.3. Changing a_1 and a_2 we in the base model we see how small shift we are able to detect using our approach.

Figures 3 and 4 show the situation for $n_1 = 75$ and $n_2 = 25$. analogously as above, Figure 3 presents the data and Figure 4 the values of corresponding statistics Z_k and Z_k^*. Notice, that the values a_1 and a_2 were chosen in such a way that the test statistics T_n^* and T_n practically attain the 5% asymptotic critical value. The permutation critical values based on 100 000 permutations are also shown.

Finally, we shrinked both base further as shown in Figure 5. Here again $n_1 = 75$ and $n_2 = 25$. However, the values a_1 and a_2 were chosen in such a way that the test statistic T_n^* attains the 5% permutational critical value based on 100 000 permutations. The results are summarized in Figures 5 and 6.

Notice, that while for the real data the shift between both baseline straight lines is equal to 170.6mm, in the situation shown in Figure 3 it is equal 40 mm and, finally, in the situation shown in Figure 5 only 30.6 mm.

Aside the situation presented above we prepared several others ones combining different types of changes, different sample sizes etc. Based on our experience with various procedures for detection changes, we would suggest to use the permutational principle whenever applicable because one can distinguish much more subtle changes.

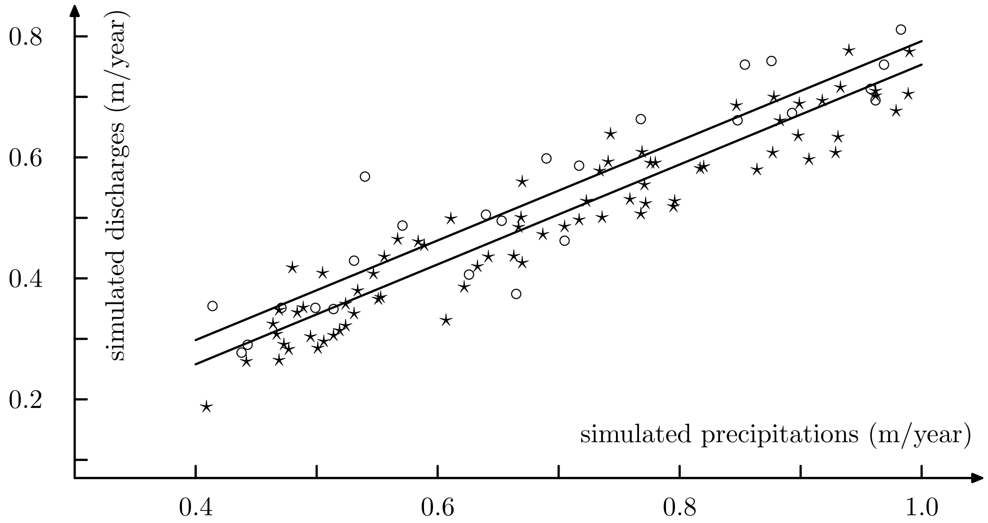

Figure 3. Simulated data and model.

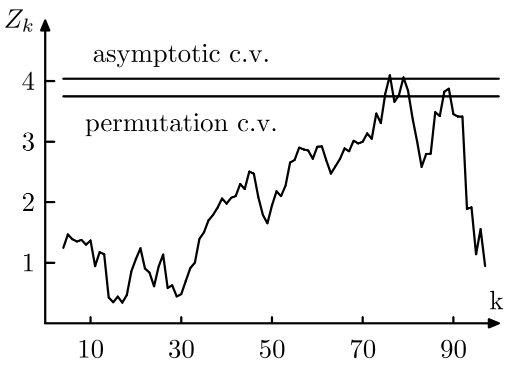

Figure 4a. Statistics Z_k.

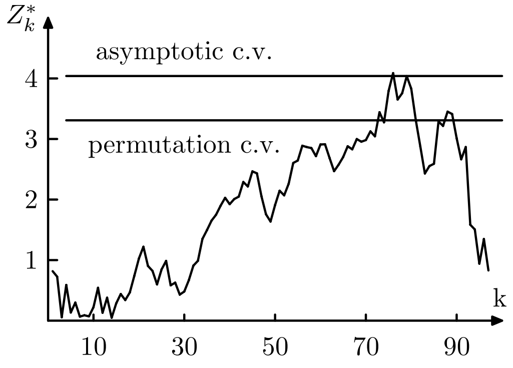

Figure 4b. Statistics Z_k^*.

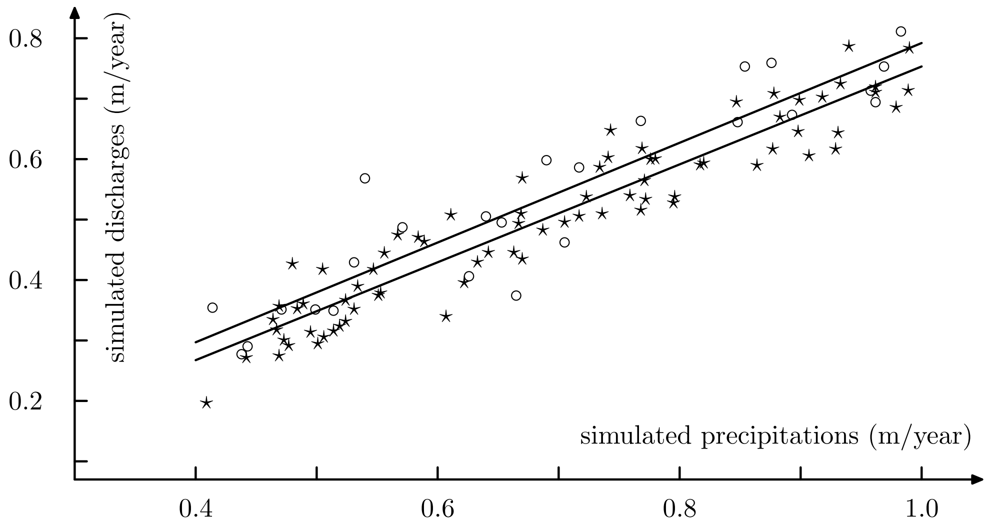

Figure 5. Simulated data and model.

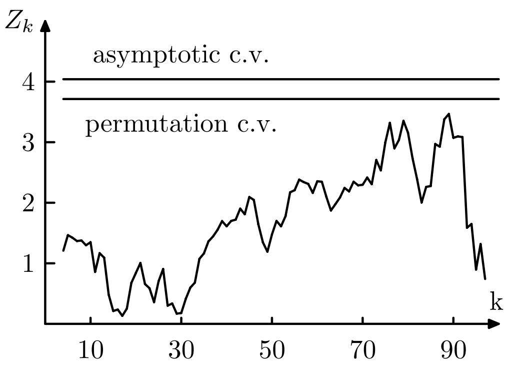

Figure 6a. Statistics Z_k.

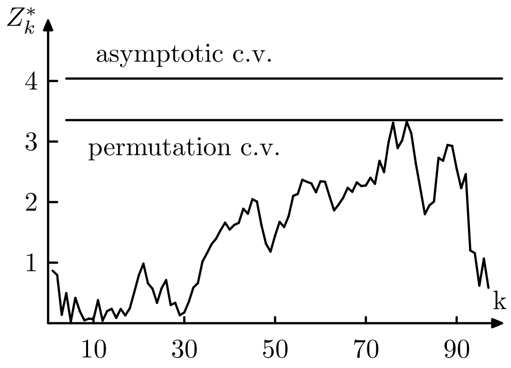

Figure 6b. Statistics Z_k^*.

Our final remark concerns the comparison of the test statistics T_n and T_n^*. Based on our experience with both real and simulated data, we can conclude that in the case of a change in the shift parameter only both statistics give more or less the same results. Therefore, taking into account the computational simplicity of T_n^*, we would prefer it to the statistic T_n. On the other hand, for more complicated situations as, e.g., the change in several parameters, the test statistic T_n seems to be more appropriate despite its calculational complexity.

5 Proofs

The crucial step of the proofs of Theorems 3.1 − 3.2 relies on the fact that given $\boldsymbol{Y}_n$ the partial sums $\boldsymbol{S}_{kn}(\boldsymbol{R})$ and $S_{kn}^*(\boldsymbol{R})$, $k = 1, \ldots, n$, can be viewed as vectors of linear rank statistics and therefore the proofs reduce to treating functionals of linear rank statistics.

More precisely, we use the results on approximations of functionals of the rank statistics

$$V_{kn} = \sum_{i=1}^{k} c_{in} a_n(R_i), \quad k = 1, \ldots, n, \tag{5.1}$$

by functionals of weighted sums of independent random variables

$$Z_{kn} = \sum_{i=1}^{k} c_{in}\Big(a_n\big(\lfloor nU_i \rfloor + 1\big) - \overline{a}_n\big(\boldsymbol{U}\big)\Big), \quad k = 1, \ldots, n, \tag{5.2}$$

where $\boldsymbol{U} = \big(U_1, \ldots, U_n\big)^T$ is a sample of the size n from an uniform distribution on $(0,1)$, $\boldsymbol{R} = \big(R_1, \ldots, R_n\big)^T$ are corresponding ranks and

$$\overline{a}_n(\boldsymbol{U}) = \frac{1}{n} \sum_{i=1}^{n} a_n\big(\lfloor nU_i \rfloor + 1\big). \tag{5.3}$$

The following theorem plays the key role in the proofs of Theorems 3.1 − 3.2.

Theorem 5.1. *Let $\boldsymbol{U} = \big(U_1, \ldots, U_n\big)^T$ be a sample of size n from uniform distribution on $(0,1)$ and let $\boldsymbol{R} = \big(R_1, \ldots, R_n\big)^T$ be the corresponding ranks. Let scores $a_n(i),\ i = 1, \ldots, n$, satisfy*

$$\sum_{i=1}^{n} a_n(i) = 0, \quad \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_n^2(i) > 0, \quad \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} |a_n(i)|^{2+\Delta_2} < \infty \quad (5.4)$$

with some $\Delta_2 > 0$. Let constants c_{in}'s satisfy

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} c_{in}^2 < \infty. \qquad (5.5)$$

Then, as $n \to \infty$,

$$\max_{1<k<n} \Big\{\sqrt{\tfrac{n}{k(n-k)}}\,\big|\,V_{kn} - Z_{kn}\,\big|\,\Big\} = O_P\Big(n^{-\nu_1}\Big) \qquad (5.6)$$

with some $\nu_1 > 0$ and, moreover, for $0 < \alpha < 1/2$,

$$\max_{1<k<n} \left\{\left|\,\frac{1}{\sqrt{n}}\Big(\frac{k(n-k)}{n^2}\Big)^{-1/2+\alpha}\big(V_{kn} - Z_{kn}\big)\right|\right\} = O_P\Big(n^{-\nu_2}\Big) \qquad (5.7)$$

with some $\nu_2 > 0$.

Proof: See Hušková (2002). □

Proof of Theorem 3.1: We apply Theorem 5.1 with

$$a_n(i) = \widehat{e}_{in}, \quad i = 1, \ldots, n, \quad \text{and} \quad c_{ij} = x_{ij,n}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p, \quad (5.8)$$

where $x_{ij,n}$ is the j-th component of the vector $\boldsymbol{x}_{in}$. Clearly, since the assumptions A.1.−A.3. the regression constants $x_{ij,n}$'s have properties requested in Theorem 3.1. Concerning the assumptions on the scores, clearly $\sum_{i=1}^{n} \widehat{e}_{in} = 0$. Finally, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_{in}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(e_i - \boldsymbol{x}_{in}^T \boldsymbol{C}_{nn}^{-1}\sum_{j=1}^{n} \boldsymbol{x}_{jn} e_j + \boldsymbol{x}_{in}^T \boldsymbol{\delta}_n \cdot I\{i > m\} - \boldsymbol{x}_{in}^T \boldsymbol{C}_{nn}^{-1}\boldsymbol{C}_{mn}^{o}\boldsymbol{\delta}_n\right)^2 \\ &= A_{n1} + A_{n2} - 2A_{n3} + A_{n4} + 2A_{n5},\end{aligned}$$

where

$$A_{n1} = \frac{1}{n}\sum_{i=1}^{n} e_i^2, \quad A_{n2} = A_{n3} = \frac{1}{n}\Big(\sum_{i=1}^{n} \boldsymbol{x}_{in} e_i\Big)^T \boldsymbol{C}_{nn}^{-1}\Big(\sum_{i=1}^{n} \boldsymbol{x}_{in} e_i\Big),$$

$$A_{n4} = \frac{1}{n}\boldsymbol{\delta}_n^T \boldsymbol{C}_{mn}\boldsymbol{C}_{nn}^{-1}\boldsymbol{C}_{mn}^{o}\boldsymbol{\delta}_n \quad \text{and}$$

$$A_{n5} = \frac{1}{n}\sum_{i=1}^{n}\left(e_i - \boldsymbol{x}_{in}^T\boldsymbol{C}_{nn}^{-1}\sum_{j=1}^{n}\boldsymbol{x}_{jn}e_j\right)\left(\boldsymbol{x}_{in}^T\boldsymbol{\delta}_n \cdot I\{i > m\} - \boldsymbol{x}_{in}^T\boldsymbol{C}_{nn}^{-1}\boldsymbol{C}_{mn}^o\boldsymbol{\delta}_n\right).$$

Standard tools give, as $n \to \infty$,

$$A_{n1} \xrightarrow{\mathcal{P}} \sigma^2, \quad A_{n2} = A_{n3} = O_P\big(n^{-1}\big), \quad A_{n4} = \frac{(n-m)m}{n^2}\,\boldsymbol{\delta}_n^T\boldsymbol{C}\boldsymbol{\delta}_n\big(1 + o(1)\big)$$

and

$$A_{n5} = O_P\big(||\boldsymbol{\delta}_n||n^{-1/2}\big),$$

which immediately implies that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_{in}^{\,2} = \sigma^2 + \frac{(n-m)m}{n^2}\,\boldsymbol{\delta}_n^T\boldsymbol{C}\boldsymbol{\delta}_n + o_p(1).$$

Therefore, the first part of the assumptions (5.4) is satisfied. The remaining part is an easy consequence of the following inequality, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\big|\,\widehat{e}_{in}\,\big|^{2+\Delta} \le D\left(\frac{1}{n}\sum_{i=1}^{n}\big|\,e_i\,\big|^{2+\Delta} + ||\boldsymbol{\delta}_n||^{2+\Delta}\right) = O_P(1),$$

with some positive constants D and $\Delta \le \min(\Delta_1,\, \Delta_2)$.

Hence, Theorem 5.1 can be applied in our situation and we receive, after few standard steps, that for given $\boldsymbol{Y}_n$

$$\big|\big|\,T_n(\boldsymbol{R}) - T_n(\boldsymbol{U})\,\big|\big| = O_P\big(n^{-\nu_1}\big),$$

where

$$T_n(\boldsymbol{U}) = \max_{p<k<n-p}\ \Big\{\widehat{\sigma}_n^{\,-2}\big(\boldsymbol{S}_{kn}(\boldsymbol{U})\big)^T\boldsymbol{C}_{kn}^{-1}\boldsymbol{C}_{nn}\boldsymbol{C}_{kn}^{o-1}\boldsymbol{S}_{kn}(\boldsymbol{U})\Big\},$$

$$\boldsymbol{S}_{kn}(\boldsymbol{U}) = \boldsymbol{V}_{kn}(\boldsymbol{U}) - \boldsymbol{W}_{kn}(\boldsymbol{U}),$$

$$\boldsymbol{V}_{kn}(\boldsymbol{U}) = \sum_{i=1}^{k}\boldsymbol{x}_{in}\widehat{e}_{(\lfloor nU_i\rfloor+1)n} - \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\sum_{j=1}^{n}\boldsymbol{x}_{jn}\widehat{e}_{(\lfloor nU_j\rfloor+1)n},$$

$$\boldsymbol{W}_{kn}(\boldsymbol{U}) = \left(\sum_{i=1}^{k}\boldsymbol{x}_{in} - \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\sum_{i=1}^{n}\boldsymbol{x}_{in}\right)\overline{e}_n(\boldsymbol{U}), \quad k = 1,\ldots,n,$$

and

$$\overline{e}_n(\boldsymbol{U}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_{(\lfloor nU_i\rfloor+1)n}.$$

Notice that due to the assumption A.1. $\big(\sum_{i=1}^{k}\boldsymbol{x}_{in}$ is the first column of $\boldsymbol{C}_{kn}$ and $\sum_{i=1}^{k}\boldsymbol{x}_{in}^T$ is the first row of $\boldsymbol{C}_{kn}\big)$, we realize that

$$\sum_{i=1}^{k} \boldsymbol{x}_{in} - \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\sum_{i=1}^{n} \boldsymbol{x}_{in} = \boldsymbol{0}, \quad k = 1, \ldots, n,$$

and hence

$$\boldsymbol{W}_{kn}(\boldsymbol{U}) = \boldsymbol{0}.$$

Moreover, since

$$E\left(\widehat{e}_{(\lfloor nU_i \rfloor+1)n} \,\middle|\, \boldsymbol{Y}_n\right) = \frac{1}{n}\sum_{i=1}^{k} \widehat{e}_{in} = 0,$$

we find that $T_n(\boldsymbol{U})$ can be rewritten in the form

$$T_n(\boldsymbol{U}) = \max_{p<k<n-p} \left\{ \widehat{\sigma}_n^{-2} \boldsymbol{V}_{kn}(\boldsymbol{U})^T \boldsymbol{C}_{kn}^{-1} \boldsymbol{C}_{nn} \boldsymbol{C}_{kn}^{o-1} \boldsymbol{V}_{kn}(\boldsymbol{U}) \right\},$$

where, conditionally on $\boldsymbol{Y}_n$, $\boldsymbol{V}_{kn}$'s are sums of independent random vectors with zero means and the variance matrix

$$var\left\{ \boldsymbol{V}_{kn}(\boldsymbol{U}) \,\middle|\, \boldsymbol{Y}_n \right\} = \boldsymbol{C}_{kn}\boldsymbol{C}_{nn}^{-1}\boldsymbol{C}_{kn}^{o}\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_{in}^{2}.$$

Now, the assertions (3.4) and (3.5) follows from Theorem 3.1.5 in Csörgő and Horváth (1997).

□

Proof of Theorem 3.2: We proceed analogously as in the proof of Theorem 3.1, but we apply Lemma 3.1.6 from Csörgő and Horváth (1997) and results of Chapter 4 in Csörgő and Horváth (1993). □

References

[1] Antoch J. and Hušková M. (2001). *M-estimators of structural changes in regression models.* Tatra Mt. Math. Publ. **22**, 1 – 12.

[2] Antoch J. and Hušková M. (2001). *Permutation tests for change point analysis.* Statist. Probab. Letters **53**, 37 – 46.

[3] Antoch J., Hušková M. and Jarušková D. (2002). *Off-line quality control.* In: Multivariate Total Quality Control: Foundations and Recent Advances, N. C. Lauro et al. eds., Springer-Verlag, Heidelberg, 1 – 86.

[4] Bai J. and Perron P. (1999). *Estimating and testing linear models with multiple structural changes.* J. of Econometrics **33**, 299 – 323.

[5] Csörgő M. and Horváth L. (1993). *Weighted Approximations in Probability and Statistics.* J. Wiley, New York.

[6] Csörgő M. and Horváth L. (1997). *Limit Theorems in Change-Point Analysis.* J. Wiley, New York.

[7] Horváth L. (1995). *Detecting changes in linear regression.* Statistics **26**, 189 – 208.

[8] Hušková M. (1997). *L_1-test procedures for detection of change.* In: L_1-Statistical Procedures and Related Topics, Y. Dodge ed., IMS Lecture Notes – Monograph Series **31**, 56 – 70.

[9] Hušková M. (2000). *Some invariant test procedures for detection of structural changes.* Kybernetika **36**, 401 – 414.

[10] Hušková M. (2002). *Limit theorems for linear rank statistics.* In preparation.

[11] Hušková M. and Antoch J. (2001), M-estimators of structural changes in regression models. Tatra Mt. Math. Publ. **22**, 1 – 12.

[12] Hušková M. and Slabý A. (2001). *Permutation tests for mutiple changes.* Kybernetika **37**, 3605 – 622.

[13] Jarušková D. (1996). *Change point detection in meteorological measurement.* Monthly Weather Review **124**, 1535 – 1543.

[14] Jarušková D. (1997). *Some problems with application of change point detection methods to environmental data.* Environmetrics **8**, 469 – 483.

[15] Jarušková D. (1998). *Change-point detection for dependent data and application to hydrology.* Istatistik. Journal of the Turkish Statistical Association **1**, 9 – 21.