Change point detection in mobile advertising

Nina Golyandina, Kliment Merzlyakov

Saint Petersburg State University
Mathematical faculty
Applied statistics department

Saint Petersburg
2018

Content

- Change point detection
  - What is change point detection?
  - Real world examples of change point detection
  - Reasons to detect change points
- Change point detection techniques
- Airpush cases
  - Fraud elimination
  - Trend extraction
  - Smart alerts

Change point detection

What is change point detection?

- Change point — point in time series where some significant change occurred
- Change point detection — group of methods to find change points in time series
- Change points can be two types:
  - Local — one time change
  - Global — change of time series structure

What is change point detection?

Let's formalize change point detection problem:

- Suppose $\mathrm{X} = (x_1, ..., x_N)$ — time series of length N
- $\mathbf{t} = \{t_1, t_2, ...\} \subset \{1, ..., K\}$ — set of change points indexes
- $K - 1$ — amount of changes in time series, which can be either known or unknown
- Change point detection problem — is a problem of searching such set of indexes, which would be the closest to real set of change point indexes: $\widehat{\mathbf{t}} = \{\widehat{t_1}, \widehat{t_2}, ...\}$
- To estimate change point detection algorithm quality we need tagged time series with known indexes $\widehat{\mathbf{t}} = \{\widehat{t_1}, \widehat{t_2}, ...\}$
- Quality can be estimated as sum of distance between real indexes and estimated indexes normalized by length of time series:

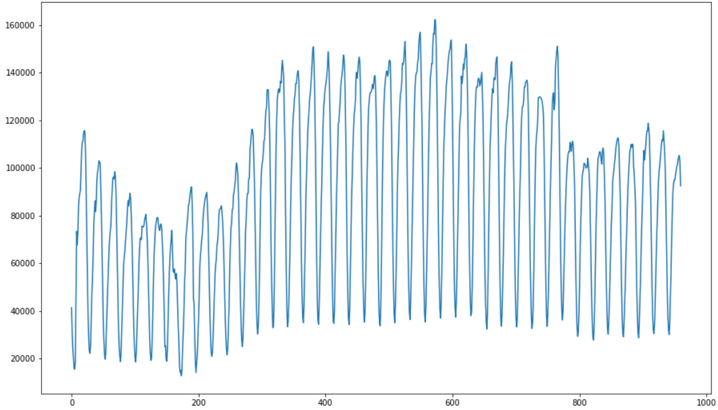
$$\sum_{i=1}^{K} \frac{|\widehat{t_i} - t_i|}{N}$$

Types of change points

- Trend change
- Mean change
- Variance change
- Single point change
- Period change
- Structure change

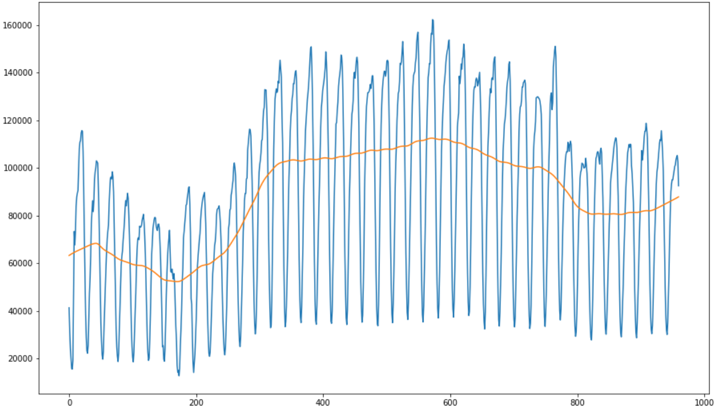
Change point examples

Common graph



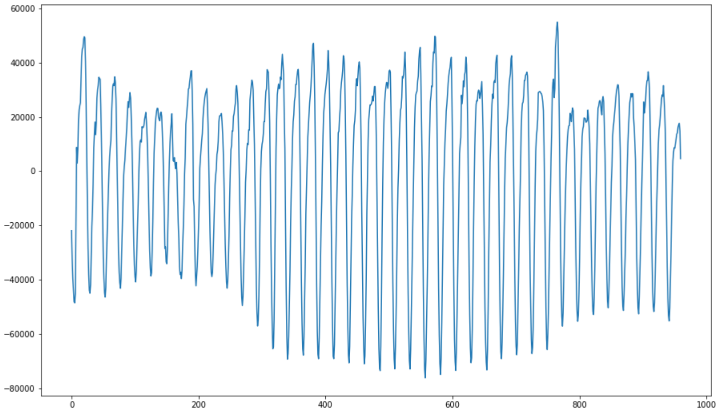
Change point examples

Common graph with trend



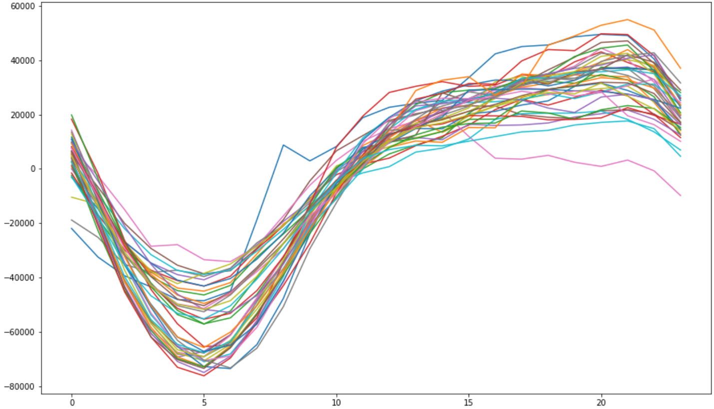
Change point examples

Common graph without trend



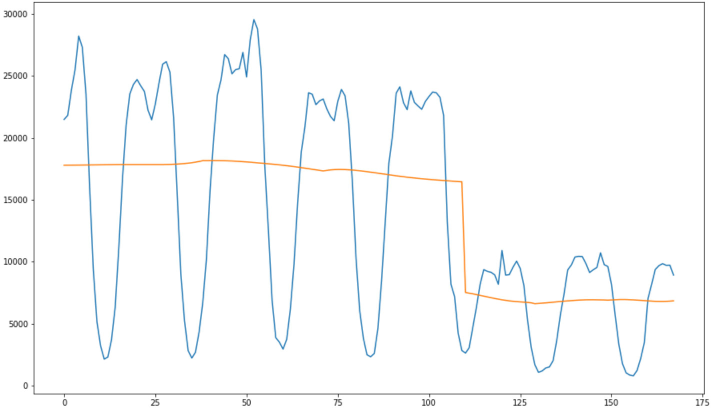
Change point examples

Common graph periodic frequency



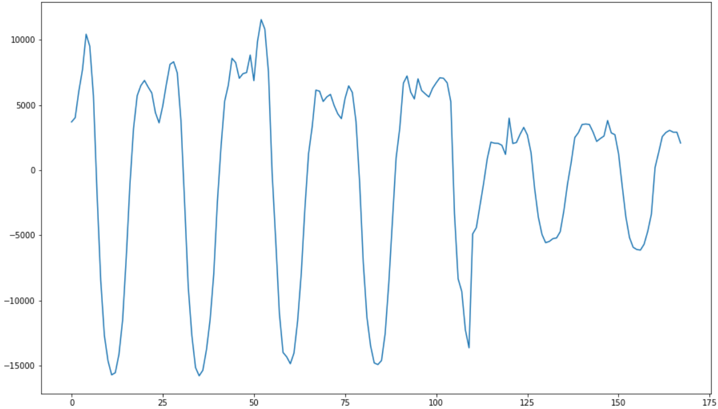
Change point examples

Mean change



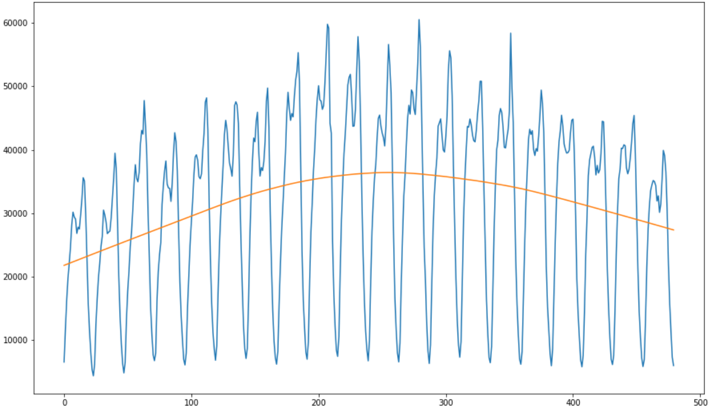
Change point examples

Variance change



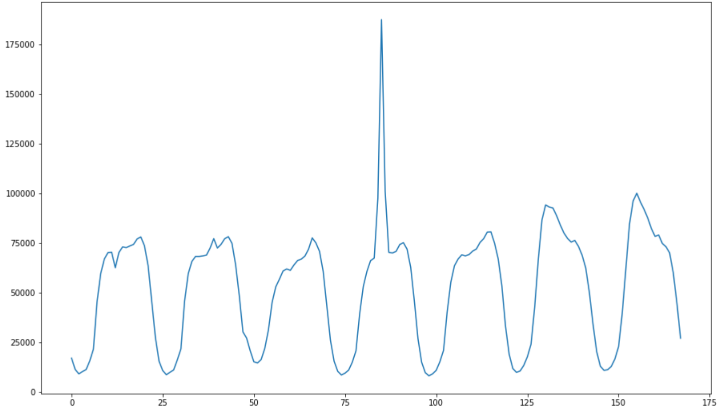
Change point examples

Trend change



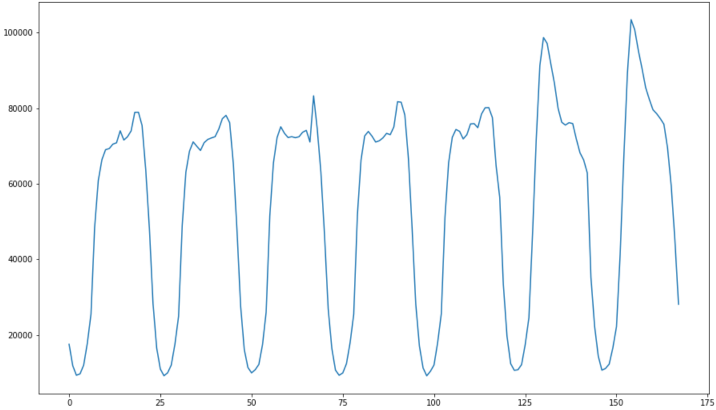
Change point examples

Point change



Change point examples

Structure change



Reasons to detect change points

- Searching issues in historical data
- Reacting on changes quickly
- Extracting trend more accurately

Change point detection techniques

SSA for change point detection

Singular spectrum analysis (SSA) method can be used to allocate change points. Let $x_1, x_2, \ldots$ be a time series and N, L, l, p, q be fixed integers so that $0 \le l < L \le N/2$ and $0 \le p < q$. The SSA change point detection algorithm is as follows. For each $n = 0, 1, \ldots$ we compute:

- the base matrix $\mathbb{X}^{(n)}$,
- the lag-covariance matrix $\mathbb{R}_n = \mathbb{X}^{(n)}(\mathbb{X}^{(n)})^\top$,
- the SVD of $\mathbb{R}_n$,
- $\nu_{n.I,p,q}$, the sum of the squared Euclidean distances between the vectors $\mathbb{X}_j^{(n)}(j = p+1, \ldots, q$ and the l-dimensional subspace $L_{n,I}$, and
- S_n, the normalized squared distance.

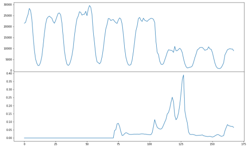
Large values of S_n indicate a change in the time series.
This method has the following parameters:

- N — window size for base matrix
- L, I — SSA window and number of eigenvectors
- p, q — window size for test matrix
- threshold

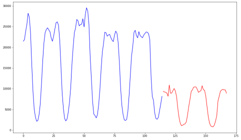
SSA in action. Mean change

Let's try to apply SSA algorithm to real time series mentioned above. With manually settled parameters.

Mean change

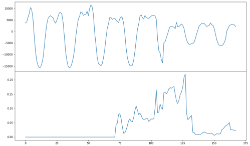


Detection $N = 48$, $L = 24$, $l = 5$, $p = 49$, $q = L$, threshold $= 0.15$

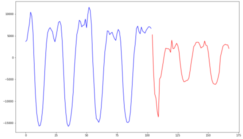


Works good for mean change

SSA in action. Variance change

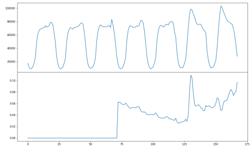


$N = 48$, $L = 24$, $l = 5$, $p = 49$, $q = L$, threshold $= 0.15$

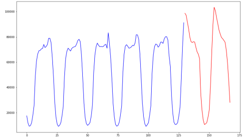


Works good for variance change

SSA in action. Period change

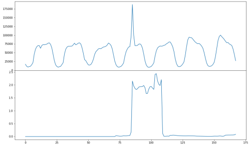


$N = 48$, $L = 24$, $l = 5$, $p = 49$, $q = L$, threshold $= 0.1$

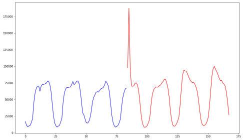


Works well for period change

SSA in action. Point change



N = 48, L = 24, I = 5, p = 49, q = L, threshold

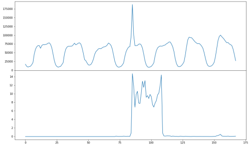


= 0.1

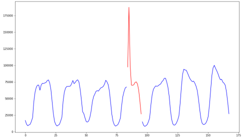
Doesn't work for point change. The reason is that we always move window from left to the right.

SSA in action. Point change

Let's try to reverse the time series each time we find change point.



$N = 48$, $L = 24$, $l = 5$, $p = 49$, $q = L$, threshold $= 0.1$

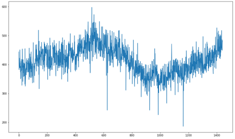


Using this upgrade we make SSA algorithm working for point change as well

Airpush cases

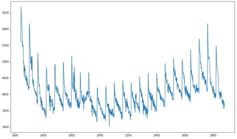
Airpush cases. Fraud elimination

- Apps minutely requests data
- Red flag: strong pattern.
- Can be a automated bots behind pattern



Clean application

Fraud application



Airpush cases. Fraud elimination

- Goal: to be able to find such apps automatically
- We can reach this goal using frequency analysis

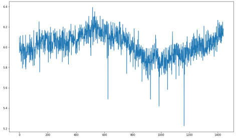
The framework can be described as follows:

1. Apply logarithm to time series to stabilyze amplitude
2. Remove trend (low frequent part) from time series
3. Apply Fourier transform to time series
4. Estimate the distribution of periodogram values
5. Compare distributions of each application with a distribution of white noise (which is exponential) using Kullback–Leibler divergence
6. If divergence $>$ threshold, then application is marked as suspicious

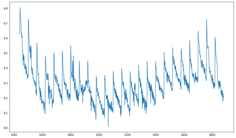
Airpush cases. Fraud elimination

1. Apply logarithm

Clean application



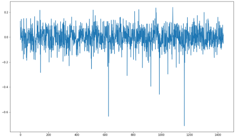
Fraud application



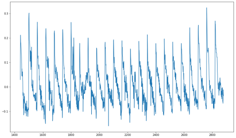
Airpush cases. Fraud elimination

2. Remove trend

Clean application



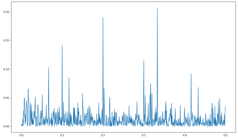
Fraud application



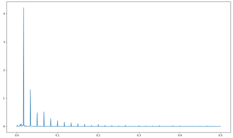
Airpush cases. Fraud elimination

3. Apply Fourier transform

Clean application



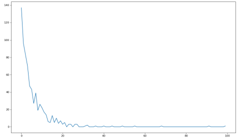
Fraud application



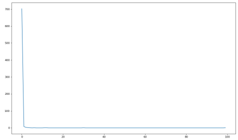
Airpush cases. Fraud elimination

4. Estimate the distribution of periodogram values

Clean application



Fraud application



Airpush cases. Fraud elimination
5. Compare distributions

- Clean app score: 0.09
- Fraud app score: 1.87