*Change-point Problems*
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

# NONPARAMETRIC METHODS IN CHANGE-POINT PROBLEMS: A GENERAL APPROACH AND SOME CONCRETE ALGORITHMS

BY BORIS S. DARKHOVSKI
*Russian Academy of Sciences*

A general approach to change-point problems is proposed. This approach is based upon two ideas. The first idea is that any change-point problem can be reduced to the problem of detection of changes in the mean value of some new sequences. The second idea is that the nonparametric family of Kolmogorov-Smirnov type statistics can be used for change-point detection in these sequences. This general approach is implemented in two cases: (a) the problem of gradual change-point detection, and (b) change-point detection in two-phase regression model.

**1. Introduction.** Change-point problems attract considerable interest nowadays. Much was done in the field of parametric change-point estimation (see Shaban (1980) and Krishnaiah and Miao (1988) for bibliography).

In this report we are dealing with nonparametric change-point detection methods. These methods do not use a priori information about data and are the most useful for applications (see Csörgő and Horváth (1988) for a review).

A general approach to change-point problems is proposed. This approach is based upon two ideas. The first idea is that any change-point problem can be reduced to the problem of detection of changes in the mean value of some new sequences. The second idea is that the nonparametric family of Kolmogorov-Smirnov type statistics can be used for change-point detection in these sequences.

This general approach is implemented in two cases: (a) the problem of gradual change-point detection, (b) change-point detection in two-phase regression model. For more details see Brodski and Darkhovski (1993).

**2. Main Ideas.** Since we are dealing with detection of different changes in probabilistic characteristics of random sequences, we need a general formulation of this problem. Any property of a random sequence is determined by

---

AMS 1991 Subject Classification: Primary 62G05; Secondary 62M10.

Key words and phrases: Dependence, estimation, gradual change, Kolmogorov-Smirnov, mixing, optimality, two-phase regression.

its distribution. Therefore a change in characteristics is, in general, a change in distribution. It follows from here that a sequence with change-points is in fact "glued" from pieces of sequences with different distributions.

In the general case, any distribution of a random sequence is determined by a set of finite-dimensional distributions satisfying the concordance conditions. Therefore any change in probabilistic characteristics is a result of a certain change in arbitrary finite-dimensional distributions. We conclude from here that the following model of a single change-point is realistic enough.

Suppose that $X = \{x_n\}_{n=1}^{n=\infty}$, $Y = \{y_n\}_{n=1}^{n=\infty}$ are two strictly stationary random sequences. Let $1 \equiv \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_k$, $k < \infty$ and

$$F'_{\tau_0,\tau_1,\cdots,\tau_k}(u_0, u_1, \cdots, u_k) \doteq F'(\cdot) = P\{x_{\tau_0} \le u_0, x_{\tau_1} \le u_1, \cdots, x_{\tau_k} \le u_k\}$$

$$F''_{\tau_0,\tau_1,\cdots,\tau_k}(u_0, u_1, \cdots, u_k) \doteq F''(\cdot) = P\{y_{\tau_0} \le u_0, y_{\tau_1} \le u_1, \cdots, y_{\tau_k} \le u_k\}.$$

Suppose that $F'(\cdot) \neq F''(\cdot)$ and consider a sequence $Z = \{z_n\}_{n=1}^{n=\infty}$ defined in the following way

$$z_n = \begin{cases} x_n, & n \le n_1 \\ v_n, & n_1 < n < n_2 \\ y_n, & n_2 \le n \end{cases}$$

where $\{v_n\}$ is an arbitrary random sequence. We say that $Z$ is a *sequence with a change-point* (with a change-point in a $(k+1)$-dimensional distribution function). Here $n_1$ and $n_2$ are the moments of the beginning and the end of a change, and an interval $(n_1, n_2)$ is the length of a transition process. By analogy one can create a model of more than one change (for this, several stationary sequences have to be "glued").

This model can be used both for aposteriori detection of change-point moments (i.e., by all information obtained) and for sequential detection (i.e., on-line with observations), but we concentrate now on aposteriori detection.

The multitude of change-point problems which arise here is as large as the set of probabilistic characteristics of a stationary random sequence. Full description of such a sequence can be done only with the help of all finite dimensional distributions. Therefore, for the purpose of theoretical analysis and for effective practical application, it is necessary to extract one basic situation from this multitude of problems to which other situations can be easily reduced.

Let us formulate this basic problem and the method of such reduction; it will be the first idea of our approach.

We say that a sequence $Z$ with a change-point is considered in the *basic situation* if a change in mathematical expectation of $Z$ has occurred, i.e.,

$$Ez_n = \begin{cases} a', & n \le m_1 \\ a_n, & m_1 < n < m_2 \\ a'', & m_2 \le n \end{cases}$$

where $a' \neq a''$, $\{a_n\}$ is an arbitrary sequence.

An algorithm that detects change of this type is called the *basic one.* Generally speaking, in the basic situation other characteristics of a random sequence might change (not only the expectation), but we shall consider them here as nuisance parameters.

Now we show that detection of changes in arbitrary distribution functions (in the model described above) can be reduced to detection of changes in the basic situation.

Suppose that the functions $F'(\cdot)$ and $F''(\cdot)$ are such that $\|F' - F''\| \ge 2\epsilon$, where $\epsilon > 0$, $\|\cdot\|$ is the sup-norm. Consider the vector $\xi_t = (z_{t+\tau_0}, z_{t+\tau_1}, \cdots, z_{t+\tau_k})$, $t \ge 0$. Then the distribution function of the vector $\xi_t$ is $F'(\cdot)$ before and $F''(\cdot)$ after the change-point. Consider a partition of $\mathbb{R}^{k+1}$ into $r < \infty$ nonintersecting subsets $\{A_j\}$, $j = 1, \cdots, r$, and the random vector $\eta_t$ such that $\eta_t = \sum_{j=1}^{r} a_j\, I(\xi_t \in A_j)$, $a_j \in \mathbb{R}^{k+1}$ where $I(\cdot)$ = indicator function.

It is well known that for any $\epsilon > 0$ subsets $\{A_j\}$ and vectors $\{a_j\}$ can be found such that $\|F'(\cdot) - F_{\eta_t}\| \le \epsilon/2$. In the same way the function $F''$ can be approximated, and without loss of generality we can assume that the approximating $\{A_j\}$ and $\{a_j\}$ are the same as for $F'$. Then for $t_1 \le n_1 - \tau_k \doteq m_1$ and $t_2 \ge n_2 - \tau_0 = n_2 - 1 \doteq m_2$ we have

$$\|F_{\eta_{t_1}} - F_{\eta_{t_2}}\| \ge \|F'(\cdot) - F''(\cdot)\| - \|F_{\eta_{t_1}} - F_{\xi_{t_1}}\| - \|F_{\eta_{t_2}} - F_{\xi_{t_2}}\| \ge \epsilon$$

i.e., distributions of the vector $\eta$ (with a finite number of values) before and after the change-point differ no less than $\epsilon$. It means that if we introduce $r$ new sequences $\{v_t^j\}$, $1 \le j \le r$, $v_t^j = I(\xi_t \in A_j)$ then at least one of these sequences has a change in the mean value no less than $\epsilon$. Therefore if the basic algorithm can detect $\epsilon$-changes of the mean value, then one can detect $2\epsilon$-changes in arbitrary distributions of the initial sequence $Z$.

This approach enables us to limit ourselves to the analysis of the basic algorithm. Two situations might appear in the solution of practical change-point problems. If it is apriori known that a certain distribution function of a random sequence changes, then a sufficient number of these indicator sequences must be constructed and the basic algorithm has to be applied to

each of them. If there is no apriori information on what distribution function changes, then we shall use the following recurrent selection procedure: first, we check whether the one-dimensional distribution function changes, then, the same for two-dimensional distributions, etc. This procedure has to be stopped if apriori requirements on quality of detection are satisfied.

The approach proposed makes it possible to involve any information on the character of changes and to apply the same algorithm of detection. Let us know that some probabilistic characteristic $U_t$ of a random sequence $Z$ changes. There exists a function $\phi(z_t, z_{t+\tau_1}, \cdots, z_{t+\tau_k})$ such that $U_t = E\,\phi(z_t, z_{t+\tau_1}, \cdots, z_{t+\tau_k})$. If we consider a new sequence $v_t = \phi(z_t, z_{t+\tau_1}, \cdots, z_{t+\tau_k})$ then this sequence has a change in the mean. For example, $U_t = Ez_t z_{t+\tau}$. Then $v_t = z_t z_{t+\tau}$. We can detect such a change with the help of the basic algorithm.

Now we consider the general statement of the basic aposteriori change-problem. We shall put this problem as the problem of vector parameter estimation in a scheme of series.

Let $\nu = (\nu_1, \nu_2, \cdots, \nu_k)$, $k \geq 1$, be an unknown vector parameter such that $0 \equiv \nu_0 < \alpha \leq \nu_1 \leq \nu_2 \leq \cdots \leq \nu_k \leq \beta < \nu_{k+1} \equiv 1$. Here $\alpha, \beta$ are apriori constants, $\alpha < 0.5 < \beta$.

Let $\phi_\nu(t)$, $t \in [0,1]$ be a parametric family of determined functions. On a probability space $(\Omega, F, P_\nu)$ consider two families of random sequences:

$$\{X^N\}, X^N = \{x^N(n)\}_{n=1}^N, \quad N > [\alpha^{-1}] \vee [(1-\beta)^{-1}]$$

and

$$\{\xi^{(i)}\}, \quad i \in J \doteq \{1, 2, \cdots, k+1\}$$

$$\xi^{(i)} = \{\xi^{(i)}(n)\}_{n=1}^\infty, \quad E_\nu \xi^{(i)}(n) \equiv 0.$$

Suppose that if $[\nu_{i-1}N] < n \leq [\nu_i N]$, $i \in J$, then

$$x^N(n) = \phi_\nu(n/N) + \xi^{(i)}(n).$$

The problem is to estimate the change-point moments $\tau_i = [\nu_i N]$, $i = 1, \cdots, k$ for the random sequence $X^N$. The character of the disorder is determined by the function $\phi_\nu(t)$. Our goal is to construct a consistent estimate of the vector parameter $\nu$ (and consequently of moments $\tau_i$) by the realization $X^N$.

The second idea of our approach consists of using the family of Kolmogorov-Smirnov type statistics for solving this problem.

We mean the family of the form

$$Y_N(n) = [(1 - n/N)(n/N)]^{\delta} \left[ \frac{1}{n} \sum_{k=1}^{n} x^N(k) - \frac{1}{N-n} \sum_{k=n+1}^{N} x^N(k) \right], \quad (1)$$
$$0 \le \delta \le 1, \quad 1 \le n \le N-1$$

and some derivatives of such statistics. Why is it useful? It turns out that such statistics are asymptotically optimal in the following sense.

We have proved (see Darkhovski (1989)) that under some regularity conditions the inequality holds (for the simplest case $\nu \in \mathbb{R}^1$):

$$\lim_{N\to\infty} N^{-1} \ln\{\inf_{\hat{\nu}} \sup_{\alpha\le\nu\le\beta} P_\nu(|\hat{\nu} - \nu| \ge \epsilon)\} \ge -2\epsilon \min(\rho(P_1, P_2), \rho(P_2, P_1))$$

where $\rho(\cdot,\cdot)$ = Kullback-Leibler distance between distributions before and after the disorder, and infimum is sought on the set of all estimates. (Analogous inequalities take place in the general case.) But on the other hand, for statistics of type (1), it can be shown that for the estimator $\hat{\nu}$ based on these statistics the inequality occurs:

$$P_\nu\{|\hat{\nu} - \nu| \ge \epsilon\} \le L_1 \exp(-L_2(\epsilon)N),$$

if the following assumptions hold:

a) Cramér condition,

b) $\phi$-mixing condition.

(Analogous estimates were proved by Carlstein (1988) for the so-called mean dominant norm statistics. Our family of statistics belongs to this class.) So we see that the statistics of type (1) give asymptotically optimal estimates in order.

**3. Concrete Algorithms.** Let us consider a gradual disorder case. In this case, $k = 2$, $\nu = (\nu_1, \nu_2)$,

$$\phi_\nu(t) = \begin{cases} 0, & \text{if } 0 \le t \le \nu_1 \\ g(t), & \text{if } \nu_1 \le t \le \nu_2 \\ a \ne 0, & \text{if } \nu_2 \le t \le 1 \end{cases}$$

where $g(t)$ is a Lipschitz function that is sandwiched between two monotone increasing functions.

Consider the following statistic

$$Y_N(n,m) = n^{-1}\sum_{k=1}^{n} x^N(k) - (N-n-m)^{-1}\sum_{k=n+m+1}^{N} x^N(k),$$
$$1 \le n,\ 1 \le m,\ n+m \le N-1.$$

For a description of the method of estimation it is convenient to consider continuous time. A continuous random field $y_N(t,s)$ is built on the set $\{(t,s): 0 \le t \le 1,\ 0 \le s \le 1,\ t+s \le 1\}$. The values of $y_N$ in nodes of the grid $(n/N, m/N)$ are equal to the values of the statistic $Y_N(n,m)$. The values of $y_N$ in other points are built by linear interpolation. The field $y_N(t,s)$ will be considered on the set $M = \{(t,s): \alpha \le t \le \beta,\ 0 \le s \le \beta-\alpha-\delta,\ \delta > 0,\ t+s \le \beta\}$.

Now we introduce some notation. Suppose that $M$ is a compact in $\mathbb{R}^2$, $h(t,s) \in C(M)$, $R = \max_{(t,s)} h(t,s)$. Let $S = \mathrm{Pr}_s M$ ($\mathrm{Pr}_s$ is an orthogonal projection operator). For every $s \in S$ define

$$T(s) = \{t : (t,s) \in M\}.$$

In our problem the set $S = [0, \beta-\alpha-\delta]$ and the set $T(s)$, $s \in S$, has the form $T(s) = [\alpha, \beta - s]$.

For every $s \in S$, $\mathcal{X} \ge 0$, define the set of $\mathcal{X}$-maximums of the function $h(\cdot, s)$ on $T(s)$:

$$A_{\mathcal{X}}(h;s) = \{\tilde{t} \in T(s) : R - \mathcal{X} \le h(\tilde{t}, s)\}.$$

On the set $S$ define the function

$$d_{\mathcal{X}}(h;s) = \mathrm{diam}\ A_{\mathcal{X}}(h;s)$$

where diam $A$ is the diameter of the set $A$, i.e., the value $\sup_{(x,y)\in A} \|x-y\|$ (by definition, diam $\emptyset = 0$).

For every $\lambda \ge 0$ we define the set of $\lambda$-minimums of the function $d_{\mathcal{X}}(h;s)$ on $S$:

$$U(\lambda, \mathcal{X}; h) = \{\tilde{s} \in S : d_{\mathcal{X}}(h;\tilde{s}) \le \inf_{s\in S} d_{\mathcal{X}}(h;s) + \lambda\}.$$

On the space of continuous functions $C(M)$ we define the families (by parameters $\mathcal{X}, \lambda$) of functionals:

$$\Phi_{\lambda,\mathcal{X}}(h) = \sup\{s \in S : s \in U(\lambda, \mathcal{X}; h)\}$$
$$\Psi_{\lambda,\mathcal{X}}(h) = \max\{t : t \in A_{\mathcal{X}}(h; \Phi_{\lambda,\mathcal{X}}(h)\}, \text{ if } A_{\mathcal{X}}(\cdot) \ne \emptyset$$
$$\Psi_{\lambda,\mathcal{X}}(h) = \max\{t : t \in T(\Phi_{\lambda,\mathcal{X}}(h))\}, \text{ if } A_{\mathcal{X}}(\cdot) = \emptyset.$$

The following values are to be the estimates $\hat{\nu}(N)$, $\hat{\nu}(N)$ of parameters $\nu_1$, $\nu_2$:

$$\hat{\nu}_1(N) = \Psi_{\lambda,\mathcal{X}}(|y_N(t,s)|)$$
$$\hat{\nu}_2(N) = \hat{\nu}_1(N) + \Phi_{\lambda,\mathcal{X}}(|y_N(t,s)|).$$

Thus, to construct the estimates $\hat{\nu}_1(N)$, $\hat{\nu}_2(N)$, the following is required:

a) for a certain $\mathcal{X} > 0$, on the segment $s \in [0, \beta - \alpha - \delta]$ to compute the function $d_{\mathcal{X}}(|y_N|; s)$ – the diameter of the set of $\mathcal{X}$-maximums of the function $|y_N(\cdot; s)|$ on $T(s)$;

b) for a certain $\lambda > 0$, on $[0, \beta - \alpha - \delta]$ to build the set of $\lambda$-minimums of the function $d_{\mathcal{X}}(|y_N|; s)$ and to find the supremum of the set – the value $\Phi_{\lambda,\mathcal{X}}(|y_N|)$;

c) to find the maximum of the set $A_{\mathcal{X}}(|y_N|; \Phi_{\lambda,\mathcal{X}}(|y_N|))$ of $\mathcal{X}$-maximums of the function $|y_N(\cdot; \Phi_{\lambda,\mathcal{X}}(|y_N|))|$ on the set $T(\Phi_{\lambda,\mathcal{X}}(|y_N|))$ and to assume this value to be the estimate $\hat{\nu}_1(N)$. Then the estimate $\hat{\nu}_2(N)$ is defined by the previous formula.

For every $\epsilon > 0$ let

$$G_\epsilon = \{\tilde{\nu} : \|\tilde{\nu} - \nu\| \leq \epsilon\}.$$

THEOREM. *Suppose that the following conditions are fulfilled:*

*1)* $\sup E_\nu \exp t\xi^{(i)}(n) < \infty$, $i = 1, 2, 3$, $|t| \leq H$.

*2)* $\phi$*-mixing condition for* $\xi = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$.

*Then for all small enough* $\epsilon > 0$ *there exist* $\lambda(\epsilon) > 0$, $\mathcal{X}(\epsilon) > 0$, *such that the estimate* $\hat{\nu}_{\lambda(\epsilon),\mathcal{X}(\epsilon)}(N)$ *converges to the set* $G_\epsilon$ $P_\nu$*-almost surely.*

Now let us settle on another problem: the change-point problem for regression models. Let us consider the following situation. Let, in the general scheme, the function $\phi_\nu(t) \in C^k[0,1]$, $k = 0, 1, \cdots$. We shall say that the *functional change-point* is taking place if $\phi_\nu^{(k+1)}(\cdot)$ is continuous everywhere except the point $t = \nu$ and at this point it has the first type discontinuity and there exists the left and the right value of this derivative. This definition is the natural generalization of the case when $\phi_\nu(t) = a$, $t \leq \nu$, $\phi_\nu(t) = b$, $t > \nu$, $a \neq b$, i.e., the simple abrupt disorder problem.

A simple example of this situation is the change-point in the linear regression model when $\phi_\nu(\cdot)$ has the form:

$$\phi_\nu(t) = \begin{cases} \alpha t + \beta, & t \leq \nu \\ \gamma t + (\alpha - \gamma)\nu + \beta, & t \geq \nu \end{cases} \qquad \alpha \neq \gamma.$$

Now we formulate one condition under which the functional change-point will be considered. Put

$$r_{k+1} = \int_0^1 \phi_\nu^{(k+1)}(s)ds.$$

CONDITION (i). *The function* $\phi_\nu^{(k+1)}(\cdot) - r_{k+1}$ *changes its sign on the segment* $[0,1]$ *at the single point* $t = \nu$.

It is easy to check that this condition is fulfilled in linear regression in particular.

For a description of the change-point detection method, fix $\epsilon$, $0 < \epsilon < 1$, and let us introduce the following notation:

$$\Delta^{(1)}x^N(n) = \epsilon^{-1}\{x^N(n + [(\epsilon N/2)]) - x^N(n - [(\epsilon N/2)])\}$$
$$\Delta^{(k)}x^N(n) = \epsilon^{-1}\{\Delta^{(k-1)}x^N(n + [(\epsilon N/2)]) - \Delta^{(k-1)}x^N(n - p\epsilon N/2)])\}, \quad k = 2,3,\cdots.$$

The decision statistic is of the form ($N_1 = N - (k+1)[(\epsilon N/2)]$, $n_1 = 1 + (k+1)[(\epsilon N)/2]$, $S = N_1 - n_1 + 1$):

$$Y_N(n) = S^{-2}(n - n_1 + 1)(S - n + n_1 - 1)\Big((n - n_1 + 1)^{-1}\sum_{i=n_1}^{n}\Delta^{(k+1)}x^N(i) - (N_1 - n)^{-1}\sum_{i=n+1}^{N_1}\Delta^{(k+1)}x^N(i)\Big), \quad n = n_1, n_1+1, \cdots, N_1 - 1.$$

The estimate $\hat{n}$ of the change-point is an arbitrary point of the set

$$\arg \max_{n_1 \le n \le N_1 - 1} |Y_N(n)|$$

and the estimate of $\nu$ is $\hat{\nu} = \hat{n}/S$.

THEOREM. *Suppose the conditions of the previous Theorem and Condition (i) are fulfilled. Then for every* $\epsilon > 0$, *the estimate* $\hat{\nu}$ *converges to the* $\epsilon$*-neighborhood of* $\nu$ $P_\nu$*-a.s.*

There are some generalizations of this problem, namely when we have several change-points and when more complicated functions are considered. These problems will be considered in other papers.

**Acknowledgement.** The author is grateful to the referee for useful comments.

## REFERENCES

BRODSKI, B. and DARKHOVSKI, B. (1993). *Nonparametric Methods in Change-Point Problems.* Kluwer Academic Publishers: The Netherlands.

CARLSTEIN, E. (1988). Nonparametric change-point estimation. *Ann. Statist.* **16**, 188–197.

CSÖRGŐ, M. and HORVÁTH, L. (1988). Nonparametric methods for change-point problems. In *Handbook of Statistics*, Vol. 7 (P. Krishnaiah and C. R. Rao, eds.), 403–425. Elsevier: The Netherlands.

DARKHOVSKI, B. (1989). Nonparametric methods in change-point problems. In *Statistics and Control of Stochastic Processes* (A. N. Shiryaev, ed.), 57–70. Nauka: Moscow.

KRISHNAIAH, P. and MIAO, B. (1988). Review about estimation of change points. In *Handbook of Statistics*, Vol. 7 (P. Krishnaiah and C. R. Rao, eds.), 375–402. Elsevier: The Netherlands.

SHABAN, S. (1980). Change-point problem and two-phase regression: An annotated bibliography. *Internat. Statist. Rev.* **48**, 83–93.

INSTITUTE FOR SYSTEMS ANALYSIS
RUSSIAN ACADEMY OF SCIENCES
9 PROSPEKT 60-LETYA OKTIABRIA
117312 MOSCOW B-312, RUSSIA