Regression-based detection and analysis of a change point

Karel Kupka, (kupka@trilobyte.cz), Milan Meloun (milan.meloun@upce.cz)
TriloByte Statistical Software, CQR: Center for Quality and Reliability

Abstract

In this paper we discuss statistical tools for detection of change of an analytical signal caused by unexpected change of analyte concentration. To detect change it is advantageous to use methods based on cumulative sums and especially the CUSUM control chart technique was found to be useful. In a noisy signal however, it is hard to decide when in the time series the change took place and to locate the source or possible responsibility of the change. Here, we suggested a nonlinear regression model with a change-point parameter. The advantage of the described procedure is its capability to locate the time of the change with better accuracy and with statistical confidence intervals at a given confidence level.

Introduction

Change point detection is a very important problem in quality control and in any situation where one is to find, prove and analyse usually unexpected change in process. It is especially important when this change may lead to loss, damage, or dangerous situations. In analytical signal analysis we can distinguish two situations: (1) detection of instability of the measuring process itself due to technical malfunction or methodological mistakes in the analytical procedure or instrument and (2) change in the analytical signal due to change of analyte concentration in monitored process (such as waste water, environment, technological processes).

Theoretical

Cusum control charts

Methods of statistical process control based on cumulative sums of deviations (cusum charts) have been used widely to diagnose departure from the target or mean value in normally distributed $N(\mu, \sigma^2)$ manufacturing process signals. In 1956 the method was adopted by Page in cusum control charts, which was further modified by Lucas in Shewhart-like cusum control charts. Compared to classical Shewhart charts, the Cusum is much more sensitive to process shift, where process mean μ_0 changes by a shift d at a time t_0. For $d = \sigma$ the cusum charts needs only 8 measurements to detect the shift compared to 44 measurements in a Shewhart chart. Detailed statistical procedures for cusum have been described elsewhere. The values plotted in the cusum chart are S_H and S_L,

$$S_H(i) = \max\left[0, d_i - k + S_H(i-1)\right]$$
$$S_L(i) = \max\left[0, k - d_i + S_L(i-1)\right] \qquad (1)$$

where k is the sensitivity of the chart to the shift in multiples of σ, usually k is set between 0.3 and 1. If any of the values S_H or S_L exceed dessision bounds ±4 a shift of at least $k\sigma$ is detected. Advantage of this procedure is relatively fast detection of the jump change in the process mean,

$$\mu = \begin{Bmatrix} \alpha & \text{for} & x < x_0 \\ \alpha + \beta & \text{for} & x \geq x_0 \end{Bmatrix} \quad (2)$$

In case of slow graduate (say, linear or quadratic) deviation of the mean,

$$\mu = \begin{Bmatrix} \alpha & for & x < x_0 \\ \alpha + \beta(x - x_0) & for & x \geq x_0 \end{Bmatrix} \quad (3)$$

or

$$\mu = \begin{Bmatrix} \alpha & for & x < x_0 \\ \alpha + \beta(x - x_0)^2 & for & x \geq x_0 \end{Bmatrix} \quad (4)$$

the cusum charts guarantee fast detection of the change in the time when the deviation exceed $k\sigma$, however, it is not possible to detect x_0 when the process change took place. The knowledge of the time x_0 may often be crucial for diagnosing the cause of the problem, specialy in cases when the departure of mean resulted in damages or injuries. For such cases, we suggested to use conditioned regression models.

Conditioned regression models

Let us have a continuous regression model in the form

$$\mu = \begin{Bmatrix} f_1(x) & for & x < x_0 \\ f_2(x) & for & x \geq x_0 \end{Bmatrix} \quad (5)$$

with $f_1(x_0) = f_2(x_0)$.

For certain family of functions we find x_0 using nonlinear regression. Estimated parameter x_0 (using eg. least squres criterion) have then asyptotically normal distribution with variance taken from Hessian matrix of the second derivatives of the objective square sums function S(**α**),

$$\sigma^2(\hat{x}_0) = \sigma_R^2 \frac{\partial^2 S}{\partial x_0^2}, \quad (6)$$

where σ_R^2 is the residual variance. Therefore it is possible to estimate the confidence interval of x_0 in which the process change took place with desired probability.

Results

The above described methods were applied to the data from simmulated processes A, B, C, where in process A a jump change (2) with $\beta = 0.3\sigma$ took place, in process B was a linear trend with $\beta = 0.03$ according to (3) and in process C there is a quadratic growth (4) with $\beta = 0.0015$. All three processes had $x_0 = 43$ and $\alpha = 5$. The data are shown on Fig. 1a, b, c. Cusum control charts (1) indicated the shift at process S, however failed for gradually growing μ in processes B and C. The delay in process shift detection is sumarized in the following table and Fig. 2-4.

Table 1 Delays in detecting process change by Cusum control chart

Change type	No of measurements	Detected change time	True change time
A (jump)	11	54	43
B (linear)	29	72	43
C (quadratic)	29	72	43

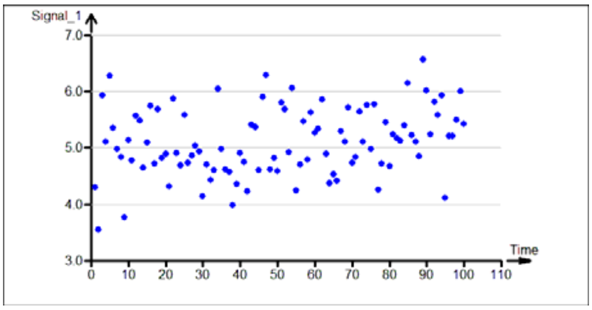

Fig 1a Signal A

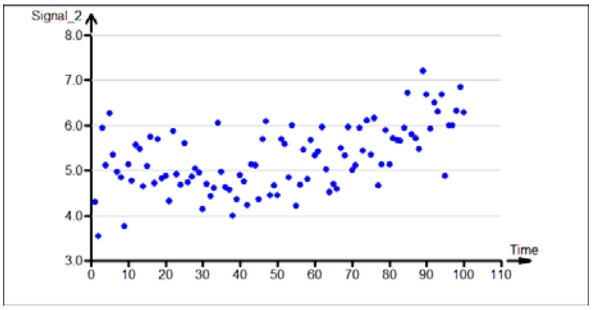

Fig. 1b Signal B

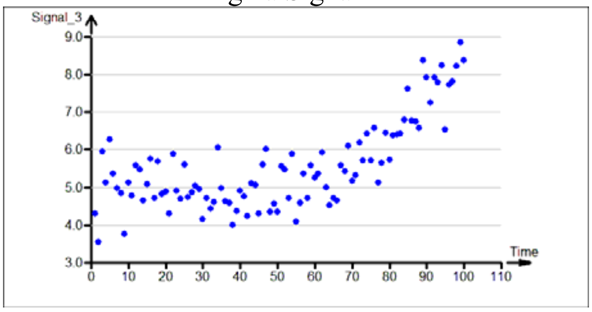

Fig 1c Signal C

Cusum charts in Figs 2 – 4 show responses (1) to signals A, B, C with k = 0.2 (upper row of plots) and k = 0.4 (lower row of plots). It can be seen that the reaction is significantly better for jump change. In case of gradually growing process mean the response is about three times slower and it is not possible to backtrace the original time of the change.

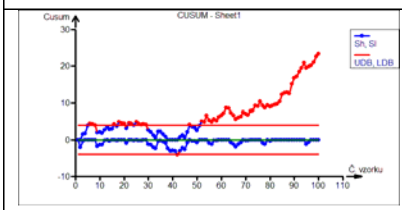

Fig 2a

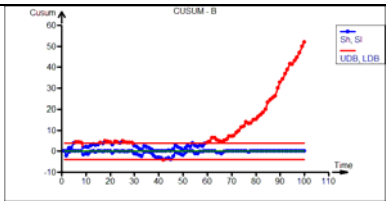

Fig 3a

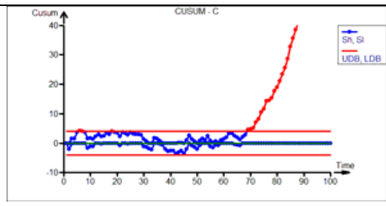

Fig 4a

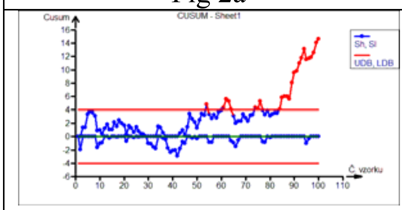

Fig 2b

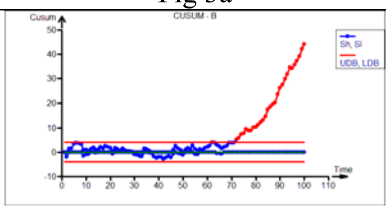

Fig 3b

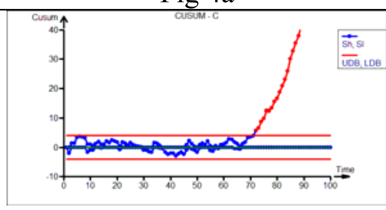

Fig 4b

With the regression models (2 – 4) it was possible to estimate the time of change (or change point) statistically using nonlinear regression. Estimates of the regression parameters are given in Tables 2, 3. Confidence intervals of the change point estimate (parameter P_3) contain the true value $x_0 = 43$ in both models.

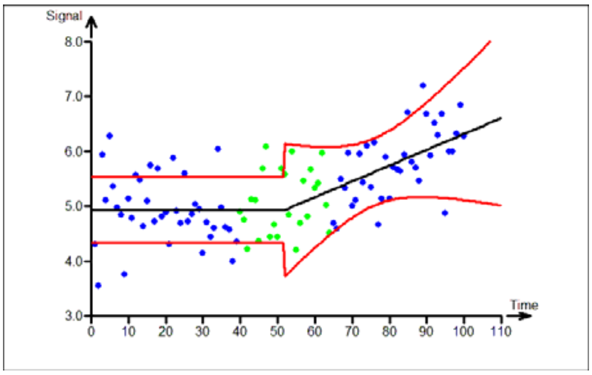

Fig 5

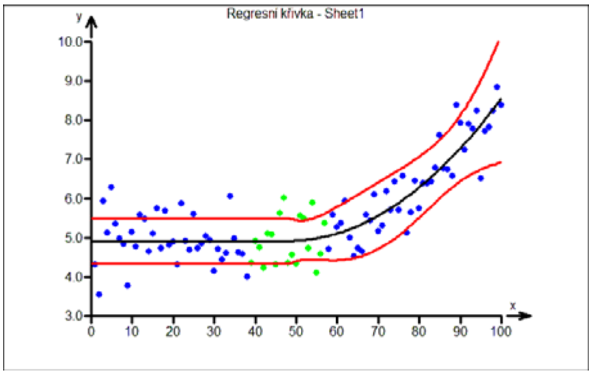

Fig 6

Table 2

Linear trend analysis in QCExpert nonlinear regression module Model: [Signal_2] ~ LT([Time],p3)*p1+GE([Time],p3)*(p1-p2*p3+p2*[Time])				
	Parametr estimate	Std Deviation	Lower CI	Upper CI
P_1	4.9304	0.0806	4.7705	5.0903
P_2	0.0289	0.0058	0.0173	0.0404
P_3	51.8678	6.287	39.3899	64.3457

Table 3

Quadratic trend analysis in QCExpert nonlinear regression module Model: [Signal_3] ~ LT([Time],p3)*p1+GE([Time],p3)*(p1+p2*([Time]-p3)^2)				
	Parameter estimate	Std Deviation	Lower CI	Upper CI
P_1	4.9188	0.0782	4.7637	5.0739
P_2	0.0014	0.0003	0.0008	0.0019
P_3	48.2574	4.4528	39.4198	57.095

Conclusion

The conditioned regression models suggested in this paper may offer a tool to detect the true time of critical change of process that could lead to damage or injury with statistical confidence intervals. The intervals are important in assigning cause to the change, and/or personal responsibility respectively. The suggested procedure is relatively simple with no need of extensive programming. The only tool needed to perform the analysis is a nonlinear regression package capable of inserting conditioned regression models.