Antonis Kokossis, Michael C. Georgiadis, Efstratios N. Pistikopoulos (Eds.)
PROCEEDINGS OF THE 33rd European Symposium on Computer Aided Process Engineering
(ESCAPE33), June 18-21, 2023, Athens, Greece
http://dx.doi.org/10.1016/B978-0-443-15274-0.50258-4

# Online Monitoring and Robust, Reliable Fault Detection of Chemical Process Systems

Zheyu Jiang[a]*

[a]*Oklahoma State University, 420 Engineering North, Stillwater, Oklahoma, USA 74078*
**Corresponding author: zheyu.jiang@okstate.edu*

## Abstract

Nowadays, large amounts of data are continuously collected by sensors and monitored in chemical plants. Despite having access to large volumes of historical and online sensor data, industrial practitioners still face several challenges in effectively utilizing them to perform process monitoring and fault detection, because: 1) fault scenarios in chemical processes are naturally complex and cannot be exhaustively enumerated or predicted, 2) sensor measurements continuously produce massive arrays of high-dimensional big data streams that are often nonparametric and heterogeneous, and 3) the strict environmental, health, and safety requirements established in the facilities demand uncompromisingly high reliability and accuracy of any process monitoring and fault detection tool. To address these challenges, in this paper, we introduce a robust and reliable chemical process monitoring framework based on statistical process control (SPC) that can monitor nonparametric and heterogeneous high-dimensional data streams and detect process anomalies as early as possible while maintaining a pre-specified in-control average run length. Through an illustrative case study of the classical Tennessee Eastman Process, we demonstrate the effectiveness of this novel chemical process monitoring framework.

**Keywords**: Process monitoring, fault detection, statistical process control, CUSUM, big data streams

## 1. Introduction

Digitalization is transforming chemical and process industries. Modern chemical plants are equipped with sophisticated digital tools and infrastructures, including numerous sensors and advanced distributed control systems (DCSs), which continuously monitor the plants' equipment performance, manufacturing processes, and mass, energy, and information flows. Together, these sensors generate massive arrays of online data streams that are often *nonparametric* (i.e., data streams do not necessarily follow any specific distribution) and *heterogeneous* (i.e., data streams do not necessarily follow the same distribution). Over the past decades, a number of algorithmic approaches have been developed to effectively utilize the large volumes of historical and online sensor data for reliable online process monitoring and early fault detection. Among them, dimensionality reduction techniques, such as principal component analysis (Jackson and Mudholkar, 1979; Fezai et al., 2018) and partial least squares regression (Geladi and Kowalski, 1986), are the most popular ones in the literature (Russell et al., 2000). These dimensionality reduction-based approaches assume that the statistics characterizing the in-control profiles must also span the subspace defining the out-of-control states or faults (Woodall et al., 2004). In other words, to use the features (e.g., principal components) obtained from historical in-control process data (known as Phase I) for online monitoring (known as Phase II), one must ensure there is no profile shift during online monitoring in some otherwise undetectable direction. However, this assumption is not guaranteed as the chemical process dynamics typically are quite complex and out-of-control states cannot be fully enumerated or anticipated a priori. Another shortcoming of dimensionality reduction-based methods is that operators and process engineers often have a hard time interpreting the results obtained because the features are in the reduced space, which does not have a one-to-one mapping to the original big data streams. Furthermore, since the number of possible distillation fault scenarios can be quite large, monitoring only the most significant subset of features can cause significant error, as the fault may not be noticeable in the selected features.

More recently, advancements in machine learning, such as support vector machine (Onel et al., 2018) and artificial neural network (Heo and Lee, 2018), offer new pathways toward chemical process monitoring and fault detection. Nevertheless, state-of-the-art machine learning-based approaches still face problems such as overfitting and poor predictive accuracy. For example, while most of the published machine learning methods perform well on training and validation sets, their fault detection accuracies rarely exceed 95% in test sets. Considering the strict EHS requirements on plant site and the severity of consequences in case of fault detection failure, such predictive accuracies are unacceptable and can be catastrophic. Furthermore, machine learning methods do not scale well for new fault scenarios that have not been encountered before. In summary, existing distillation process monitoring and fault detection frameworks are inadequate and unsuitable to address practical, sophisticated data stream characteristics and fault scenarios encountered in chemical process industries.

In this work, we introduce a generic chemical process monitoring and fault detection framework featuring nonparametric and heterogeneous big data streams. This framework can detect process mean shifts or anomalies as early as possible while maintaining a user-specified false alarm rate (or in-control average run length, IC-ARL). Specifically, we adopt and simplify the quantile-based statistical process control (SPC) framework that generalizes the proven and reliable multivariate cumulative sum (CUSUM) control charts for nonparametric, heterogeneous big data systems (Ye and Liu, 2022). To demonstrate its effectiveness in chemical process monitoring, we apply this new framework to the classic problem of Tennessee Eastman Process (Downs and Vogel, 1993) and compare the fault detection performance with PCA and SVM-based approaches for the first time.

## 2. Recent Advancements in Multivariate SPC

We use $X(t) = (X_1(t), \cdots, X_p(t))$ to denote the measurement of $p$ data streams over the observation time $t = 1,2, \cdots$. We assume that the local statistic $X_j(t)$ is i.i.d. across time $t$ for every $j$. Note that the i.i.d. assumption of the data streams is often satisfied when $X_j(t)$ measures the residual value (Zou et al., 2015). Also, we emphasize that the independency across different data streams is not required. To combine these individual local statistics into a single global monitoring statistic for fault detection, Tartakovsky et al. (2006) and Mei (2010) proposed using the maximum and the sum of local statistics to form the global monitoring statistic, respectively. Mei (2011) further proposed another global monitoring scheme known as the "top-$r$ approach" based on the sum of the largest $r$ local statistics. Unfortunately, all these methods were developed under the assumption that all data streams follow a normal distribution, which is rarely encountered in chemical and process industries. To address this issue, various nonparametric multivariate CUSUM procedures (e.g., Liu et al., 2015) have been developed following the pioneering work of Qiu and Hawkins (2001, 2003). Although these multivariate CUSUM methods relax the normality assumption, they still assume that all data streams follow the same distribution, thereby limiting its applicability to monitoring homogeneous data streams. Earlier this year, Ye and Liu (2022) proposed a quantile-based nonparametric SPC algorithm to construct the local statistic for each data stream. The idea is to incorporates the ordering information of in-control data measurements by categorizing them into a number of quantiles (e.g., 10 or 15) for each data stream in Phase II (Qiu and Li, 2011). Next, in Phase II, online measurements are categorized into the learnt quantiles to generate a quantile-based distribution, which is compared with the quantile-based distribution of in-control data for the detection of any mean shift. Thus, this new quantile-based SPC approach can effectively monitor heterogeneous data streams, and it does not require any prior knowledge about the fault scenarios. In the next section, we present a simplified formulation of the original quantile-based SPC framework of Ye and Liu (2022).

## 3. Formulation and Methodology

Due to space limitations, we only highlight the key results in this quantile-based SPC framework. First, in Phase I, for each data stream $j = 1, \cdots, p$, in-control measurements are ordered and partitioned into $d$ quantiles: $I_{j,1} = (-\infty, q_{j,1}]$, $I_{j,2} = (q_{j,1}, q_{j,2}]$, …, $I_{j,d} = (q_{j,d-1}, +\infty)$, such that each quantile contains exactly $\frac{1}{d}$ of the in-control measurements. Therefore, one can define cumulative intervals as $CI_{j,i}^{+} = [q_{j,i}, +\infty)$ and $CI_{j,i}^{-} = (-\infty, q_{j,i}]$ for $i = 1, \cdots, d-1$. This information is then used in Phase II monitoring. Specifically, in Phase II, a vector $\boldsymbol{Y}_j(t) = (Y_{j,1}(t), \cdots, Y_{j,d}(t))$ is defined for each data stream $j$, where $Y_{j,q}(t) = \mathbb{I}\{X_j(t) \in I_{j,q}\}$ and $q = 1, \cdots, d$. Here, $\mathbb{I}\{X_j(t) \in I_{j,q}\}$ is the indicator function that equals 1 when $X_j(t) \in I_{j,q}$ and 0 otherwise. Correspondingly, one can further define two vectors, $\boldsymbol{A}_j^{+}(t) = (A_{j,1}^{+}(t), \cdots, A_{j,d-1}^{+}(t))$ and $\boldsymbol{A}_j^{-}(t) = (A_{j,1}^{-}(t), \cdots, A_{j,d-1}^{-}(t))$, for each data stream $j$, such that $A_{j,i}^{+}(t) = \mathbb{I}\{X_j(t) \in CI_{j,i}^{+}\}$ and $A_{j,i}^{-}(t) = \mathbb{I}\{X_j(t) \in CI_{j,i}^{-}\}$. One can show that $A_{j,i}^{+}(t) = 1 - \sum_{k=1}^{i} Y_{j,k}(t)$ and $A_{j,i}^{-}(t) = \sum_{k=1}^{i} Y_{j,k}(t)$. And $\mathbb{E}(A_{j,i}^{+}(t)) = 1 - i/d$ and $\mathbb{E}(A_{j,i}^{-}(t)) = i/d$ for $j = 1, \cdots, p$ and $i = 1, \cdots, d$. Therefore, detecting the mean shifts in the distribution of $X_j(t)$ is equivalent to detecting shifts in the distribution of $A_{j,i}^{+}(t)$ and $A_{j,i}^{-}(t)$ with respect to their expected values. Specifically, $A_{j,i}^{+}(t)$ (resp. $A_{j,i}^{-}(t)$) is more sensitive to upward (resp. downward) mean shift (Ye and Liu, 2022). Thus, to detect upward ($+$) and downward ($-$) mean shifts, the multivariate CUMSUM procedure developed of Qiu and Hawkins (2001, 2003) was adopted by defining $C_j^{\pm}(t)$ as:

$$
\begin{aligned}
C_j^{\pm}(t) = \Big[\Big(\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\Big) - \Big(\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t-1) + \mathbb{E}\big(\boldsymbol{A}_j^{\pm}\big)\Big)\Big]^T \\
\cdot \operatorname{diag}\Big(\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t-1) + \mathbb{E}\big(\boldsymbol{A}_j^{\pm}\big)\Big)^{-1} \\
\cdot \Big[\Big(\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\Big) - \Big(\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t-1) + \mathbb{E}\big(\boldsymbol{A}_j^{\pm}\big)\Big)\Big],
\end{aligned}
\tag{1}
$$

where $\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t)$ and $\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t)$ are two $(d-1)$-dimensional vectors with $\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t) = \boldsymbol{S}_j^{\pm,\mathrm{exp}}(t) = \boldsymbol{0}$ if $C_j^{\pm}(t) \leq k$, whereas $\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t) = \frac{C_j^{\pm}(t)-k}{C_j^{\pm}(t)}\Big(\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t-1) + \boldsymbol{A}_j^{\pm}(t)\Big)$ and $\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t) = \frac{C_j^{\pm}(t)-k}{C_j^{\pm}(t)}\Big(\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t-1) + \mathbb{E}\big(\boldsymbol{A}_j^{\pm}\big)\Big)$ if $C_j^{\pm}(t) > k$. With this, the local statistic $W_j^{+}(t)$ (resp. $W_j^{-}(t)$) for detecting upward (resp. downward) mean shift is: $W_j^{\pm}(t) = \Big(\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t) - \boldsymbol{S}_j^{\pm,\mathrm{exp}}(t)\Big)^T \cdot \operatorname{diag}\Big(\boldsymbol{S}_j^{\pm,\mathrm{exp}}(t)\Big)^{-1} \cdot \Big(\boldsymbol{S}_j^{\pm,\mathrm{obs}}(t) - \boldsymbol{S}_j^{\pm,\mathrm{exp}}(t)\Big)$, which is shown to be equivalent to $W_j^{\pm}(t) = \max\{0, C_j^{\pm}(t) - k\}$ (Qiu and Hawkins, 2001). Here, $k$ is a pre-computed allowance parameter that restarts the CUSUM procedure by resetting the local statistic back to 0 if there is no evidence of upward or downward mean shift after a while (Xian et al., 2021). To detect both upward and downward mean shifts in a data stream $j$, we simply define a two-sided local statistic $W_j(t) = \max\{W_j^{+}(t), W_j^{-}(t)\}$ (Li, 2020). And the initial condition is $W_j(0) = W_j^{+}(0) = W_j^{-}(0) = 0$ for all $j = 1, \cdots, p$. Finally, we ranklist $W_j(t)$ for the current time $t$ based on its magnitude: $W_{(1)}(t) \geq W_{(2)}(t) \geq \cdots \geq W_{(p)}(t)$, where $W_{(j)}(t)$ denotes the $j^{\text{th}}$ largest estimated local statistic. With this, we generalize the top-$r$ approach (Mei, 2011) to determine the stopping time $T$ for raising an alarm and declaring the system is out of control for monitoring heterogeneous data streams: $T = \inf\{t > 0 : \sum_{(j)=1}^{r} W_{(j)}(t) \geq h\}$, where $r$ is typically much less than $q$ (Mei, 2011), and $h$ is a constant threshold value related to false alarm rate (Liu and Shi, 2013). A commonly used $h$ corresponds to the false alarm (Type-I error) rate of no more than 0.0027 (classic $3\sigma$ limit). Overall, this quantile-based SPC framework developed by Ye and Liu (2022) offers strong statistical justifications and great flexibility as process engineers can customize the choice of $h$ based on the severity of potential failures.

## 4. Case Study: Tennessee Eastman Process

The Tennessee Eastman Process (TEP, see Figure 1) is an extensively used benchmark case for comparative assessment of process monitoring algorithms. It consists of a reactor, a product condenser, a separator, and a stripping column. The TEP takes four feed streams (streams 1-4) and partially converts them into desired products G and H and byproduct F. As illustrated in Figure 1, the model contains 11 manipulated and 41 measured variables, as well as 28 predefined fault scenarios to choose from. To generate in-control and out-of-control process data, we utilize the MATLAB/Simulink-based GUI developed by Anderson et al. (2022). In total, 50 hours of in-control process data were generated and collected in Phase I for quantile learning and threshold value $h$ determination.

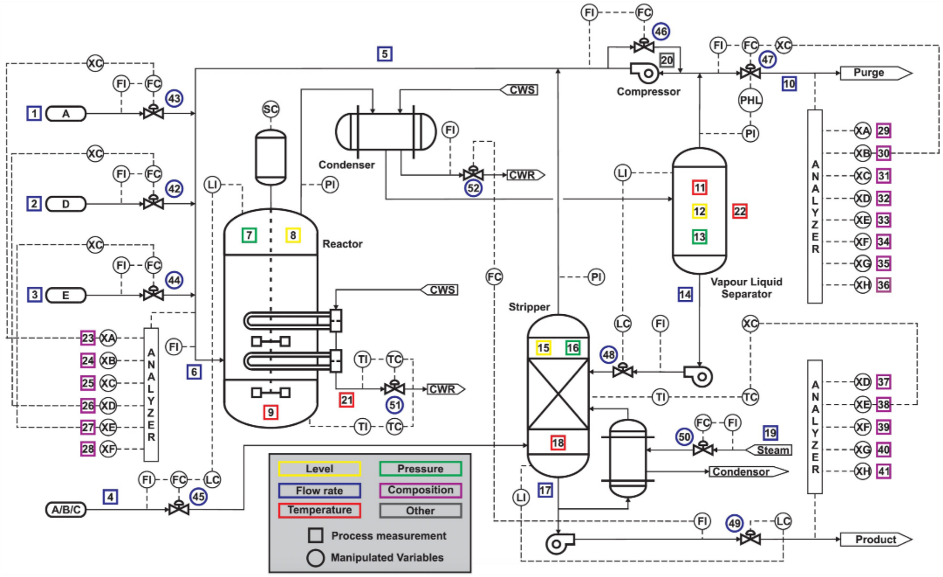

Figure 1. Schematic of Tennessee Eastman Process (source: Ma et al., 2020)

As an illustrative example, we perform process monitoring and fault detection of three representative faults, namely IDV 2, 3, and 13, as summarized in Table 1. In Phase II, we first run the process at normal operations and collect 2 hours of in-control data, followed by starting the fault and operate the process and collect 3 hours of out-of-control measurements. We compare our SPC approach with two other process monitoring algorithms commonly used in chemical industries, namely PCA and SVM. Specifically, we utilize the open source Pyphi package developed by García-Muñoz at Eli Lilly that performs multivariate PCA and Hotelling's $T^2$ analyses (López-Negrete et al., 2010). We select and keep all principal components whose eigenvalues are greater than 1, thereby leading to a total of 15 principal components four all four fault scenarios. For SVM, we adopt the method and comparison results of Onel et al. (2018).

Table 1. Summary of fault scenarios considered in this comparison study.

| **Fault No.** | **Description** | **Fault Type** |
|---|---|---|
| IDV 2 | Stream 4 composition of B (with constant A/C ratio) | Step |
| IDV 3 | Temperature in stream 2 | Step |
| IDV 13 | Reaction kinetics | Slow drift |

The comparison results of fault detection speed and the corresponding false alarm rate of all three monitoring frameworks, quantified by how many additional observations (after fault is introduced) are needed for each algorithm before it realizes the process's out-of-control status and raises an alarm, are tabulated in Table 2. As we can see, among the three monitoring frameworks, quantile-based SPC framework yields the fastest fault detection speed in all three fault scenarios, while maintaining the lowest false alarm rate. Furthermore, quantile-based SPC framework only stores and uses quantile information ($I_{j,i}$) for online monitoring and fault detection and is thus very computationally efficient. This result is exciting, given that a lower false alarm rate will lead to reduced fault detection speed due to more conservative monitoring behavior. While more thorough and extensive comparison studies are still ongoing, preliminary results presented in this work clearly demonstrate the effectiveness and attractiveness of this novel SPC framework in effective online monitoring of big data streams and robust, reliable fault detection.

Table 1. Summary of fault detection speed (characterized by out-of-control run length) and false alarm rate using three monitoring frameworks.

| **Fault No.** | **SPC** | **PCA-$T^2$** | **SVM** |
|---|---|---|---|
| IDV 2 | 125 (0.27%) | 216 (0.5%) | 180 (0.8%) |
| IDV 3 | 95 (0.27%) | 366 (0.5%) | 16815 (83%) |
| IDV 13 | 128 (0.27%) | 1131 (0.5%) | 675 (12.7%) |

## 5. Conclusion

In this work, we present a novel, powerful chemical process monitoring and fault detection framework for nonparametric, heterogeneous big data streams. In particular, the heterogeneity nature of this framework is enabled by recent advancements in SPC, such as the quantile-based multivariate CUSUM (Ye and Liu, 2022). We also compare the performance of this SPC framework with two benchmark process monitoring algorithms, PCA-$T^2$ and nonlinear SVM, in the classic example of Tennessee Eastman Process. Compared with existing dimensionality reduction or machine learning based approaches, the SPC-based framework possesses several advantages, including high reliability and accuracy with customizable, precisely controlled false alarm rate, guaranteed detection of process anomalies or mean shifts, significantly faster fault detection speed, unique capabilities to handle nonparametric and heterogeneous big data streams, low computational costs, etc.

## References

E.B. Andersen, I.A. Udugama, K.V. Gernaey, A.R. Khan, C. Bayer, M. Kulahci, 2022, An easy to use GUI for simulating big data using Tennessee Eastman process, *Quality and Reliability Engineering International*, 38, 1, 264-282.

J.J. Downs, E.F. Vogel, 1993, A plant-wide industrial process control problem, *Computers and Chemical Engineering*, 17, 3, 245-255.

R. Fezai, M. Mansouri, O. Taouali, M.F. Harkat, N. Bouguila, 2018, Online reduced kernel principal component analysis for process monitoring. *Journal of Process Control*, 61, 1-11.

P. Geladi, B.R. Kowalski, 1986, Partial least-squares regression: a tutorial. *Analytica Chimica Acta.*, 185, 1-17.

S. Heo, J.H. Lee, 2018, Fault detection and classification using artificial neural networks, IFAC-PapersOnLine, 51, 18, 470-475.

J.E. Jackson, G.S. Mudholkar, 1979, Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21, 341-349.

K. Liu, J. Shi, 2013, Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a Bayesian network, *IIE Trans.*, 45, 6, 630-643.

K. Liu, Y. Mei, J. Shi, 2015, An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring, *Technometrics*, 57, 3, 305-319.

J. Li, 2020, Efficient global monitoring statistics for high-dimensional data, *Qual. Rel. Eng. Int.*, 36, 1, 18-32.

R. López-Negrete, S. García-Muñoz, L.T. Biegler, 2010, An efficient nonlinear programming strategy for PCA models with incomplete data sets, 24, 301-311.

L. Ma, J. Dong, K. Peng, 2020, A novel key performance indicator oriented hierarchical monitoring and propagation path identification framework for complex industrial processes, *ISA Trans.*, 96, 1-13.

Y. Mei, 2010, Efficient scalable schemes for monitoring a large number of data streams, *Biometrika*, 97, 2, 419-433.

Y. Mei, 2011, Quickest detection in censoring sensor networks, *Proc. IEEE Int. Symp. Inf. Theory Proc.*, 2148-2152.

M. Onel, C.A. Kieslich, E.N. Pistikopoulos, 2018, A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process, *AIChE Journal*, 65, 3, 992-1005.

P. Qiu, D. Hawkins, 2001, A Rank-Based Multivariate CUSUM Procedure, *Technometrics*, 43, 2, 120-132.

P. Qiu, D. Hawkins, 2003, A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions, *Journal of the Royal Statistical Society: Series D*, 52, 2, 151-64.

P. Qiu, Z. Li, 2011, On nonparametric statistical process control of univariate processes, *Technometrics*, 53, 4, 390-405.

E. Russell, L.H. Chiang, R.D. Braatz, 2000, Data-driven methods for fault detection and diagnosis in chemical processes. Springer, 10-45.

A.G. Tartakovsky, B.L. Rozovskii, R.B. Blažek, H. Kim, 2006, Detection of intrusions in information systems by sequential change-point methods", *Statist. Methodol.*, 3, 3, 252-293.

W.H. Woodall, D.J. Spitzner, D.C. Montgomery, S. Gupta, 2004, Using Control Charts to Monitor Process and Product Quality Profiles, *Journal of Quality Technology*, 36, 3, 309-320.

X. Xian, C. Zhang, S. Bonk, K. Liu, 2021, Online monitoring of big data streams: A rank-based sampling algorithm by data augmentation, *Journal of Quality Technology*, 53, 2,135-153.

H. Ye, K. Liu, 2022, A Generic Online Nonparametric Monitoring and Sampling Strategy for High-Dimensional Heterogeneous Processes, *IEEE Trans. Auto. Sci. Eng.*, 19, 3, 1503-1516.

C. Zou, W. Jiang, Z. Wang, X. Zi, 2015, An efficient on-line monitoring method for high-dimensional data streams, *Technometrics*, 57, 3, 374-387.