# CHANGE POINT PROBLEM

## *Introduction*

In a number applications one has to analyze data obtained during a longer time period indicating that possible statistical models might change once or several times during the observational period.

The problem is to decide whether the model is not changing (null hypothesis) or whether the model changes once or more times (alternative hypotheses). Eventually, the problem to divide the time ordered data into segments in such a way that in each *segment* shows a kind of stacionarity or homogenity or, in other words, the data in these segments can be described by relatively simple statistical models.

*Possible applications*

(i) *meteorology, climatology, hydrology or environmental studies* : certain characterics (temperature, content of water impurity, water discharges) are recorded, except a random fluctuation they might indicate some systematic changes or heterogenity. The question: do human activities (urbanization and deforestation) cause some changes in the studied characteristics (global warming)?

*Possible applications*

(i) *meteorology, climatology, hydrology or environmental studies* : certain characterics (temperature, content of water impurity, water discharges) are recorded, except a random fluctuation they might indicate some systematic changes or heterogenity. The question: do human activities (urbanization and deforestation) cause some changes in the studied characteristics (global warming)?

(ii) *Econometrics time series* It typically reduces to identification of structural breaks; e.g., to identify the time point when the dependence of gross domestic products on some explanatory variables has changed, breaks in financial time series (stock indices, share prices, foreign exchange rates).

(iii) *Statistical quality control in industry* some quantitative or qualitative characteristics of products might changed; the problem is to identify time of the change or, in other words, to find proper segment(s). Often early detection of a change is desired.
(iv) Similar type of data (or time series) can found in *biology, medicine, etc.*, .e.g. reaction on medical treatment

MODELS

(1) *Change in distribution*

Observations $Y_1, \ldots, Y_n$ obtained in time ordered points $t_1 < \cdots < t_n$ are independent, $X_i$ has the distribution $F_i$, $i = 1, \ldots n$ and the testing problem:

$$H_0 : F_1 = \ldots = F_n, \qquad (.1)$$

$$H_1 : \text{there exists } m < n \text{ such that} \qquad (.2)$$

$$F_1 = \ldots = F_m \neq F_{m+1} = \ldots = F_n.$$

$m$ unknown parameter; for $m$ known it reduces to two-sample problem.

(2) *Change in location*

Observations $Y_1, \ldots, Y_n$ obtained in time ordered points $t_1 < \cdots < t_n$

$$Y_i = \mu + \delta I\{i > m\} + e_i,\ i = 1, \ldots, n, \qquad (.3)$$

where $1 \leq m \leq n, \mu, \delta \neq 0$ are unknown parameters, $I\{A\}$ denotes the indicator of a set $A$, the distribution of the error terms $e_i$'s satisfies:

(A.1) $e_1, \ldots, e_n$ are i.i.d. random variables with $Ee_i = 0$, $0 < var\, e_i = \sigma^2$ and $E|e_i|^\nu < \infty$ with some $\nu > 2$.

(2) *Change in location*

Observations $Y_1, \ldots, Y_n$ obtained in time ordered points $t_1 < \cdots < t_n$

$$Y_i = \mu + \delta I\{i > m\} + e_i,\ i = 1, \ldots, n, \qquad (.3)$$

where $1 \leq m \leq n, \mu, \delta \neq 0$ are unknown parameters, $I\{A\}$ denotes the indicator of a set $A$, the distribution of the error terms $e_i$'s satisfies:

(A.1) $e_1, \ldots, e_n$ are i.i.d. random variables with $Ee_i = 0$, $0 < var\, e_i = \sigma^2$ and $E|e_i|^\nu < \infty$ with some $\nu > 2$.

The testing problem

$$H_0 : m = n \quad \text{against} \quad H_1 : m < n. \qquad (.4)$$

and estimation of $m$. Variants: some the parameters that can be known: $\mu$, $\delta$, $\sigma$ and the distribution of $e_i$'s.

(3) *Change in linear regression*

Observations $Y_1, \ldots, Y_n$ obtained in time ordered points $t_1 < \cdots < t_n$

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + I\{i > m\} \boldsymbol{x}_i^T \boldsymbol{\delta} + e_i,$$

$e_1, \ldots, e_n$ are i.i.d. satisfying (A.1)

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ ... regression constants, design points

The testing problem

$$H_0 : m = n \quad \text{against} \quad H_1 : m < n.$$

and estimation of $m$.

(4) *Change in regression*

Observations $Y_1, \ldots, Y_n$ obtained in time ordered points $t_1 < \cdots < t_n$

$$Y_i = r_j(t_i, x_i, \theta_j) + \sigma_j(t_i, x_i, \theta_j)e_i$$

$$m_{j-1} < i \leq m_j, \quad j = 1, \ldots, q,$$

$0 = m_0 < m_1 < \cdots < m_q = n$ unknown parameters (usually called change points)

$\theta_j,\ j = 1, \ldots, q$ unknown parameters

$x_i,\ i = 1, \ldots, n$ known regression constants

$r_j(.)$ regression functions

$\sigma_j(.)$ variance functions

$e_1, \ldots, e_n$ random errors

Typical testing problem: continuity of regression function

## *Terminology and methods*

$m$ is called *change point*

*Terminology and methods*

$m$ is called *change point*

---

change point problem
disorder problem
structural breaks
switching regime, etc.

---

*Terminology and methods*

$m$ is called *change point*

---

change point problem
disorder problem
structural breaks
switching regime, etc.

---

Many variants and generalizations: dependent observations, change in variance,....

---

*Terminology and methods*

$m$ is called *change point*

---

change point problem
disorder problem
structural breaks
switching regime, etc.

---

Many variants and generalizations: dependent observations, change in variance,....

---

*Possible approaches* to construction of test statistics and estimators
parametric & nonparametric
Bayesian & nonBayesian& pseudoBayesian (only $m$ is

*Parametric & nonparametric*
(i) likelihood ratio test and its modifications
(ii) robust ($M$-tests)
(iii) nonparametric (rank statistics, empirical distribution functions, $U$-statistics

---

*Parametric & nonparametric*
(i) likelihood ratio test and its modifications
(ii) robust ($M$-tests)
(iii) nonparametric (rank statistics, empirical distribution functions, $U$-statistics

---

*Bayesian & nonBayesian*
choice of prior is a problem
inference made on posterior distribution
sometimes only $m$ assumed to be random-

*Likelihood ratio principle*

Assume that $Y_1, \ldots, Y_n$ are independent r.v.'s

$Y_i$ has density $f(x; \boldsymbol{\theta}_i),\ i = 1, \ldots, n$

$H_0 : \boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_n$

$H_1 : \boldsymbol{\theta}_1 = \ldots = \boldsymbol{\theta}_m \neq \boldsymbol{\theta}_{m+1} = \ldots = \boldsymbol{\theta}_n$

$m, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$-parameters

$\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$

likelihood ratio : $\dfrac{\text{density under} H_1}{\text{density under} H_0}$

the unknown parameters are replaced by maxlikelihood estimators under respective assumptions

*Normal distribution case*

We concentrate on the change in location, assuming that $Y_1, \ldots, Y_n$ have normal distribution with *known variance* $\sigma^2$.

$$\max_{1 \le k < n} \Lambda_{k,n}$$

$$\Lambda_{k,n} = \max_{\mu,\delta} \left( 2\pi\sigma^2 \right)^{-n/2} \exp\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{k} (Y_i - \mu)^2 \}$$

$$\times \exp\{ -\frac{1}{2\sigma^2} \sum_{i=k+1}^{n} (Y_i - \mu - \delta)^2 \} \Big)$$

$$\times \Big( \max_{\mu} \Big( (2\pi\sigma^2)^{-n/2} \exp\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \} \Big) \Big)^{-1}$$

After some calculation we get

$$T_{n1}(\mu) = 2\log\left(\max_{1\le k<n}\left(\Lambda_{k,n}\right)\right)$$

$$T_{n1}(\mu) = \max_{1<k<n}\left\{\sqrt{\frac{n}{k(n-k)}}\frac{1}{\sigma}\left|\sum_{i=1}^{k}(Y_i-\overline{Y}_n)\right|\right\}^2, \tag{.5}$$

where $\overline{Y}_n = (\sum_{i=1}^{n} Y_i)/n$.

After some calculation we get

$$T_{n1}(\mu) = 2\log\left(\max_{1\le k<n}\left(\Lambda_{k,n}\right)\right)$$

$$T_{n1}(\mu) = \max_{1<k<n}\left\{\sqrt{\frac{n}{k(n-k)}}\frac{1}{\sigma}\Big|\sum_{i=1}^{k}(Y_i-\overline{Y}_n)\Big|\right\}^2, \quad (.5)$$

where $\overline{Y}_n = (\sum_{i=1}^{n} Y_i)/n$.

$$T_{n1}(\mu) = \max_{1<k<n}\left\{\sqrt{\frac{k(n-k)}{n}}\frac{1}{\sigma}|\overline{Y}_k - \overline{Y}_k^o|\right\}^2$$

Modifications:

$$T_{n2}(\mu,\beta) = \sup_{0<t<1} \frac{\left|\sum_{i=1}^{\lfloor (n+1)t \rfloor}(Y_i - \overline{Y}_n)/\sqrt{n}\right|}{(t(1-t))^{\beta}\sigma}, \quad (.6)$$

$\beta \in \langle 0, 1/2)$, $\lfloor a \rfloor$ denotes the integer part of $a$, often $\beta = 0$.

$$T_{n3}(\mu,\gamma) = \int_0^1 \frac{\left|\sum_{i=1}^{\lfloor (n+1)t \rfloor}(Y_i - \overline{Y}_n)/\sqrt{n}\right|}{(t(1-t))^{\gamma}\sigma} dt, \quad (.7)$$

$\gamma < 3/2???$- Bayesian like procedure
all test statistics are functionals of

$$\sum_{i=1}^{k}(Y_i - \overline{Y}_n), \quad k = 1, \ldots, n$$

*case of* $\sigma^2$ *unknown*, normally distributed random variables

$$T_{n1,\sigma}(\mu) = n \log \left( \max_{1<k<n} \frac{\widehat{\sigma}_k^2 k/n + \widehat{\sigma}_k^{o2}(n-k)/n}{\widehat{\sigma}_n^2} \right) \tag{.8}$$

where

$$\overline{Y}_k = (\sum_{i=1}^{k} Y_i)/k, \quad \overline{Y}_k^o = (\sum_{i=k+1}^{n} Y_i)/(n-k)$$

$$\widehat{\sigma}_k^2 = \frac{1}{k}\sum_{i=1}^{k}(Y_i - \overline{Y}_k)^2, \quad \widehat{\sigma}_k^{o2} = \frac{1}{n-k}\sum_{i=k+1}^{n}(Y_i - \overline{Y}_k^o)^2$$

test statistic for change in mean and/or variance, normally distributed random variables

$$T_{n1}(\mu,\sigma) = 2\log\left(\max_{1<k<n} \frac{\left(\widehat{\sigma}_k\right)^k \left(\widehat{\sigma}_k^o\right)^{n-k}}{\left(\widehat{\sigma}_n\right)^n}\right) \qquad (.9)$$

*Some particular cases:*
change in mean (location) $\mu, \delta, \sigma$ known,$m$ unknown, normally distributed observations

$$\max_{1<k<n} \frac{\delta}{\sigma^2} \sum_{i=k+1}^{n} (Y_i - \mu - \delta/2) \qquad (.10)$$

*Critical regions*

Large values of the test statistic $T_{n,1}(T_{n,2}(\beta)....)$ indicate $H_o$ does not hold;

Critical region corresp.to level $\alpha$:

$$T_{n,1} \geq d_n(\alpha)$$

where $d_n(\alpha)$ is determined in a way that the test has level $\alpha$, only approximations available

asymptotic distribution and resampling methods.

Both possibilities will be discussed.

Behavior under alternatives

We focus on $T_{n,1}(\mu), T_{n,2}(\mu, \beta), T_{n,3}(\mu, \beta)$

## *Test statistics for change in location*

$$T_{n1}(\mu) = \max_{1<k<n} \left\{ \sqrt{\frac{n}{k(n-k)}} \frac{1}{\sigma} \left| \sum_{i=1}^{k} (Y_i - \overline{Y}_n) \right| \right\}^2,$$

$$T_{n2}(\mu, \beta) = \sup_{0<t<1} \frac{\left| \sum_{i=1}^{\lfloor (n+1)t \rfloor} (Y_i - \overline{Y}_n)/\sqrt{n} \right|}{(t(1-t))^{\beta}\sigma},$$

$$T_{n3}(\mu, \gamma) = \int_0^1 \frac{\left| \sum_{i=1}^{\lfloor (n+1)t \rfloor} (Y_i - \overline{Y}_n)/\sqrt{n} \right|}{(t(1-t))^{\gamma}\sigma} dt,$$

Large values of the test statistics indicate that $H_o$ is not true. Critical regions with level $\alpha$ are of the form:

$$T_{n1}(\mu) \geq t_{n,1}(\alpha)$$
$$T_{n2}(\mu, \beta) \geq t_{n,2}(\alpha)$$
$$T_{n3}(\mu, \gamma) \geq t_{n,3}(\alpha)$$

Problem: approximations for $t_{n,1}(\alpha), t_{n,2}(\alpha), t_{n,3}(\alpha)$

Possibilities:

(i) limit distribution under $H_o$

(ii)resampling

(iii) Bonferroni inequality ($n$ small)

(iii) *Bonferroni inequality* ($n$ small)
$(H_o, H_{1k}), k = 1, \ldots, n-1$, $H_{1k}$- change after $k$th observation, $H_1 = \cup_k H_{1k}$
$(H_o, H_{1k})$ is a two-sample problem, test statistics known
For $(H_o, H_1)$ the null hypothesis is rejected if rejected for at least one of the problems $(H_o, H_{1k})$,
if the test for $(H_o, H_{1k})$ has level $\alpha_k, k = 1, \ldots$ then for $(H_o, H_1)$ the test has level $\leq \sum_{k=1}^{n-1} \alpha_k$.
Bonferroni: $A_1, \ldots, A_m$ — events on the same probability space

$$P(\cup_{k=1}^{n-1} A_k) \leq \sum_{k=1}^{n-1} P(A_k)$$

$A_k$– rejection region for $(H_o, H_{1k})$

*Limit behavior under* $H_o$

We assume location model with $\nu = 4$ and with $H_o$ (all holds true even for $\nu > 2$). Normality is not needed. Notice that the processes

$$V_{n,1}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor (n+1)t \rfloor} \frac{(Y_i - \mu)}{\sigma}, \quad t \in (0,1),$$

converge in distribution to a Wiener process $\{W(t), t \in (0,1)\}$

*Limit behavior under $H_o$*

We assume location model with $\nu = 4$ and with $H_o$ (all holds true even for $\nu > 2$). Normality is not needed. Notice that the processes

$$V_{n,1}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor (n+1)t \rfloor} \frac{(Y_i - \mu)}{\sigma}, \quad t \in (0,1),$$

converge in distribution to a Wiener process $\{W(t), t \in (0,1)\}$

and the processes

$$V_{n,2}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor (n+1)t \rfloor} \frac{(Y_i - \overline{Y}_n)}{\sigma}, \quad t \in (0,1),$$

converge in distribution to the Brownian bridge $\{B(t), t \in$

*Theorem 1* (Darling, Erdős , 1953) Under $H_0$ for all $t$

$$\lim_{n\to\infty} P_{H_0}\Big(a(\log n)\, T_{n1}(\mu) \leq t + b_1(\log n)\Big) = \exp\{-2e^{-t}\}, \tag{.11}$$

where

$$a(t) = \sqrt{2\log t}, \quad \log t > 0, \tag{.12}$$

$$b_1(t) = 2\log t + \frac{1}{2}\log\log t - \frac{1}{2}\log(\pi), \quad \log t > 0. \tag{.13}$$

*Theorem 2* Under $H_0$ for all $x$

$$\lim_{n\to\infty} P_{H_0}\Big( T_{n2}(\mu,\beta) \le x\Big) = P\Big( \sup_{0<t<1} \frac{|B(t)|}{(t(1-t))^{\beta}} \le x\Big), \tag{.14}$$

and

$$\lim_{n\to\infty} P_{H_0}\Big( T_{n3}(\mu,\gamma) \le x\Big) \tag{.15}$$

$$= P\Big( \int_{0<t<1} \frac{|B(t)|}{(t(1-t))^{\gamma}} dt \le x\Big),$$

where $\beta \in \langle 0, 1/2)$, $\gamma < 3/2???$, $\{B(t); t \in (0,1)\}$ is a Brownian bridge.

*REFERENCES*

[1] Antoch J. and Hušková M. (1999). *Estimators of changes*. In: Asymptotics, Nonparametrics and Time Series, Subir Ghosh ed., Marcel Dekker, New York, 533 – 578.

[2] Antoch J., Hušková M. and Jarušková D. (2001). *Off-line quality control*. In: Multivariate Total Quality Control: Foundation and Recent Advances, Lauro N. C. et al. eds., Springer-Verlag, Heidelberg, 1 – 86.

[3] Brodsky B.E. and Darkhovsky B.S. (2000). *Non-Parametric Statistical Diagnosis; Problems and Methods*. Kluwer Academic Publishers, Dordrecht.

[4] Bai J. and Perron P. (1998). *Estimating and testing linear models with multiple structural changes*. Econometrica **66**, 47 – 78.

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