Data-driven diagnostics of variability during changeover in biopharmaceutical freeze-drying

Gianluca Lombardini,a Mohamed Rami Gaddem,a Stephanie Knueppel,b Sara Badr,a Hirokazu Sugiyama,a\*

aDepartmentof Chemical System Engineering, The University of Tokyo, Tokyo 113-8656, Japan

bEngineering, Science & Technology, F. Hoffmann – La Roche Ltd., Wurmisweg, 4303 Kaiseraugst, Switzerland

\*sugiyama@chemsys.t.u-tokyo.ac.jp

Abstract

Changeover processes in biopharmaceutical freeze-drying are crucial to ensure a sterile environment for drug product manufacturing. Leak tests aim to avoid contamination in the freeze-drying chamber using pressure measurements over time. However, process disturbances such as temperature fluctuations can lead to elevated pressure measurements and other process variabilities, which in turn can lead to false alarms. This work investigates the main sources of variability during leak testing by analyzing industrial, noisy batch process data with a relatively low number of batches using principal component analysis and a sparse formulation thereof. It is demonstrated, that for such datasets, adding a sparsity constraint to principal component analysis improves consistency and simplifies interpretability of the results. This work highlights the challenges of using data-driven methodologies with real industrial data and shows the need for developing more tailored solutions to handle noisy, drifting, and imbalanced datasets.

**Keywords**: Freeze-drying, leak detection, industrial data, data-driven diagnostics, false alarms

* 1. Introduction

Pharmaceutical freeze-drying is an integral unit operation allowing drug products to be stored over extended time periods. As with all pharmaceutical manufacturing processes, product quality is of paramount importance. Apart from production process and product quality control, changeover processes such as cleaning and sterilization in place (CIP/SIP) or leak testing are essential for assuring high-grade, contaminant-free products. Robust monitoring and control strategies are needed for the safe and efficient operation of pharmaceutical freeze-drying processes, which are already time-, energy-, and cost-intensive with a high potential for quality-impairing faults. However, advanced monitoring strategies for changeover processes are scarce. Leak testing, for example, is performed to measure the amount of external air entering the freeze-dryer, possibly contaminating the sterile environment. Many pharmaceutical companies base leak testing solely on the pressure increase over a specified timeframe. However, pressure increase under vacuum is subject to a multitude of factors, such as temperature fluctuations or evaporation of remaining water in the chamber. These factors lead to so-called internal or virtual leaks, which elevate pressure and potentially cause false alarms (Sahni et al., 2022). Failed tests due to false alarms are often passed through simple repetition without any physical modifications to the equipment. Even so, repetition increases downtime and hence decreases the availability of the machine for production. Virtual leaks are usually characterized by non-linear pressure profiles over time. The only available model to separate virtual from real leaks so far was proposed by Calzavara et al. (2021), which follows a mechanistic approach.

In the last few decades, an extensive amount of data-driven process monitoring, and fault diagnosis systems have been proposed (Yin et al., 2014). Many of these methods, however, are tested and validated on simulated datasets but can be difficult to apply to real industrial datasets (Ji and Sun, 2022). As more data from real pharmaceutical manufacturing plants is being stored, the possibilities to propose and test appropriate monitoring methods are increasing. For that, however, characteristics of industrial data, such as individual machine characteristics, process variabilities, low sample size, non-relevant process variables, and noise must be accounted for. Furthermore, it has been shown that in addition to data drifts, shifts in data baselines can occur over time, especially before and after equipment maintenance (Zürcher et al., 2022), which is an additional source of variability.

This work focuses on identifying the main sources of variability from a low sample size, noisy, and high dimensional dataset of two parallelized, industrial freeze-dryers within the same drug product filling line. Two multivariate methods, namely multiway principal component analysis (PCA) and a sparse formulation thereof (SparsePCA), are compared. It is demonstrated that including multivariate methods in leak test evaluation can reduce false alarms while maintaining leak detection reliability in comparison to currently employed empirical approaches.

* 1. Methods
     1. Data Preparation

Data preparation was carried out according to Figure 1. Almost two years of process data were available. First, the sensor measurements were extracted from a data historian, and appropriate sensors were selected based on discussions with process experts and by analyzing the measuring frequency of each sensor. Based on the nature of the data compression algorithm used to store process sensor measurements, sensors with an average of less than one data point per ten minutes during leak testing were considered irrelevant and discarded, resulting in J = 28 variables. To obtain variable profiles over time for each batch (interchangeably used with leak test), sensor measurements were interpolated every full minute to obtain K = 225 time points. For freeze-dryer 1 (FD 1), I1 = 57 batches were extracted and I2 = 48 for freeze-dryer 2 (FD 2), giving a total of I = 105 batches. In the second step, the three-dimensional array *X* was batch-wise unfolded to obtain the data matrix *X*. Prior to transforming the data by PCA and SparsePCA, *X* was centered by the mean and scaled by the standard deviation.

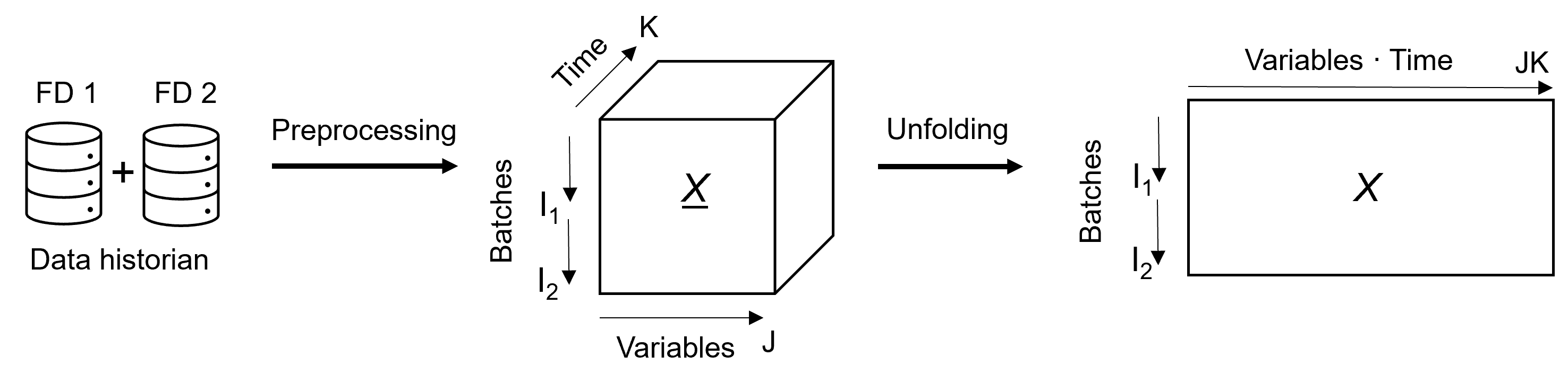


Figure 1: Data preparation workflow was conducted in two main steps. First, process data was extracted from a data historian and preprocessed into an array *X* of size I×J×K, where I = I1 + I2. Then, the array is batch-wise unfolded into a matrix *X* of size I×JK (or 105×6’300).

* + 1. Leak Test Classification

Real leaks were identified based on maintenance records, while virtual leaks were identified as runs that featured failed tests with non-linear pressure profiles that were passed upon simple repetition. In total, six classes of leak tests were defined: FD 1, year 1; FD 1, year 2; FD2, year 1; FD 2, year 2; Virtual leaks; Real leaks.

* + 1. PCA and SparsePCA

PCA decomposes a data matrixinto C principal components with score matrix with score vectors [ *t*1  *t*2  *… t*C ], a loading matrix and a residual error:

|  |  |
| --- | --- |
| . | (1) |

The decomposition can be formulated as an optimization problem according to

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

where and *F* refers to the Frobenius norm. The problem can be easily solved by singular value decomposition. For SparsePCA, many formulations are available. The one used in this work is formulated as an optimization problem using the Scikit-learn implementation in Python (Pedregosa et al., 2011) with an *l*1 penalty on the components,

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

where the sparsity coefficient *α* equals 0.1 in this work. SparsePCA has been known and previously applied for process monitoring (Luo et al., 2017) and in other fields, e.g., image analysis (Sjöstrand et al., 2007).

From *P* with elements *,* a sum of loadings value for each variable *j* for each component *c* is defined as

|  |  |
| --- | --- |
|  | (4) |

to simplify extracting dominating variable contributions for each component.

* + 1. Silhouette Score

To quantify and compare the ability of PCA and SparsePCA to cluster batches according to freeze-dryer and other factors, the average silhouette score (Rousseeuw, 1987),

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

is used, where *ai* is the mean intra-class distance of sample *i* and *bi* is the mean nearest-class distance of sample *i*. *S* ranges from 1 to -1, where 1 is the best value, meaning good clustering, and -1 the worst value.

* 1. Results and Discussion
     1. Leak Test Variability

In Figure 2, the first six component scores are shown for PCA and SparsePCA. When the results were compared with each other, two distinct differences were observed:

1. *t*1 vs. *t*2: The leak tests form four distinct clusters according to FD and year. SparsePCA results in a silhouette score of *S* = 0.86. Ordinary PCA does not achieve such clear separation, with a lower silhouette score of *S* = 0.59.
2. *t*5 vs. *t*6: SparsePCA separates the failed leak tests due to a real leak from all other leak tests on a single score vector *t*5 with *S* = 0.89. Using ordinary PCA, the best separation of the tests during a real leak and all other leak test was found using both *t*5 and *t*6, where S = 0.60. This combination gives a better separation of the real leaks than *t*3 and *t*4.

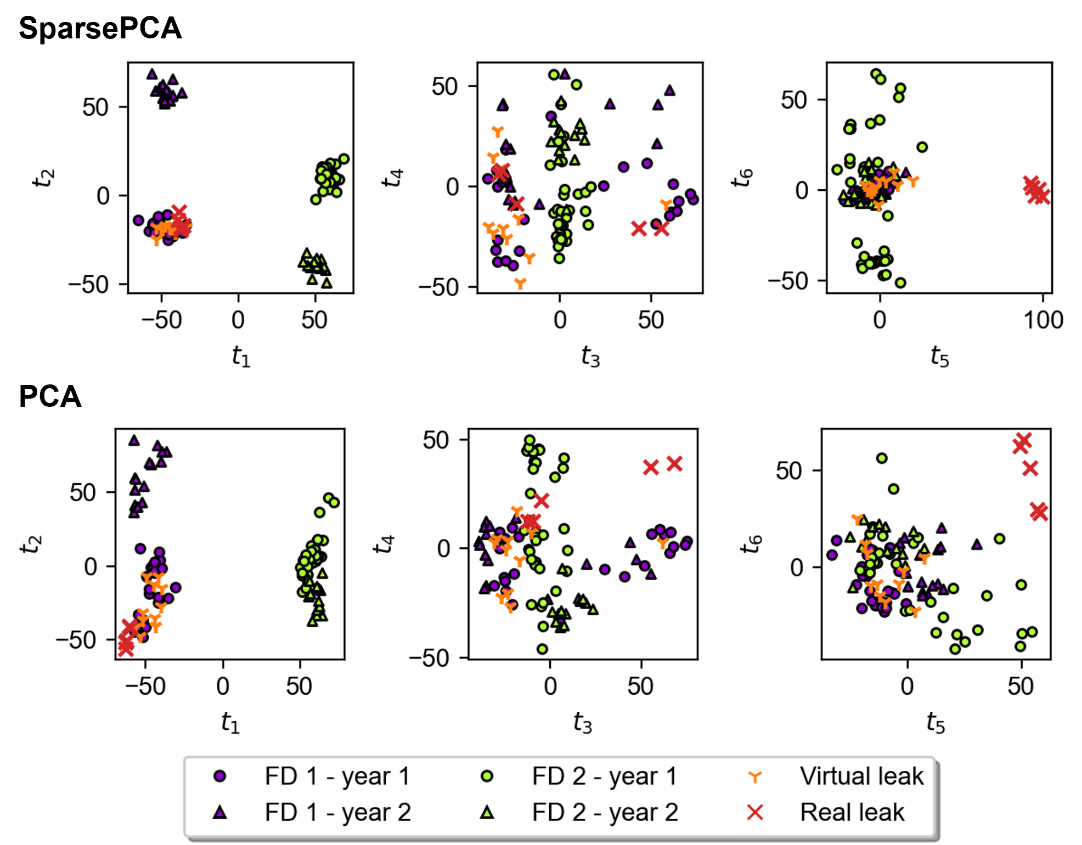


Figure 2: The first six principal components are shown in score plots for SparsePCA and PCA. The leak testing batches are labeled according to freeze-dryer (FD) and year. Failed batches due to virtual or real leaks are labeled individually.

The score plots showed that the SparsePCA results are much more useful for identifying and analyzing variability between leak tests than results from ordinary PCA. Interpretation was simplified due to clustering of the leak tests with higher silhouette scores. The first component scores *t*1 divided the leak tests according to machine-to-machine variability. The second component scores *t*2 captured year-to-year variability, caused by, e.g., year-end equipment maintenance and sensor recalibration. The fifth component score *t*5 separated tests during a real leak, which allowed for determining leak location by investigating the sum of loadings plot in Figure 3. The plot points out five dominating variables of mainly two groups. Variables 10, 11 and 17 represent cooling rod measurements. Variables 23 and 24 are the pressure measurements of two identical capacitance manometers. Closer inspection of the loading plots of variable 24 (*p*5,24,*k*) revealed that the pressure increase over batch time *k* was, as expected, higher for the batches during a real leak, as more external air entered the freeze-dryer. Variable 17 is the opening percentage of the control valve for cooling by liquid nitrogen of one of the cooling rods in the condenser. The loading plot over batch time *k* (*p*5,17,*k*) shows that the valve opening was higher during leak testing in case of a real leak. Maintenance reports indeed confirmed that a leak was found in the condenser.

A graph of different sizes and numbers

Description automatically generated with medium confidence

Figure 3: The sum of loadings for each variable for component 5 is shown on the left. On the right, individual loadings of component 5 are shown for variables number 17 (top) and 24 (bottom).

The simplified interpretability of SparsePCA can be explained by the fact that sparsity constraints on loading coefficients are enforced, such that small coefficients are forced to zero. This effect reduces noise introduced by a large number of variables in the dataset with relatively few batches used here. It has been shown that the consistency of PCA, or in simpler terms, the sensitivity to outliers and noise, is improved by including sparsity constraints in a high dimensional and low sample size context (Zou and Xue, 2018).

* + 1. Leak Detection

The results above show that in FD 1 in year 1 many leak tests failed due to real and virtual leaks, while FD 2 operated smoothly. The leak rate threshold for the leak test to pass was increased from year 1 to year 2 to the upper limit recommended by Sahni et al. (2022). This led to no false alarms in year 2 for FD 1.

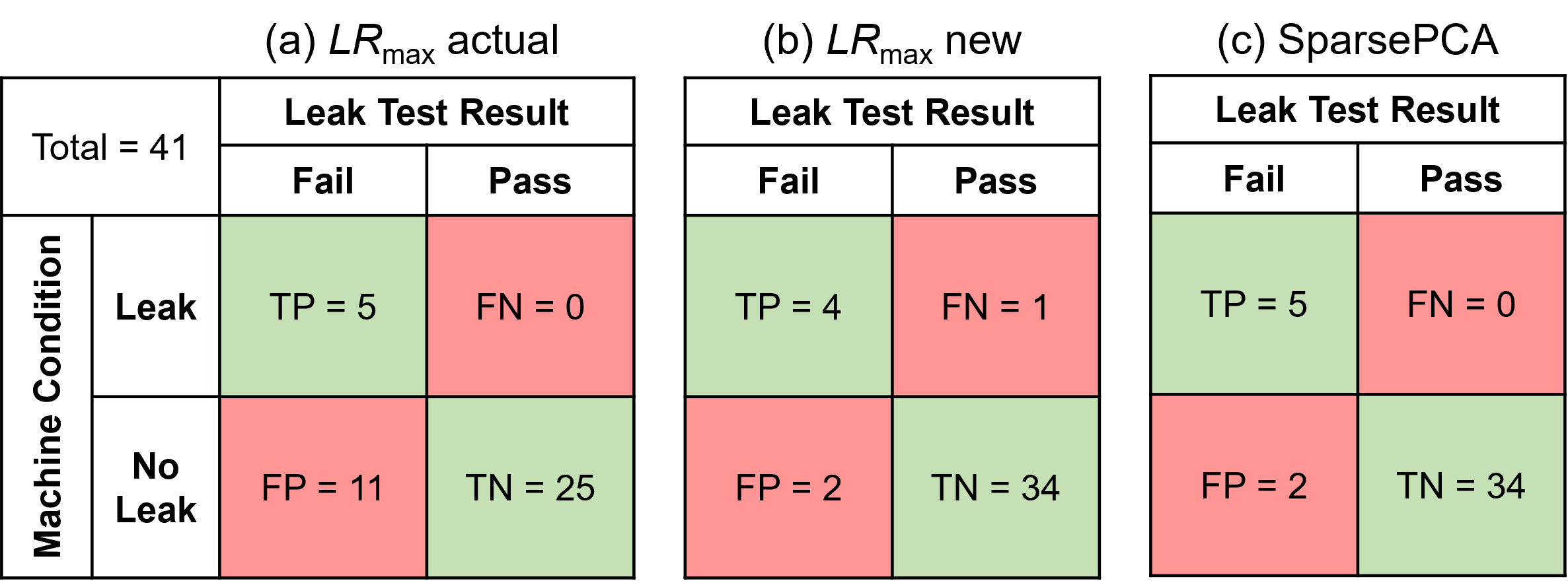


Figure 4: Confusion matrices are shown for leak tests of FD 1 during year 1 for (a) the actual maximum leak rate *LR*max, (b) for the upper recommended limit and (c) based on SparsePCA.

However, if the recommended limit were retrospectively applied to leak tests in FD1 during year 1, the false alarms (FP) would have decreased, but one undetected leak would have occurred (FN) (Figure 4(b) compared to 4(a)). Hence, an apparent trade-off between reducing downtimes caused by false alarms and leak detection reliability exists if leak testing is solely based on leak rate thresholds. Here, multivariate statistical process control could offer an additional decision metric for improved pass/fail decisions. The multivariate methods used above were able to distinguish between real and virtual faults. The confusion matrix in Figure 4(c) shows that by applying those results retrospectively, all tests during a real leak are detected (TP) without increasing the false alarm rate (FP). Of course, production operators need to be informed about potential leaks immediately, which requires an online monitoring model with appropriate sensitivity. Moreso, the model should consider other sources of process variability, such as machine-to-machine and year-to-year. Further reducing the sample size of an already small sample-sized high dimensional dataset might cause limitations, because the dataset could be too small to capture normal operating conditions. The effective integration of data from multiple operating periods and parallel units without compounding system noise will be investigated in future work.

* 1. Conclusion

In conclusion, this research underscores the imperative need for customized approaches when dealing with industrial and noisy datasets. The comparative analysis between SparsePCA and traditional PCA has demonstrated the distinct advantages of SparsePCA, particularly in scenarios characterized by low sample size and high dimensionality. This finding underscores the importance of adapting analytical techniques to the specific challenges posed by datasets common in industrial settings. Finally, the incorporation of multivariate statistics into leak testing has been identified to enhance reliability while managing false alarm rates effectively.

References

E. Sahni, B. Van Meervenne, S. Schneid, M. Dekner, S. Bedi, X. Tang, D. A. Hamilton, O. McGarvey, M. Frei, N. Zinfollino, E. V. Velez, M. Gosmer, D. Hill, 2022, Lyophilizer Leak Rate Testing – An Industry Survey and Best Practice Recommendation, Journal of Pharmaceutical Sciences, 111, 2714-2718.

G. Calzavara, L. Consolini, G. Ferrari, 2021, Leak Detection and Classiﬁcation in Pharmaceutical Freeze-Dryers: an Identiﬁcation-Based Approach, 2021 60th IEEE Conference on Decision and Control (CDC), 1568-1573.

S. Yin, S. X. Ding, X. Xie, H. Luo, 2014, A Review on Basic Data-Driven Approaches for Industrial Process Monitoring, IEEE Transaction on Industrial Electronics, 61, 6418-6428.

C. Ji, W. Sun, 2022, A Review on Data-Driven Process Monitoring Methods: Characterization and Mining of Industrial Data, Processes, 10, 335.

P. Zürcher, S. Badr, S. Knüppel, H. Sugiyama, 2022, Data-driven equipment condition monitoring and reliability assessment for sterile product manufacturing: Method and application for an operating facility, Chemical Engineering Research and Design, 188, 301-314.

F. Pedregosa et al., 2011, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825-2830.

P. J. Rousseeuw, 1987, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53-65.

L. Luo, S. Bao, J. Mao, D. Tang, 2017, Fault Detection and Diagnosis Based on Sparse PCA and Two-Level Contribution Plots, Industrial and Engineering Chemistry Research, 56, 225-240.

K. Sjöstrand, E. Rostrup, C. Rydberg, R. Larsen, C. Studholme, H. Baezner, J. Ferro, F. Fazekas, L. Pantoni, D. Inzitari, G. Waldemar, 2007, Sparse decomposition and modeling of anatomical shape variation, IEEE Transactions on Medical Imaging, 26, 1625-1635.

H. Zou, L. Xue, 2018, A Selective Overview of Sparse Principal Component Analysis, Proceedings of the IEEE, 106, 1311-1320.