Change point in trending regression

Marie Hušková

Charles University, Prague

ROBUST 2010
joint work with *A. Aue, L.Horváth, J. Picek*

Outline

1. Introduction

Outline

Outline

1. Introduction
2. Regression model-formulation
3. Some asymptotics

Outline

Outline

Outline

Outline

• Procedures on stability of statistical models - structural breaks, disorder, stability, segmented regression, switching regression, change point problem, etc.

• Observations $Y_1, \ldots, Y_n$ obtained at the ordered time points $t_1 < \cdots < t_n$ such that
$Y_1, \ldots, Y_{k^*}$ —- model I
$Y_{k^*}, \ldots, Y_n$ —- model II
k^* — **change point** — unknown

• The problem: to detect (to test H_0: no change & H_1: there is a change), to identify k^* (to estimate k^*) and to estimate the model before and after the change.

• Many variants – multiple changes, abrupt changes, gradual changes, changes in various parameters, changes in distributions, independent observations, dependent observations.

• Construction of tests and estimators — various approaches as in most of the statistical problems.

• Theoretical problems and theoretical problems.

• Applications (meteorology, climatology, hydrology or environmental studies, econometric time series, statistical quality control, etc.)

Outline

Regression model- formulation

$Y_1, \ldots, Y_n$ are observed at time points $t_1 < \cdots < t_n$:

$$
\begin{aligned}
Y_i \quad &= \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \qquad & i = 1 \ldots, k^* \\
&= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\delta} + e_i, \qquad & i = k^* + 1 \ldots, n,
\end{aligned}
$$

$e_1, \ldots, e_n$ — random errors – i.i.d., zero mean, nonzero variance σ^2 and finite $E|e_i|^{2+\Delta}$ with some $\Delta > 0$

$\boldsymbol{\beta}, \boldsymbol{\delta} \neq \mathbf{0}$ — parameters

k^* .... change point

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ – p-dim. design points (random or nonrandom):

nontrending regression: $\frac{1}{n}\sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T \approx \frac{k}{n}\mathbf{C}, \ k \leq n$

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ – p-dim. design points (random or nonrandom):

nontrending regression: $\frac{1}{n}\sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T \approx \frac{k}{n}\mathbf{C}, \; k \leq n$

trending regression: $\mathbf{x}_i = \mathbf{h}(i/n), i = 1, \ldots, n$, h smooth nonconstant vector function

Main problems:

H_0 : no jump in regression & H_1: at most one jump

estimators of change points

Test statistics

$$T_n = \max_{p \le k < n-p} \left\{ \left(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0 \right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0 \right) \right\}$$

- $\widehat{\boldsymbol{\beta}}_k$ — LSE of $\boldsymbol{\beta}$ based on $Y_1, \dots, Y_k$
- $\widehat{\boldsymbol{\beta}}_k^0$ — LSE of $\boldsymbol{\beta}$ based on $Y_{k+1}, \dots, Y_n$
- $\widehat{\boldsymbol{\Sigma}}_k^{-1}$ is an estimator of the variance matrix of $\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0$

Equivalently

$$T_n = \max_{p \le k < n-p} \left\{ \mathbf{S}_k^T \mathbf{C}_k^{-1} \mathbf{C}_n (\mathbf{C}_k^0)^{-1} \mathbf{S}_k \frac{1}{\widehat{\sigma}_n^2} \right\},$$

$\mathbf{S}_k = \sum_{i=1}^k \mathbf{x}_i \widehat{e}_i, \quad k = 1, ..., n,$
$\widehat{e}_i = Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n, \ i = 1, \dots, n$ – - residuals

$\mathbf{C}_k = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^T, \quad \mathbf{C}_k^0 = \mathbf{C}_n - \mathbf{C}_k$

Test statistics

$$T_n = \max_{p \le k < n-p} \left\{ \left(\widehat{\beta}_k - \widehat{\beta}_k^0 \right)^T \widehat{\mathbf{\Sigma}}_k^{-1} \left(\widehat{\beta}_k - \widehat{\beta}_k^0 \right) \right\}$$

- $\widehat{\beta}_k$ — LSE of β based on $Y_1, \dots, Y_k$
- $\widehat{\beta}_k^0$ — LSE of β based on $Y_{k+1}, \dots, Y_n$
- $\widehat{\mathbf{\Sigma}}_k^{-1}$ is an estimator of the variance matrix of $\widehat{\beta}_k - \widehat{\beta}_k^0$

Equivalently

$$T_n = \max_{p \le k < n-p} \left\{ \mathbf{S}_k^T \mathbf{C}_k^{-1} \mathbf{C}_n (\mathbf{C}_k^0)^{-1} \mathbf{S}_k \frac{1}{\widehat{\sigma}_n^2} \right\},$$

$\mathbf{S}_k = \sum_{i=1}^k \mathbf{x}_i \widehat{e}_i, \quad k = 1, ..., n,$
$\widehat{e}_i = Y_i - \mathbf{x}_i^T \widehat{\beta}_n, \; i = 1, \dots, n$ – - residuals

$\mathbf{C}_k = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^T, \quad \mathbf{C}_k^0 = \mathbf{C}_n - \mathbf{C}_k$
$\widehat{\sigma}_n^2$ is an estimator of $var\, e_i = \sigma^2$

Critical regions

$$T_n > c_n(\alpha)$$

$c_n(\alpha)$ — critical value

α — level

Approximation of the critical values:

(i) limit distribution of T_n under H_0;

(ii) resampling methods (bootstrap)

Estimator $\widehat{k}^*$ of the change point k^* defined as such $\widehat{k}^*$ it maximizes w.r.t. k

$$\left\{(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)\right\}$$

Outline

Test statistics:

$$T_n = \max_{p \le k < n-p} \left\{ \left(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0 \right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \left(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0 \right) \right\}$$

$$T_n(\varepsilon) = \max_{\varepsilon n \le k \le (1-\varepsilon)n} \left\{ \dots \right\}, \quad 0 < \varepsilon < 1/2$$

nontrending and polynomial regression:

$$\lim_{n\to\infty} P(a(\log n)(T_n)^{1/2} \le t + b_p(\log n))$$

$$= \exp\{-2\exp\{-t\}\},\ t \in R^1,$$

$a(y) = (2\log y)^{1/2},$

$b_p(y) = 2\log y + \frac{p}{2}\log\log y - \log(2\Gamma(p/2)), y > 1,$

$\mathbf{x}_i = (1, i/n, (i/n)^2, \dots, (i/n)^{p-1})^T, \quad i = 1, \dots, n$

nontrending regression

$$T_n(\varepsilon) \to^d \sup_{\varepsilon<t<1-\varepsilon} \left\{ \frac{\sum_{i=1}^p B_i^2(t)}{t(1-t)} \right\}$$

$\{B_j(t); t \in (0,1)\}$, $j = 1, ..., p$, — independent Brownian bridges

trending regression $\mathbf{x}_i = \mathbf{h}(i/n)$

$$T_n(\varepsilon) \to^d \sup_{\varepsilon<t<1-\varepsilon} \mathbf{S}^T(t)\mathbf{C}(t)\mathbf{C}^{-1}(1)\mathbf{C}^0(t)\mathbf{S}(t)$$

$$\mathbf{S}(t) = \int_0^t \mathbf{h}(x)dB(x) - \mathbf{C}(t)\mathbf{C}^{-1}(1)\int_0^1 \mathbf{h}(x)dB(x), \quad t \in [0,1]$$

with $\{B(x),\ x \in [0,1]\}$ being a Brownian bridge,
$\mathbf{C}(t) = \lim_{n\to\infty} \frac{1}{n}\mathbf{C}_{\lfloor nt \rfloor}$.

Outline

Trending regression

Assumptions

- (A.1) The sequence $(e_i : i \geq 1)$ satisfy
$$E[e_i] = 0, \qquad E[e_i^2] = \sigma^2 > 0 \tag{1}$$
- (A.2) There are independent standard Brownian motions $(W_{1,n}(t) : t \geq 0)$ and $(W_{2,n}(t) : t \geq 0)$ such that
$$\max_{1 \leq k \leq n/2} \frac{1}{k^{1/\nu}} \left| \sum_{i=1}^{k} e_i - \tau W_{1,n}(k) \right| = \mathcal{O}_P(1) \qquad (n \to \infty) \tag{2}$$
and
$$\max_{n/2 < k < n} \frac{1}{(n-k)^{1/\nu}} \left| \sum_{i=k+1}^{n} e_i - \tau W_{2,n}(n-k) \right| = \mathcal{O}_P(1) \qquad (n \to \infty) \tag{3}$$
with some $\nu > 2$ and $\tau > 0$.

- (A.3) The components of $\mathbf{h}(.)$ are continuous on $[0,1]$. The matrices $\int_0^t \mathbf{h}(x)\mathbf{h}^T(x)dx$ and $\int_t^1 \mathbf{h}(x)\mathbf{h}^T(x)dx$ are regular for all $t \in (t^0, 1-t^0)$ for all $t_0 \in (0, 1/2)$.
- (A.4) There are p linearly independent p-dimensional vectors $\mathbf{a}_{01}, \ldots, \mathbf{a}_{0p}$ and nonnegative $0 \le \gamma_{01} < \ldots < \gamma_{0p}$ such that

$$\lim \sup_{t \to 0_+} \frac{1}{t^{\gamma_{0p}+1}} \left| \mathbf{h}(t) - \sum_{\ell=1}^{p} \mathbf{a}_{0\ell} t^{\gamma_{0\ell}} \right| < \infty. \tag{4}$$

- (A.5) There are p linearly independent p-dimensional vectors $\mathbf{a}_{11}, \ldots, \mathbf{a}_{1p}$ and nonnegative $0 \le \gamma_{11} < \ldots < \gamma_{1p}$ such that

$$\lim \sup_{t \to 1_-} \frac{1}{t^{\gamma_{1p}+1}} \left| \mathbf{h}(t) - \sum_{\ell=1}^{p} \mathbf{a}_{1\ell} t^{\gamma_{1\ell}} \right| < \infty. \tag{5}$$

Particular cases:

• *polynomial regression*

$$\mathbf{h}(t) = (t^{\gamma_1}, \ldots, t^{\gamma_p})^T,\ t \in [0,1], \quad 0 \le \gamma_1 < \ldots < \gamma_p$$

• *harmonic regression*

$$\mathbf{h}(t) = (\cos(2\pi t\omega_1), \sin(2\pi t\omega_1), \ldots, \cos(2\pi t\omega_p), \cos(2\pi t\omega_p))^T,\ t \in [0,1]$$

$\omega_1, \ldots, \omega_p$ known.

• *independent* $e_1, \ldots, e_n$: $\tau = \sigma$.

• *dependent* $e_1, \ldots, e_n$: the estimator with flat top kernel used:

$$\widehat{\tau}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(\widehat{e}_j - \overline{e}_n)^2 + \frac{2}{n}\sum_{j=1}^{q_n} w_j \sum_{i=1}^{n-j}(\widehat{e}_i - \overline{e}_n)(\widehat{e}_{i+j} - \overline{e}_n)$$

$\widehat{e}_j = Y_j - \mathbf{x}_j^T\widehat{\beta}_n, \quad \overline{e}_n = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i$

$$w_j = 1I\{1 \le j \le q_n/2\} + 2(1 - j/q_n)I\{q_n/2 < j \le q_n\}$$

Outline

Proof

Recall notation:

$T_n = \max_{p \le k < n-p} V_k(\tau^2/\widehat{\tau}_n^2)$

$V_k = \mathbf{S}_k^T \mathbf{C}_k^{-1} \mathbf{C}_n (\mathbf{C}_k^0)^{-1} \mathbf{S}_k / \tau^2$

$\mathbf{S}_k = \sum_{i=1}^k \mathbf{x}_i \widehat{e}_i, \quad k = 1, ..., n,$

$\widehat{e}_i = Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n, \ i = 1, \ldots, n$ – - residuals

$\mathbf{C}_k = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^T, \quad \mathbf{C}_k^0 = \mathbf{C}_n - \mathbf{C}_k$

$\mathbf{x}_i = \mathbf{h}(i/n)$

Main step of the proof

- The limit behavior of T_n is the same as $\max(T_{n0}, T_{n1})$

$$T_{n0} = \max_{p \le k < ns_n} V_k$$

$$T_{n1} = \max_{p \le n-k < ns_n} V_k$$

for some $s_n \to 0_+$, T_{n0} and T_{n1} are asymptotically independent.

- T_{n0} has the same limit distribution as

$$T_{n0}^0 = \max_{p \le k < ns_n} \mathbf{Z}_k^T \mathbf{M}_k^{-1} \mathbf{Z}_k / \tau^2$$

$\mathbf{Z}_k = \sum_{i=1}^k \mathbf{x}_i e_i, \quad \mathbf{M}_k = \sum_{i=1}^k \mathbf{x}_i \mathbf{x}_i^T.$

- T_{n0}^0 will not change if $\mathbf{x}_i$ is replaced by $\mathbf{x}_i^0 = \mathbf{A}\mathbf{x}_i$ with arbitrary nonsingular $p \times p$ matrix $\mathbf{A}$.

- $\max_{1 \le k \le n/2} |\mathbf{Z}_k - \tau \sum_{j=1}^k \mathbf{x}_j (W_{1,n}(j) - W_{1,n}(j-1))| = o_P(n^{-\kappa})$ for some $\kappa > 0$.

- After some steps we get that T_{n0}^0 has the same distribution as

$$\max_{k_n \le k < ns_n} \frac{1}{k\tau} \mathbf{Q}_k^T \mathbf{R}_k^{-1} \mathbf{Q}_k$$

with $\mathbf{Q}_k = \sum_{i=1}^k ((i/k)^{\gamma_{01}}, \dots, (i/k)^{\gamma_{0p}})^T e_i$
$\mathbf{R}_k = \left(\frac{1}{k}\sum_{j=1}^k (j/k)^{\gamma_{0v}+\gamma_{0r}}\right)_{v,r=1}^p.$

- Finally, we get that T_{n0}^0 has the same distribution as

$$\sup_{k_n \le t \le ns_n} \mathbf{Q}(t)^T \mathbf{V}(t)^{-1} \mathbf{Q}(t)$$

$$\mathbf{Q}(t) = \left(\frac{1}{t^{\gamma_{0\ell}+1/2}} \int_0^t x^{\gamma_{0\ell}} dW_{n,1}(x);\ \ell = 1, \dots, p\right)$$

$$\mathbf{V}(t) = \left(\frac{1}{t^{\gamma_{0\ell}+\gamma_{0\ell'}+1}} \int_0^t x^{\gamma_{0\ell}+\gamma_{0\ell'}} dx;\ \ell, \ell' = 1, \dots, p\right)$$

- After exponential transformation $\mathbf{Q}(t)^T \mathbf{V}(t)^{-1} \mathbf{Q}(t)$ is a norm of stationary Gaussian process results of Albin, Piterbarg, Jaruskova etc. are applied.

Outline

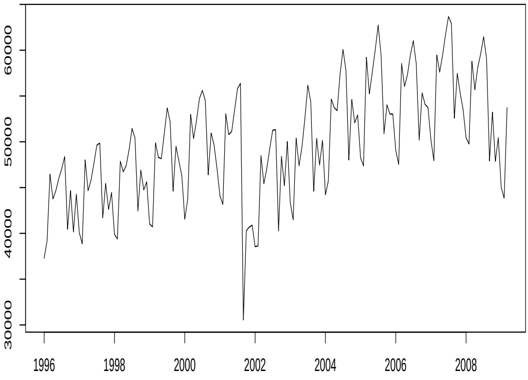


Figure 1: Monthly air carrier traffic in the United Stated from January 1996 to March 2009.

Simulation

$$Y_j = 4 + \frac{6j}{n} + \frac{3}{2}\cos(\frac{2\pi j}{12n}) + \frac{3}{2}\sin(\frac{2\pi}{12n}) + \frac{9}{10}\cos(\frac{2\pi}{4n} + \frac{1}{2}\sin(\frac{2\pi}{4n}) + e_j,$$

$j = 1, \ldots, 200$

e_j –either AR(1) or MA(1), normal distr.

Change point $k^* = 100$ either in the intercept or in one harmonic regressor

1000 repetitions

critical value obtained through circular block bootstrap

$H_A^{(1)}$ change in intercept, AR (1) or MA(1)
$H_A^{(1)}$ change in one regressor, AR (1) or MA(1)

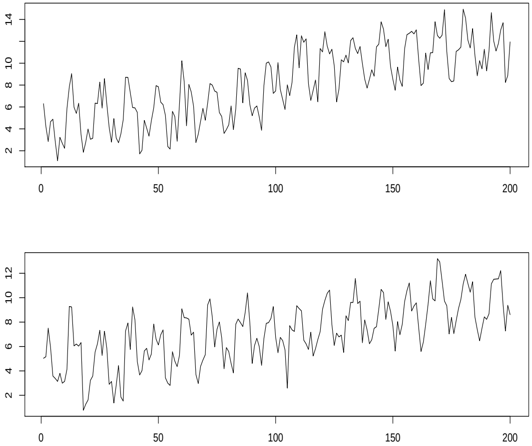


Figure 4: Time series plots of the processes under $H_A^{(1)}$ with $\delta = 2$ (upper panel) and $H_A^{(2)}$ (lower

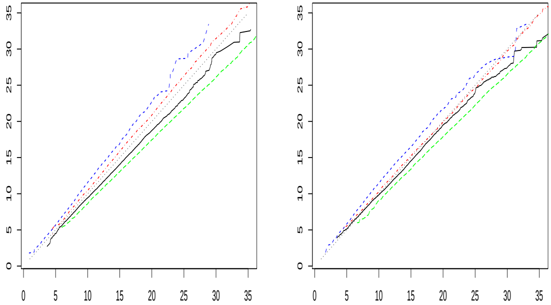


Figure 2: Circular bootstrap distribution versus finite sample distribution for AR(1) innovations (left) and MA(1) innovations (right) with scalings $\varphi^2 = 2$ (- - -), 1.5 (·-·-·), 1.0 (—), 0.5 (- - -).

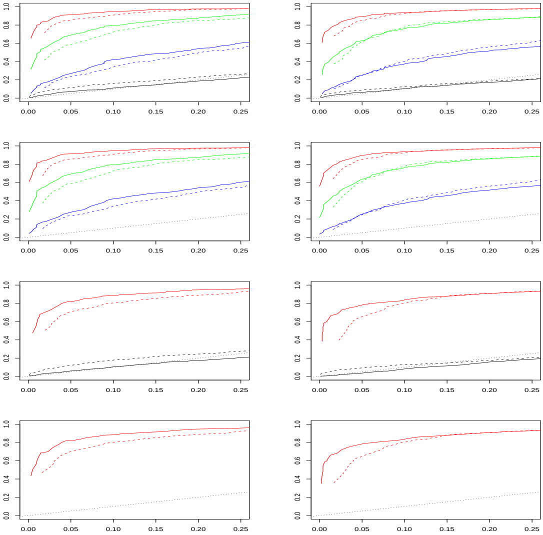
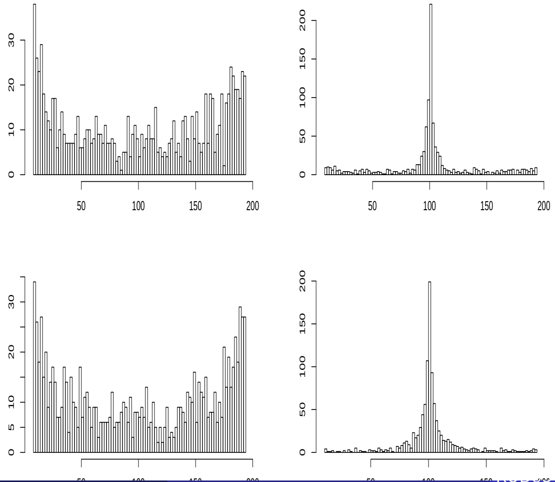
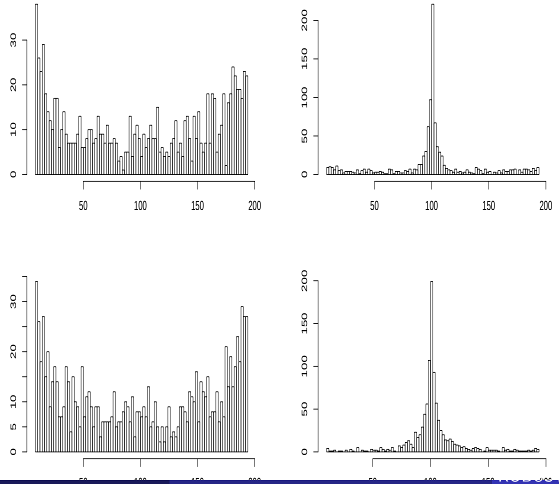


Figure 3: Size-power curves for $H_A^{(1)}$ (upper half) and $H_A^{(2)}$ (lower half) for the asymptotic modification (first and third line) and the the circular bootstrap (second and fourth line) with AR(1) innovations (left) and MA(1) innovations (right).

17


30
20
10
0
50
100
150
200
200
150
100
50
0
50
100
150
200
30
20
10
5
0
200
150
100
50
0


30
20
10
0
50
100
150
200
200
150
100
50
0
50
100
150
200
30
20
10
5
0
200
150
100
50
0

Data

Monthly air traffic data

model through the root of data, $n = 159$

$$Y_j = \beta_0 + \beta_1 j/n + \sum_{\ell=1}^{q} \Big(\beta_\ell^c \cos(2\pi\omega_{ell} j/n) + \beta_\ell^{(s)} \sin(2\pi\omega_{ell} j/n) \Big) + e_j$$

$q = 4$, $\omega_1 = 2/160$, $\omega_3 = 13/160$, $\omega_3 = 40/160$, $\omega_1 = 80/160$

ω_2 – annual cycle

ω_3 – quarterly cycle

ω_4 —two months cycle

$\widehat{k}^* = 69$

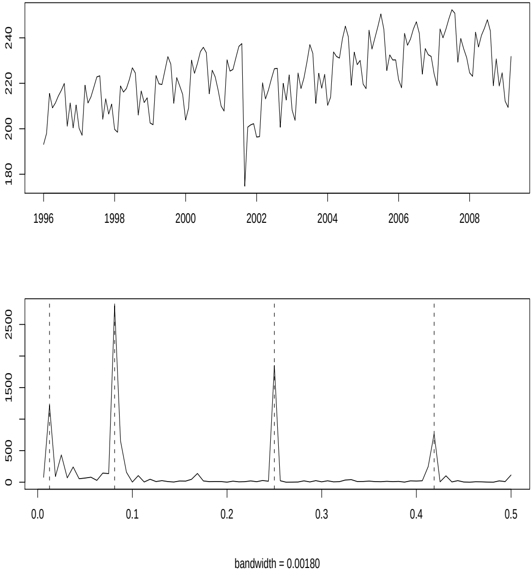


Figure 6: Square root transformation of the monthly air carrier traffic data (upper panel) and its

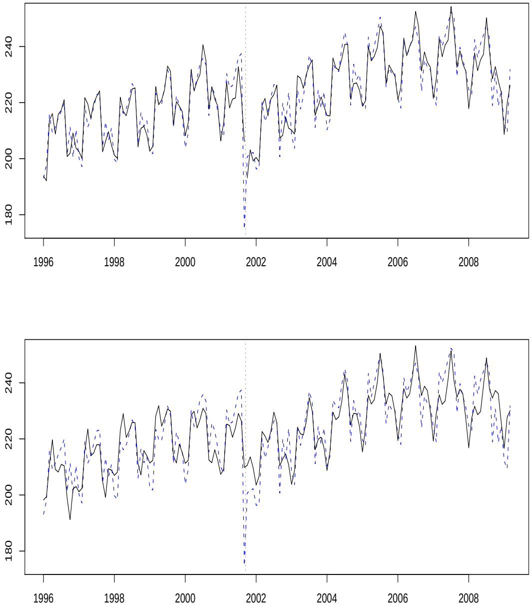


Figure 7: The fitted model based on the proposed data segmentation procedure (upper panel) and