Feature Space Monitoring for Smart Manufacturing via Statistics Pattern Analysis

Q. Peter He[a*], Jin Wang[a*], Devarshi Shah[a]

[a]*Department of Chemical Engineering, Auburn University, Auburn AL 36849, USA*
qhe@auburn.edu; wang@auburn.edu

Abstract

Statistical process monitoring (SPM) is an important component in the long-term reliable operation of any system and its importance can only become greater in the era of smart manufacturing (SM). Previously we proposed statistics pattern analysis (SPA) based on the idea of using various statistics to quantify process characteristics, and monitoring these statistics instead of process variables themselves to perform process monitoring. In this work we provide a comprehensive review on recent progresses made in SPA framework, which underpins a roadmap of SPM we outlined recently. Both sample-wise feature extraction and variable-wise feature extraction are discussed, with new applications in both fault detection and diagnosis, and soft sensor development. Specifically, we provide the first systematic examination on the SPA's capability in handling process characteristics including dynamics, nonlinearity and data non-Gaussianity; and compare its performance to representative state-of-the-art SPM methods to highlight the enhanced capability of feature-based monitoring. In addition, the performance of SPA is tested using the benchmark industrial simulator TEP for fault detection and diagnosis, plus a wet lab and an industrial case studies for soft sensor development. Finally, the advantages and potential limitations of SPA in addressing the new challenges presented by smart manufacturing big data are discussed.

Keywords: statistical process monitoring, fault detection, fault diagnosis, soft sensor, smart manufacturing, statistics pattern analysis.

1 Introduction

The goal of process monitoring is to detect the onset and identify the root cause of any change that causes a manufacturing environment to deviate from its desired operation. Process monitoring is an important component and key enabler for the long-term reliable operation of any process (Severson et al., 2016). It has been recognized that the importance of process monitoring can only become greater when the controlled systems are getting more complex, equipped with more sensors, operated under non-steady state, controlled at tighter margins, and getting yet closer to autonomous operations in the era of smart manufacturing (He and Wang, 2018).

Principal component analysis (PCA), partial least squares (PLS), and their variants based multivariate SPM methods, which are the state-of-the-arts for industrial applications, have limitations when applied to the processes that are not operated at steady-state. This is mainly due to the underlying assumption for the PCA and PLS based SPM methods that the process data are assumed to be independent and identically distributed (i.i.d.) samples drawn from a multivariate Gaussian distribution. When a process exhibits strong dynamics, nonlinearity or non-Gaussian distribution, to name a few, the performance of these SPM methods may deteriorate significantly, depending on how well the normal process operation data can be approximated by a multivariate Gaussian distribution.

To address these limitations, many SPM methods, such as kernel PCA (KPCA), and independent component analysis (ICA), have been developed recently. However, as discussed in Section 2, these methods have their own challenges such as the difficulty in selecting parameters (He and Wang, 2017; Qin, 2012). On the other hand, a feature based method that we recently proposed, termed statistics pattern analysis (SPA), has demonstrated superior performance in different applications such as fault detection, fault diagnosis and soft sensor or virtual metrology for both batch and continuous processes (He and Wang, 2018, 2011; Stone et al., 2017; Wang and He, 2010). Since then, many extensions and variants of SPA have been proposed by others(Deng et al., 2016; Deng and Tian, 2013; He and Xu, 2016; Ma et al., 2011; Ning et al., 2014; Song et al., 2015; Zhang et al., 2015). Recently, Rendall et al. (2017) proposed feature oriented batch analytics platform where SPA was cited as a representative feature-based approach.

In this work, we provide a comprehensive review on the recent progresses made on the SPA framework, which includes features other than statistics of process variables for process monitoring, and examine its capability in addressing various challenges discussed above, *i.e.*, process dynamics, nonlinearity and non-Gaussianity. In addition, we expand the sample-wise feature extraction (*i.e.*, various features are computed using a group of samples of a given variable obtained at different time instants) to variable-wise feature extraction (*i.e.*, various features are computed using a group of variables sampled at the same time instant), and demonstrate its usefulness in soft sensor development for spectroscopic data analysis. In this work, several new case studies, both simulated and industrial, are provided to compare the performance of SPA with other representative monitoring and soft sensor methods. The rest of the paper is organized as the following: Section 2 provides an update of a road map for SPM and briefly reviews the SPA framework; Section 3 examines the performance of SPA in fault detection, fault diagnosis and soft sensing using simulated and industrial case studies, and compares SPA with several representative methods in addressing process dynamics, nonlinearity and non-Gaussianity. Section 4 discusses how SPA can help address the challenges of big data generated in smart manufacturing and Section 5 draws conclusions.

2 Statistics Pattern Analysis (SPA) Framework

Recently we presented a road map to capture the development of SPM in the last century (He and Wang, 2018), which divides the development of SPM into three generations: 1st generation: statistical process control (SPC); 2nd generation: multivariate statistical process monitoring (MSPM); and 3rd generation: yet to be properly defined and named. With recent developments in the field, it becomes clear that the 3rd generation SPM can be broadly categorized as feature space monitoring (FSM), in contrast to the monitoring of the original variable space of the 1st and 2nd generations. The updated road map is shown in Figure 1. The road map suggests that the key enabler of the successes achieved by a new generation of SPM methods was to utilize additional information that the previous generation did not. Specifically, for the 1st generation, SPC enabled the elimination of unnecessary adjustments made to the process through making use of the mean (1st order statistics) and variance (2nd order statistics) of product quality variable(s), therefore significantly reduced process variation and improved product quality. For the 2nd generation methods, besides the information utilized by SPC, the variance/covariance of both product quality variables and process variables were monitored as well. The added information allows the detection of those faults that cannot be detected by SPC methods, therefore enables significantly improved monitoring performance. For the 3rd generation methods, which is termed feature space

monitoring or FSM, where features that capture the key process characteristics, such as nonlinearity, non-Gaussianity which cannot be captured by variance/covariance matrix of process data, are to be used for monitoring the process. Because of the additional information included in FSM, it is expected to further improve the monitoring performance compared to the 2nd generation SPM methods. As discussed later in a greater detail, the ever-increasing prevalence of big data with 4V challenges, *i.e.*, Volume, Velocity, Variety and Veracity (Zikopoulos et al., 2012), has also necessitated the transition from the original space monitoring paradigm to the feature space monitoring paradigm.

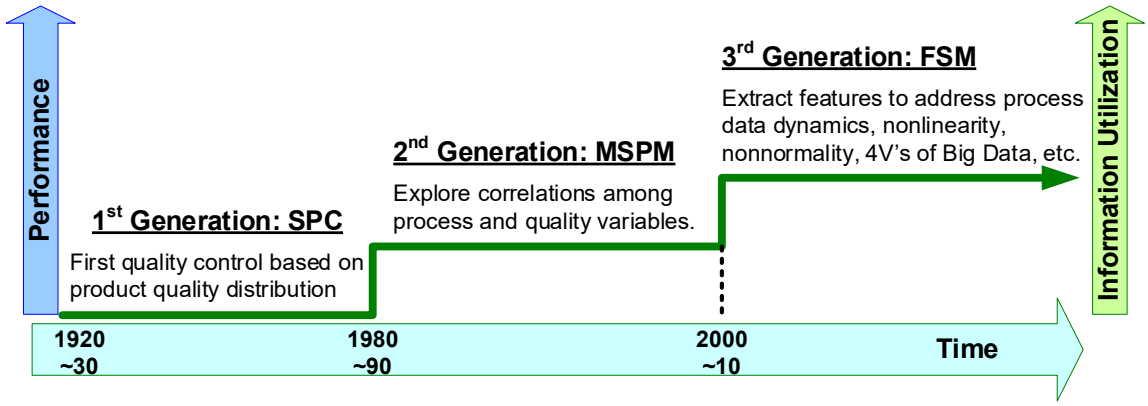


Figure 1. An updated roadmap of statistical process monitoring (SPM) technology

It is worth noting that various methods have been developed to address process dynamics, nonlinearity and non-Gaussianity, which can be viewed as the transition between 2nd and 3rd generation methods. Here we briefly review the representative ones. To address process dynamics, time lag shifting methods, including dynamic PCA (DPCA) (Ku et al., 1995) and its variants are the most commonly used. However, it has been suggested that such time lagging may not fully eliminate the dynamics among augmented samples, and results in auto-correlated scores in the principal subspace. To address process nonlinearity, kernel based methods, such as kernel PCA (KPCA) and its variants have been studied extensively (Lee et al., 2004a). In additions, neural networks and principal curves based methods have been reported as well. However, it has been reported that the performances of KPCA-based methods can be (highly) sensitive to the choice of kernel and tuning parameters; even for the same kernel but different parameters, the resulted features could be significantly different. It has been further suggested that there is no proper way to guide the choice of kernel and determination of parameters other than cross-validation. In additions, often times the mapped features are still nonlinearly correlated (Wang et al., 2013), which limits the application of these methods. To address process data non-Gaussianity, independent component analysis (ICA) has been proposed (Lee et al., 2004b). ICA is an alternative linear decomposition to PCA that was originally proposed as a blind source separation technique to separate a multivariate signal into additive subcomponents by assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other. Although ICA can perform decomposition on non-Gaussian data better than PCA, it is not intended to improve the Gaussianity of its projected components. Instead, its control limits are usually obtained using empirical or data-driven techniques such as kernel density estimation (Lee et al., 2004b).

To provide an alternative approach to address process dynamics, different types of nonlinearity and non-Gaussianity, as well as the 4V challenges of big data, we proposed statistics pattern

analysis (SPA) based on the idea of using various statistics, including 1st, 2nd and higher-order statistics (HOS) of process variables, to quantify process characteristics. In the SPA framework, these statistics, instead of process variables themselves, are monitored to perform process monitoring (He and Wang, 2011; Wang and He, 2010). In other words, SPA models the variance-covariance structure of the process statistics, instead of the variance-covariance structure of the process variables used in PCA and PLS based methods. By utilizing the additional information other than the first and second moments utilized in the 2nd generation methods, we expect to achieve enhanced monitoring performance, which has been validated in many simulated and industrial case studies. Figure 2 shows the schematic plot of how SPA can be applied for continuous and batch process fault detection by sample-wise feature extraction (*i.e.*, statistics estimation).

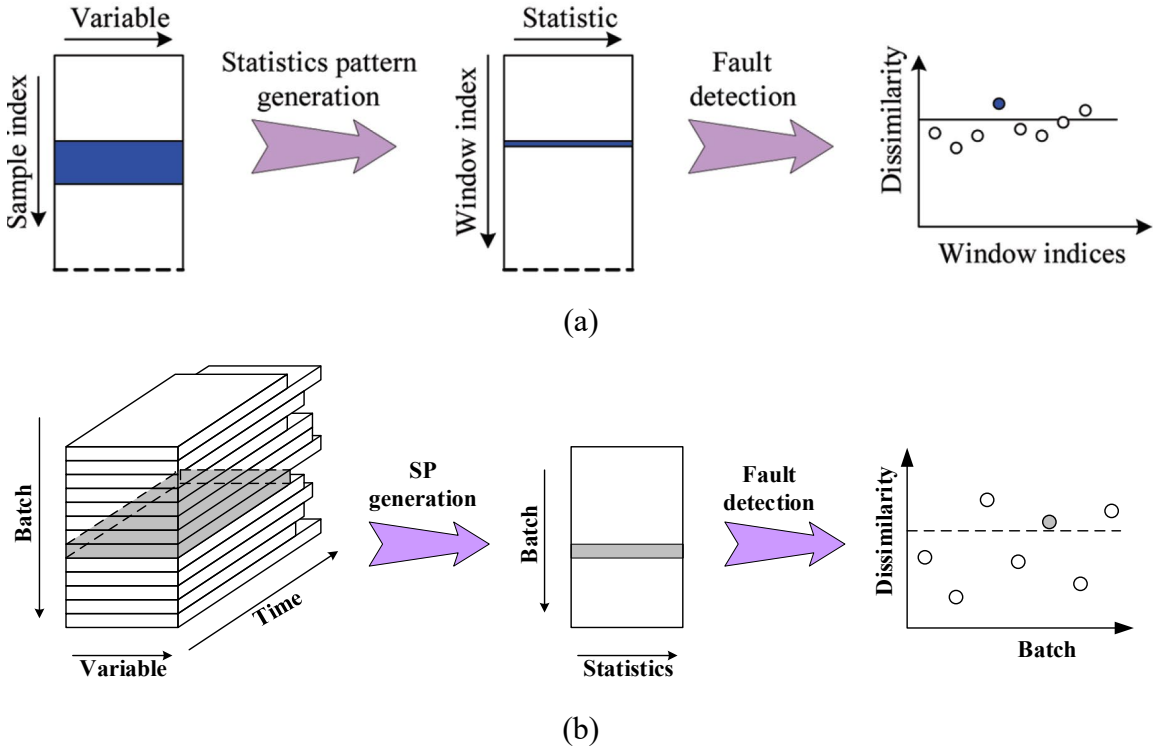


Figure 2: SPA based fault detection for continuous (a) and batch (b) processes.

For continuous processes, various statistics are computed using a window-based approach as shown in Figure 2 (a), *i.e.*, using all samples in a window to compute the statistics for a given variable. The obtained statistics of different variables, also known as statistics patterns (SPs), are monitored for fault detection. Although window-based approaches usually cause detection delay, the detection delay of SPA is actually comparable to PCA-based approaches. This is due to the following two reasons: (1) the significantly reduced normal process variation in the statistics making the model sensitive to small changes; (2) although some statistics, such as median, are insensitive and have delayed responses to a process change, other statistics, such as minimum, maximum and skewness, are very sensitive and respond rapidly to it. Therefore, detection delay is actually not an issue for window-based SPA. For batch processes, as shown in Figure 2 (b), SPs are computed for each batch or each step if multiple steps are involved in a batch; then the obtained

SPs are being monitored. Monitoring statistics for each batch/step not only enables improved monitoring performance, but also eliminates the preprocessing steps caused by different batch/step durations, such as trajectory alignment and warping, and making SPA attractive for automatic implementation. Various simulated and real industrial case studies have demonstrated that SPA offers a highly promising platform for SPM (Galicia, 2012; Galicia et al., 2012; He and Wang, 2018, 2011; Wang and He, 2010).

Remarks 1: Both statistics and non-statistics based features can be easily incorporated into the SPA framework.

For statistical features, to capture key process characteristics, various process statistics other than mean and variance/covariance of process variables can be included. For example, skewness and kurtosis can be included to quantify the non-Gaussianity of the process data, autocorrelation and cross-correlation can be included to capture process dynamics. In addition, non-statistical process features can be easily included for process monitoring, such as process knowledge based landmark features (Wold et al., 2009); profile-driven features (Rendall et al., 2017); geometry based features (Wang et al., 2015).Therefore, the exact form or the number of statistics/features in SPA varies with applications, and the performance of SPA can be optimized through feature selection. Meanwhile, as we show in the following case studies, the performance of SPA for process monitoring is quite robust with respect to feature selection, and it can provide superior monitoring performance even without feature selection or optimization.

Remarks 2: The SPA framework can be implemented either sample-wise or variable-wise.

Because many process variables are highly correlated with each other, variable-wise feature extraction allows us to capture the comprehensive picture of the process or system while significantly reducing the number of variables to be included in the model. This is particularly true for various spectroscopic data, and we demonstrate such SPA implementation using soft-sensor development as an example.

3 SPA in Addressing Current Challenges in SPM

As mentioned before, several key characteristics that limit the success of 2nd generation SPM methods, such as process dynamics, nonlinearity, and process data non-Gaussianity, are ubiquitous for manufacturing processes, and will become even more so for smart manufacturing. In this section, we examine the capability of SPA in addressing these challenges, with applications in two major areas of process monitoring: fault detection and diagnosis, and soft sensor development; and we compare the performance of SPA with several representative methods using multiple simulated and industrial case studies.

3.1 Stirred tank heater: an illustrative example

In this section, we use a simulated example to examine SPA's capability in handling process dynamics, nonlinearity and non-Gaussianity, and compare it with representative methods that were developed to address these challenges.

The simulated case study is a stirred tank heater (Bequette, 1998) as shown in Figure 3 (a), where the objective is to raise the temperature of the inlet stream to a desired value. Detailed descriptions and model assumptions can be found in (Bequette, 1998).

The material and energy balances yield the following two modelling equations:

$$\frac{dT}{dt} = \frac{F}{V}(T_i - T) + \frac{UA(T_j - T)}{V\rho c_p} \quad (1)$$

$$\frac{dT_j}{dt} = \frac{F_j}{V_j}\left(T_{ji} - T_j\right) - \frac{UA(T_j - T)}{V_j \rho_j c_{pj}} \quad (2)$$

where A is the heat transfer area, U heat transfer coefficient, c_p heat capacity, F volumetric flowrate, ρ density, T temperature, t time, V volume, subscripts i, j, and ji for inlet, jacket, jacket inlet, respectively. To excite the system so that various characteristics such as nonlinearity, dynamics and non-Gaussianity are amplified during normal operation, a sinusoidal disturbance is injected into the jacket flowrate:

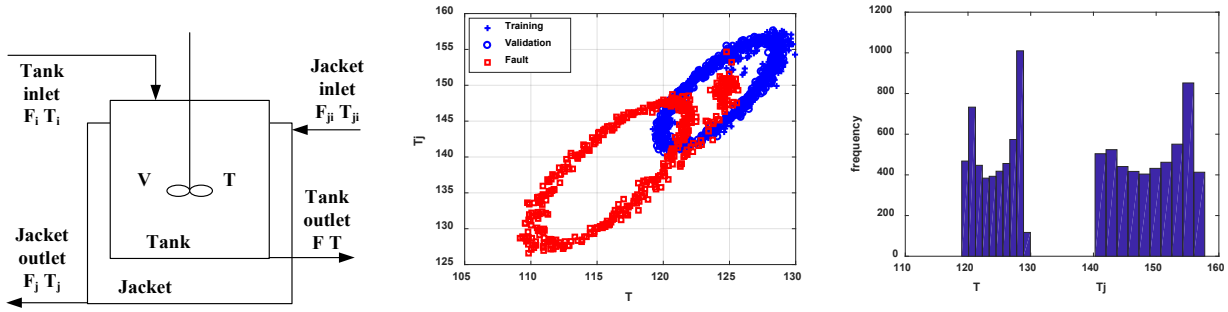
$$F_j = F_{js} + 0.5sin\left(\frac{t \cdot \pi}{10}\right) + n(0, 0.01) \quad (3)$$

where steady-state $F_{js} = 1.5$, and $n(0, 0.01)$ represents white noise with zero mean and standard deviation 0.01.

To examine the effectiveness of different algorithms in process monitoring, a fault was introduced into the process where a leakage in jacket inlet stream is introduced at a given time point.

$$F_{jf} = F_j - 0.3 \quad (4)$$

The normal data are divided into training and validation subsets. The behaviors of normal and faulty operations are shown in Figure 3 (b) where the tank temperature T is plotted against the jacket temperature T_j. The histograms of T and T_j (Figure 3 (c)) show that the observations obtained from this dynamic and nonlinear process are highly non-Gaussian, which might hinder the effectiveness of many 2nd generation methods. It is worth noting that the disturbance and fault introduced are rather large for illustrative purposes, as shown in Figure 3(b) and (c).

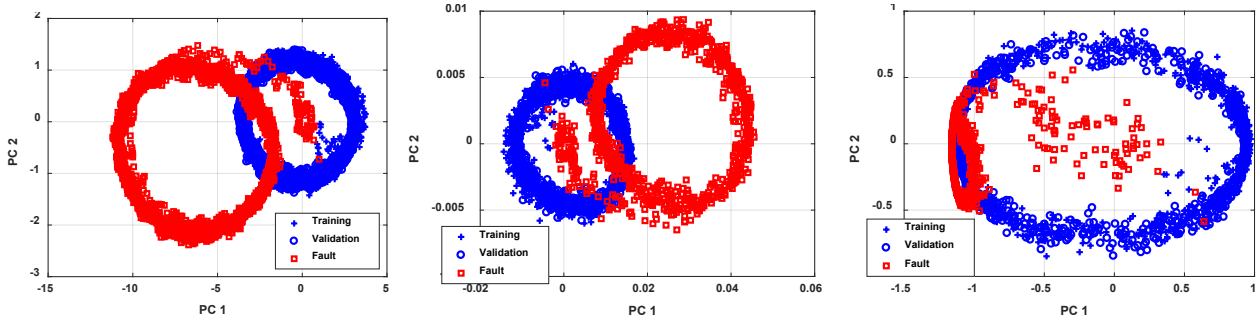


(a) (b) (c)

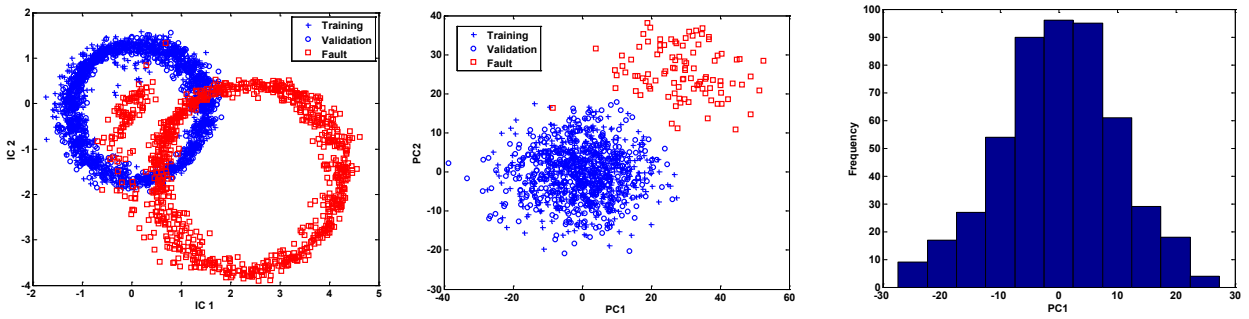
Figure 3. (a) Jacketed stirred tank heater (Bequette, 1998); (b) Process behaviors under normal (blue pluses and circles) and faulty (red squares) operations; (c) Histograms of T and T_j.

In this illustrative example, we choose DPCA as the representative method for addressing process dynamics, KPCA for addressing process nonlinearity, with consideration of two kernels – Gaussian kernel and sigmoid kernel, and ICA for addressing process non-Gaussianity. In this case study, the main purpose is to examine how effective each modeling approach is in terms of eliminating/transforming certain challenging process characteristics, and to what extent each method can differentiate faulty behavior from normal operation. To do so, we examine the principal or dominant subspace captured by the model while ignoring the residual or excluded components. To make sure key characteristics are captured by the principal subspace, we optimize the number of PC while ensure that at least 95% of variance are captured.

Figure 4 plots the principal component (PC) scores obtained by projecting the (augmented) samples or features onto the first two dimensions of principal subspace determined by each model. As can be seen from Figure 4 (a), despite the scaling effect (changing ellipses into circles), DPCA scores look very similar to the original data as shown in Figure 3 (b), with highly auto-correlated samples, which indicates that augmenting time shifted samples do not effectively remove process dynamics. Similarly, the KPCA results for a Gaussian kernel, Figure 4(b), looks very similar to the original data, despite different orientations. The KPCA result with sigmoid kernel, Figure 4(c), shows clear deviation from the process data, however, the circular behavior of both the normal and faulty data are preserved, which confirms that kernel transformation does not effectively remove the process nonlinearity, at least for this case study. Figure 4(d) shows the first two IC's using the default nonlinearity setting of FastICA algorithm (Bingham and Hyvärinen, 2000) where the circular behavior of both normal and faulty data remains, which suggest non-Gaussian distribution remains as well. For both KPCA and ICA results, process dynamics remains in the feature space (for KPCA) or independent component space (for ICA), as clearly shown in the trajectory of fault development and high correlated scores. SPA results are shown in Figure 4 (e) and (f). Figure 4(e) shows the score plot of the statistics extracted from the data. It can be seen that the nonlinear dynamic data under normal operation has been transformed into a multivariate Gaussian distributed scores, which is also validated in the histogram of the first score of the statistics in Figure 4 (f). In other words, the SPA transformation has completely transformed the dynamic, nonlinear, and non-Gaussian original process measurements into independent Gaussian distributed statistics. In addition, it is important to note that only SPA method can clearly separate the cluster of faulty samples from the normal samples, as a result of effectively addressing the process dynamics, nonlinearity, and non-Gaussianity.



(a) (b) (c)



(d) (e) (f)

Figure 4. Principal or dominant subspace captured by various methods, (a) DPCA; (b) KPCA – Gaussian kernel; (c) KPCA – sigmoid kernel; (d) ICA; (e) SPA; and histogram of PC1 from SPA (f) for normal data

This illustrative example clearly demonstrates the effectiveness of SPA in addressing these common process characteristics/challenges, which have contributed to its superior performances in fault detection and diagnosis for both batch and continuous processes. In the next section, we use TEP to further demonstrate this point.

3.2 Tennessee Eastman Process – a realistic benchmark simulated process for fault detection and diagnosis

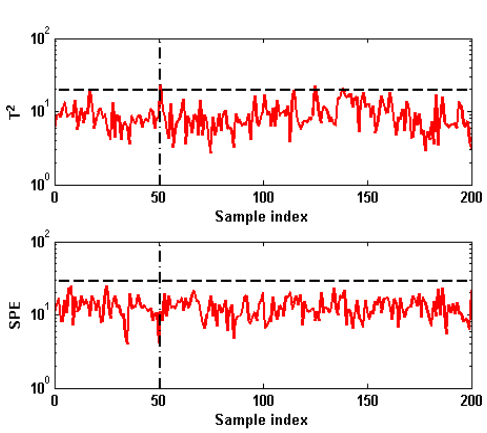
In process monitoring, once a fault is detected, fault diagnosis usually follows to identify the root cause of the fault, so that appropriate actions can be taken to address the situation. Due to its importance to process operations, fault diagnosis has drawn significant interests (Qin, 2012). However, because of the physical/chemical principles that govern the process (such as mass and energy balance) and the feed-back control loops, a fault often propagates to variables other than the root cause variable, the so-called "smearing effect", and makes fault diagnosis highly challenging. The process dynamics, nonlinear and non-Gaussianity could further deteriorate the fault diagnosis performance of different algorithms, as illustrated in this section.

In this case study, we use the benchmark Tennessee Eastman Process (TEP) simulation to demonstrate the fault detection and diagnosis performance of SPA. The TEP simulator has been widely used by the process systems engineering community as a realistic example to compare various monitoring and control approaches (Kano et al., 2002; Ku et al., 1995; Russell et al., 2000). The process consists of five major unit operations: a reactor, a product condenser, a vapor-liquid separator, a recycle compressor, and a product stripper. Four reactants A,C,D and E plus the inert B are fed to the reactor to generate products G and H, as well as byproduct F through two exothermic reactions. More details of the process description and simulation set up can be found in (Downs and Vogel, 1993).

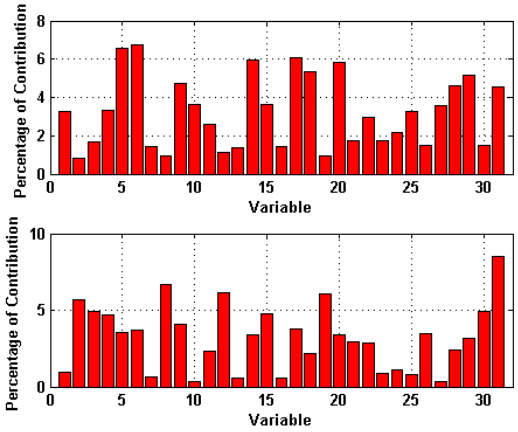
For comparison, two linear methods, PCA and DPCA, and two nonlinear methods, KPCA and ICA, are applied to detect and diagnose the faults. Among these algorithms, PCA, DPCA, ICA and SPA use contribution plots to identify the root cause variable of the fault, *i.e.*, the variables with large contributions are likely the root cause of the fault, while KPCA using a reconstruction method for fault diagnosis, *i.e.*, the variables with significantly reduced reconstruction index are most likely the root cause. All methods have been optimized and the optimal detection and

diagnosis performance are presented here. It is worth noting that the KPCA method was found to be quite sensitive to the model parameters; in addition, a good set of detection settings may not provide a good diagnosis and vice versa. Also, it is important to note that for the SPA based fault diagnosis, two subsets of the contribution plots are provided to show the contributions from variable mean and standard deviation. The contributions from other statistics such as auto- and cross-correlations are small, therefore were not shown.

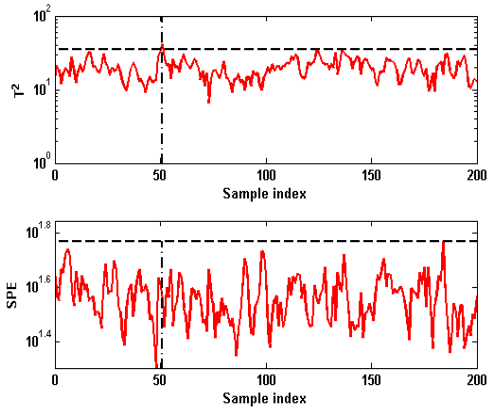
Totally 20 faults were included in the TEP simulator (Ricker, 1996). For SPA, the detection results for all faults was presented in Wang and He (2010), and the diagnosis results for all faults was provided in the Appendix. As shown in the Appendix, for detected faults, the contribution plots from SPA usually provides a clean diagnosis, with limited or no "smearing" effect. In terms of comparison, generally speaking for the faults that can be easily detected, all five methods were effective in pinpointing the major fault-contributing process variables. However, for faults that are difficult to detect, the fault diagnosis performances of different methods are quite different, where only SPA-based fault diagnosis is able to correctly diagnose the faults. Here we use fault 5 and 12 to illustrate this.



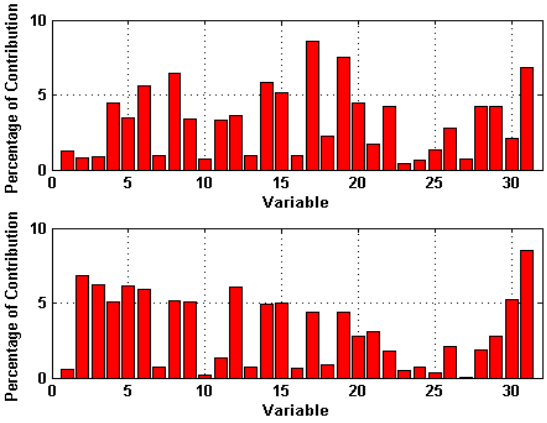
(a)



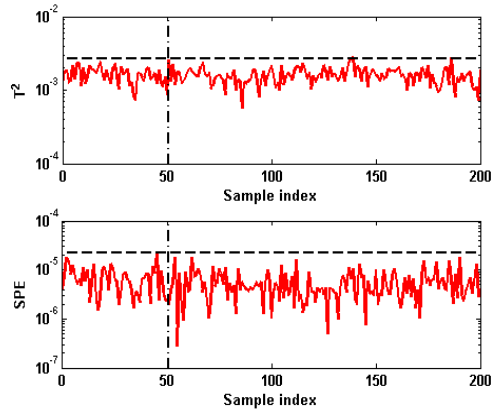
(b)



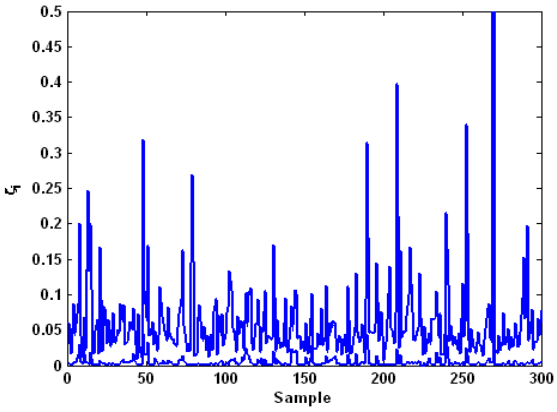
(c)



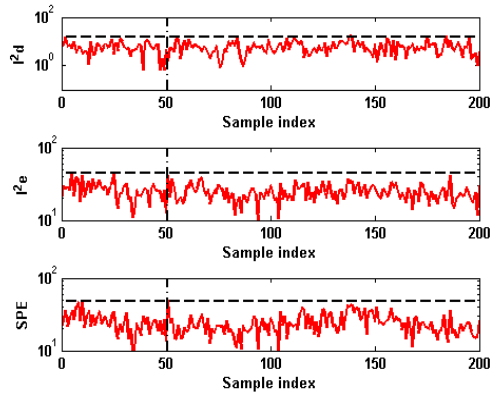
(d)



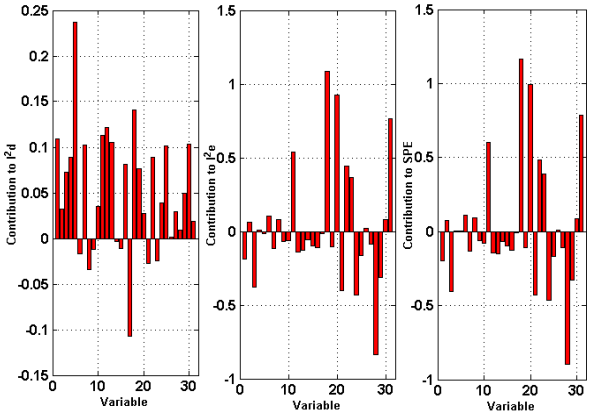
(e)



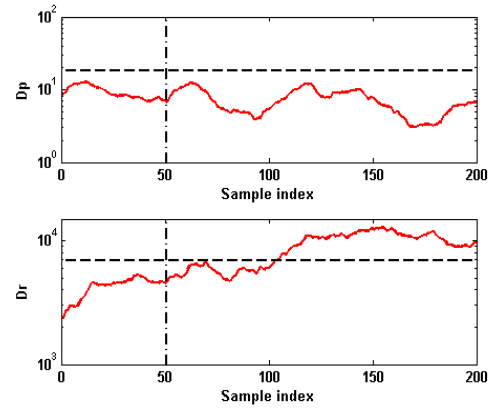
(f)



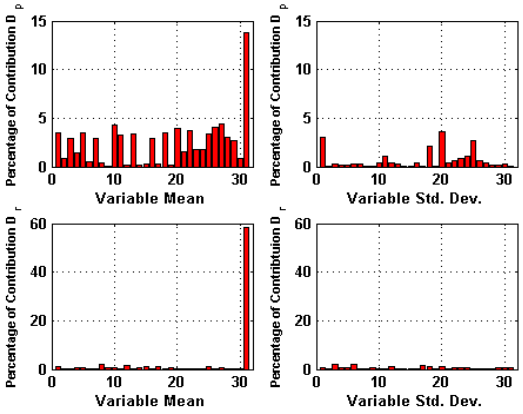
(g)



(h)



(i)



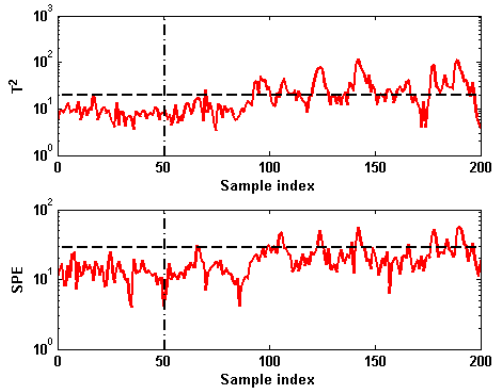
(j)

Figure 5. Detection and diagnosis of fault 5. PCA: (a) fault detection (top T^2, bottom SPE) and (b) diagnosis using contribution plots; DPCA: (c) fault detection (top T^2, bottom SPE) and (d) diagnosis using contribution plots; KPCA: (e) fault detection (top T^2, bottom SPE) and (f) diagnosis using reconstruction index ζ_i; ICA: (g) fault detection (top I_d^2, middle I_d^2, bottom SPE) and (h) diagnosis using contribution plots; SPA: (i) fault detection (top D_p, similar to T^2, bottom D_r, similar to SPE) and (j) diagnosis using contribution plots.

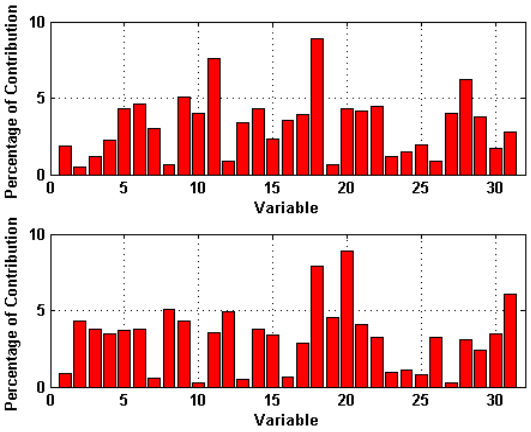
Fault 5 was caused by a step change to the inlet temperature of the condenser cooling water. Fault detection and diagnosis results from different methods are shown in Figure 6, which indicates that only SPA was able to detect the fault consistently through D_r; while other methods failed to detect the fault. In terms of fault diagnosis, since the condenser cooling water temperature is not measured, it is expected that variable 31, the condenser cooling water flow rate, would be adjusted by the feedback controller, and should be identified as the root cause of the fault. Because PCA, DPCA, KPCA and ICA were not able to detect the fault, with fault detection rate lower than 5%, it is expected that these methods are able to correctly diagnose the fault either. As shown in Figure 6, the contribution plots from PCA, DPCA and ICA all show a wide range of process variables contributing to the fault, confirming our expectation. In the case of KPCA, Figure 6 (f) shows that variable 31 has a significant drop in the reconstruction index, and would be identified as the root cause variable, although KPCA did not identify this fault (with fault detection rate of 3% and 0.4% by T^2 and SPE indices respectively). On the other hand, SPA was able to clearly isolate the root cause to be the variable 31, moreover, it correctly identified that it was the shift in mean, not variance, that caused the faulty behavior.

Fault 12 was caused by introducing random variation in condenser cooling water inlet temperature. For fault 12, all methods were able to detect the fault, although PCA, DPCA, KPCA and ICA's performance were not as consistent as SPA, as shown in Figure 6. Again, since cooling water temperature was not measured, and the fault was a random variation in the cooling water temperature, the fault will not trigger a shift in cooling water flow rate. Instead, the random variation in condenser cooling water inlet temperature (not measured) would likely affect the downstream separator right after the condenser, which is closest to the actual fault location. In other words, the introduced fault would lead to random variation in the separator temperature (measured, variable 11), which should be identified as the root cause. Figure 6 compares the fault diagnosis performance of all 5 methods. From Figure 6, it is clear that although all methods were able to detect the fault, only SPA was able to correctly diagnose the fault; moreover, SPA further indicate that it was the change in the variance of variable 11, not the change in mean, that contributes to fault 12.

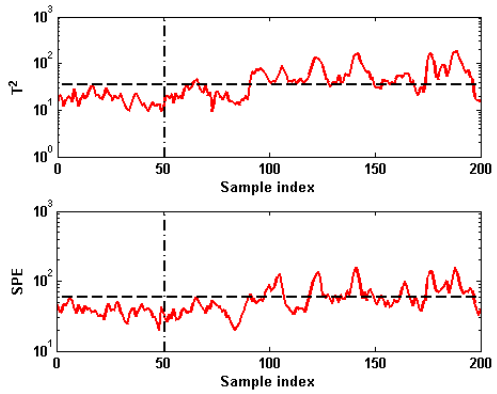
This realistic, simulated case study demonstrates that for large scale, complex dynamic processes, SPA can effectively address various challenges associated with process dynamics, nonlinearity and non-Gaussianity, therefore delivery improved fault detection and diagnosis results compared to some existing SPM methods.


T²
Sample index
SPE
Sample index

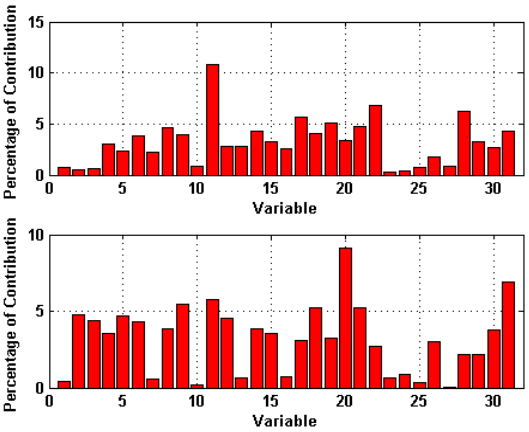
(a)


Percentage of Contribution
Variable
Percentage of Contribution
Variable

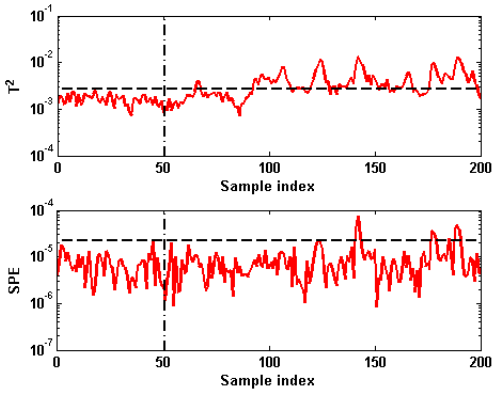
(b)


T²
Sample index
SPE
Sample index

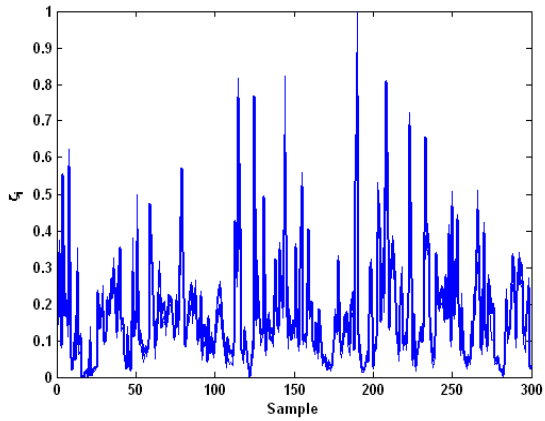
(c)


Percentage of Contribution
Variable
Percentage of Contribution
Variable

(d)


T²
Sample index
SPE
Sample index

(e)


ζ
Sample

(f)

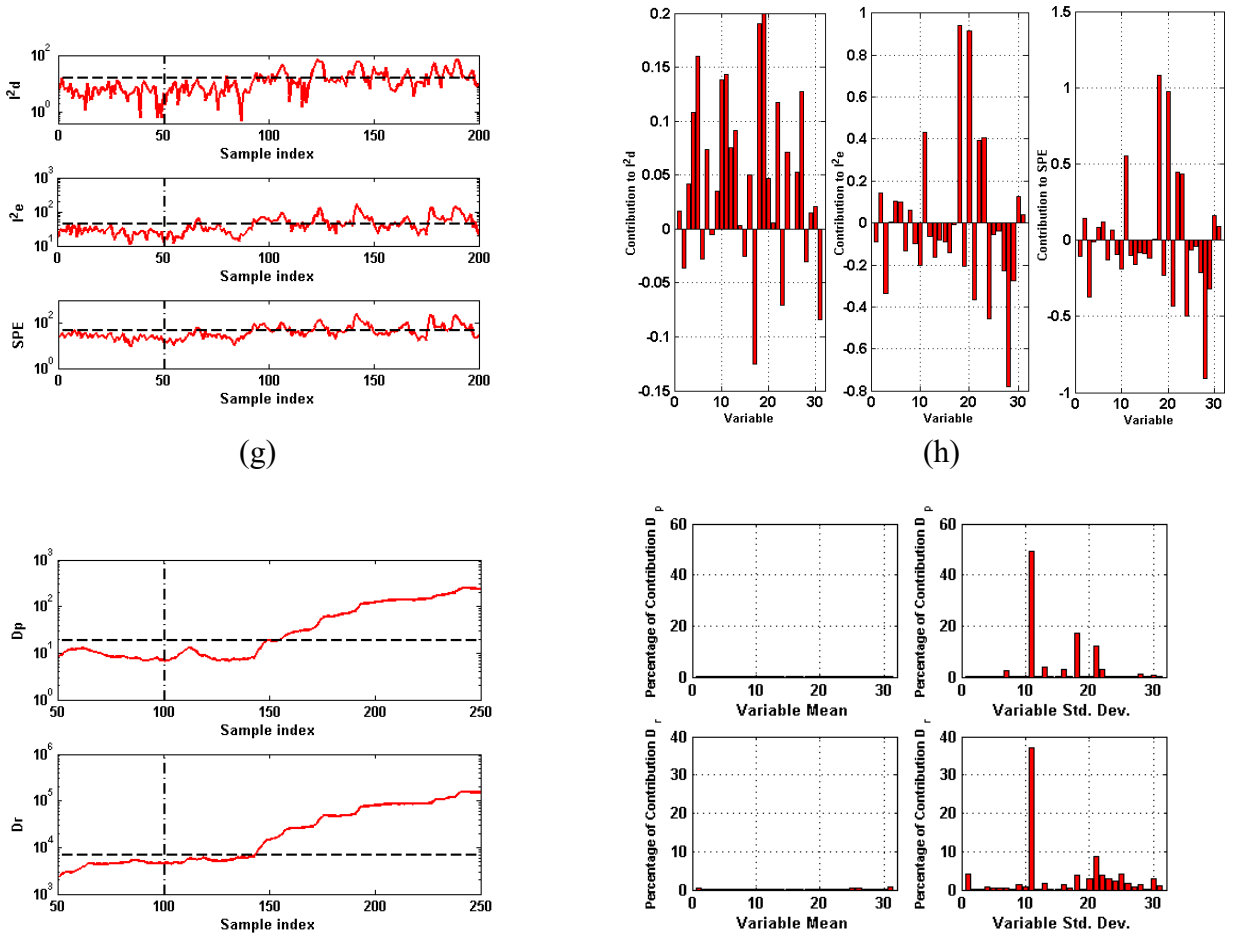


Figure 6. Detection and diagnosis of fault 12. PCA: (a) fault detection (top T^2, bottom SPE) and (b) diagnosis using contribution plots; DPCA: (c) fault detection (top T^2, bottom SPE) and (d) diagnosis using contribution plots; KPCA: (e) fault detection (top T^2, bottom SPE) and (f) diagnosis using reconstruction index ζ_i; ICA: (g) fault detection (top I_d^2, middle I_d^2, bottom SPE) and (h) diagnosis using contribution plots; SPA: (i) fault detection (top D_p, similar to T^2, bottom D_r, similar to SPE) and (j) diagnosis using contribution plots.

3.3 Spectroscopic data analysis: Soft sensor development

Besides fault detection and diagnosis, soft sensor is another important research area in process monitoring. By correlating the easily measured secondary variables with the primary variables, one application of soft sensor is to provide information on those important but hard-to-measure variables, such as product quality variables. Another application of soft sensor is to provide prediction on infrequently measured process variables (such as concentration of different components) so that prompt control actions can be taken. In the last few decades, spectroscopic techniques such as near-infrared (NIR) and UV/Vis spectroscopies have gained wide applications. Beyond their traditional applications in analytical chemistry, spectroscopic techniques are applied in many different fields, including biotechnological, pharmaceutical, petrochemical, and

agricultural and food industries (Gendrin et al., 2008; Karoui and De Baerdemaeker, 2007; Meher et al., 2006).

Because the spectroscopic readings at different wavelengths are highly correlated, it has been shown that variable selection could significantly improve a soft sensor's prediction performance and reduce the model complexity (Wang et al., 2015). Although many successful applications have been reported, such variable selection methods do have their limitations, such as (high) sensitivity to the choice of training data, and deteriorated performance when testing on new samples. One possible reason for these limitations is the removal of useful wavelengths or segments of wavelengths during the calibration process, which resulted in "tilted" model to overfit or capture the noise or unknown disturbances contained in the calibration data. As a result, the model prediction performance may deteriorate significantly when the model is extrapolated or applied to new samples. In fact, this limitation is not unique to spectroscopic chemometric models, it is true to all data-driven soft sensor models, which is in essence a balance between model accuracy and robustness.

To address this limitation, we proposed a feature-based soft sensor approach utilizing SPA. As shown in Figure 7, instead of selecting certain wavelengths or wavelength segments, the SPA-based soft sensor considers the whole spectrum which is divided into segments, and extracts different features over each spectrum segment to build the soft sensor. It is important to note that SPA-based fault detection and diagnosis extracts features sample-wise, *i.e.*, features are computed using consecutive measurements of a same variable; while the SPA-based soft sensor extracts features variable-wise, *i.e.*, features are computed using a segment of adjacent wavelengths of the same sample, which significantly reduces the number of input variables. Therefore, the SPA model contains the complete information from the full spectrum without any selection or removal, but significantly reduces the dimension of input variables by using the summarizing features instead of individual wavelengths. We expect such an approach would offer improved robustness without sacrificing model accuracy.

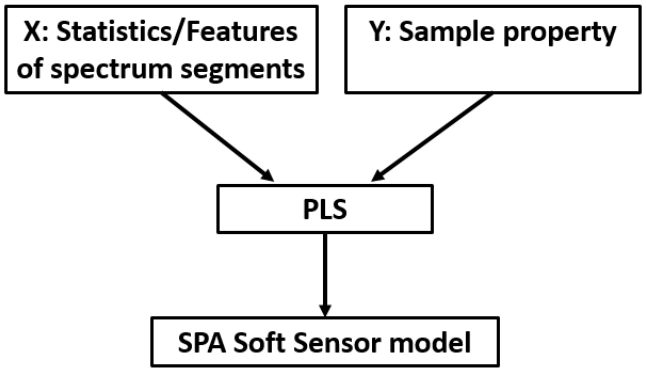
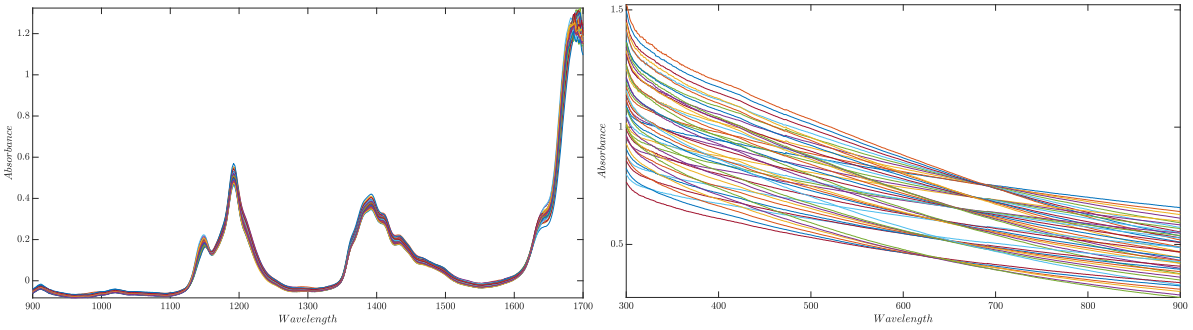


Figure 7. Schematic of SPA-based soft sensor

In this section, we use two case studies to demonstrate the versatile applicability and the performance of the SPA-based soft sensors. The data sets used in the case studies are the following, with the corresponding sample spectra given in Figure 8.

1. **Gasoline dataset:** This dataset consists of 60 samples of NIR absorbance spectra and corresponding octane numbers. Wavelength range is 900nm to 1700nm at 2nm interval. More details of the dataset can be found in (Kalivas, 1997).
2. **Coculture dataset:** This dataset consists of 47 samples of UV/Vis absorbance spectra of *E.coli* and *S. cerevisiae* coculture with known individual cell mass concentration. In this dataset, spectra were clearly separated into 6 groups. Wavelength range is 300nm to 900nm at 1nm interval. Detailed description of the dataset and the experimental design can be found in (Stone et al., 2017).



(a) (b)

Figure 8. NIR spectra of gasoline (a) and UV/Vis spectra of coculture datasets (different colors refer to different samples)

The performance of the SPA-based soft sensor is compared with a full PLS soft sensor utilizing all variables, and two representative variable selection methods: a shrinkage method (least absolute shrinkage and selection operator, or Lasso) (Tibshirani, 1996) and an interval based variable selection method (synergy interval PLS, or SiPLS) (Nørgaard et al., 2000). For consistent and fair comparison across different datasets, the datasets were divided into training, validation and test sub-sets in consistent proportions. Details of the data division for both datasets are given in Table 1. Literature (Xu and Liang, 2001) and our experiences suggest that such division of training and validation (*i.e.*, ~55% vs. ~45%) results in models that are generally without overfitting issues.

Table 1 Division of data into training, validation and test subsets

Dataset	Training (%)	Validation (%)	Test (%)	Total (%)
Gasoline	27 (45.0%)	21 (35.0%)	12 (20.0%)	60 (100%)
Coculture	21 (44.7%)	16 (34.0%)	10 (21.3%)	47 (100%)

To systematically test SPA-based soft sensor and compare its performance with full PLS, Lasso and SiPLS based soft sensors, we follow a Monte Carlo validation and testing (MCVT) procedure, which is an adapted Monte Carlo cross-validation (MCCV) (Xu and Liang, 2001). In addition, the following MCVT-based indices are proposed to assess the performance of different soft sensor approaches.

- Normalized root mean squared error ($NRMSE$) as percentage of the measurement range:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})_i^2}}{(y_{max}-y_{min})} \times 100\% \quad (1)$$

- Average $NRMSE$ ($\overline{NRMSE}$):

$$\overline{NRMSE} = \frac{\sum_{i=1}^{M} NRMSE_i}{M} \qquad (2)$$

- Standard deviation of NRMSE (σ_{NRMSE})

$$\sigma_{NRMSE} = \sqrt{\frac{\sum_{i=1}^{M} (NRMSE_i - \overline{NRMSE})}{M-1}} \qquad (3)$$

- Normalized mean prediction error (NMPE) as percentage of the measurement mean:

$$NMPE = \frac{\sum_{i=1}^{n} (y - \hat{y})_i}{\sum_{i=1}^{n} y_i} \times 100\% \qquad (4)$$

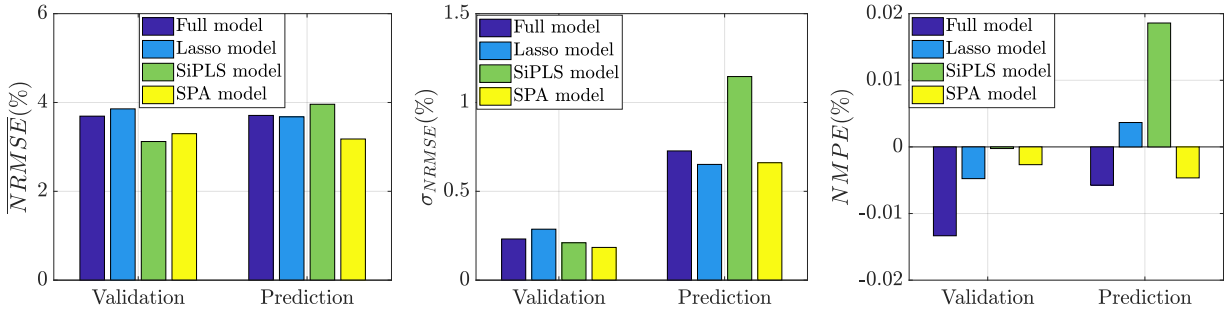
where n is the total number of validation (n_V) or prediction (n_P) samples in each MC run, and M is the total number of MC runs during validation (M_V) or prediction (M_P).

After optimization, the average model sizes in terms of number of variables/features in the final soft sensor models over 25 MC predictions are listed in Table 2. It can be seen that all models with variable selection are substantially smaller than the full model and SPA has the smallest model size in the coculture case study (where sample spectra do not contain a peak) and the second smallest model size in the gasoline case study (where sample spectra contain several peaks).

Table 2 Average number of variables/features of different soft sensors over 25 MC runs

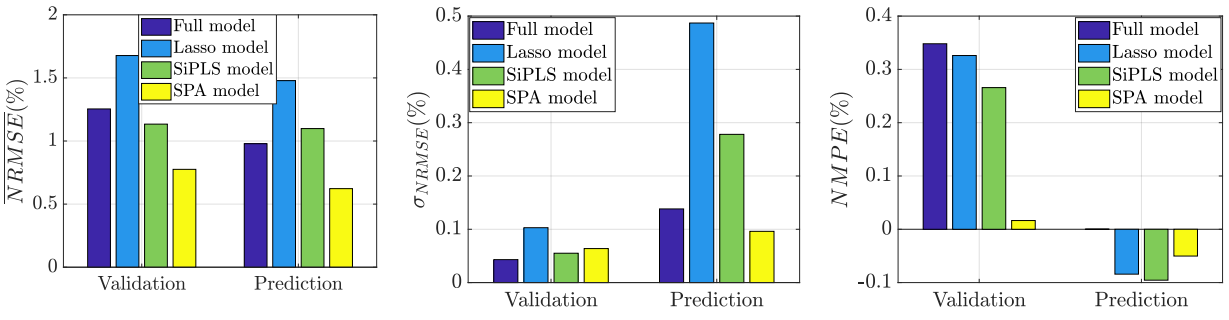
Dataset	Full PLS	SiPLS	LASSO	SPA
Gasoline	401	84	14	30
Co-cult (E. coli)	601	129	102	34
Co-cult (S. cerevisiae)	601	138	109	28

Figures 8-10 compare the performance indices of the full PLS model, Lasso, SiPLS and SPA for both gasoline and coculture data sets. These results show that in general SPA shows superior performance than the other methods. More importantly, we want to note that SPA performs especially better at extreme or boundary regions, as illustrated in Figure 12, which compares the model prediction from 4 different soft sensors and actual measurements. The red ellipse highlights the samples that SPA predictions are significantly better than other approaches.



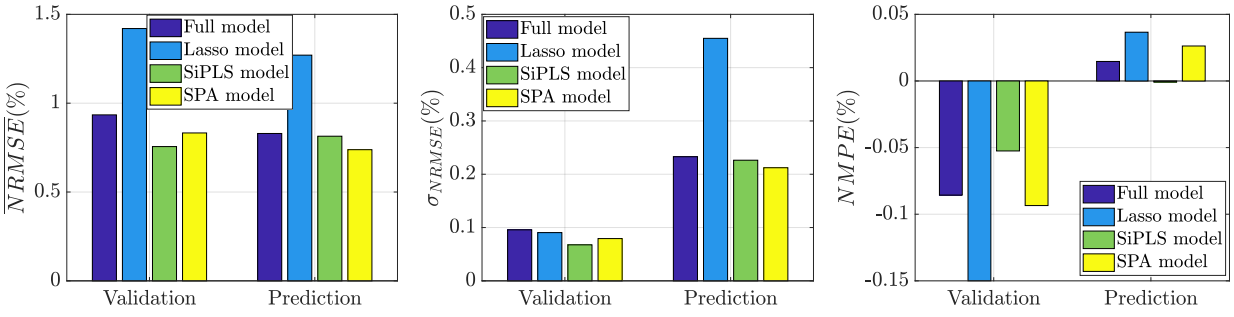
(a) (b) (c)

Figure 9. Comparison of soft sensors using gasoline data: (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $NMPE$



(a) (b) (c)

Figure 10. Comparison of soft sensors using coculture data (*E. coli*): (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $NMPE$



(a) (b) (c)

Figure 11. Comparison of soft sensors using coculture data (*S. cerevisiae*): (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $NMPE$

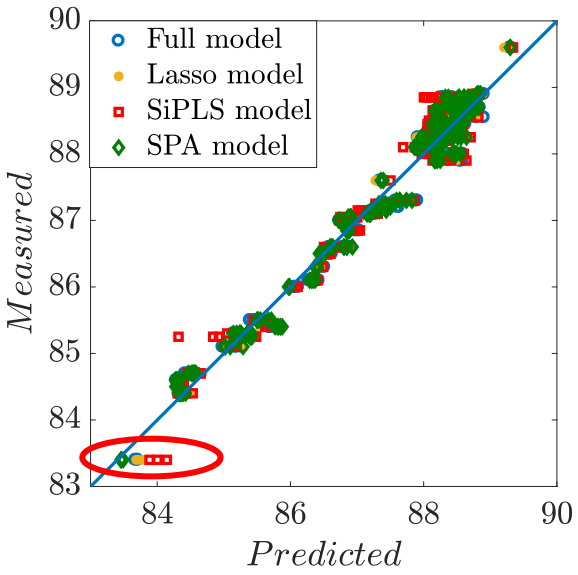


Figure 12. Comparison of predicted vs. measured octane numbers from different soft sensors using the gasoline data. The red ellipse highlights the region where SPA-based soft sensor performs significantly better than the full PLS, Lasso and SiPLS based soft sensors.

4 Addressing the 4V Challenges of Big Data

Smart manufacturing (SM) and big data generated from SM have drawn increased attention in the SPM community in the past few years (Qin, 2014; Severson et al., 2016). As detailed in (He and Wang, 2018), SPA offers some advantages in addressing the 4V challenges of big data, *i.e.*, Volume, Velocity, Variety and Veracity (Zikopoulos et al., 2012). In reducing the number of observations, which is one aspect of Volume, for batch processes sample-wise feature extraction of SPA-based method can reduce an entire batch (or batch step) into batch (or batch step) features; for continuous processes, window-based SPA approach is efficient in significantly reducing number of observations. In reducing the number of variables, which is the other aspect of Volume, variable-wise feature extraction in SPA-based models has been used to extract features from optical emission spectroscopy (OES) (Suthar et al., 2018), NIR and UV-Vis spectra, which effectively reduces number of variables to significantly smaller number of features. For data variety, SPA can help as statistics extracted from different data sources can be conveniently integrated together. For Veracity, SPA is advantageous as data uncertainty will have much less impact on extracted features (*e.g.*, statistics) than variable themselves. Finally, because SPA can significantly reduce problem size in both time/sample-wise and variable-wise, and it often eliminates data pre-processing, SPA has the potential to be used for monitoring real-time streaming data (*i.e.*, Velocity).

5 Discussions and Conclusions

In this work, we use multiple case studies to examine the capabilities of SPA-based methods in addressing the existing challenges of statistical process monitoring, including process dynamics, nonlinearity, and non-Gaussianity. Using an illustrative example, we demonstrate that SPA can transform dynamic, nonlinear process data that exhibit strong non-Gaussianity into multivariate Gaussian distributed features that capture key process characteristics. The extracted features can then be conveniently modelled by multivariate statistical methods such as PCA and PLS for various applications, such as fault detection and diagnosis, and soft sensor or virtual metrology. The superior performance of the SPA-based method in fault detection and diagnosis is further demonstrated using the benchmark TEP case study. Because SPA-based fault diagnosis method links the root cause of a fault to different variable statistics via contribution plots, it provides extra information in addition to identifying the major fault-contributing variable(s), such as whether the fault is due to a change in the variable mean or variance. For soft sensor applications, the performance of SPA-based soft sensor is tested using a lab UV/Vis dataset and an industrial NIR dataset, which confirmed its superior and robust performance, especially at extreme or boundary regions.

It is worth noting that SPA is not without its limitations, such as the amount of data required to estimate various statistics, trade-off between robustness and sensitivity, which will be studied more systematically in the future. In addition, how to identify the key features to be extracted, as well as under what conditions that SPA would outperform or underperform other methods need further investigation.

6 Acknowledgements

Financial supports from National Science Foundation, NSF-CBET #1547124 (He), NSF-CBET #1805950 (He), and NSF-CBET #1547163 (Wang and Shah) are greatly appreciated.

References

Bequette, B.W., 1998. Process dynamics: modeling, analysis, and simulation. Prentice Hall PTR Upper Saddle River, NJ.

Bingham, E., Hyvärinen, A., 2000. A fast fixed-point algorithm for independent component analysis of complex valued signals. Int. J. Neural Syst. 10, 1–8.

Deng, X., Tian, X., 2013. Nonlinear process fault pattern recognition using statistics kernel PCA similarity factor. Neurocomputing 121, 298–308.

Deng, X., Tian, X., Chen, S., Harris, C.J., 2016. Statistics local fisher discriminant analysis for industrial process fault classification, in: Control (CONTROL), 2016 UKACC 11th International Conference On. pp. 1–6.

Downs, J.J., Vogel, E.F., 1993. A plant-wide industrial process control problem. Comput. Chem. Eng. 17, 245–255.

Galicia, H., 2012. Advanced monitoring and soft sensor development with application to industrial processes. Auburn University.

Galicia, H.J., He, Q.P., Wang, J., 2012. Statistics Pattern Analysis based fault detection and diagnosis, in: CPC VIII Conference.

Gendrin, C., Roggo, Y., Spiegel, C., Collet, C., 2008. Monitoring galenical process development by near infrared chemical imaging: One case study. Eur. J. Pharm. Biopharm. 68, 828–837.

He, F., Xu, J., 2016. A novel process monitoring and fault detection approach based on statistics locality preserving projections. J. Process Control 37, 46–57.

He, Q.P., Wang, J., 2018. Statistical process monitoring as a big data analytics tool for smart manufacturing. J. Process Control 67, 35–43. https://doi.org/10.1016/j.jprocont.2017.06.012

He, Q.P., Wang, J., 2017. Statistical Process Monitoring: recent developments and future prospects for smart manufacturing, in: Proceedings of 2018 International Symposium on Process Systems Engineering.

He, Q.P., Wang, J., 2011. Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. AIChE J. 57, 107–121. https://doi.org/10.1002/aic.12247

Kalivas, J.H., 1997. Two data sets of near infrared spectra. Chemom. Intell. Lab. Syst. https://doi.org/10.1016/S0169-7439(97)00038-5

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R., Bakshi, B.R., 2002. Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem, in: Computers and Chemical Engineering. https://doi.org/10.1016/S0098-1354(01)00738-4

Karoui, R., De Baerdemaeker, J., 2007. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. Food Chem. 102, 621–640.

Ku, W., Storer, R.H., Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. Chemom. Intell. Lab. Syst. 30, 179–196.

Lee, J.-M., Yoo, C., Choi, S.W., Vanrolleghem, P.A., Lee, I.-B., 2004a. Nonlinear process monitoring using kernel principal component analysis. Chem. Eng. Sci. 59, 223–234.

Lee, J.-M., Yoo, C., Lee, I.-B., 2004b. Statistical process monitoring with independent component analysis. J. Process Control 14, 467–485.

Ma, H., Hu, Y., Shi, H., 2011. Statistics kernel principal component analysis for nonlinear process fault detection, in: Intelligent Control and Automation (WCICA), 2011 9th World Congress On. pp. 431–436.

Meher, L.C., Sagar, D.V., Naik, S.N., 2006. Technical aspects of biodiesel production by transesterification---a review. Renew. Sustain. energy Rev. 10, 248–268.

Ning, C., Chen, M., Zhou, D., 2014. Hidden Markov model-based statistics pattern analysis for multimode process monitoring: an index-switching scheme. Ind. Eng. Chem. Res. 53, 11084–11095.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-

infrared spectroscopy. Appl. Spectrosc. https://doi.org/10.1366/0003702001949500
Qin, S.J., 2014. Process data analytics in the era of big data. AIChE J. 60, 3092–3100.
Qin, S.J., 2012. Survey on data-driven industrial process monitoring and diagnosis. Annu. Rev. Control 36, 220–234.
Rendall, R., Lu, B., Castillo, I., Chin, S.-T., Chiang, L.H., Reis, M.S., 2017. A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. Ind. Eng. Chem. Res. 56, 8590–8605.
Ricker, N.L., 1996. Decentralized control of the Tennessee Eastman challenge process. J. Process Control 6, 205–221.
Russell, E.L., Chiang, L.H., Braatz, R.D., 2000. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. Chemom. Intell. Lab. Syst. 51, 81–93.
Severson, K., Chaiwatanodom, P., Braatz, R.D., 2016. Perspectives on process monitoring of industrial systems. Annu. Rev. Control 42, 190–200.
Song, Y., Jiang, Q., Yan, X., 2015. Fault diagnosis and process monitoring using a statistical pattern framework based on a self-organizing map. J. Cent. South Univ. 22, 601–609.
Stone, K.A., Shah, D., Kim, M.H., Roberts, N.R.M., He, Q.P., Wang, J., 2017. A Novel Soft Sensor Approach for Estimating Individual Biomass in Mixed Cultures. Biotechnol. Prog.
Stone, K.S., Shah, D., He, Q.P., Wang, J., 2017. A Novel Soft Sensor for Estimating Individual Biomass in Mixed Cultures, in: Proceedings of 2017 American Control Conference. pp. 4797–4802.
Suthar, K., Shah, D., Wang, J., Peter He, Q., 2018. Feature-based Virtual Metrology for Semiconductor Manufacturing, in: Computer Aided Chemical Engineering. https://doi.org/10.1016/B978-0-444-64241-7.50342-6
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 267–288.
Wang, J., He, Q.P., 2010. Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis. Ind. Eng. Chem. Res. 49, 7858–7869. https://doi.org/10.1021/ie901911p
Wang, R.C., Edgar, T.F., Baldea, M., Nixon, M., Wojsznis, W., Dunia, R., 2015. Process fault detection using time-explicit Kiviat diagrams. AIChE J. 61, 4277–4293.
Wang, Z., He, Q.P., Wang, J., 2015. Comparison of variable selection methods for PLS-based soft sensor modeling. J. Process Control 26, 56–72. https://doi.org/10.1016/j.jprocont.2015.01.003
Wang, Z., Xu, J., Gao, D., Fu, Y., 2013. Multiple empirical kernel learning based on local information. Neural Comput. Appl. 23, 2113–2120.
Wold, S., Kettaneh-Wold, N., MacGregor, J.F., Dunn, K.G., 2009. Batch process modeling and MSPC.
Xu, Q.-S., Liang, Y.-Z., 2001. Monte Carlo cross validation. Chemom. Intell. Lab. Syst. 56, 1–11. https://doi.org/10.1016/s0169-7439(00)00122-2
Zhang, H., Tian, X., Deng, X., Cai, L., 2015. A local and global statistics pattern analysis method and its application to process fault identification. Chinese J. Chem. Eng. 23, 1782–1792.
Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., Corrigan, D., others, 2012. Harness the Power of Big Data The IBM Big Data Platform. McGraw Hill Professional.