Tests for multiple changes in linear regression models

Miriam Marušiaková, Marie Hušková

Department of Statistics
Charles University in Prague

ROBUST 2010, Králíky, 2.2. 2010

Outline

Introduction

Change point analysis

- Important topic in statistical and econometric research
  - Interesting theoretical problems
  - Applications in many fields
- Data sample - the model might change during the observational period

Problems

1. Test whether change(s) occured or not
2. Estimate the number of changes and their locations

Hypothesis testing

Test procedures

- Max-type test statistics (CUSUM procedures)
- MOSUM type tests
- Sum-type or Bayesian type statistics
- F-type tests (Bai and Perron, 1998), M-type tests

Approximations to critical values

- Limit null distribution
- Residual bootstrap (Antoch and Hušková, 2001)

Permutation principle

- Test statistic - expressible through the partial sums of residuals; under H_0 - partial sums of errors
- If e_i i.i.d. $\Rightarrow (e_1, \ldots, e_n)'$ and $(e_{R_1}, \ldots, e_{R_n})'$ have the same distribution, where
  $\boldsymbol{R} = (R_1, \ldots, R_n)'$ is a random permutation of $(1, \ldots, n)'$ such that
  $$P(\boldsymbol{R} = \boldsymbol{r}) = \frac{1}{n!}$$
- Permutation version of the test statistic:
  errors e_i replaced by permuted residuals $\hat{e}_{R_i}$
- We study conditional distribution of the permutation test statistic, given the observations $\boldsymbol{Y}$

How do permutation tests work?

1. From the data sample we calculate the test statistic F_n.
2. We generate a random permutation $\boldsymbol{R} = \boldsymbol{r}$ of $(1, \ldots, n)'$.
3. We calculate a permutational version $F_n(\boldsymbol{r})$ of our test statistic for $\boldsymbol{r}$.
4. We repeat the last two steps N times.
5. We obtain the empirical distribution of $F_n(\boldsymbol{R})$ and calculate its empirical α-quantile c_α.
6. We reject the null hypothesis if $F_n > c_\alpha$.

Outline

Linear regression model

$$
\begin{aligned}
Y_i &= \boldsymbol{x}_i' \boldsymbol{\beta}_1 + e_i && i = 1, \dots, t_1 \\
Y_i &= \boldsymbol{x}_i' \boldsymbol{\beta}_2 + e_i && i = t_1 + 1, \dots, t_2 \\
&\vdots \\
Y_i &= \boldsymbol{x}_i' \boldsymbol{\beta}_{m+1} + e_i && i = t_m + 1, \dots, n,
\end{aligned}
$$

- unknown change points $t_1, \dots, t_m$
- unknown parameters $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jq})'$
- regressors $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})'$
- errors $e_1, \dots, e_n$

Non-trending regression and independent errors

(A1) e_i are i.i.d. with $Ee_1 = 0$, $Ee_1^2 = \sigma^2$, $E|e_1|^{2+\Delta} < \infty$, $\Delta > 0$.

(A2) e_i are independent of $\boldsymbol{x}_j$ for all i and j.

(A3)

$$\frac{1}{n}\sum_{i=1}^{\lfloor ns \rfloor} \boldsymbol{x}_i \boldsymbol{x}_i' \stackrel{def}{=} \frac{1}{n}\boldsymbol{C}_{\lfloor ns \rfloor} \xrightarrow{P} s\boldsymbol{C} > 0 \text{ uniformly in } s,\ 0 < s \leq 1.$$

(A4) For all $\eta > 0$, there exists a $D > 0$, such that for all large n,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^{2+\Delta} > D\right) < \eta.$$

(A5) The true change points satisfy

$$t_j^0 = \lfloor n\lambda_j^0 \rfloor, \quad 0 = \lambda_0^0 < \lambda_1^0 < \ldots < \lambda_{m+1}^0 = 1.$$

Autoregressive model of order p

$$\boldsymbol{x}_i' = (1, Y_{i-1}, Y_{i-2}, \dots, Y_{i-p})$$

(B1) e_i are i.i.d. with $Ee_1 = 0$, $Ee_1^2 = \sigma^2$ and $E|e_i|^4 < \infty$.

(B2) $Y_1, \dots, Y_p$ are independent of $e_{p+1}, \dots, e_n$ and the roots of $t^p - \beta_{j1} t^{p-1} - \dots - \beta_{jp}$ are $|t| < 1, \quad j = 1, \dots, m+1$.

(B3) The initial values $Y_1, \dots, Y_p$ satisfy:

$$(Y_p - \mu_1, \dots, Y_1 - \mu_1)' = \sum_{j=0}^{\infty} \boldsymbol{B}^j \boldsymbol{e}_{p-j},$$

$$\boldsymbol{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \boldsymbol{I}_{p-1} & & 0 \end{pmatrix}; \; \boldsymbol{e}_k = (e_k, 0, \dots, 0)'; \; \mu_1 = EY_i, \, i \le t_1.$$

Autoregressive models

Limit behavior of regressors

- under H_0 $(m = 0)$

$$\frac{\boldsymbol{C}_{\lfloor ns \rfloor}}{n-p} = \frac{1}{n-p} \sum_{i=p+1}^{\lfloor ns \rfloor} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{P} s\boldsymbol{C} > 0,$$

- under H_A $(m = k)$

$$\frac{\boldsymbol{C}_{\lfloor ns \rfloor}}{n-p} \xrightarrow{P} \boldsymbol{Q}_s = \sum_{j=1}^{k+1} \left(\min\{s, \lambda_j^0\} - \lambda_{j-1}^0 \right) \boldsymbol{C}^{(j)} I_{\{s \geq \lambda_{j-1}^0\}},$$

where $\boldsymbol{C}^{(j)} > 0,\ j = 1, \ldots, k+1$ and $s \in (0, 1]$.

Trending regression

(C1) e_i are i.i.d. with common symmetric pdf F.

(C6) The regressors $\boldsymbol{x}_i = \boldsymbol{h}(i/n)$ satisfy:
$h_1(x) = 1,\ 0 \le x \le 1,$
$h_2(\cdot), \ldots, h_q(\cdot)$ are continuously diff. functions on $[0, 1]$
and as $n \to \infty$,

$$\frac{1}{n}\boldsymbol{C}_{\lfloor ns \rfloor} = \frac{1}{n}\sum_{i=1}^{\lfloor ns \rfloor} \boldsymbol{h}(i/n)\boldsymbol{h}'(i/n) \to \boldsymbol{C}(s) = \int_0^s \boldsymbol{h}(x)\boldsymbol{h}'(x)dx > 0$$

uniformly in $s \in (0, 1]$, $\boldsymbol{C}(s)$ is strictly increasing in s.

Trending regression (cont.)

The score function ψ and $\lambda(t) = -\int \psi(e-t)dF(e), \quad t \in R$:

(C2) ψ is non-decreasing, antisymetric: $\psi(x) = -\psi(-x)$.

(C3) $\sigma^2(\psi) = \int \psi^2(x)dF(x) \in (0,\infty)$.
$a > 0$, $D_1 > 0$, $D_2 > 0$

$$\int (\psi(x-s_2) - \psi(x-s_1))^2 \, dF(x) \leq D_1|s_2 - s_1|^a$$

$$|s_j| \leq D_2, \; j = 1, 2.$$

(C4) $\lambda'(\cdot)$ exists and is Lipschitz in a neighbourhood of zero, $\lambda(0) = 0$ and $\lambda'(0) > 0$.

(C5) Either
  (i) ψ is bounded, or
  (ii) $\psi(x) = x$ and $\int |x|^{2+\Delta} dF(x) < \infty$.

Outline

F-type test statistic

Test for no change versus k (maximum M) changes

$$
\begin{aligned}
F_n^{\varepsilon}(k,q) &= \max_{(t_1,\dots,t_k)\in T_{\varepsilon,k}} F_n(t_1,\dots,t_k;q),\\
DF_n^{\varepsilon}(M;q) &= \max_{k=1,\dots,M} \max_{(t_1,\dots,t_k)\in T_{\varepsilon,k}} F_n(t_1,\dots,t_k;q),
\end{aligned}
$$

where

$$
\begin{aligned}
F_n(t_1,\dots,t_k;q) &= \frac{1}{kq}\frac{SSR_0 - SSR_k(\boldsymbol{t})}{\hat{\sigma}^2_{n,k}(\boldsymbol{t})}\\
\hat{\sigma}^2_{n,k}(\boldsymbol{t}) &= SSR_k(\boldsymbol{t})/(n-(k+1)\,q) \xrightarrow{P} \sigma^2 \quad \text{under } H_0\\
\boldsymbol{t} &= (t_1,\dots,t_k)'\\
T_{\varepsilon,k} &= \{(t_1,\dots,t_k):\ t_{j+1}-t_j \geq \lfloor n\varepsilon \rfloor,\ \forall j=0,\dots,k\}
\end{aligned}
$$

F-type test - equivalent expression

$$
\begin{aligned}
SSR_0 - SSR_k(\boldsymbol{t}) &= \sum_{j=1}^{k+1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_i \right)' \boldsymbol{C}_{t_{j-1},t_j}^{-1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_i \right) \\
&\stackrel{H_0}{=} - \left(\sum_{i=1}^{n} \boldsymbol{x}_i e_i \right)' \boldsymbol{C}_n^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i e_i \right) \\
&\quad + \sum_{j=1}^{k+1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i e_i \right)' \boldsymbol{C}_{t_{j-1},t_j}^{-1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i e_i \right)
\end{aligned}
$$

$$
\hat{e}_i = Y_i - \boldsymbol{x}_i' \boldsymbol{C}_n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i Y_i \stackrel{H_0}{=} e_i - \boldsymbol{x}_i' \boldsymbol{C}_n^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i e_i
$$

M-estimation

$$\sum_{i=1}^{n} \rho\left(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) = \min!, \quad \text{which usually leads to}$$

$$\sum_{i=1}^{n} \psi\left(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) x_{ij} = 0,\ j = 1, \ldots, q, \quad \psi = \rho'$$

Jurečková and Sen (1996)

$$\hat{\boldsymbol{\beta}}_n(\psi) - \boldsymbol{\beta} = \boldsymbol{C}_n^{-1} \frac{1}{\lambda'(0)} \sum_{i=1}^{n} \boldsymbol{x}_i \psi(e_i) + o_p(n^{-1/2}).$$

$$\hat{\boldsymbol{\beta}}_n(\psi) \sim AN\left(\boldsymbol{\beta}, \frac{\sigma^2(\psi)}{(\lambda'(0))^2} \boldsymbol{C}_n^{-1}\right)$$

M-type test statistic

$$F_n^{\varepsilon}(\psi, k, q) = \max_{(t_1,\dots,t_k)\in T_{\varepsilon,k}} F_n(\psi, t_1, \dots, t_k; q)$$

$$F_n(\psi, t_1, \dots, t_k; q) = \frac{1}{kq\ \tilde{\sigma}^2_{n,k}(\psi, \boldsymbol{t})} \sum_{j=1}^{k+1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_i(\psi) \right)' \boldsymbol{C}^{-1}_{t_{j-1},t_j} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_i(\psi) \right)$$

$$\tilde{\sigma}^2_{n,k}(\psi, \boldsymbol{t}) - \sigma^2(\psi) = o_p(1), \quad \hat{e}_i(\psi) = \psi\left(Y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n(\psi) \right),\ i = 1, \dots, n$$

Non-trending regression and AR models

Let $\boldsymbol{W}(\cdot)$ be a vector of q independent standard Wiener processes on $[0,1]$. Under H_0 ($m=0$) and given assumptions, as $n \to \infty$,

$$DF_n^{\varepsilon}(M;q) \quad \xrightarrow{d} \quad \max_{1 \le k \le M} \sup_{(\lambda_1,\dots,\lambda_k)\in\Lambda_{\varepsilon,k}} F(\lambda_1,\dots,\lambda_k;q)$$

with

$$F(\lambda_1,\dots,\lambda_k;q) \stackrel{def}{=} \frac{1}{kq}\sum_{j=1}^{k} \frac{\|\lambda_j \boldsymbol{W}(\lambda_{j+1}) - \lambda_{j+1}\boldsymbol{W}(\lambda_j)\|^2}{\lambda_j\lambda_{j+1}(\lambda_{j+1}-\lambda_j)}$$

and the supremum is taken over the set

$$\Lambda_{\varepsilon,k} = \{(\lambda_1,\dots,\lambda_k) :\ \lambda_{j+1}-\lambda_j \ge \varepsilon,\ \forall j = 0,\dots,k\}.$$

Trending regression

$$DF_n^{\varepsilon}(\psi, M, q) \xrightarrow{d} \max_{1\le k\le M} \left\{ \sup_{(\lambda_1,\ldots,\lambda_k)\in\Lambda_{\varepsilon,k}} F_{\boldsymbol{h}}(\lambda_1,\ldots,\lambda_k;q) \right\},$$

where $F_{\boldsymbol{h}}(\lambda_1,\ldots,\lambda_k;q) = \dfrac{1}{kq}\Bigg\{ -\boldsymbol{W_h}(1)'\boldsymbol{C}(1)^{-1}\boldsymbol{W_h}(1) +$

$$\sum_{j=1}^{k+1} \left(\boldsymbol{W_h}(\lambda_j) - \boldsymbol{W_h}(\lambda_{j-1})\right)' \left(\boldsymbol{C}(\lambda_j) - \boldsymbol{C}(\lambda_{j-1})\right)^{-1} \left(\boldsymbol{W_h}(\lambda_j) - \boldsymbol{W_h}(\lambda_{j-1})\right) \Bigg\},$$

$$\boldsymbol{W_h}(t) = \int_0^t \boldsymbol{h}(x)dW(x),$$

$$\Lambda_{\varepsilon,k} = \left\{(\lambda_1,\ldots,\lambda_k) :\ \lambda_{j+1} - \lambda_j \ge \varepsilon,\ \forall j = 0,\ldots,k\right\}.$$

Permutation version of F-type test statistic

$$DF_n^{\varepsilon}(M, q, \boldsymbol{R}) = \max_{1 \le k \le M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon,k}} F_n(t_1, \ldots, t_k; q, \boldsymbol{R})$$

with

$$F_n(t_1, \ldots, t_k; q, \boldsymbol{R}) = \frac{1}{kq\ \hat{\sigma}_n^2} \left[-\left(\sum_{i=1}^{n} \boldsymbol{x}_i \hat{e}_{R_i} \right)' \boldsymbol{C}_n^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i \hat{e}_{R_i} \right) + \sum_{j=1}^{k+1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_{R_i} \right)' \boldsymbol{C}_{t_{j-1},t_j}^{-1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_i \hat{e}_{R_i} \right) \right],$$

where $\hat{e}_{R_1}, \ldots, \hat{e}_{R_n}$ are permuted L_2-residuals.

Permutation version of M-type test statistic

$$DF_n^{\varepsilon}(\psi, M, q, \boldsymbol{R}) = \max_{1 \le k \le M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon,k}} F_n(\psi, t_1, \ldots, t_k; q, \boldsymbol{R})$$

with

$$F_n(\psi, t_1, \ldots, t_k; q, \boldsymbol{R})$$
$$= \frac{1}{kq\ \hat{\sigma}_n^2(\psi)} \Bigg[- \left(\sum_{i=1}^{n} \boldsymbol{h}(i/n) \hat{e}_{R_i}(\psi) \right)' \boldsymbol{C}_n^{-1} \left(\sum_{i=1}^{n} \boldsymbol{h}(i/n) \hat{e}_{R_i}(\psi) \right)$$
$$+ \sum_{j=1}^{k+1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{h}(i/n) \hat{e}_{R_i}(\psi) \right)' \boldsymbol{C}_{t_{j-1},t_j}^{-1} \left(\sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{h}(i/n) \hat{e}_{R_i}(\psi) \right) \Bigg],$$

where $\hat{e}_{R_1}(\psi), \ldots, \hat{e}_{R_n}(\psi)$ are permuted M-residuals.

Theorem

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ follow the model with $m \geq 0$ changes. Then under the considered assumptions, for arbitrary $x \in \mathbb{R}$, as $n \rightarrow \infty$,

$$P\left(\max_{1 \leq k \leq M} \max_{(t_1, \ldots, t_k) \in T_{\varepsilon,k}} F_n(t_1, \ldots, t_k, q, \boldsymbol{R}) \leq x \middle| \boldsymbol{Y} \right) \xrightarrow{P} P\left(\max_{1 \leq k \leq M} \sup_{(\lambda_1, \ldots, \lambda_k) \in \Lambda_{\varepsilon,k}} F(\lambda_1, \ldots, \lambda_k) \leq x \right),$$

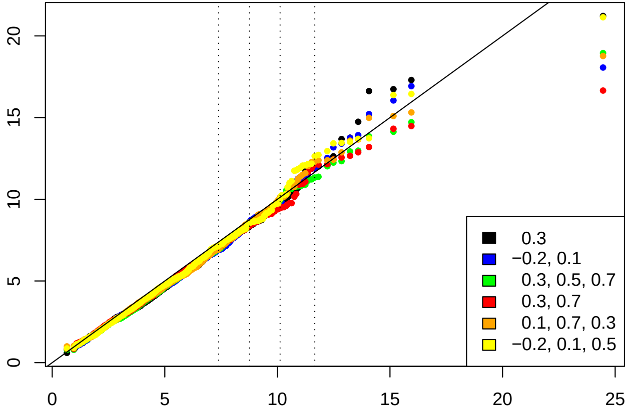
where $F(\lambda_1, \ldots, \lambda_k)$ is a random variable such that

$$F_n(t_1, \ldots, t_k, q) \xrightarrow{d} F(\lambda_1, \ldots, \lambda_k) \quad \text{under } H_0, \quad n \rightarrow \infty.$$

Outline

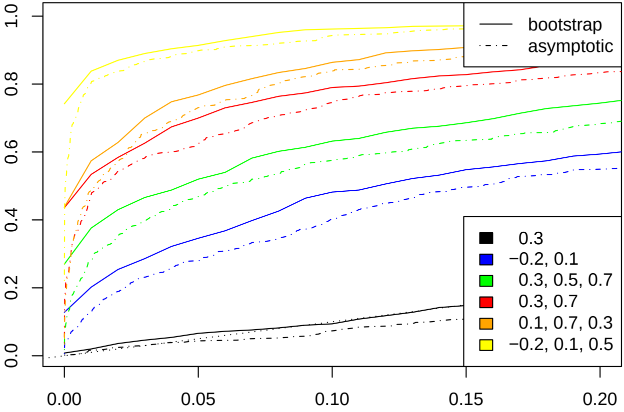
QQ plot for the null distr. of $DF_n^{\varepsilon}(q, M)$ vs. $DF_n^{\varepsilon}(q, M, \boldsymbol{R})$

$n = 180,\ M = 2,\ \varepsilon = 0.15,\ q = 1$



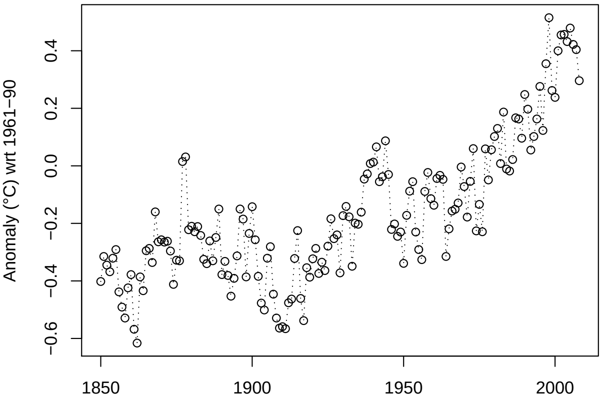
SPC plots for $DF_n^{\varepsilon}(q, M)$ with respect to perm. distr.

Size-power-curves plots; $n = 180$, $M = 2$, $\varepsilon = 0.15$, $q = 1$



Outline

Global temperature data



http://www.metoffice.gov.uk/research/hadleycentre/obsdata

Fitted models

Segmented model with linear / quadratic trend with k changes

$$\begin{aligned} Y_i &= \beta_{j0} + \beta_{j1}\left(\frac{i}{n}\right) + e_i \\ Y_i &= \beta_{j0} + \beta_{j1}\left(\frac{i}{n}\right) + \beta_{j2}\left(\frac{i}{n}\right)^2 + e_i \end{aligned}$$

$i = t_{j-1} + 1, \ldots, t_j, \qquad j = 1, \ldots, k+1, \qquad k = 1, 2, 3, 4$

$n = 159$ (years $1850, \ldots, 2008$)
$n\varepsilon = \lfloor 159 * 0.05 \rfloor = 7$ years

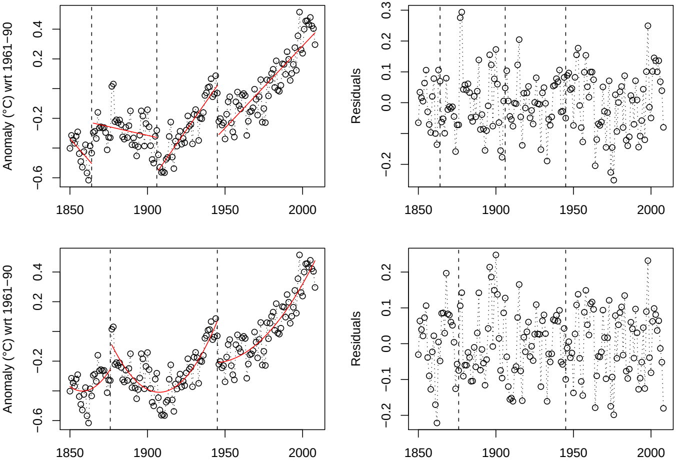
F-type tests

q	$F_n^\varepsilon(1,q)$	$F_n^\varepsilon(2,q)$	$F_n^\varepsilon(3,q)$	$F_n^\varepsilon(4,q)$	$F_n^\varepsilon(5,q)$	$DF_n^\varepsilon(5,q)$
2	49.6	47.7	40.7	37.8	33.6	49.6
3	14.5	20.6	17.4	16.0	14.3	20.6

Permutation critical values

	$q=2$			$q=3$		
	90%	95%	99%	90%	95%	99%
$F_n^\varepsilon(1,q)$	5.893	6.921	8.937	4.938	5.563	6.768
$F_n^\varepsilon(2,q)$	5.363	5.965	6.876	4.589	5.037	5.895
$F_n^\varepsilon(3,q)$	5.014	5.497	6.649	4.313	4.765	5.378
$F_n^\varepsilon(4,q)$	4.813	5.274	6.340	4.193	4.542	5.182
$F_n^\varepsilon(5,q)$	4.664	5.068	6.036	4.101	4.380	5.028
$DF_n^\varepsilon(5,q)$	6.080	6.956	8.937	5.093	5.767	6.768

Segmented models with linear / quadratic trend



Outline

Conclusion

Tests for detection of multiple changes

- F-type tests (Bai and Perron, 1998)
- Generalized M-type tests

Models

- Linear regression models with non-trending or trending regressors and independent errors
- Autoregressive models

Approximations to critical values

- Bootstrap with or without replacement

Some important publications

BAI AND PERRON (1998).
Estimating and testing linear models with multiple structural changes.
Econometrica, 66, 47–78.

HUŠKOVÁ AND PICEK (2005)
Bootstrap in detection of changes in linear regression
Sankhyā: The Indian Journal of Statistics, Vol. 67, 200 –226

KIRCH (2006)
Resampling Methods for the Change Analysis of Dependent Data
PhD thesis, University of Cologne

Software

R DEVELOPMENT CORE TEAM (2008)
R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria
ISBN 3-900051-07-0, URL http://www.R-project.org

ZEILEIS, LEISCH, HORNIK AND KLEIBER (2002)
strucchange: An R Package for Testing for Structural Change in Linear Regression Models
Journal of Statistical Software **7**, 1 – 38