Digital Twins for Process Monitoring and Anomaly Detection

Anziel Malandri,a Mehmet Mercangözb\*

aFluor Corporation, 140 Pinehurst Road, Farnborough GU14 7BF, United Kingdom

bImperial College London, Imperial College Rd, London SW7 2AZ, United Kingdom

m.mercangoz@imperial.ac.uk

Abstract

As industrial processes become more intricate over time, data-driven process monitoring has emerged as an effective approach for detecting faults within these complex systems. However, the requirement for large volumes of historical data poses a challenge when implementing data-driven process monitoring in greenfield plants lacking such data. This paper introduces a novel solution by proposing the utilisation of digital twins for the implementation of data-driven process monitoring, specifically based on principal component analysis (PCA). By employing simulations used during the design stage, synthetic training data can be generated to emulate the behaviour and correlations observed in the real process. The synthetic data can be used to train a PCA model, eliminating the need for an extensive data set to achieve reliable results. Consequently, this approach expedites the implementation of data-driven process monitoring. The effectiveness of this method is shown through case studies involving a reactive absorption process. The results show that some amount of actual process data is still required for recalibration, but the proposed approach can be built with 75% fewer samples compared to a method based purely on process data to operate with equivalent or better performance.

**Keywords**: PCA, Digital Twins, Fault Detection, Transfer Learning

* 1. Introduction

Industrial processes become progressively more complex as they expand. This growing complexity makes it increasingly impractical to depend solely on human operators to identify and manage faults and emergencies. Neglecting to spot faulty conditions promptly can result in substantial safety, environmental, and financial problems. Consequently, there is a need for the creation of automated process monitoring techniques to aid operators in addressing faulty conditions (Harrou et al., 2021). To accomplish this goal, many data-driven process monitoring and fault detection methods were formulated (Jiang et al., 2019). Nevertheless, the demand for a substantial amount of historical data presents a significant hurdle for data-driven models for greenfield chemical plants, where data may be scarce or utterly absent if the plant has not yet been commissioned. To address this issue, the present paper proposes leveraging digital twins based on design-stage models to generate synthetic data. Said data is used to train a data-driven process monitoring method based on principal component analysis (PCA).

The contribution of this paper is the investigation of digital twins to expedite the implementation of data-driven process monitoring methods when plant data is scarce or not available – and the demonstration of this approach on an experimental facility. The remaining sections of the paper are organised in the following manner: The background is presented in section 2. Section 3 describes the methodology of this implementation. Section 4 will illustrate and discuss the results of implementing the proposed method. Finally, section 5 provides conclusions and prospects for future work.

* 1. Background

This section overviews principal component analysis, the monitoring statistics to be used alongside PCA, and the Sobol sequence.

* + 1. Principal Component Analysis (PCA)

PCA is a dimensionality reduction method that can discover underlying features and correlations within a multivariate dataset. It projects the dataset into a lower dimensional subspace using singular value decomposition (SVD) (Abdi and Williams, 2010). Before SVD is applied, the input dataset , is first standardised using its mean and variance. From here on, refers to the standardised data. After which, a covariance matrix is calculated to obtain the loading matrix and using Eq. (2).

|  |  |
| --- | --- |
|  | (2) |

where is a contains the eigenvalues of the covariance matrix. The eigenvalues are equivalent to the variance of each principal component. Typically, only the first principal components () are used to construct the PCA model.

* + 1. Monitoring Statistics

This subsection will describe the monitoring statistics , SPE, and One Class Support Vector Machines (OCSVMs). PCA also provides a means to isolate the detected faults, but this will be outside the scope of the presented work.

* + - 1. Hotelling

The statistic calculates the variations solely within the PC values at each time point. Specifically, the value is determined by summing the squares of the retained PC scores and dividing this sum by the corresponding eigenvalue derived from non-faulty data as seen in Eq. (3) (Hotelling, 1933).

|  |  |
| --- | --- |
|  | (3) |

Where and are the retained loading matrix and eigenvalues of the PCA model.

* + - 1. SPE

The SPE, or Q metric, is used to detect faults within the reconstruction space, SPE is calculated by Eq. (5) (Joe Qin, 2003).

|  |  |
| --- | --- |
|  | (5) |

Where is the reconstructed value of the input .

* + - 1. OCSVM

OCSVM is a method that classifies data into groups based on a kernel function. (Wang et al., 2006). In broad terms, the OCSVM procedure typically employs a kernel function to map input data points into a higher-dimensional feature space. The differentiation between normal and anomalous data becomes more distinct and manageable in this elevated feature space.

* + 1. Sobol Sequence

The Sobol sequence is a type of quasi-random, low-discrepancy sequence frequently employed in Monte Carlo simulations for conducting sensitivity analyses. The objective of Monte Carlo simulations itself is to comprehensively explore the entire input space using a reasonably sized sample (Burhenne et al., 2011).

* 1. Methodology
     1. Experiments on the Reactive Absorption Process

The real data is generated via experiments using an experimental reactive absorption process for CO2. The normal samples are compiled from many runs with variations within a specific operating protocol. Faulty conditions are generated by (1) substantially increasing the CO2 flow rate and decreasing the air flow rate, (2) External cooling of the feed gas inlet of the absorber column, (3) Induced failures of the column outlet pump, causing flooding to occur within the column. Additionally, occasional sensor failures cause faults to occur within the normal operation protocol as well.

* + 1. Absorption Rig Model

The process is modelled using Aspen Plus V11. The input variables used are based on the controllable values of the process. The specifications of the absorption column are based on the equipment's design data to replicate the design stage's fidelity. The heat exchangers are modelled by specifying a constant heat flow to replicate the effect of instantaneous fluctuations in flow rate. The schematic of the model is illustrated in Figure 1**.**

A diagram of a machine

Description automatically generated

Figure 1. Aspen Plus model schematic

* + 1. Synthetic Data Generation

Based on the available measurements, eight variables are chosen as listed in Table 1. Measurements 1-5 are input variables, whereas 6-8 are the output variables within the Aspen Plus model. As such, five variables must be varied for the synthetic data generation. Using the Sobol sequence, 2048 data points are generated with the lower and upper bounds specified according to a 95% confidence interval of the actual measurements.

Table 1. Chosen measurements.

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Stream/Unit** | **Variable** | **Description** |
| 1 | WATER-IN | Temperature | Process water temperature |
| 2 | CO2 | Flow rate | CO2 flow rate |
| 3 | AIR | Flow rate | Air flow rate |
| 4 | GAS-COLD | Pressure | Feed gas pressure |
| 5 | LEANIN | pH | Lean solvent pH |
| 6 | RICHOUT | pH | Rich solvent pH |
| 7 | C1 | Level | Column liquid level |
| 8 | GAS-HOT | Temperature | Feed gas temperature |

* + 1. Synthetic Data Generation

The training procedure begins by standardising the data using its mean and variance. The PCA model is then computed using the SVD method to obtain each principal component's loading matrix and eigenvalues. The eigenvalues are then normalised and sorted in descending order to obtain the explained variance ratio of each principal component. Typically, the number of principal components retained corresponds with a cumulative explained variance ratio of 95%. However, for this study only the first two principal components with a lower cumulative explained variance are selected for ease of analysis. Afterwards, the training data is used to obtain the feature and reconstruction space values used to calculate the T2 and SPE thresholds and train the OCSVM.

* + 1. Recalibration

To address the mismatch in the centring of the operating points of the data and the simulation, the PCA needs to be recalibrated using incremental amounts of the actual data set over time. While previous studies recalibrate by updating the digital twin (Kubosawa et al., 2022). This study opts to recalibrate by directly updating the PCA model.

The PCA is recalibrated by first recalculating the mean and variance of the data, for standardisation. The PCA is then retrained and thresholds for T2 and SPE are recalculated. Another method is to retrain the OCSVM without updating the mean and variance.

To measure the impact of recalibration, the false alarm rate (FAR) metric is used as described in Eq. (7). The entire real data on normal conditions are fed through the PCA, classifying the data as faulty or otherwise.

|  |  |
| --- | --- |
|  | (7) |

* 1. Results and Discussion
     1. Synthetic PCA Performance

The model is tested using real data sets on normal and anomalous conditions to analyse the synthetic PCA's performance. The results are illustrated in Figure 2. The T2 performance is rather poor, where some data points in the anomalous region have lower T2 values than the normal region. On the other hand, The SPE metric distinguishes faulty conditions rather well. This issue may arise from the inherent errors of the Aspen model's calculations and the lack of statistical information from the real data set. The latter appears more significant since the data is standardised using mean and variance. Therefore, having different means will shift the PC score's centre. And because the T2 metric measures deviations in PC values away from 0, the T2 evaluation will be inaccurate. Another issue arising from a lack of statistical information is that there is no reliable way of determining the monitoring statistic's thresholds, as the 99.5% confidence intervals are a property of the real data set.

As the lack of statistical information is inherent at the design stage, a method is required to recalibrate the data set and determine thresholds.

A graph with orange and blue lines

Description automatically generatedA graph of a graph showing a blue and orange line

Description automatically generated

Figure 2. Synthetic PCA SPE (left) and T2 (right) values.

* + 1. Synthetic-PCA Recalibration

Recalibration is conducted over increments of 1000 data points of the actual data set on normal conditions. As a comparison, a PCA trained over the actual data set is also constructed. The FAR values of this incremental recalibration are illustrated in Figure 3. The results show that the FAR values for the T2 and SPE recalibration drop the fastest, requiring only 15000 data points to achieve a 0.5% FAR. At the same time, the OCSVM achieves 0.5% FAR by using all the actual data and the true PCA using around 20000 data points. The discrepancy between the T2 and SPE with OCSVM may arise because the threshold calculation for T2 and SPE uses the F and normal distributions, respectively. In contrast, OCSVM approximates a kernel map, thus creating a tighter confidence interval than T2 and SPE. Regarding the discrepancy between the synthetic and non-synthetic PCA, it may arise due to random errors within the actual data set skewing the PCA fit. Whereas the synthetic data set provides a clear correlation with a much smaller error. To further illustrate the recalibration's performance, a synthetic PCA recalibrated with 15000 data points are tested and shown in Figure 4. This model can have low FAR while still detecting the faults within the anomalous data set. Therefore, using synthetic PCA with T2 and SPE appears to be optimal. Not shown in the paper is an alternative approach where the input process data were fed to the Aspen model and the Aspen model outputs were compared with actual output process data to generate an error metric analogous to an SPE value. This approach did not perform very well as it cannot account for large deviations, which are compatible with model predictions.

* 1. Conclusions

This study introduced a concept for leveraging synthetic training data for a PCA-based process monitoring approach. Synthetic training data allows data-driven process monitoring methods to be trained on minimal real data. The proposed method was validated by testing the PCA on real data, both normal and faulty. Upon recalibration, the synthetic-trained PCA could perform on-par or better compared to a PCA using real data with 75% less training samples. This method can be extended by using higher fidelity design information such as detailed heat exchangers and valve models. More complex models such as adversarial auto-encoders can also be used to generate SPE values to account for the nonlinearity in the process. It is well-known that closed-loop controllers influence the distribution of process data. Replicating the control architecture in the simulations to generate closed-loop data and analysing the impact of synthetic closed-loop data on the performance of the proposed approach is a direction for future work.

A graph of a number of samples

Description automatically generatedA graph with blue and orange lines

Description automatically generatedA graph of a graph

Description automatically generated

Figure 3. FAR values for recalibrating T2 and SPE (left), OCSVM (centre), and non-artificial (right) PCA.

A graph of a number of different colored lines

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generated

Figure 4. SPE (left) and T2 (right) values for Synthetic PCA updated with 15000 samples.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comp Stat 2: 433--459.

Burhenne, S., Jacob, D., Henze, G.P., others, 2011. Sampling based on Sobol’sequences for Monte Carlo techniques applied to building simulations, in: Proc. Int. Conf. Build. Simulat. pp. 1816–1823.

Harrou, F., Sun, Y., Hering, A.S., Madakyaru, M., Dairi, A., 2021. Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches, Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches. Elsevier.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417.

Jiang, Q., Yan, X., Huang, B., 2019. Review and Perspectives of Data-Driven Distributed Monitoring for Industrial Plant-Wide Processes. Ind. Eng. Chem. Res. 58, 12899–12912.

Joe Qin, S., 2003. Statistical process monitoring: basics and beyond. J. Chemom. A J. Chemom. Soc. 17, 480–502.

Kubosawa, S., Onishi, T. and Tsuruoka, Y., 2022. Sim-to-real transfer in reinforcement learning-based, non-steady-state control for chemical plants. SICE Journal of Control, Measurement, and System Integration, 15, 10-23.

Wang, D., Yeung, D.S., Tsang, E.C.C., 2006. Structured one-class classification. IEEE Trans. Syst. Man, Cybern. Part B 36, 1283–1295.