Efficient Monitoring of Highly Dimensional Data Streams

S. Bersimis[1], ***K. Skarlatos***[1], ***P. Economou***[2]

[1]Department of Business Administration, University of Piraeus, Greece
[2]Department of Civil Engineering, University of Patras, Greece

The creation of big data is driven by the increasing digitization of information and the proliferation of devices that collect data. As technology advances, the volume, variety, and velocity of data generation continue to grow, leading to the emergence of big data analytics, which aims to extract valuable insights from these extensive datasets. The increasing volume of data presents several opportunities and challenges. In the context of Statistical Process Monitoring (SPM), analyzing high-dimensional data can lead to the ”curse of dimensionality,” where the data becomes sparse, making it difficult to detect patterns and anomalies. Moreover, another challenge is related to modeling and monitoring complex relationships between multiple variables. Furthermore, implementing Multivariate SPM (MSPM) in real-time settings is difficult due to the need for rapid computation and decision-making, especially when dealing with large volumes of streaming data. A possible solution is to combine MSPM approaches with Machine Learning (ML) methods, however, more challenges arise related to model interpretability, feature selection, and integration of results. In this paper, we propose a robust method which is based on a strategy coming from cluster analysis context. The proposed method is compared to established multivariate control charts based on Hotelling T-Square statistic as well as to other statistics coming from cluster analysis framework, i.e. Dunn’s Index. An extensive simulation study showed that the proposed method outperforms in terms of performance its competitors under different scenarios, when data streams consists of a large number of correlated features.

Keywords: Process monitoring, Separability index, Data streams, Machine learning, Cluster Analysis