Online changepoint test in a nonlinear expectile model

Computational and Methodological Statistics – CMStatistics 2022 | London, UK | 17–19 December, 2022

Matúš Maciak
Join work with G.Ciuperca & M.Pešta
University Lyon & Charles University
France & Czechia

CMStatistics 2022

Outline

1. **Underlying stochastic framework**

   Nonlinear conditional expectile regression model

2. **Formal statitical change point test**

   Real-time detection procedure for online data

3. **Asymptotic performance**

   Theoretical properties of the estimation and detection procedures

4. **Empirical performance**

   Real data illustrations and simulation results

The simplest shape to estimate...

- a **sparse location model** proposed by Harchaoui and Lévy-Leduc (2010)

$$Y_i = \mu_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N;$$

(Yao and Au (1989); Mammen and Van De Geer (1997); Massart (2004), Boysen (2009); Frick et al. (2014); Fryzlewicz (2014); Hyun et al. (2017); Lin et al. (2017), and others;)

The simplest shape to estimate...

- a **sparse location model** proposed by Harchaoui and Lévy-Leduc (2010)

$$Y_i = \mu_i + \varepsilon_i, \quad \text{for } i = 1, \dots, N;$$

(Yao and Au (1989); Mammen and Van De Geer (1997); Massart (2004), Boysen (2009); Frick et al. (2014); Fryzlewicz (2014); Hyun et al. (2017); Lin et al. (2017), and others;)

- a straightforward generalization for a **piece-wise linear** model

$$Y_i = a_i + b_i X_i + \varepsilon_i, \quad a_i, b_i \in \mathbb{R}, \quad i = 1, \dots, N;$$

(Feder (1975); Lerman (1980); Bosetti et al. (2008); Kim et al. (2009); Qui et al. (2009); Hudecová (2011); M. and Mizera (2016); M. (2017), Ciuperca and M. (2019), and others;)

The simplest shape to estimate...

- a **sparse location model** proposed by Harchaoui and Lévy-Leduc (2010)

$$Y_i = \mu_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N;$$

(Yao and Au (1989); Mammen and Van De Geer (1997); Massart (2004), Boysen (2009); Frick et al. (2014); Fryzlewicz (2014); Hyun et al. (2017); Lin et al. (2017), and others;)

- a straightforward generalization for a **multiple regression** model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_i + \varepsilon_i, \quad \boldsymbol{\beta}_i \in \mathbb{R}^p, \quad i = 1, \ldots, N;$$

(Leonardi and Bühlmann (2016); Qian and Su (2016); Ciuperca and M.(2018), and others)

The simplest shape to estimate...

- a **sparse location model** proposed by Harchaoui and Lévy-Leduc (2010)

$$Y_i = \mu_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, N;$$

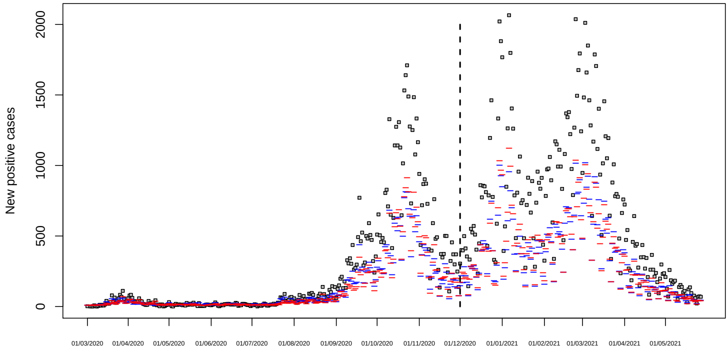
(Yao and Au (1989); Mammen and Van De Geer (1997); Massart (2004), Boysen (2009); Frick et al. (2014); Fryzlewicz (2014); Hyun et al. (2017); Lin et al. (2017), and others;)

- a straightforward generalization for a **general regression** model

$$Y_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}_i) + \varepsilon_i, \quad \boldsymbol{\beta}_i \in \mathbb{R}^p, \quad i = 1, \ldots, N;$$

for some **unknown** parametric **linear/nonlinear function** $f(\boldsymbol{x}, \boldsymbol{\beta})$
(with all kinds of different shape restrictions being possibly imposed on f);

Covid-19 cases in Prague, Czech Republic



- Covid-19 positive cases in Prague, Czech Republic
- Period from the first positive case (March 1, 2020) until end of May 2021
- Covid-19 restrictions and their role in the overall (global) population

Underlying stochastic (change point) model

- **historical data** $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \dots, m\}$, for q-dimensional $\boldsymbol{X}_i \in \mathbb{R}^q$;
- **underlying regression model** of the form

$$Y_i = f(\boldsymbol{X}_i, \beta) + \varepsilon_i, \qquad i = 1, \dots, m$$

$\hookrightarrow$ for a given nonlinear parametric function $f : \mathbb{R}^q \to \mathbb{R}$ and some **unknown vector of parameters** $\beta \in \mathbb{R}^p$ (to be estimated);

Underlying stochastic (change point) model

- **historical data** $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \ldots, m\}$, for q-dimensional $\boldsymbol{X}_i \in \mathbb{R}^q$;
- **underlying regression model** of the form

$$Y_i = f(\boldsymbol{X}_i, \beta) + \varepsilon_i, \qquad i = 1, \ldots, m$$

$\hookrightarrow$ for a given nonlinear parametric function $f : \mathbb{R}^q \to \mathbb{R}$ and some **unknown vector of parameters** $\beta \in \mathbb{R}^p$ (to be estimated);

- after historical data new **online data** $\{(Y_i, \boldsymbol{X}_i);\ i = m+1, \ldots, m+T_m\}$ are sequentially observed—in a one by one manner (for $T_m \in \mathbb{N}$);
- **analogous model as for the historical data** is supposed to hold, however

$$Y_i = f(\boldsymbol{X}_i, \beta_i) + \varepsilon_i, \qquad i = m+1, \ldots, m+T_m$$

$\hookrightarrow$ again for **unknown vectors of parameters** $\beta_i \in \mathbb{R}^p$ but some of them hypotetically different than $\beta \in \mathbb{R}^p$;

Formal statistical test of no change point

- **In the first step** the historical data $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \ldots, m\}$ are used to construct an empirical estimate $\widehat{\beta}_m$ for the unknown vector $\boldsymbol{\beta} \in \mathbb{R}^p$;

Formal statistical test of no change point

- **In the first step** the historical data $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \ldots, m\}$ are used to construct an empirical estimate $\widehat{\beta}_m$ for the unknown vector $\beta \in \mathbb{R}^p$;
- **In the second step** the online data $\{(Y_i, \boldsymbol{X}_i);\ i = m+1, \ldots, m+T_m\}$ are utilized to run a real-time change point test of the null hypothesis

$$H_0 : \quad \beta_i = \beta^0, \qquad i = m+1, \ldots, m+T_m$$

against the alternative hypothesis of the form

$$H_A : \exists k_m^0 \in \{1, \ldots, T_m\}$$

such that

$$\begin{array}{ll} \beta_i = \beta^0 & i = m+1, \ldots, m+k_m^0 \\ \beta_i = \beta^1 & i = m+k_m^0+1, \ldots, m+T_m \end{array}$$

where $\beta^0 \neq \beta^1$;

Formal statistical test of no change point

- **In the first step** the historical data $\{(Y_i, \boldsymbol{X}_i);\ i = 1, \ldots, m\}$ are used to construct an empirical estimate $\widehat{\beta}_m$ for the unknown vector $\beta \in \mathbb{R}^p$;
- **In the second step** the online data $\{(Y_i, \boldsymbol{X}_i);\ i = m+1, \ldots, m+T_m\}$ are utilized to run a real-time change point test of the null hypothesis

$$H_0 : \quad \beta_i = \beta^0, \qquad i = m+1, \ldots, m+T_m$$

against the alternative hypothesis of the form

$$H_A : \exists k_m^0 \in \{1, \ldots, T_m\}$$

such that

$$\begin{array}{ll} \beta_i = \beta^0 & i = m+1, \ldots, m+k_m^0 \\ \beta_i = \beta^1 & i = m+k_m^0+1, \ldots, m+T_m \end{array}$$

where $\beta^0 \neq \beta^1$;

- **Real-time (online)** testing procedures performed as online data arrive;

Step 1: Estimation of the parameter vector β

- **Conditional expectile** estimation in terms of the minimization problem

$$\widehat{\beta}_m = \underset{\beta \in \mathbb{R}^p}{Argmin} \sum_{i=1}^{m} \rho_\tau\Big(Y_i - f(\boldsymbol{X}_i, \beta)\Big)$$

$\hookrightarrow$ for the **expectile loss function** $\rho_\tau(x) = \left|\tau - \mathbb{I}_{\{x<0\}}\right| x^2$ for $x \in \mathbb{R}$;
(conditional expectiles known as the only coherent and elicitable risk measure)

Step 1: Estimation of the parameter vector β

- **Conditional expectile** estimation in terms of the minimization problem

$$\widehat{\beta}_m = \underset{\beta \in \mathbb{R}^p}{Argmin} \sum_{i=1}^{m} \rho_\tau \Big(Y_i - f(\mathbf{X}_i, \beta) \Big)$$

$\hookrightarrow$ for the **expectile loss function** $\rho_\tau(x) = \Big|\tau - \mathbb{I}_{\{x<0\}}\Big| x^2$ for $x \in \mathbb{R}$;
(conditional expectiles known as the only coherent and elicitable risk measure)

- **Assymptotic behaviour** of the expectile proposed estimator of $\beta^0 \in \mathbb{R}^p$:

$$\widehat{\beta}_m = \beta^0 + \mathbf{\Omega}^{-1} \frac{1}{m} \sum_{i=1}^{m} \nabla f(\mathbf{X}_i, \beta^0) g_\tau(\varepsilon_i) + o_{\mathbb{P}}(m^{-1/2})$$

for the sample size of the historical data tending to infinity, thus $m \to \infty$;

Step 1: Estimation of the parameter vector β

- **Conditional expectile** estimation in terms of the minimization problem

$$\widehat{\beta}_m = \underset{\beta \in \mathbb{R}^p}{Argmin} \sum_{i=1}^{m} \rho_\tau \Big(Y_i - f(\boldsymbol{X}_i, \beta) \Big)$$

$\hookrightarrow$ for the **expectile loss function** $\rho_\tau(x) = \Big|\tau - \mathbb{I}_{\{x<0\}}\Big| x^2$ for $x \in \mathbb{R}$;
(conditional expectiles known as the only coherent and elicitable risk measure)

- **Assymptotic behaviour** of the expectile proposed estimator of $\beta^0 \in \mathbb{R}^p$:

$$\widehat{\beta}_m = \beta^0 + \boldsymbol{\Omega}^{-1} \frac{1}{m} \sum_{i=1}^{m} \nabla f(\boldsymbol{X}_i, \beta^0) g_\tau(\varepsilon_i) + o_{\mathbb{P}}(m^{-1/2})$$

for the sample size of the historical data tending to infinity, thus $m \to \infty$;

- **Notation:**
  - $\boldsymbol{\Omega} = \mathbb{E}[h_\tau(\varepsilon)] \boldsymbol{V}(\beta^0)$ and $\boldsymbol{V}(\beta^0) = \lim_{m\to\infty} \boldsymbol{V}_m(\beta^0)$
  - $\boldsymbol{V}_m(\beta^0) = \frac{1}{m} \sum_{i=1}^{m} \nabla f(\boldsymbol{X}_i, \widehat{\beta}_m) \nabla^\top f(\boldsymbol{X}_i, \widehat{\beta}_m)$
  - $g_\tau(x) = \rho'_\tau(x)$ and $h_\tau(x) = \rho''_\tau(x)$

Step 2: Test of the null hypothesis H_0 against H_A

- **Null hypothesis:** the online data are generated under the same probabilistic model as the model generating the historical data ($\beta^0 \in \mathbb{R}^p$);
- **Alternative:** the online data are generated from a different model than the historical data however, the change is only determined within the paramter vector $\beta^1 \neq \beta^0$;

Step 2: Test of the null hypothesis H_0 against H_A

- **Null hypothesis:** the online data are generated under the same probabilistic model as the model generating the historical data ($\beta^0 \in \mathbb{R}^p$);
- **Alternative:** the online data are generated from a different model than the historical data however, the change is only determined within the paramter vector $\beta^1 \neq \beta^0$;

- **Test statistic**

$$\mathcal{T}(m) = \sup_{1 \leq k \leq T_m} \frac{\|S(m,k)\|_\infty}{z(m,k,\gamma)}$$

Step 2: Test of the null hypothesis H_0 against H_A

- **Null hypothesis:** the online data are generated under the same probabilistic model as the model generating the historical data ($\beta^0 \in \mathbb{R}^p$);
- **Alternative:** the online data are generated from a different model than the historical data however, the change is only determined within the paramter vector $\beta^1 \neq \beta^0$;

- **Test statistic**

$$\mathcal{T}(m) = \sup_{1 \leq k \leq T_m} \frac{\|S(m,k)\|_\infty}{z(m,k,\gamma)}$$

- **Notation:**
  - $S(m,k) = \boldsymbol{J}_m^{-1/2}(\widehat{\beta}_m) \sum_{i=m+1}^{m+k} \nabla f(\boldsymbol{X}_i, \widehat{\beta}_m)_\tau(\widehat{\varepsilon}_i)$
  - $\boldsymbol{J}_m(\widehat{\beta}_m) = \frac{Var[g_\tau(\varepsilon)]}{m} \sum_{i=1}^{m} \nabla f(\boldsymbol{X}_i, \widehat{\beta}_m) \nabla^\top f(\boldsymbol{X}_i, \widehat{\beta}_m)$
  - $z(m,k,\gamma) = m^{1/2}(1 + k/m)(k/(k+m))^\gamma$, for some $\gamma \in [0, \frac{1}{2})$

Change point test asymptotics

- Distinguishing for **two differrent scenarios**:
  - **Open-end procedure:** $\lim_{m \to \infty} T_m/m = \infty$
  - **Closed-end procedures:** $\lim_{m \to \infty} T_m/m = T$, for $T \in (0, \infty)$

Change point test asymptotics

- Distinguishing for **two differrent scenarios**:
  - **Open-end procedure:** $\lim_{m\to\infty} T_m/m = \infty$
  - **Closed-end procedures:** $\lim_{m\to\infty} T_m/m = T$, for $T \in (0,\infty)$

- **Distribution of $\mathcal{T}(m)$ under the null hypothesis:**

  Under some technical assumptions and the null hypothesis validity

$$\mathcal{T}(m) = \sup_{1\leq k\leq T_m} \frac{\|S(m,k)\|_\infty}{z(m,k,\gamma)} \xrightarrow[m\to\infty]{\mathcal{D}} \sup_{0<t<L(T)} \frac{\|\boldsymbol{W}_p(t)\|_\infty}{t^\gamma}$$

  for a p-dimensional Wiener process $\{\boldsymbol{W}_p(t);\ t\in(0,\infty)\}$ and either $L(T)=1$ (open-end) or $L(T)=T/(T+1)$ (closed-end);

Change point test asymptotics

- Distinguishing for **two differrent scenarios**:
  - **Open-end procedure:** $\lim_{m\to\infty} T_m/m = \infty$
  - **Closed-end procedures:** $\lim_{m\to\infty} T_m/m = T$, for $T \in (0,\infty)$

- **Distribution of $\mathcal{T}(m)$ under the null hypothesis:**
  Under some technical assumptions and the null hypothesis validity

$$\mathcal{T}(m) = \sup_{1\le k\le T_m} \frac{\|S(m,k)\|_\infty}{z(m,k,\gamma)} \xrightarrow[m\to\infty]{\mathcal{D}} \sup_{0<t<L(T)} \frac{\|W_p(t)\|_\infty}{t^\gamma}$$

  for a p-dimensional Wiener process $\{W_p(t);\ t\in(0,\infty)\}$ and either $L(T)=1$ (open-end) or $L(T)=T/(T+1)$ (closed-end);

- **Consistency of the test (behaviour of $\mathcal{T}(m)$ under the alternative:**
  Under the alternative hypothesis and $m^{1/2}\|\beta^0-\beta^1\|_2\to\infty$ for $m\to\infty$

$$\mathcal{T}(m) \xrightarrow[m\to\infty]{\mathbb{P}} \infty.$$

Simulation setup

Motivated by Choi, S.H., Kim, H.K., Lee, Y. (2003). Nonlinear asymmetric least squares estimators. Journal of the Korean Statistical Society 32(1), 47 – 64.

- Gompertz curve: $f(x, \beta) = \exp\{-\beta_1 e^{-\beta_2 x}\}$, for $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top \equiv (10, 5)^\top$
- Estimation of $\beta \in \mathbb{R}^2$ by the iterative grid search algorithm
- Historical data: $m \in \{20, 50, 200\}$; Online data: $T_m \in \{10, m/2, m \log m\}$
- Various change point scenarios wrt. to β^1 and the change point location

Under the null hypothesis

Distribution	m	$\widehat{\beta}_1$ | Std.Er. $\beta_1 = 10.00$	$\widehat{\beta}_2$ | Std.Er. $\beta_2 = 5.00$	Scenario 1 $T_m = 10$	Scenario 2 $T_m = m/2$	Scenario 3 $T_m = m\log(m)$
$N(0,1)$	20	10.52 *(2.832)*	5.32 *(1.825)*	7.74 %	7.74 %	7.52 %
	50	10.32 *(2.857)*	5.20 *(1.382)*	4.92 %	6.08 %	5.64 %
	200	10.24 *(2.766)*	5.03 *(0.808)*	5.08 %	5.58 %	6.54 %
$N(1,1)$	20	10.18 *(2.889)*	5.61 *(2.020)*	5.76 %	5.76 %	6.28 %
	50	10.27 *(2.869)*	5.39 *(1.638)*	4.08 %	3.26 %	4.46 %
	200	10.30 *(2.833)*	5.08 *(0.977)*	4.40 %	4.26 %	4.88 %
$L(0,1)$	20	10.50 *(2.842)*	5.32 *(1.797)*	7.84 %	7.84 %	9.86 %
	50	10.37 *(2.850)*	5.18 *(1.374)*	4.08 %	5.12 %	7.38 %
	200	10.29 *(2.769)*	5.02 *(0.799)*	5.06 %	4.90 %	5.58 %

Table 1 Simulation results under the null hypothesis and $\tau = 1/2$ (conditional mean). The parameter estimates are reported with the corresponding standard errors (in parentheses) over 5000 Monte Carlo simulations. Relative proportions of false rejections are given for three different scenarios for T_m reflecting the open-end and closed-end procedures. The nominal level of all the tests is always set to be $\alpha = 0.05$.

Under the alternative hypothesis

Dist.	m	k_m^0	Scenario 1 $T_m = 10$		Scenario 2 $T_m = m/2$		Scenario 3 $T_m = m\log(m)$	
$N(0,1)$	20	$k_m^{(1)}$	2.79 %	[0.77 | 0.78]	2.79 %	[0.77 | 0.78]	2.44 %	[0.72 | 0.71]
		$k_m^{(2)}$	12.80 %	[0.31 | 0.22]	12.80 %	[0.31 | 0.22]	20.40 %	[0.19 | 0.12]
	50	$k_m^{(1)}$	4.10 %	[0.73 | 0.76]	4.10 %	[0.76 | 0.78]	5.20 %	[0.93 | 0.93]
		$k_m^{(2)}$	13.30 %	[0.33 | 0.11]	14.60 %	[0.44 | 0.46]	35.30 %	[0.32 | 0.25]
	200	$k_m^{(1)}$	6.52 %	[0.87 | 0.89]	8.99 %	[0.80 | 0.82]	21.76 %	[0.83 | 0.85]
		$k_m^{(2)}$	16.70 %	[0.46 | 0.44]	47.60 %	[0.46 | 0.49]	96.40 %	[0.15 | 0.10]
$N(1,1)$	20	$k_m^{(1)}$	2.51 %	[0.71 | 0.67]	2.51 %	[0.71 | 0.67]	2.42 %	[0.62 | 0.55]
		$k_m^{(2)}$	11.20 %	[0.37 | 0.39]	11.20 %	[0.37 | 0.39]	23.30 %	[0.29 | 0.22]
	50	$k_m^{(1)}$	4.20 %	[0.83 | 0.83]	4.30 %	[0.83 | 0.83]	5.00 %	[0.85 | 0.89]
		$k_m^{(2)}$	12.20 %	[0.43 | 0.46]	13.30 %	[0.49 | 0.33]	32.30 %	[0.48 | 0.58]
	200	$k_m^{(1)}$	5.52 %	[0.82 | 0.84]	7.20 %	[0.55 | 0.55]	21.71 %	[0.83 | 0.85]
		$k_m^{(2)}$	14.21 %	[0.42 | 0.44]	38.00 %	[0.48 | 0.49]	93.80 %	[0.18 | 0.15]
$L(0,1)$	20	$k_m^{(1)}$	2.64 %	[0.77 | 0.78]	2.64 %	[0.77 | 0.78]	2.29 %	[0.74 | 0.78]
		$k_m^{(2)}$	11.50 %	[0.36 | 0.22]	11.50 %	[0.36 | 0.22]	21.20 %	[0.24 | 0.16]
	50	$k_m^{(1)}$	3.95 %	[0.75 | 0.78]	4.10 %	[0.78 | 0.79]	4.99 %	[0.81 | 0.86]
		$k_m^{(2)}$	12.90 %	[0.43 | 0.44]	13.20 %	[0.48 | 0.50]	37.70 %	[0.30 | 0.21]
	200	$k_m^{(1)}$	6.70 %	[0.75 | 0.67]	6.93 %	[0.80 | 0.79]	24.04 %	[0.81 | 0.83]
		$k_m^{(2)}$	15.60 %	[0.40 | 0.47]	37.90 %	[0.46 | 0.47]	96.40 %	[0.18 | 0.12]

Table 2 Empirical powers of the proposed real-time changepoint test

Covid-19 positive cases in Prague

Motivated by Chen, D.G., Chen, X., Chen, J.K. (2020.) Reconstructing and forecasting the covid-19 epidemic in the united states using a 5-parameter logistic growth model. Global Health Research and Policy 32(1), 1 – 7.

- Gompertz model $f(x, \beta) = K \exp\{-\beta_1 e^{-\beta_2 x}\}$, for $\beta = (\beta_1, \beta_2, K)^\top \in \mathbb{R}^3_+$
- Historical data: $m = 1275$; Online data: $T_m = 176$
- The null hypothesis rejected on the third day of the online data
  (test statistics $\mathcal{T}(m) = 3.4211$ with the critical value $c_{0.95}((\gamma) = 2.4260$ and $\gamma = 0.1$)

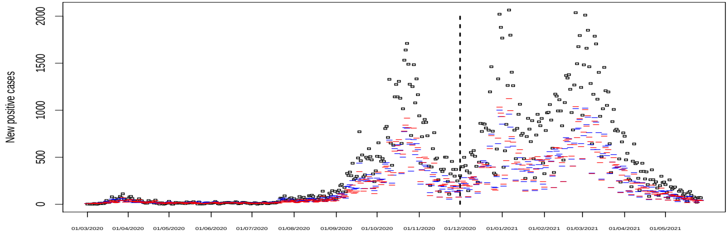
Covid-19 positive cases in Prague

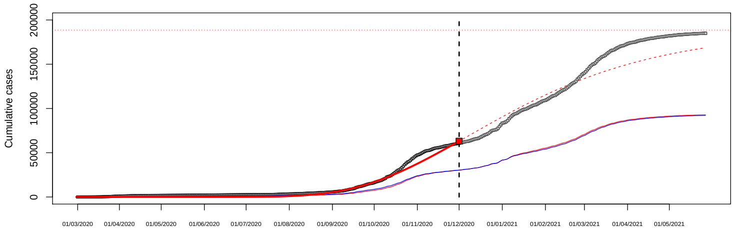
Motivated by Chen, D.G., Chen, X., Chen, J.K. (2020.) Reconstructing and forecasting the covid-19 epidemic in the united states using a 5-parameter logistic growth model. Global Health Research and Policy 32(1), 1 – 7.

- Gompertz model $f(x, \beta) = K \exp\{-\beta_1 e^{-\beta_2 x}\}$, for $\beta = (\beta_1, \beta_2, K)^\top \in \mathbb{R}_+^3$
- Historical data: $m = 1275$; Online data: $T_m = 176$
- The null hypothesis rejected on the third day of the online data (test statistics $\mathcal{T}(m) = 3.4211$ with the critical value $c_{0.95}((\gamma) = 2.4260$ and $\gamma = 0.1$)

Estimation method & Data		β_1	β_2	K	Objective function
Historical data (until 01/12/2020)					
Symmetric least squares	$(\tau = 0.50)$	36.04	0.0129	187 811	3.00×10^6
Expectile method	$(\tau = 0.11)$	37.97	0.0129	188 576	0.78×10^6
Re-estimation after the change detection (04/12/2020)					
Symmetric least squares	$(\tau = 0.50)$	35.38	0.0130	180 174	3.27×10^6
Expectile method	$(\tau = 0.19)$	36.04	0.0129	179 718	1.92×10^6
All available data (until 26/05/2021)					
Symmetric least squares	$(\tau = 0.50)$	20.17	0.0096	256 970	10.9×10^6
Expectile method	$(\tau = 0.26)$	20.18	0.0095	255 032	8.53×10^6

Covid-19 positive cases in Prague





Real-time change point detection in a nonlinear expectile model

- **Nonlinearity of the model**
  (very good model flexibility while preserving simple interpretation)
- **Conditional expectile estimation**
  (robustness wrt. assymetric error distributions, risk modeling)
- **Consistent change point detection**
  (distribution of the null hypothesis does not depend on the functional form of the underlying model nor the unknown parameters)
- **Straightforward applicability**
  (some caution is needed when using different functional models)

Thank you for your attention!