Online Multiscale Method for Change Detection in Automated Data-Quality Monitoring

Ronglong Fang [1]
in collaboration with Markus Diefenthaler[2] Abdullah Farhat [1] Holly Szumila-Vance[2] Yuesheng Xu [1]

[1]Old dominion university, Norfolk, VA, USA.

[2]Jefferson lab, Newport news, VA, USA

Streaming Readout X: Jefferson Lab, May 17–19

Overview

Why online automated data-quality monitoring is important?

1. Increase reliability of the data

   The quality of data analysis depends on the quality of data. In most data science project, data scientists need to clean data first. Monitor data is one way to clean data.

2. Find and fix issues on time.

   The online data-quality monitoring can find problems in data while data taking. Hence, the problem in data can be found on time. In experimental physics, it possible takes a year or longer to obtain data. So the stable and believable of the detector is very important.

To deal with sequential data, our goal is :

1. detect when a change occurs.
2. detect what kind of change occurs, e.g., sudden change occurs or linear gradual change occurs.
3. determine which data to keep and which to drop.
4. update models when significant changes is detected.

Our approach

1. Using multiscale basis to represent the data.
   (a) If no change happens, the coefficient of the multiscale bases is close to zeros.
   (b) If change happens, the coefficients of the multiscale basis is far away from zero

   change in raw data set $\rightarrow$ outlier in coefficients set

2. Detect outlier in the coefficients set.

3. If the coefficient is outlier in the coefficients set, then change happens in the support of the coefficient.

Multiscale representation for the raw data

For a fixed scale parameter k and the data set

$$\left[d_j, d_{j+1}, \ldots, d_{j+2^k-1}\right],$$

the coefficient for this data is defined as

$$a_j^k = \frac{1}{2^{2k}} \sum_{i=0}^{2^k-1} d_{j+i}\psi_k(x_i), \tag{1}$$

where ψ_k is the k-th level basis, the $x_i := \frac{i}{2^{2k}} + \frac{0.5}{2^{2k}}$ are the corresponding discrete nodes, $\left[d_j, d_{j+1}, \ldots, d_{j+2^k-1}\right]$ the is the support of the coefficient a_j^k.

Requirements for the multiscale basis

Let's use ψ defined on $[0,1]$ to denote the basis function. The support of ψ is defined as

$$\text{supp}\,(\psi) := \{x \in [0,1] : \psi(x) \neq 0\}$$

The basis function ψ has the following properties :

1. Vanishing moment [1] property of order n, that is

$$\int_0^1 \psi(x) x^j dx = 0, \ j = 0, 1, \ldots, n-1. \tag{2}$$

2. Local support, i.e., the support of ψ is a subset of $[0,1]$.

The choice of basis

The first one is piecewise constant test function with order 1 vanishing moment.

$$\psi(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1/2 \\ -1, & 1/2 < x \leqslant 1 \end{cases} \tag{3}$$

The second one is piecewise linear test function with order 2 vanishing moment.

$$\psi(x) = \begin{cases} 1 - 4x, & x \in [0, \frac{1}{2}] \\ 4x - 3, & x \in (\frac{1}{2}, 1] \end{cases} \tag{4}$$

To obtain the local support test functions (different level test functions), we can shrink and shift the original test function (the details can be found in the book [1]).

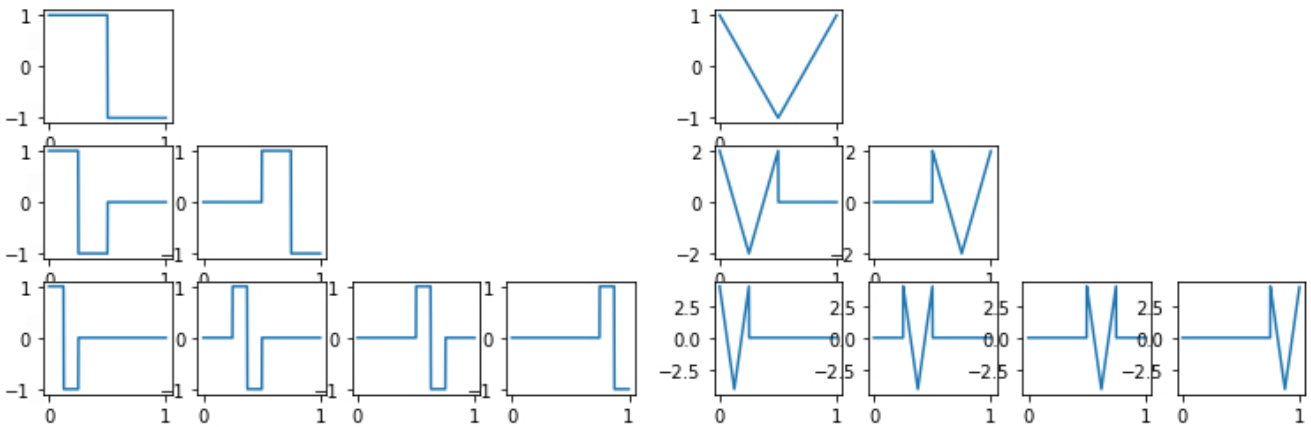

Figure 1: Local support test function with level 1, 2, 3.

Online multiscale structure

The structure of the online multiscale algorithm can be described as following

$$
\begin{aligned}
&\underbrace{d_0, d_1}_{a_0^1} \\
&\overbrace{d_0, d_1, \underbrace{d_2, d_3}_{a_2^1}}^{a_0^2} \\
&d_0, d_1, d_2, d_3, \underbrace{d_4, d_5}_{a_4^1} \\
&\underbrace{d_0, d_1, d_2, d_3, \overbrace{d_4, d_5, \underbrace{d_6, d_7}_{a_6^1}}^{a_4^2}}_{a_0^3}
\end{aligned}
\tag{5}
$$

Online multiscale change detection algorithm

Algorithm 1: Online multiscale change detection

Data: sequential data $d_0, d_1, \cdots, d_i, \cdots$; test function ψ; the minimum scale k_{min} and the maximum scale k_{max}. W denotes the empty data set, $W :=$ {add the first $2^{k_{max}}$ data}; Let $a := \{a_{k_{min}}, a_{k_{min}+1}, \cdots, a_{k_{max}}\}$ denotes the coefficient for each scale and each element is an empty set; let CInter $:= \{\text{CInter}_{k_{min}}, \text{CInter}_{k_{min}+1}, \cdots, \text{CInter}_{k_{max}}\}$ denotes the change interval.

Result: The change interval in the data set CInter.

$\ell = 2^{k_{max}-k_{min}}$,
for $j = \ell, \ell+1, \cdots$ **do**
  $W := W \cup \{$ the latest $2^{k_{min}}$ data $\}$,
  for $m = k_{min}, k_{min}+1, \ldots, k_{max}$ **do**
    if $floor\left(\frac{len(W)}{2^m}\right) - floor\left(\frac{j}{2^{m-k_{min}}}\right) > 0$ **then**
      Let $s_0, s_1, \ldots, s_{2^m-1}$ to denote the latest 2^m new data, calculate the coefficients for this data set

$$a_m = a_m \cup \{a_{\text{new}}\}$$

      where $a_{new} = \frac{1}{2^{2m}} \sum_{i=0}^{2^m-1} \psi_m(x_i) s_i$.
      Detect whether a_{new} is an outlier.
      if a_{new} *is an outlier in the the set* a_m **then**
        store the begin index and the end index of the corresponding data to CInter_m
      end
    end
  end
end

Remark

1. *In each scale, we check weighted the summation* (1). *If the summation is small, then we conclude no change happens in its support. If the summation is big (outlier), then we conclude some change (sudden change or linear change) happens its support.*

2. *For the sequential data, we begin with the small scale, and when we obtain more data, we increase the scale. The design of the algorithm is suitable for monitoring the sequential data.*

3. *If the change detected in a small scale, then we can conclude the change happens in a small set, which means the detection results is more accurate. So we start with accurate scale, and then detect for next scale.*

Outlier detection

Once we obtain the coefficient, we need to detect the outlier in the coefficient set. We use the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ of the coefficient set to define outlier, i.e.,

$$\text{Out}_t := \{a : a \notin [\hat{\mu} - t\hat{\sigma}, \hat{\mu} + t\hat{\sigma}]\} \tag{6}$$

where t is the threshold need to be predefined. For larger threshold t, our results is more reliable, but the accuracy is lower. For small threshold t, the results is less reliable, but the accuracy is greater.

GEM data

1. GEM data is used to track and reconstruct the momentum of the proton.

2. The data we use the the channel 'moudle 1 strip 300 axis 1' from the data set 'hit_0_bbgem_304'. The data is obtain from SAMPA system and GEM detector in Jlab, and it is used to study cosmic ray.

3. The size of data is 30006912.

4. We add sudden change and linear gradual change artificially. The goal is to detect the change in the data set.

Original Data and changed data

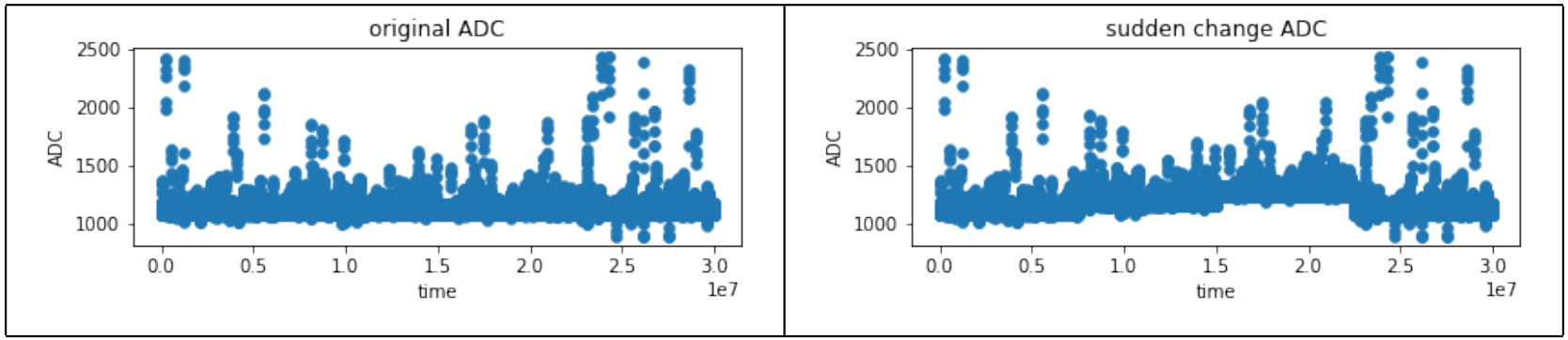

Figure 2: Original data and sudden changed data

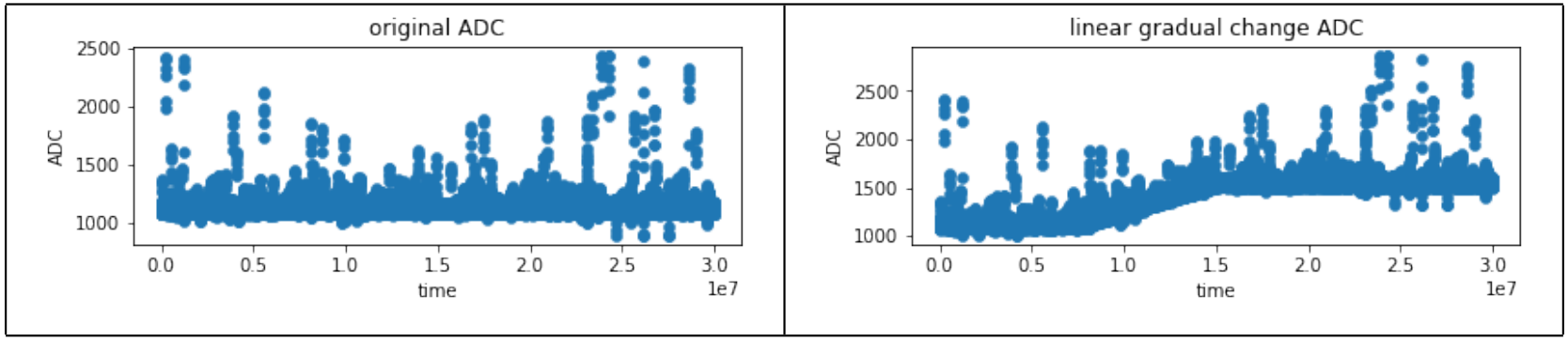

Figure 3: Original data and linear gradual changed data

Sudden change result

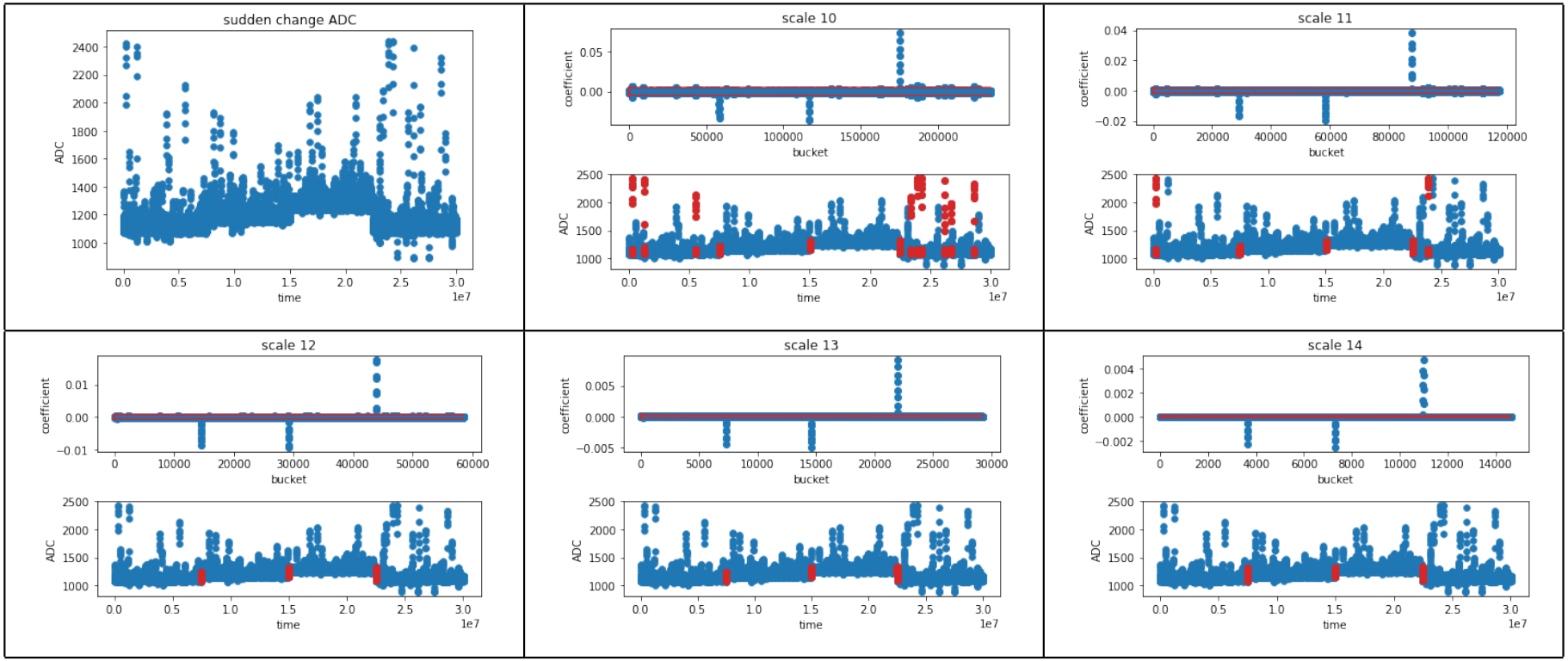

Figure 4: Sudden change results : scale 10-14, threshold 10.

The multiscale representation magnifies the sudden change and shrinks the noise in the raw data set. When we increase the scale, the results is more reliable, however, the accuracy goes down.

Sudden change result

1. the coefficient represent the summation of (1). For the sudden change, we choose the piecewise constant wavelet test function (3). The upper red and low red bound is determined by the mean and standard deviation of the coefficient.
2. If the coefficient lies out of the bound, we conclude change happens. The red point in the raw data set is the changeable data interval and the blue point the unchangeable data interval.
3. In the scale 10-11, most peaks has been detected in the algorithm. In scale 12-14, the sudden changes has been detected. So the accuracy for sudden is $2^{12} = 4096$ to $2^{14} = 16384$.

Gradual change results

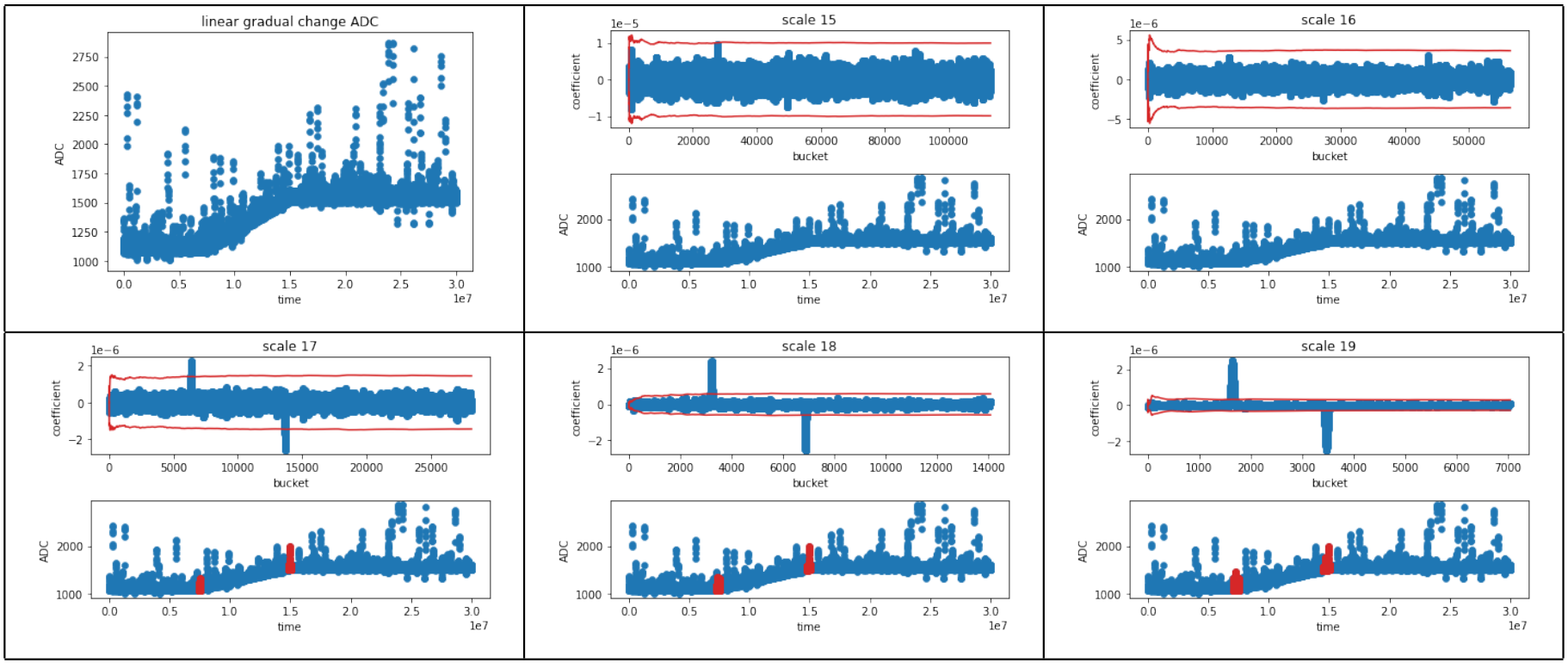

Figure 5: linear gradual change results : scale 15-19

For the linear gradual change, we got a similar results. The gradual change is been detected in a higher level.

SBS data

1. The SBS data is used to track proton form factors for high Q2.

2. There are 8 run data set, each runs store the different type of data.

3. In each run, the length of data is 200001. Hence, the total length is 1600008, see Table 1.

4. The goal is to detect the difference between different runs.

The type of data

run	current [uA]	target/type	HV
13420	0	cosmic	nominal
13598	1	LD2	nominal
13599	3	LD2	nominal
13600	3	LD2	reduced-1500V
13601	5	LD2	nominal
13602	7	LD2	nominal
13603	7	LD2	reduced-1500V
13604	7	LD2	0V

Table 1: run list

SBS data

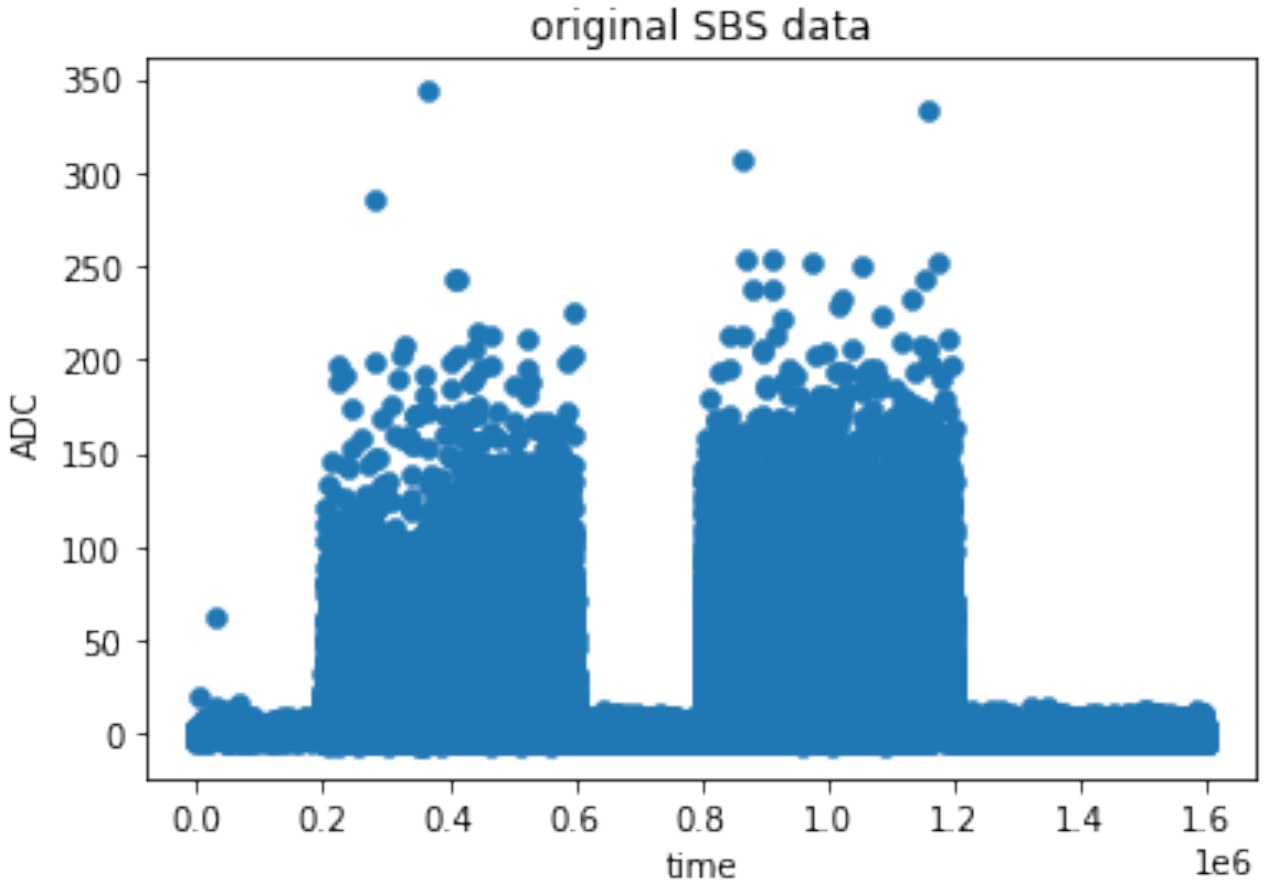

Figure 6: SBS data

The difference between different can be viewed as change.

Detect results

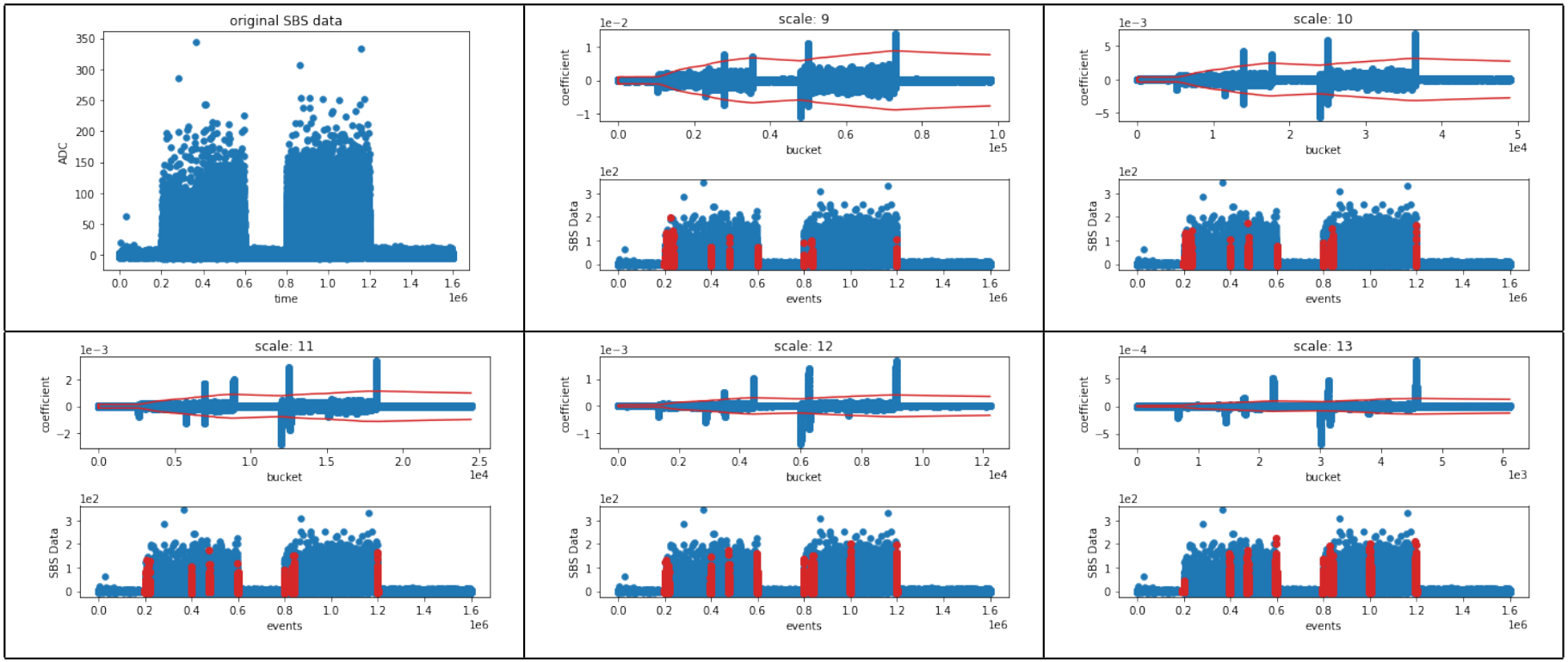

Figure 7: SBS data results : scale 15-19, threshold 10.

Summary

1. Using multiscale basis to represent the raw data set.

   change in raw data set $\rightarrow$ outlier in coefficient set

   In the representation, the change in the data has been magnified and the noise in the data has been shrinked.

2. Develop online multiscale method, which is suitable for the sequential data.

Thank you


Z. Chen, C. A. Micchelli, and Y. Xu, *Multiscale methods for Fredholm integral equations*, vol. 28, Cambridge University Press, 2015.