Long-term Constant Relation Analysis of Variables Based on Bayesian Optimization and SSA and application to the monitoring of non-stationary process

Jingzhi Rao,a Chengyu Han,a Chong Liu,a Wei Sun,a,\* Jingde Wang, a

aCollege of Chemical Engineering, Beijing University of Chemical Technology, North Third Ring Road 15, Chaoyang District, Beijing, 100029, China

\*Corresponding Author’s E-mail: sunwei@mail.buct.edu.cn

Abstract

Stationary subspace analysis (SSA) as a long-term constant relationship analysis method has been widely used in non-stationary process monitoring, where traditional multivariate statistical process monitoring (MSPM) is not applicable in that most of them are based on the assumption of the process is stationary. It is considered that the data is a linear summation of stationary and non-stationary sources in SSA and the long-term constant relation between non-stationary variables can be effectively extracted. In the SSA steps, the data should be manually segmented before input, and the number of the stationary sources also needs to be manually determined. If the parameters are not set appropriately, the monitoring results will be compromised. Bayesian optimization (BO) is an effective parameter tuning approach in machine learning, due to the ability to find good points in a search space without many function evaluations. Thus, an SSA-BO non-stationary process monitoring strategy is proposed in this paper. Firstly, the input data is normalized with z-score and divided into $N$ epochs. Then the number of the stationary sources $d\_{s}$ is determined, so that the dimension of projection matrices in stationary and non-stationary subspaces can be determined and they are initialized respectively. The two projection matrices are optimized by conjugate gradient descend. BO are used to search for the optimal combination of parameters thus obtaining the optimal projection matrix with an improved objective function. Finally, monitoring statistics and control limits are constructed with the projections of the original variables in the stationary subspace to enable the non-stationary process monitoring. The proposed method is validated by a numerical case and an industrial case, and the monitoring results demonstrate that optimal parameters combination could be searched by BO.

**Keywords**: long-term constant relation, non-stationary, fault detection

* 1. Introduction

Process monitoring is as an effective method to ensure the stability and safety of real-time process operation. A large amount of operation data collected by Distributed Control System (DCS) provides abundant support for data-driven process monitoring methods (Cheng Ji et al., 2022). Multivariate Statistical Process Monitoring (MSPM) has gained considerable attention due to the fact that MSPM doesn’t require considerable prior process knowledge and their practical implementations are straightforward. The basic idea of represented MSPM method Principal Component Analysis (PCA), is to project the process data into a low-dimensional subspace that contains the most variance of the original data and accounts for correlations among different variables. Most of MSPM method are based on the assumption that the process is operated in a pre-defined normal state (Scott, D et al., 2020), which is usually stationary. However, non-stationary variables could exist in large-scale and complex chemical processes, which could be a result of equipment aging, adjustments in normal plans, and external disturbances (Cheng Ji et al., 2022). It brings huge difficulty for MSPM to achieve satisfactory process monitoring performance. To further develop process monitoring methods for non-stationary processes is of great concern.

Many researches have been made to address the issues of non-stationary process monitoring, and representative methods include difference strategies, model adaptive updating strategies and long-term constant relationship analysis. Long-term constant relationship analysis is considered as an effective way for handling non-stationary processes, by which the long-term constant relationship among the non-stationary variables is extracted. Stationary subspace analysis (SSA) is a typical long-term constant relationship analysis method for extracting the stationary and non-stationary components of a high-dimensional signal which was first proposed by Bunau et al (Von Bünau, P et al., 2009). There are two key parameters when establishing the SSA model, number of epochs the data divided $N$ and number stationary sources $d\_{s}$, which need to be determined manually and will affect the performance of SSA monitoring model. In this regard, optimization algorithms can be hybridized with SSA to tune automatically the parameters, resulting in the optimal parameters combination. Bayesian Optimization (BO) is good choice in optimization of parameters of machine learning algorithms, which has been shown to outperform other prior art global optimization algorithms on a number of challenging optimization benchmark functions (D.R. Jones et al., 2001). Therefore, in order to obtain better monitoring results, BO is adopted to optimize the two key parameters of SSA.

* 1. Theory and method
		1. Stationary Subspace Analysis

SSA is a blind source separation approach that factorizes the observed signal $x(t)$ into stationary and non-stationary source based on the Eq. (1):

$x\left(t\right)=As\left(t\right)=\left[A^{s},A^{n}\right]\left[\begin{matrix}s^{s}(t)\\s^{n}(t)\end{matrix}\right]$ (1)

where $A$ is an invertible matrix.$s^{s}(t)$ is the stationary sources and $s^{n}(t)$ is the non-stationary source. The goal of SSA is to separate the stationary sources and non-stationary sources by estimating a demixing matrix.

$P=A^{-1}=\left[\begin{matrix}P^{s}\\P^{n}\end{matrix}\right]$ (2)

$\left[\begin{matrix}s^{s}(t)\\s^{n}(t)\end{matrix}\right]=A^{-1} x\left(t\right)=\left[\begin{matrix}P^{s}x\left(t\right)\\P^{n}x\left(t\right)\end{matrix}\right]$ (3)

where $P^{s}$ and $P^{n}$ are the stationary and non-stationary projection matrices.

The specific steps of SSA are as follows:

1. The process data are divided into $N$ consecutive and nonoverlapping epochs, $[X\_{1},X\_{2}……X\_{N}]$. For any projection matrix $P$, it is possible to obtain the mean $μ\_{s,i}=P^{s}μ\_{i}$ and covariance matrix $Σ\_{s,i}=P^{s}Σ\_{i}$ of stationary sources in each epoch, thus obtaining the distribution $Norm(μ\_{s,i},Σ\_{s,i})$.
2. The distance between the stationary sources and the standard normal distribution is calculated in each epoch which is measured by the Kullback-Leibler divergence $D\_{KL}$. $D\_{KL}$ are summed over each epoch to construct an objective function.

$f(P^{s})=\sum\_{i}^{N}D\_{KL}[Norm(μ\_{s,i},Σ\_{s,i})‖Norm\left(0,I\right)]$ (4)

corresponds to the following optimization objective:

$min\sum\_{i}^{N}D\_{KL}[Norm(μ\_{s,i},Σ\_{s,i})‖Norm(0,I)]$ (5)

$s.t. P^{s}(P^{s})^{T}=I$ (6)

The problem is usually solved by the gradient descend method to obtain the optimal stationary projection matrix $P^{s}$ and stationary sources $P^{s}x\left(t\right)$

1. Similarly, an objective function can be constructed as follows to obtain the optimal non-stationary projection matrix $P^{n}$ and stationary sources $P^{n}x\left(t\right)$:

$g(P^{n})=\sum\_{i}^{N}D\_{KL}[Norm(μ\_{n,i},Σ\_{n,i})‖Norm\left(0,I\right)]$ (7)

corresponds to the following optimization objective:

$max\sum\_{i}^{N}D\_{KL}[Norm(μ\_{n,i},Σ\_{n,i})‖Norm(0,I)]$ (8)

$s.t. P^{n}(P^{n})^{T}=I$ (9)

* + 1. Bayesian Optimization

The goal of BO is to minimize or maximize an objective function in a bounded area. The idea of constructing the BO objective function is as follows:

$D\_{KL}$ between an n-dimensional multivariate series and the standard normal distribution can be calculated as follows:

$D\_{KL,M}=\frac{1}{2}[\left‖μ\_{p}\right‖^{2}-logdet\left(Σ\_{p}\right)+Tr\left(Σ\_{p}\right)-n]$ (9)

Summed $D\_{KL}$ of one of the n-dimensional multivariate series can be calculated with Eq. (10):

$D\_{KL,S}=\frac{1}{2}[\sum\_{i=1}^{n}\left‖μ\_{i}\right‖^{2}-\sum\_{i=1}^{n}logdet\left(Σ\_{i}\right)+\sum\_{i=1}^{n}Tr\left(Σ\_{i}\right)-\sum\_{i=1}^{n}1]$ (10)

where $\left‖μ\_{p}\right‖^{2}$ is the sum of squares of the means of each series and $Tr\left(Σ\_{p}\right)$ is the sum of diagonal elements of the original multivariate series.

$\left‖μ\_{p}\right‖^{2}=\sum\_{i=1}^{n}\left‖μ\_{i}\right‖^{2}$ (11)

$Tr\left(Σ\_{p}\right)=\sum\_{i=1}^{n}Tr\left(Σ\_{i}\right)$ (12)

$error= D\_{KL,M}-D\_{KL,S}=\frac{1}{2}\left[\sum\_{i=1}^{n}logdet\left(Σ\_{i}\right)-logdet\left(Σ\_{p}\right)\right]=\frac{1}{2}log\left(\frac{Σ\_{1}Σ\_{2…}Σ\_{n}}{det\left(Σ\_{p}\right)}\right)$ (13)

where $[Σ\_{1}Σ\_{2…}Σ\_{n}]$ are the diagonal elements of $Σ\_{p}$.Thus the average$ D\_{KL}$for each of the n-dimensional multivariate series can be calculated:

$\frac{D\_{KL,S}}{n}=\frac{D\_{KL,M}-error}{n}$ (14)

Considering the stationary source and non-stationary source, the optimization goals of BO is:

$Max:log(\frac{D\_{KL,M,n}-error\_{n}}{D-d\_{s}}/\frac{D\_{KL,M,s}-error\_{s}}{d\_{s}})$ (15)

$s.t. N\geq \frac{D-d\_{s}}{2}+2$ (16)

where $D\_{KL,M,s}$ and $D\_{KL,M,n}$ are the overall $D\_{KL}$ of stationary source and non-stationary source.$d\_{s}$ is the number of stationary source and $D$ is the dimensions of the original signals. The goal of the BO is to find the optimized parameter combination ($N, d\_{s} )$ to optimize the performance of SSA.

* + 1. Monitoring steps

The SSA-BO process monitoring method is divided into two sections, offline modeling and online monitoring. In offline modeling, the input data is normalized with z-score and divided into $N$ epochs and the number of the stationary sources $d\_{s}$ is determined, so that the projection matrices in stationary and non-stationary subspaces can be initialized. The two projection matrices are optimized by conjugate gradient descend. BO are used for searching for the optimal combination of parameters thus obtaining the optimal projection matrix with an improved objective function. Finally, Mahalanobis distance $MS$ and control limits $L$ are constructed with the projections of the original variables in the stationary subspace to enable the non-stationary process monitoring.

In online monitoring, the inputted data are projected into the stationary subspace with the projection matrix obtained in offline modeling. $MS$ is calculated with the stationary series and if $MS$ exceed $L$, the system will trigger an alarm.

* 1. Cases and results
		1. Numerical case

In order to verify the effectiveness of the above objective function, the 5-dimensional stationary source $s^{s}\left(t\right)=$[$s\_{1t}^{s},s\_{2t}^{s}……s\_{5t}^{s}]$ , the 5-dimensional non-stationary source$ s^{n}\left(t\right)=$[$s\_{1t}^{n},s\_{2t}^{n}……s\_{5t}^{n}]$and$ A\in R^{10×10}$ are generated randomly, thus obtaining $X=A·[s^{s}\left(t\right),s^{n}\left(t\right)]^{T}=(x\_{1t},x\_{2t}…x\_{10t})$. The total number of samples is 3000, the training dataset includes 2000 samples and the rest 1000 samples are divided into the test dataset.

The faults are introduced to $x\_{7t}$ at 500th samples, where a random walk process is introduced to the data causing the deviation from its original trend.

Fault Detection Rate (FDR) and False Alarm Rate (FAR) are applied to evaluate the process monitoring performance. FDR and FAR of SSA with different parameter combinations are shown in Figure 1.



Figure 1 monitoring performance of SSA with different parameters combinations

It can be observed that with $d\_{s}$ increases, FDR will also increase. However, large $d\_{s}$ will lead to high FAR, which means worse performance. In addition, the number of epochs $N$ will have a certain effect on the monitoring results. Thus, BO is performed to search the optimized parameters, the initial number of random searches is set to 20, the number of iterations is 50, and the search interval for $d\_{s}$ is (1,9), the search interval for $N$ is (2,10). The optimized parameters result, $N=4.434≈4$, $d\_{s} =5.198≈5$, indicates that number of stationary sources of original signals could be obtained by BO. The monitoring results of SSA with the parameters above are in Figure 2.



Figure 2 monitoring results of SSA with optimized parameters

It can be observed that when the fault occurs, SSA triggers an alarm and FDR is 0.994, FAR is 0.006, which indicates that the parameters searched by BO is effective.

* + 1. Industrial case

In the catalytic reforming process unit of a petrochemical company, the pressure drop at the hot end of the heat exchanger often increases abnormally, which pose a significant safety risk if left unaddressed. A section of historical data with abnormal rise of pressure drop is selected, including 3000 samples, and the sampling frequency is 1 minute, each sample consists of 21 variables.



Figure 3 Variation trends of some key variables

As shown in Figure 3, the pressure drop at the hot end of the heat exchanger follows a similar non-stationary trend to that of the circulating hydrogen feed rate. However, around the 2450th sample point, an abnormal increase in the pressure drop at the hot end of the heat exchanger is observed because of the different trends from circulating hydrogen feed rate and stationary naphtha feed rate, which suggests that the fault occurs.

To validate the proposed method, the first 2000 samples are used for training the model, while the remaining samples are the test dataset to verify the model. The results of BO are in Table 1.

Table 1 Bayesian Optimization results of industrial case

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Target | Allowed | $$N$$ | $$d\_{s}$$ |
| 1 | 8.238 | True | 9.506 | 14.69 |
| 2 | 15.24 | False | 2.002 | 6.744 |
| … | … | … | … | … |
| **56** | **11.34** | **True** | **12.03** | **9.217** |
| … | … | … | … | … |
| 69 | 8.169 | True | 12.40 | 14.22 |
| 70 | 10.47 | False | 5.526 | 6.069 |

The optimized parameters, $N=12.029≈12$, $d\_{s} =9.217≈9$. The monitoring results of SSA with different parameters are in Figure 4.



 Figure 4 monitoring results of SSA with different parameters

It can be observed from Figure 4 that SSA with the optimized parameters triggers an alarm at 459th sample points, while there are large number of false alarms or miss alarms in the results of SSA with other parameters, which indicates that the parameters searched by BO is effective.

* 1. Conclusion

In this work, a process monitoring strategy based on SSA-BO is proposed. BO is applied to search for the optimal combination of parameters in SSA thus obtaining the optimal projection matrix with a modified objective function and improved monitoring statistics, and control limits are constructed with the projections of the original variables in the stationary subspace to enable the non-stationary process monitoring. The strategy proposed is also applied in a numerical case and an industrial case. The results show that the parameters of SSA searched by BO is effective for non-stationary process monitoring and it can trigger an early alarm of the faults, while SSA with other parameters results in large number of false alarms or miss alarms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 22278018).

References

Ji, C., Sun, W., 2022, A review on data-driven process monitoring methods: Characterization and mining of industrial data. Processes,10,2, 335.

Scott, D., Shang, C., Huang, B.,Huang, D.,2022, A holistic probabilistic framework for monitoring nonstationary dynamic industrial processes. IEEE Transactions on Control Systems Technology, 29, 5, 2239-2246.

Ji, C., Ma, F., Wang, J., & Sun, W.,2022, Early Identification of Abnormal Deviations in Nonstationary Processes by Removing Non-Stationarity. In Computer Aided Chemical Engineering, 49, 1393-1398.

Von Bünau, P., Meinecke, F. C., Király, F. C., Müller, K. R.,2009, Finding stationary subspaces in multivariate time series. Physical review letters, 103, 21, 214101.

D.R. Jones.,2001, A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 21, 4, 345–383.