Control Engineering Practice 127 (2022) 105304

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

An interpretable unsupervised Bayesian network model for fault detection and diagnosis

Wei-Ting Yang [a,*], Marco S. Reis [b], Valeria Borodin [c], Michel Juge [d], Agnès Roussy [c]

[a] Department of Data Science and Analytics, BI Norwegian Business School, 0484 Oslo, Norway
[b] Univ Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II - Pinhal de Marrocos, 3030-790 Coimbra, Portugal
[c] Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-42023, Saint-Etienne, France
[d] Department of Data Science, STMicroelectronics 190 Avenue Célestin Coq, 13106 Rousset, France

ARTICLE INFO

Keywords:
Bayesian Network (BN)
eXplanatory Artificial Intelligence (XAI)
Fault detection
Fault diagnosis
Interpretable machine learning

ABSTRACT

Process monitoring is a critical activity in manufacturing industries. A wide variety of data-driven approaches have been developed and employed for fault detection and fault diagnosis. Analyzing the existing process monitoring schemes, prediction accuracy of the process status is usually the primary focus while the explanation (diagnosis) of a detected fault is relegated to a secondary role. In this paper, an interpretable unsupervised machine learning model based on Bayesian Networks (BN) is proposed to be the fundamental model supporting the process monitoring scheme. The proposed methodology is aligned with the recent efforts of eXplanatory Artificial Intelligence (XAI) for knowledge induction and decision making, now brought to the scope of advanced process monitoring. A BN is capable of combining data-driven induction with existing domain knowledge about the process and to display the underlying causal interactions of a process system in an easily interpretable graphical form. The proposed fault detection scheme consists of two levels of monitoring. In the first level, a global index is computed and monitored to detect any deviation from normal operation conditions. In the second level, two local indices are proposed to examine the fine structure of the fault, once it is signaled at the first level. These local indices support the diagnosis of the fault, and are based on the individual unconditional and conditional distributions of the monitored variables. A new labeling procedure is also proposed to narrow down the search and identify the fault type. Unlike many existing diagnosis methods that require access to faulty data (supervised diagnosis methods), the proposed diagnosis methodology belongs to the class that only requires data under normal conditions (unsupervised diagnosis methods). The effectiveness of the proposed monitoring scheme is demonstrated and validated through simulated datasets and an industrial dataset from semiconductor manufacturing.

1. Introduction

In today's highly complex manufacturing processes, a variety of process monitoring systems are employed that leverage the large amounts of available process data, to rapidly detect and correct deviations from Normal Operating Conditions (NOC). The standard implementation of a process monitoring scheme consists of two phases. The first phase aims at assessing the stability of the process (Phase I—stability assessment), using a reference dataset sufficiently representative from the process under NOC. If the process is declared to be stable at the end of Phase I, then Phase II takes place (Phase II—implementation of the monitoring scheme), where the focus is now to rapidly detect any abnormality or fault in the system (stage 1) and then proceed with its diagnosis (stage 2) in order to identify and isolate the root case. The first stage is also known as *fault detection*, while the second stage is often referred as *fault diagnosis*.

The objective of this paper is to explore the potential of Bayesian Networks to process monitoring, extending their current detection capabilities, complemented with more in-depth and systematic diagnosis and analysis tools. The proposed approach considers a causal structure using both process data and domain knowledge provided by Subject-matter-experts (SMEs). This hybrid causal model is then used for both fault detection and fault diagnosis. The diagnosis outcome can be displayed in a graphical form, which clearly depicts the fault location, greatly facilitating interpretation. Apart from the interpretability, the proposed process monitoring approach proves to be competitive against conventional multivariate methods operating under the same assumptions, while offering several advantages in terms of interpretation and

[*] Corresponding author.
E-mail addresses: wei-ting.yang@bi.no (W.-T. Yang), marco@eq.uc.pt (M.S. Reis), valeria.borodin@emse.fr (V. Borodin), michel.juge@st.com (M. Juge), roussy@emse.fr (A. Roussy).

https://doi.org/10.1016/j.conengprac.2022.105304
Received 1 December 2021; Received in revised form 3 June 2022; Accepted 2 August 2022
Available online 19 August 2022

causal diagnosis. The proposed approach can work with either data-rich or data-poor situations. Many of the more complex methods highly rely on large amounts of training data to obtain a high accuracy, such as deep learning approaches. However, such data volume may not always be available in practice. This is usually the case of process data. For instance, the collected data of a specific product can be sparse in a high-mix low-volume production line, or the data from a new launched product can be of limited size. Furthermore, most diagnosis approaches based on BNs require past faulty observations to extract the faulty patterns (supervised diagnosis techniques), heavily depending on the existence of such information which is not always available. By contrast, the proposed approach only requires NOC data to establish the structure, define the control limits and conduct fault diagnosis (i.e., is an unsupervised diagnosis technique).

This article is organized as follows. Section 2 gives an overview of the existing literature on fault detection and monitoring approaches. Theoretical background on Bayesian networks is briefly provided in Section 3. In Section 4, the details of the proposed BN-based monitoring scheme are introduced and explained. In Section 5, two case studies (one simulated and another industrial from a semiconductor fabrication facility) are then investigated to validate the performance of the proposed approach, both in terms of accuracy and interpretability. Finally, the main contributions of this article are summarized in Section 6, together with perspectives of future work.

2. Related background

2.1. A brief overview of fault detection and diagnosis

A large number of univariate and multivariate Statistical Process Control (SPC) approaches have been developed to detect process/operation/sensor upsets (Nomikos & MacGregor, 1995; Qin, 2014; Rato, Delgado, Martins, & Reis, 2020; Rato, Reis, Schmitt, Hubert, & De Ketelaere, 2016; Reis, Rendall, Rato, Martins, & Delgado, 2021; Reis & Saraiva, 2006). For high-dimensional systems, Statistic Process Control based on Principal Component Analysis (PCA), PCA-SPC, is the most well-known and widely applied methodology (MacGregor & Kourti, 1995; Rato et al., 2016; Reis, 2019). PCA-SPC is a state-of-the-art monitoring technique for large-scale linear and static systems, that estimates and monitors the manifold structure of data in an efficient way, by decomposing the variability in two parts: variability in the PCA subspace and the residual variability. The two monitoring statistics of PCA-SPC are the Squared Prediction Error (SPE) and the Hotelling's T^2 statistics of the PCA scores (Hotelling, 1947; Jackson & Mudholkar, 1979). The SPE monitors the residual variability, while the Hotelling's T^2 monitors the in-plane variability (or structured variability). Together with their associated control limits, these statistics are directly involved in the detection stage to monitor whether any observations are out of control (ooc). More reviews and discussions regarding PCA-based monitoring methods can be found in the literature (MacGregor & Kourti, 1995; Qin, 2012; Reis, Gins, & Rato, 2019). Several methods incorporating other machine learning algorithms have also been proposed in the literature, such as K-Nearest Neighbor (He & Wang, 2007), one-class support vector machines (Mahadevan & Shah, 2009), and artificial neural networks (Samanta, Al-Balushi, & Al-Araimi, 2003). In recent years, deep learning approaches have also been used for fault detection (Lv, Wen, Bao, & Liu, 2016; Sun, Paiva, Xu, Sundaram, & Braatz, 2020; Zhang, Jiang, Li, & Yang, 2018). If the past faulty observations are available, deep neural networks can be used to extract fault patterns including complex nonlinear dynamics (Siegel, Pratt, Sun, & Sarma, 2018). Convolutional neural networks were also proposed to further consider both spatial and temporal effects (Ge et al., 2021; Wu & Zhao, 2018).

In general, fault detection methods aim to compress the multivariate data into a small number of features or monitoring indices that still capture the overall variability patterns. Once a fault is detected through these indices, the next step is to identify the root cause so that the necessary corrective actions can take place and the process malfunction is repaired. One of the popular diagnostic approaches for a PCA-based monitoring method is the contribution plot, which provides potential connections between the out-of-control signal on a control chart and the original variables that may have caused it (Alcala & Qin, 2011; Miller, Swanson, & Heckler, 1998). However, it is well-known that this method suffers from the *smearing-out* problem (Van den Kerkhof, Vanlaer, Gins, & Van Impe, 2013; Reis et al., 2019). This means that, if a variable is faulty, the contributions will also involve other variables highly correlated with the faulty one, even though they are perfectly fine. To overcome the smearing-out problem, several methodologies were proposed. If a sufficiently rich historical dataset with fault labels are available, a more conclusive diagnosis can be sometimes obtained. For instance, Raich and Çinar (1997) proposed a PCA-based discriminant framework, where a similarity index is used to associate abnormal observations to faulty clusters. Many supervised learning approaches have been investigated as well, such as Fisher Discriminant Analysis (Chiang, Jiang, Zhu, Huang, & Braatz, 2015), support vector machine (Widodo & Yang, 2007), neural networks (Venkatasubramanian & Chan, 1989). Alternatively, structured approaches have also been proposed (Bauer & Thornhill, 2008; Rato & Reis, 2015a, 2015b, 2017; Reis et al., 2019). Other diagnostic models based purely on prior knowledge have been also extensively studied but will not be discussed in this paper, as they follow a different rational. A more comprehensive review can be found in the literature (see e.g., Li, Li, Gu, & Chen, 2020; Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003).

Analyzing closely the existing literature, it is possible to verify that a wide range of data-driven approaches has been developed and applied in many different fields, including for process monitoring, that are strongly oriented towards maximal accuracy, but neglect the process structure of existing knowledge of the underlying casual connectivity. However, many applications require both accuracy and explanatory capabilities, to ensure a correct interpretation of the situation and to take the right corrective actions. The scope of this problem goes beyond industrial processes, affecting also the medical domain (Holzinger, Biemann, Pattichis, & Kell, 2017), judgments involving human rights (Rudin, 2019), and decisions in safety-critical tasks (Varshney & Alemzadeh, 2017). Therefore, the increasing demand for interpretability has driven more attention to what is now known as, *eXplainable Artificial Intelligence* (XAI) (Arrieta et al., 2020; Gilpin et al., 2018; Guidotti et al., 2018). More discussion regarding the definitions of interpretability or explainability can be found in the literature (Arrieta et al., 2020; Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019). In brief terms, an explainable AI system does not aim to only learn the patterns from data, but also to provide human-understandable results. A variety of XAI approaches have been discussed in recent works (Gilpin et al., 2018; Guidotti et al., 2018). The categories of various explainable ML approaches are, according to the systematization proposed by Guidotti et al. (2018), referred as: black-box explanation and transparent-box design. The first category provides post-hoc explanations for a black-box ML approach by incorporating other interpretable models, such as local classification models to check if a feature (or an object) exists in an image (Ribeiro, Singh, & Guestrin, 2016). The second category adopts inherent transparent models, such as regression models, decision trees, or Bayesian networks, etc. Arrieta et al. (2020), to provide an explanation to the pattern observed.

2.2. Interpretable approaches for fault detection and diagnosis

Enhancing the interpretability of data-driven manufacturing solutions certainly helps the progress of Industry 4.0, with more intelligent and informative systems. Following the taxonomy of Guidotti et al. (2018), let us now examine and analyze more closely the existing process monitoring approaches regarding their interpretability. Classic multivariate approaches, such as PCA, and PLS, have been widely implemented in process monitoring and proved their effectiveness in fault detection. Nevertheless, relationships found in data are non-causal, and the results are often lacking interpretation. These approaches usually require incorporating other techniques, such as contribution plots or Signed Directed Graph, to give more diagnosis insights (Qin, Valle, & Piovoso, 2001; Vedam & Venkatasubramanian, 1999). This analysis flow from detection to diagnosis is similar to the post-hoc explanation approach. In recent years, instead of detection-oriented approaches, several methodologies incorporating causal connections have been developed to address their diagnosis capability (Reis & Gins, 2017). Chiang et al. (2015) proposed two indicators, modified distance and modified causal dependency, incorporating a causal map to identify unknown, known, and multiple faults. Yang, Shah, and Xiao (2012) employed a Signed Directed Graph and determined the propagation path based on the signs of arcs and the signs of nodes. Bayesian Network (BN) is also one of the popular methods among these causal approaches (Cai, Huang, & Xie, 2017). Yang and Lee (2012) considered a Bayesian network based on several discretized sensor variables, and these variables consist of different states: normal, warning, or error. By entering quality data in the evidence node, the faults can be isolated by analyzing the posterior probabilities of other nodes. Mori, Mahalec, and Yu (2014) developed a process monitoring scheme based on Bayesian networks, where the final structure is determined by several sub-networks and may include close loops. A likelihood index was employed for detection, and the propagation path is determined by the approximate integration of conditional probabilities. Conditional Gaussian networks have also been employed to improve the efficiency of fault diagnosis (Lou, Li, Atoui, & Jiang, 2020; Verron, Tiplica, & Kobi, 2010). Lou et al. (2020) used a discrete variable to represent the status of a process, including: normal operation, known fault types occurred in the past, and unknown faults. The posterior probability of the status is computed in real-time, for each new incoming observation. By comparing the value with the statistical limit, a fault type can thus be proposed.

Analyzing the available literature, there is a lack of approaches with the flexibility to integrate both existing knowledge and data induced knowledge, in a unique and coherent framework. Bayesian Networks offer this possibility, together with a clear probabilistic interpretation of its outcomes. These features are explored in this work, but first we provide a short introduction to its structure, estimation algorithms and properties.

3. Bayesian networks

In this paper, Bayesian Networks (BN) are employed as a foundation of the proposed process monitoring method. This section aims to provide the theoretical background on BN, including a general terminology, a description of the learning procedure, and some relevant properties.

3.1. Basics

A Bayesian network is a probabilistic model expressing the conditional dependencies of a set of variables through a Directed Acyclic Graph (DAG) (Pearl, 2014). Let $\mathbf{X} = [X_1, X_2, \dots, X_m]$ be a data matrix with n samples and m variables. The graph, denoted by $G = (V, E)$, is composed of a set $V = \{V_1, V_2, \dots, V_m\}$ of nodes and a set E of arcs. Each node $V_k \in V$ represents the random variables X_k in $\mathbf{X}$, $k \in \{1, 2, \dots, m\}$. An arc $e \in E$ describes the cause–effect relationships existing between variables as an asymmetric dependency. The mapping notations used in this paper are summarized in Appendix A. The absence of an arc implies the existence of conditional independence between the corresponding variables. An example is shown in Fig. 1, where X_1 is a parent node of X_3 and X_3 is a child node of X_1. This dependency can be described as X_1 causes X_3 while X_3 cannot cause X_1.

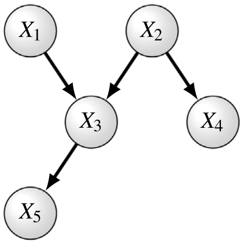


Fig. 1. An illustration of a simple Bayesian network.

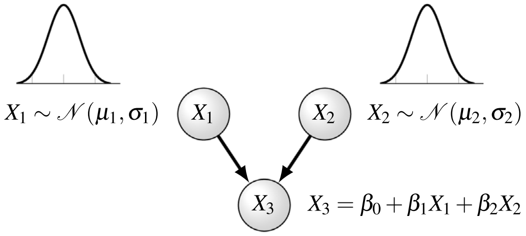

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$
X_1
X_2
$X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$
X_3
$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Fig. 2. An example of a Gaussian Bayesian network and the local distribution of each node.

A Bayesian network satisfies the Markov condition: each node is conditionally independent of its non-descendants, given its parents. Due to the Markov condition, the joint probability can be expressed in a product form: $p(X_1, X_2, \dots, X_m) = \prod_{k=1}^{m} p(X_k | \mathbb{X}_{pa(k)})$, where $\mathbb{X}_{pa(k)}$ is the set of parent nodes of X_k and $p(X_k | \mathbb{X}_{pa(k)})$ is the conditional probability of X_k given $\mathbb{X}_{pa(k)}$.

In this paper, we assume that all variables follow Gaussian distributions, and relationships among variables are linear. The linear hypothesis can be expanded after its performance is established and the benefits confirmed against benchmark methods operating under the same frame of assumptions. Still, linearity is expected to work well under small/moderate deviations from the nominal operation, where non-linearity, if present, does not dominate the shape of the response surface. Any node can be thus expressed via a multiple linear regression model involving its causal parents (see Fig. 2), which greatly simplifies the associated computations.

3.2. Structure learning

Learning the structure of Bayesian networks can be a complex and computationally intensive because the cardinality of the set of possible networks is usually enormous. There are two main categories of approaches for learning the graphical structure from data: constraint-based and score-based (Scutari, 2009). Constraint-based algorithms identify the conditional independencies of all variables through statistical tests that determine if an arc exists or not. The procedure starts with a fully connected undirected graph, and then determines the conditional independencies of each pair of variables given a subset of other variables. Many algorithms have been proposed, such as inductive causation algorithm (Verma & Pearl, 1992), PC algorithm which is named after its inventors (Spirtes, Glymour, Scheines, & Heckerman, 2000), and glow-shrink algorithm (Margaritis, 2003). The outcomes of constraint-based algorithms are affected by the testing order, and some algorithms can be inefficient when dealing with a large number of variables (Margaritis, 2003).

Score-based algorithms firstly score each possible graphical structure based on how well it describes the observed data, and the structure with the highest score is selected. Many scoring methods are available for structure learning (Campos, 2006). In this paper, the Akaike

Information Criterion (AIC) is chosen as the score function:

$$\texttt{score}_{AIC}(G, \mathbf{X}) = log(\hat{L}) - d_G \tag{1}$$

where $\hat{L} = p(\mathbf{X}|G, \hat{\theta}_G)$ is the maximum value of the likelihood function, $\hat{\theta}_G$ is the maximum likelihood estimate and d_G is the model complexity.

The objective of the score-based algorithms is to find an optimal structure that maximizes the score. In the case of Gaussian Bayesian Networks, the model complexity is the number of estimated coefficients. However, finding an optimal structure is known to be NP-hard (Chickering, 2002). The standard approach to solve this problem is to perform a heuristic search. Many heuristic search algorithms have been proposed for leaning the BN structure (Chickering, 2002; Elidan, Ninio, Friedman, & Schuurmans, 2002), but some are complicated and hard to implement. The simplest search algorithm—Hill Climbing (HC), can be a practical choice in terms of the trade-off between effectiveness and efficiency (Teyssier & Koller, 2012).

In addition to identifying the arcs from data, the algorithm also provides the flexibility to integrate pre-defined directions on specific arcs. Based on the domain knowledge, the arcs that present known causalities can be defined as a whitelist, and the arcs that present infeasible causalities will be defined as a blacklist (Scutari, 2009; Yang, Blue, Roussy, Pinaton, & Reis, 2020). During the structure learning procedure, any movement against either whitelist or blacklist will be viewed as a violation.

3.3. Properties of DAGs

Reachability in graph theory refers to the ability to go from one node to another node through a path. Assuming a pair of nodes (V_i, V_j), V_i can reach V_j if there exists a path that starts with V_i and ends with V_j, denoted by $V_i \prec V_j$. Let $\mathscr{R} = \{V_i \prec V_j\}$ be the set of relations on V, where $(V_i, V_j) \in V \times V$. $\mathscr{R}$ indicates the reachability relation of G. Topological sorting of graph G is to find a permutation T_G of V according to the precedence relation $\mathscr{R}$. For each pair of nodes $(V_i, V_j) \in V \times V$, if $V_i \prec V_j \in \mathscr{R}$, then V_i must precede V_j in T_G (Knuth & Addison-Wesley, 1997), i.e. V_i is the ancestor of V_j. Any DAG has at least one permutation T_G by topological sorting. The topological sorting problem is essentially equivalent to arranging the nodes of a directed graph into a straight line, so that for any node $V_k \in V$ the ancestors of V_k must be in front of V_k in T_G.

In this work, the proposed diagnosis procedure follows the order in a topologically sorted permutation. Hence, the information of the ancestor can be taken into account when abnormal nodes are analyzed.

4. Process monitoring based on Bayesian networks

In this paper, we propose a novel process monitoring scheme based on Bayesian Networks (BN). Before introducing the details of the approach, an overview is provided in Fig. 3. The proposed approach is based on an interpretable machine learning model dedicated to understanding and interpreting the outcomes of the monitoring methodology. By modeling the process system with a BN, the connections among variables are established and can be visualized in a graphical form (see Fig. 3(a)). As causal relations defined by SME can be included in the structure construction, we can be sure that the dependencies are also fully consistent with the existing domain knowledge.

In what follows, the details of the proposed metrics are introduced in Section 4.1. In Section 4.2, different types of faults are discussed and are presented in a network form. The analysis based on the local metrics is covered as well. The proposed BN-based online monitoring procedure is consolidated in Section 4.3 (see Fig. 3(b) and (c)).

4.1. BN-based monitoring metrics

Let $\mathbf{X} = [X_1, X_2, \ldots, X_m]$ be a data matrix with n samples and m variables. Each variable X_k presents a process variable (or sensor variable) in a process system, $k = \{1, 2, \ldots, m\}$. As explained in Section 3, the relations between the variables in $\mathbf{X}$ can be presented by a Bayesian network G. A process system modeled by G can be monitored through the proposed metrics. Note that the granularities of these metrics can be determined by setting up the desired batch size. Let us define n_b as the size of the batch, and $\mathbf{X}$ can be rewritten as $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_B \end{bmatrix}$, where $\mathbf{X}_b \in \mathbb{R}^{n_b \times m}$, $b = \{1, 2, \ldots, B\}$. The metrics are computed for each $\mathbf{X}_b \in \mathbb{R}^{n_b \times m}$, $b \in \{1, 2, \ldots, B\}$. In the rest of this paper, an *individual mode* refers to a monitoring scheme when $n_b = 1$, and a *batch mode* implies that $n_b > 1$. The local metrics are firstly introduced, as the global metric is simply a summation of the local metrics.

Unconditional local likelihood index L^u_k. As mentioned in Section 3, we assume all the variables follow Gaussian distribution in this study. The unconditional log-likelihood of variable X_k of batch b is defined as:

$$L^u_{k,b}(\hat{\mu_k}, \hat{\sigma_k}; \mathbf{X}_b) = -\frac{n_b}{2} ln(2\pi) + \frac{n_b}{2} ln(\hat{\sigma_k}^2) - \frac{1}{2\hat{\sigma_k}} \sum_{i=1}^{n_b} (x_{ik} - \hat{\sigma_k})^2 \tag{2}$$

where $\hat{\mu_k}$ is the sample mean of X_k and $\hat{\sigma_k}$ is the sample standard deviation of X_k based on a reference dataset. In the rest of this paper, the unconditional local likelihood refers to the unconditional log-likelihood of a variable. A low value of $L^u_{k,b_{new}}$ of a new batch b_{new} indicates that the current process variable is different from the original distribution. Note that the unconditional local likelihood index is independent of the BN structure, as the causalities between variables are not taken into account.

Conditional local likelihood index L^c_k. Suppose the distribution of variable X_k is affected by its parents $\mathbb{X}_{pa(k)}$. To evaluate the local change that excludes the effects of parents, we propose a conditional local likelihood index L^c_k, which examines the conditional distribution of X_k instead of the original distribution. As described in Section 3, any node in G can be expressed as a linear regression model involving its causal parents $X_k = \alpha_k + \mathbf{X}_{pa(k)}\beta_k + \epsilon_k$, where $\mathbf{X}_{pa(k)}$ is the matrix of parent variables and ϵ_k is the error term. The conditional distribution of $X_k|\mathbf{X}_{pa(k)}$ is equivalent to the distribution of ϵ_k. The conditional log-likelihood index of X_k of batch b, i.e. log the likelihood of ϵ_k of batch b, can be written as:

$$\begin{aligned} L^c_{k,b}(\hat{\theta}_{X_k|\mathbf{X}_{pa(k)}}; \mathbf{X}_b) &= L^c_{k,b}(\hat{\mu}_{\epsilon_k}, \hat{\sigma}_{\epsilon_k}; \mathbf{X}_b) \\ &= -\frac{n_b}{2} ln(2\pi) + \frac{n_b}{2} ln(\hat{\sigma}^2_{\epsilon_k}) - \frac{1}{2\hat{\sigma}_{\epsilon_k}} \sum_{i=1}^{n_b} (\epsilon_{ik} - \hat{\mu}_{\epsilon_k})^2 \end{aligned} \tag{3}$$

where ϵ_{ik} is the residual of sample i in batch b, $\hat{\mu}_{\epsilon_k}$ and $\hat{\sigma}_{\epsilon_k}$ are the estimated statistics of ϵ_k of the reference dataset. In this paper, we use the unconditional local likelihood to represent the unconditional log-likelihood of a variable. Since L^c_k excludes the effects caused by causal parents, a low likelihood implies that the underlying model has changed, either α_k or β_k, and such change leads to large ϵ_{ik}.

Global likelihood index L^g. The global index measures the overall stability of a process by looking at the joint distribution of a network G. As the joint distribution of G can be decomposed into the local distribution of individual variables, the likelihood function can be expressed as the product of local likelihood. The global likelihood index of batch b is defined as the sum of the log local likelihood function:

$$L^g_b(\theta_G; \mathbf{X}_b) = \sum_{k=1}^{m} L^c_{k,b}\left(\hat{\theta}_{X_k|\mathbf{X}_{pa(k)}}; \mathbf{X}_b\right) \tag{4}$$

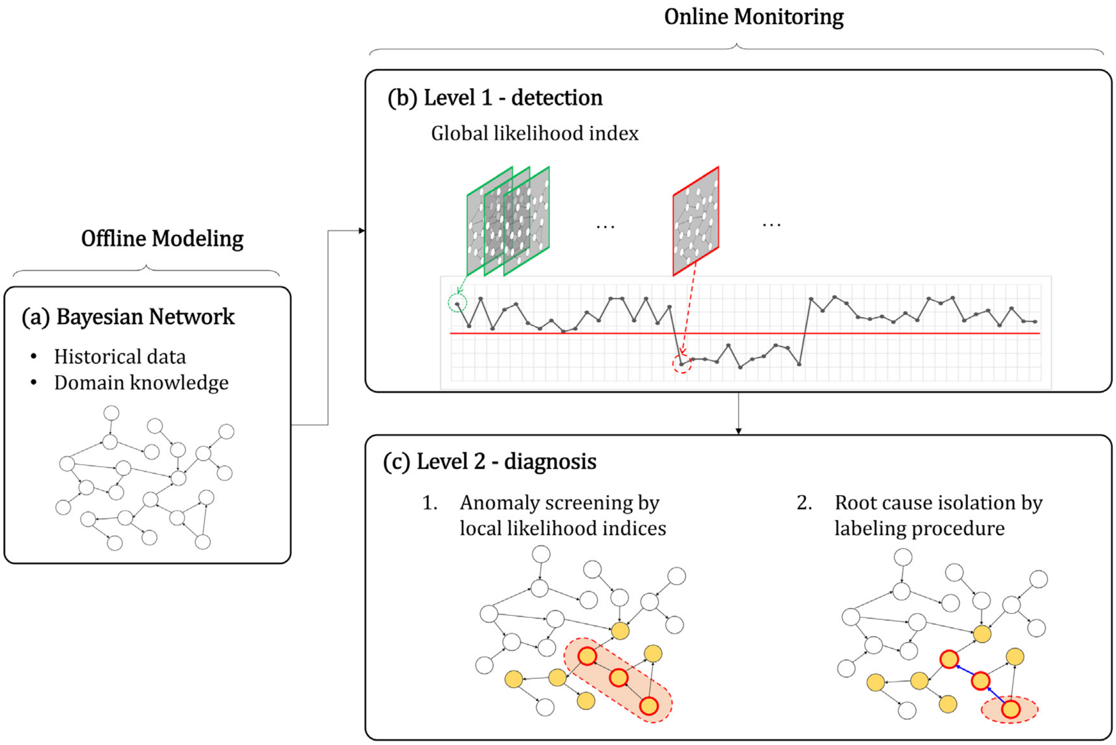


Fig. 3. Process monitoring method based on Bayesian networks.

The global likelihood index L_b^g can be used to check if the new batch b is similar to the reference data and if the underlying model can be well presented by G. By doing so, the task of monitoring multiple variables in a process system can be simplified to monitoring a single index.

Control limits for statistics. Statistical control charts are derived to monitor the statistics: L_k^u, L_k^c, and L^g. The control limits of these three metrics are obtained following the same procedure. The procedure consists of applying κ-fold cross-validation to avoid over-optimistic limits, that eventually lead to inflated false alarm rates. In each iteration, training data from $\kappa - 1$ folds are used for learning parameters, and the likelihood metric L is computed in the κth left-out fold, where $L = \{L_k^u, L_k^c, L^g\}$. Assume there are B_k batches in the κth left-out fold. After all iterations, $\kappa \times B_k$ metrics are collected $\ell = [L_1, L_2, \dots, L_{\kappa\times B_k}]$, and set ℓ is used to determine the control limits.

To establish the appropriate control limits for anomaly detection, kernel density estimation (Silverman, 1986), a non-parametric method, is employed to estimate the probability density function of each likelihood index, L_k^u, L_k^c, and L^g. The kernel density estimation of L is defined as:

$$\hat{f}_{KDE} = \frac{1}{nh} \sum_{j=1}^{\kappa\times B_k} \mathscr{K}\left(\frac{L - \ell_j}{h}\right) \tag{5}$$

where $\mathscr{K}$ represents the kernel function and h is bandwidth. The control limits can be determined by $\int_{-\infty}^{LCL} \hat{f}_{KDE}(\ell)dx = \alpha$, where α is the pre-defined Type I error. As high likelihood implies that the new batch is close to the reference distribution, only Lower Control Limit (LCL) is needed for fault detection. The control limits of the metrics, L_k^u, L_k^c, and L^g, are denoted by h_k^u, h_k^c, and h^g, respectively.

4.2. Types of faults: Network representation

In general, the detected faults can be further categorized into two main groups: **(i)** *process faults:* an anomaly induced by a change in the process, or **(ii)** *sensor faults:* a bias due to incorrect readings from a faulty sensor. In this section, the emerging patterns in a network for these faults are discussed. To simplify the illustration, a simple linear system presented by a BN is used to discuss the different types of faults (see Fig. 4). The unconditional local distribution of variable X_k is denoted by $f_{X_k}^u$ and conditional local distribution of variable X_k is denoted by $f_{X_k}^c$. The parameters have been estimated from a historical dataset.

Ideally, a manufacturing process should remain stable, under a state of statistical process control. However, various factors during manufacturing may affect the stability of a process and lead to abnormal changes. These causes can be aging components, disparate material suppliers, or inconsistent operations. Such changes in the process can be captured by leveraging extensive sensor readings, where the readings may perform either a slow drift, a rapid shift, or a growing dispersion. In this study, we focus on three types of faults, namely a *correlation change* in the process (emulating a process fault, as variables are expected to lose their normal operating conditions associations in these circumstances), a *step change* due to an operation perturbation, and a *sensor bias* due to a malfunctioning sensor.

Process fault: Correlation change. Assume that the correlation between nodes X_1 and X_2 has changed from β_2 to β_2' over a period of time. Suppose the distribution of X_1 remains the same, but a change in β_2 alters the center and spread of the distribution of X_2, i.e. $f_{X_2}' \neq f_{X_2}$, as illustrated in Fig. 5. In this context, for a new batch b, the unconditional local likelihood index $L_{2,b}^u$ of X_2 will be lower than the level of its normal condition, i.e. $L_{2,b}^u < h_2^u$ if the magnitude is significant. Since the conditional local likelihood index $L_{2,b}^c$ is computed based on the residuals ϵ_2 according to (3), given an outdated parameter β_2, one can expect a low $L_{2,b}^c$ as well. In this paper, we use different colors to visualize an abnormal (faulty) node. A node filled-in yellow indicates that its unconditional likelihood is below the threshold, i.e., $L_{k,b}^u < h_k^u$. A node highlighted with a red rim indicates that its conditional likelihood is below the threshold, i.e., $L_{k,b}^c < h_k^c$ (see Fig. 5). The control limits h_k^u and h_k^c are obtained through formula (5). In other words, a yellow node without a red rim means that the given variable has a significantly different value than at the original level, but by removing the influence of its parent, this variable has no other additional anomaly. On the

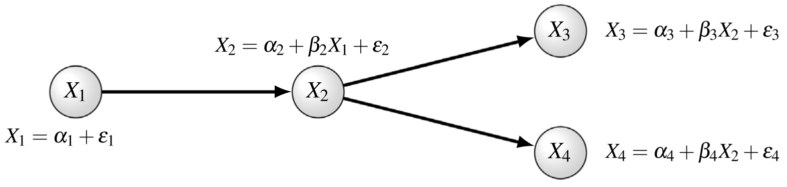

$X_2 = \alpha_2 + \beta_2 X_1 + \varepsilon_2$
X_3
$X_3 = \alpha_3 + \beta_3 X_2 + \varepsilon_3$
X_1
X_2
$X_1 = \alpha_1 + \varepsilon_1$
X_4
$X_4 = \alpha_4 + \beta_4 X_2 + \varepsilon_4$

Fig. 4. An example of a linear process system represented by a Bayesian network.

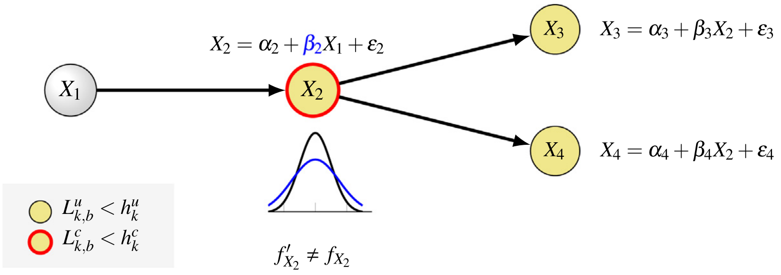

$X_2 = \alpha_2 + \beta_2 X_1 + \varepsilon_2$
X_3
$X_3 = \alpha_3 + \beta_3 X_2 + \varepsilon_3$
X_1
X_2
X_4
$X_4 = \alpha_4 + \beta_4 X_2 + \varepsilon_4$
○ $L^u_{k,b} < h^u_k$
○ $L^c_{k,b} < h^c_k$
$f'_{X_2} \neq f_{X_2}$

Fig. 5. Likelihood indices of batch b given a perturbation in the relationship between nodes X_1 and X_2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

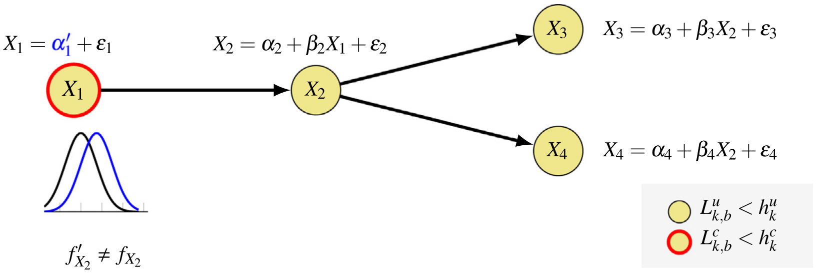

$X_1 = \alpha_1' + \varepsilon_1$
$X_2 = \alpha_2 + \beta_2 X_1 + \varepsilon_2$
X_3
$X_3 = \alpha_3 + \beta_3 X_2 + \varepsilon_3$
X_1
X_2
X_4
$X_4 = \alpha_4 + \beta_4 X_2 + \varepsilon_4$
○ $L^u_{k,b} < h^u_k$
○ $L^c_{k,b} < h^c_k$
$f'_{X_2} \neq f_{X_2}$

Fig. 6. Likelihood indices of batch b given a step-change in node X_1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

other hand, a yellow node with a red rim means that this variable is abnormal and such anomalies are not inherited from its parent.

Such change also affects the distributions of descendants of X_2, X_3, and X_4, and leads to low $L^u_{3,b}$ and $L^u_{4,b}$. But the changing magnitudes of $L^u_{3,b}$ and $L^u_{4,b}$ depends on their association with X_2. For instance, a significant change in X_2 may only cause an insignificant decreasing in $L^u_{3,b}$ if β_3 is small. However, the conditional indices, $L^c_{3,b}$ and $L^c_{4,b}$, should remain at their ordinary level if the parameters are the same.

Operation perturbation: Step-change. Suppose that a step change has occurred in X_1, where $\alpha_1' = \alpha_1 + \delta$. Consequently, one can expect that there will be a shift in the distribution of X_1 as shown in Fig. 6. Similar to a correlation change, such process change will propagate to the descendants of X_1. Assume the impacts caused by X_1 are significant for all the descendants. For a new batch b, its unconditional local likelihood indices of the descendants of X_1 are abnormal (see Fig. 6). Nevertheless, since the models of descendants are not changed, their conditional local likelihood indices should exhibit regular, i.e. $L^c_{k,b} > h^c_k$. In this work, we assume that a step-change only occurred in the root nodes, which are responsible for injecting variability and drifting in the system.

Sensor fault. To monitor the stability of the process, we highly rely on the massive sensor reading data. However, a malfunctioning sensor will give a misleading result. A process monitoring mechanism should take into account the case of incorrect sensor readings. In this faulty scenario, we assume the sensor of process variable X_2 is not functioning, and its readings present a bias with the magnitude δ, while the actual underlying process remains at the normal level. In this case, both $L^u_{2,b}$ and $L^c_{2,b}$ of batch b show abnormal as these indices are computed based on the incorrect sensor readings (see Fig. 7). The unconditional local likelihood L^u_b of its descendants should stay normal because the actual distribution of X_2 is the same. However, index L^c_b of descendants of X_2 can be irregular because the effects from parents are computed incorrectly. In other words, an unfilled node with a red rim indicates that the variable is normal, but the sensor reading of its parents may be faulty.

In brief terms, changes in process or disturbances caused by operations, are propagated to the descendants and can be observed by looking at the node marked in yellow (see Figs. 5–6). On the other hand, the impact of a sensor bias will only show up in the root cause node as shown in Fig. 7. The conditional distribution alarm, marked by a red rim, aims to narrow the suspicious region by excluding the abnormal nodes caused by their parents. By analyzing these local likelihood indices, together with a graphic visualization, we are able to spot the faulty region and get more information about the occurred fault. A diagnosis approach based on this information is introduced in the next section.

4.3. Online monitoring

With a learned BN and the proposed metrics, the proposed BN-based process monitoring can be implemented online. The global likelihood index L^g (4), is used for fault detection (Level 1 - detection), and the other two local metrics (2)–(3) are used for targeting the root cause (Level 2 - diagnosis). The flow chart of online detection and diagnosis is shown in Fig. 8. Details of each step are described in the following subsections.

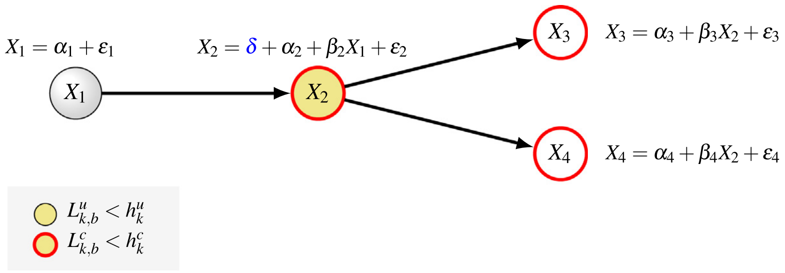

$X_1 = \alpha_1 + \varepsilon_1$
X_1
$X_2 = \delta + \alpha_2 + \beta_2 X_1 + \varepsilon_2$
X_2
X_3
$X_3 = \alpha_3 + \beta_3 X_2 + \varepsilon_3$
X_4
$X_4 = \alpha_4 + \beta_4 X_2 + \varepsilon_4$
$L^u_{k,b} < h^u_k$
$L^c_{k,b} < h^c_k$

Fig. 7. Likelihood indices of batch b given a sensor bias in X_1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.1. Level 1: fault detection

As discussed in Section 4.1, the global likelihood index L^g is used to monitor the overall stability of a process system. The proposed process monitoring procedure starts with a global screening by the following two steps.

1. Computing L^g_b for a new batch b.
2. Comparing L^g_b with threshold h^g according to formula (5).

If $L^g_b < h^g$, batch b is tagged as an abnormal batch, and the next diagnosis procedure is triggered.

4.3.2. Level 2: fault diagnosis

Anomaly screening by local likelihood indices. After a fault is detected in batch b, the local distribution of all variable X_k, $k = \{1, 2, \dots, m\}$, are assessed based on their local metrics, namely: unconditional local likelihood index $L^u_{k,b}$ and conditional local likelihood index $L^c_{k,b}$. By comparing these metrics with the corresponding control limits h^u_k, h^c_k, one obtain a set of abnormal variables $\mathbb{X}_{abn}$. For each variable $X_k \in \mathbb{X}_{abn}$, $L^u_{k,b} < h^u_k$ and $L^c_{k,b} < h^c_k$. The goal of this step is to quickly filter out the set of abnormal variables, as shown in the first graph in Fig. 3(c). In the next paragraph, we describe the labeling process in detail.

Root cause isolation by labeling. In this paper, we aim to bring further diagnosis insights by labeling the abnormal variables so that the result of the analysis can be linked to the repairing action. The labeling procedure involves m' iterations, which correspond to the number of variables in $\mathbb{X}_{abn}$. In each iteration, the abnormal variables in $\mathbb{X}_{abn}$ are analyzed by checking their parents and children. At the end of each iteration, these abnormal variables are labeled with the suspicious faulty type.

Two assumptions are made to simplify the labeling procedure:

- If a pair of connected nodes is marked as abnormal, we assume the fault has occurred in the parent, and the abnormality has been propagated to the children.
- Multiple faults are not considered, i.e. only one type of fault is assumed to occur in a node.

The sequence is following the order of topological sorting T_G (see Section 3). Let T'_G be a topological sorted list, which only contains a set of abnormal variables $\mathbb{X}_{abn}$. The iteration starts with the first variable in T'_G. The labeling rules to be applied in each iteration are shown in the sub-flow chart outlined by the dotted line in Fig. 8.

- X_k **is not a root node:** If node X_k is not a root node, i.e. a node with parents, the approach checks if the anomaly in X_k is a consequence of its abnormal parents. If any variable in $\mathbb{X}_{pa(k)}$ has been diagnosed, the type of fault of X_k will be labeled with the same type as its parents. However, if the states of $\mathbb{X}_{pa(k)}$ are normal, the algorithm checks the metrics of its children $\mathbb{X}_{ch(k)}$, to get more information.

  As referred in Section 4.2, a process change in correlation between X_k and its parent will have an impact on the distribution of variable X_k and possibly on its descendants. Therefore, for any $X_c \in \mathbb{X}_{ch(k)}$, if $L^u_{c,b} < h^u_c$, we conjecture a correlation change has occurred between X_k and its parent.

  If none of the unconditional distribution of $\mathbb{X}_{ch(k)}$ is abnormal, the next rule is to check the conditional distribution of $\mathbb{X}_{ch(k)}$. If the sensor of X_k is malfunctioning, in other words, the reading of X_k is not reliable, the unconditional distribution of $\mathbb{X}_{ch(k)}$ should stay in their normal state, while their conditional distribution may be perturbed because the conditional distribution takes into account the incorrect sensor information. Thus, for any $X_c \in \mathbb{X}_{ch(k)}$, if $L^c_{c,b} < h^c_c$, we conjecture a sensor bias in X_k. For other scenarios, such as when $\mathbb{X}_{ch(k)} = \emptyset$, or both $L^u_{c,b}$ and $L^c_{c,b}$ are within control limits, the label would be: *correlation change/sensor bias*.
- X_k **is a root node:** If X_k is a root node, the metrics of $\mathbb{X}_{ch(k)}$ are used to gauge the faulty type of X_k (see Fig. 8). Assume there is a step change in X_k, and such change will alter the distribution of $\mathbb{X}_{ch(k)}$. Consequently, for any X_c in $\mathbb{X}_{ch(k)}$, if $L^u_{c,b} < h^u_c$, we conjecture a step change has occurred in X_k. Note that the difference between a step change and a correlation change is that a step change only occurred in a root node, while a correlation change is defined as the correlation with parents (see Section 4.2). If none of the unconditional distribution of $\mathbb{X}_{ch(k)}$ is abnormal, the next rule is to check the conditional distribution of $\mathbb{X}_{ch(k)}$. If the sensor of X_k is malfunctioning, the unconditional distribution of $\mathbb{X}_{ch(k)}$ should stay in the normal state, while their conditional distribution may present irregular. Thus, for any X_c in $\mathbb{X}_{ch(k)}$, if $L^c_{c,b} < h^c_c$, we conjecture a *sensor bias* in X_k. For other scenarios, such as $\mathbb{X}_{ch(k)} = \emptyset$, or both $L^u_{c,b}$ and $L^c_{c,b}$ are within control limits, the label suggested is: *step-change/sensor bias*.

After the labeling procedure, a set of variables with possible faulty types is obtained. By removing those labeled by their parents, we can get a smaller set of abnormal variables denoted by $\mathbb{X}_{label}$ (see Fig. 3). Reducing the suspicious area shall improve the quality of diagnosis and speed up the time of recovering from the fault. More supporting results can be referred as well for troubleshooting by labeling the isolated variables with the type of fault. However, we suggest analyzing these labels only when the data is sufficient. For instance, when the batch size is set to be greater than 30, the labeling procedure is conducted based on more evidence and can produce more reliable results.

5. Numerical experiments

The effectiveness of the proposed framework is evaluated with respect to the Principal Component Analysis (PCA) (see Appendix B) on two case studies. Note that the underlying data generating mechanism used in the numerical simulations refers to a linear and static system. Therefore, in these conditions, PCA-based statistical process monitoring is a suitable benchmark method, with many successful applications reported in the literature involving high-dimensional systems, including industrial applications (in fact, PCA-based process monitoring is one of rare cases of high-dimensional monitoring approaches being applied in industry). The first case study is based on a simulated dataset, with the

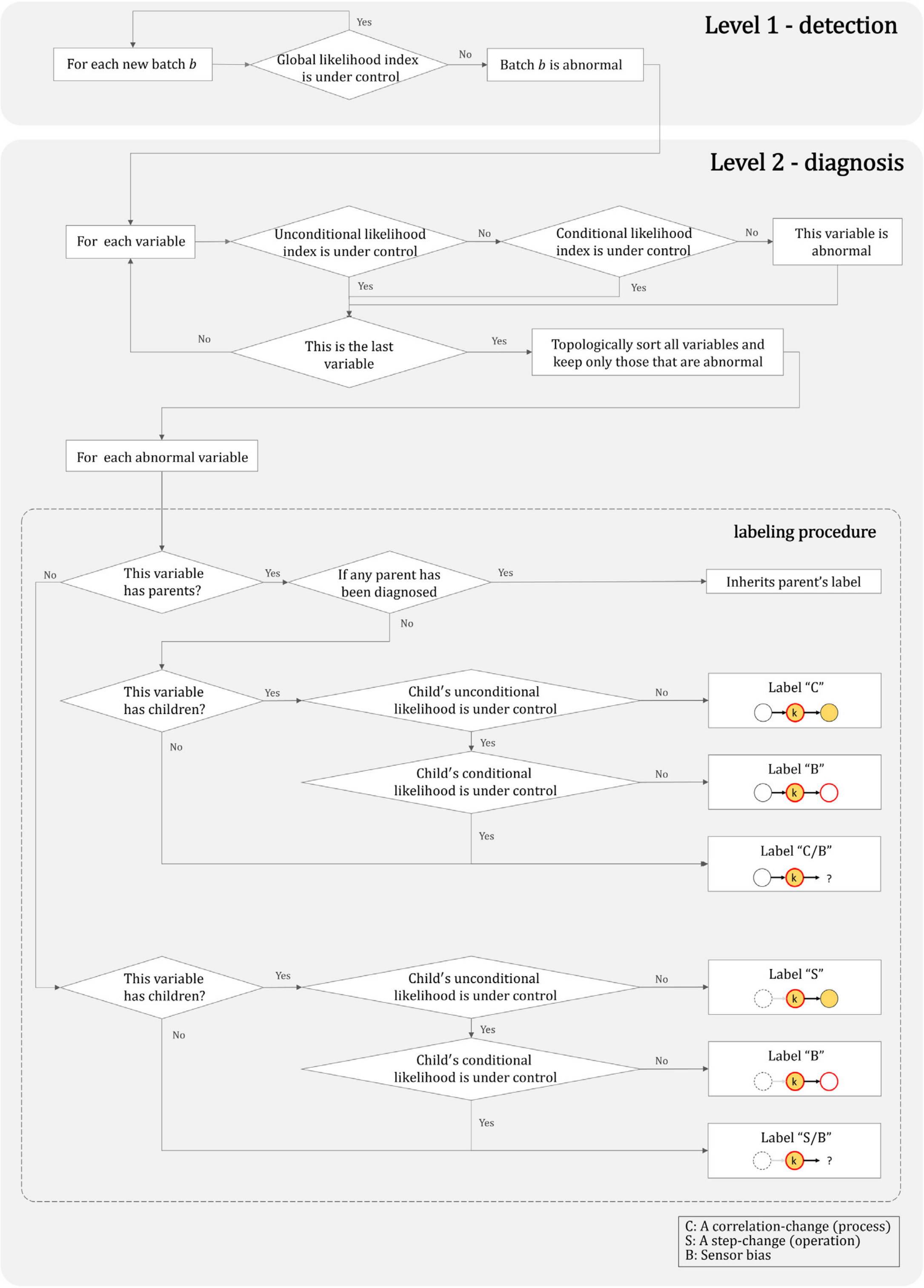

Level 1 - detection
For each new batch b
Global likelihood index is under control
Yes
No
Batch b is abnormal
Level 2 - diagnosis
For each variable
Unconditional likelihood index is under control
Conditional likelihood index is under control
This variable is abnormal
This is the last variable
Topologically sort all variables and keep only those that are abnormal
For each abnormal variable
labeling procedure
This variable has parents?
If any parent has been diagnosed
Inherits parent's label
This variable has children?
Child's unconditional likelihood is under control
Child's conditional likelihood is under control
Label "C"
Label "B"
Label "C/B"
Label "S"
Label "S/B"
k
?
C: A correlation-change (process)
S: A step-change (operation)
B: Sensor bias

Fig. 8. Flow chart of online detection and diagnosis.

goal of consolidating the properties of the proposed framework. With an explicit structure and several pre-defined faults, the performance of the proposed method can be examined without any ambiguity in a simulated scenario, opposite to what happens in the analysis of industrial data, where the root cause and the fault starting time are not always known, or are uncertain. The second case study is conducted on an industrial example, where we can assess the performance of monitoring methodologies in real-world environments, dealing with limited resources and information.

5.1. Case study 1: A simulated linear system

This case study aims to test the effectiveness of the proposed detection and diagnosis method by introducing different types of faults. To

$$\begin{aligned}
X_1 &= \delta_1 + 1.2\lambda X_8 + 0.8X_9 + \varepsilon_1\\
X_2 &= 0.6X_1 + \varepsilon_2\\
X_3 &= 0.05 + 0.22X_1 + \varepsilon_3\\
X_4 &= 1 + 0.4X_1 + \varepsilon_4\\
X_5 &= 0.062 + 0.16X_1 + \varepsilon_5\\
X_6 &= 0.6X_1 + \varepsilon_6\\
X_7 &= 0.7X_1 + \varepsilon_7\\
X_8 &= \delta_8 + \varepsilon_8\\
X_9 &= \varepsilon_9\\
X_{10} &= 1 + 0.4X_8 + \varepsilon_{10}\\
X_{11} &= 0.56 + 0.15X_8 + \varepsilon_{11}\\
X_{12} &= 0.8X_{16} + 0.51X_3 + \varepsilon_{12}\\
X_{13} &= 1.3X_3 + \varepsilon_{13}\\
X_{14} &= 1 + 0.4X_3 + \varepsilon_{14}\\
X_{15} &= 0.028 + 1.3X_8 + \varepsilon_{15}\\
X_{16} &= \varepsilon_{16}
\end{aligned}$$

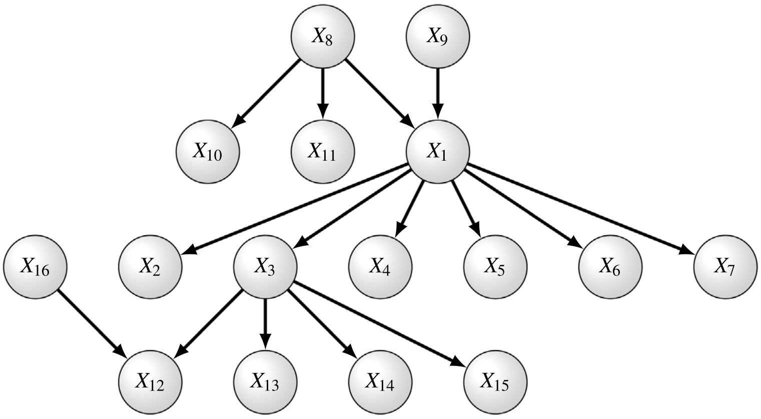


Fig. 9. Causal network and its relationships (Tamada et al., 2003).

Table 1
Generated datasets with induced faults.

Set	Induced fault	Modified factor	Change in descendants
$\mathbf{X}_1^{test}$	Process: Correlation change between X_1 and X_8	$\lambda' = 2$	True
$\mathbf{X}_2^{test}$	Operation: Step change in X_8	$\delta_8' = 2$	True
$\mathbf{X}_3^{test}$	Sensor: Reading bias of X_1	$\delta_1' = 2$	False
$\mathbf{X}_4^{test}$	–	–	–

Table 2
Fault detection ability.

Evaluation		PCA-T^2 (%)	PCA-Q (%)	BN-L^g (%)
Sensitivity (TPR)	$\mathbf{X}_1^{test}$	**14.7**	5.8	14.1
	$\mathbf{X}_2^{test}$	**6.2**	0.8	5.9
	$\mathbf{X}_3^{test}$	11.0	**100.0**	**100.0**
	Mean	10.6	35.0	**40.0**
Specificity (TNR)	$\mathbf{X}_4^{test}$	99.1	**99.3**	98.9

eliminate the bias caused by an unknown causal structure, we assume that the structure is known, i.e. the structure learning was conducted successfully, and is therefore skipped.

Data description. A causal network proposed by Tamada et al. (2003) was adopted. This network consists of 16 nodes (i.e., variables). The causal relations among these variables are presented in Fig. 9, where ϵ_i is a white noise sequence with a signal-to-noise ratio of 10 dB. The network equations also contain parameters used to generate three different faulty scenarios (Table 1), λ and δ_i. Under NOC, the multiplicative factor λ is set to be 1, and offset factors δ_i are set to be 0. As mentioned before (see Section 4.2), the fault types simulated are the following: a change in correlation between two variables (as a result of some abnormal changes in the system); a drifting in the operational conditions (due to a changes in some process inputs, e.g., raw materials, environmental conditions, inlet streams, etc.); a bias in the sensor (due to some malfunction in the measuring device).

Based on the definition described in Fig. 9, a synthetic dataset with 1000 samples is generated and denoted as $\mathbf{X}^{train}$. The dataset $\mathbf{X}^{train}$ is considered as the reference NOC data and is used to learn the PCA model and BN model parameters for the monitoring schemes.

To assess the capabilities of the proposed process monitoring scheme, a set of testing data $D^{test} = \{\mathbf{X}_1^{test}, \mathbf{X}_2^{test}, \mathbf{X}_3^{test}, \mathbf{X}_4^{test}\}$ were generated based on the settings provided in Table 1. Each dataset in D^{test} corresponds to a 1000×16 matrix. An artificial fault in $\mathbf{X}_1^{test}$ is introduced by changing the correlation between X_1 and X_8. The multiplicative factor λ' for the modified variable X_1' is set to 2 instead of 1, and the descendants of X_1 are generated based on X_1'. Similarly, a step change is inserted in the dataset $\mathbf{X}_2^{test}$ by setting the offset factor δ_8' to be 2, and its descendants are generated based on X_8'. Dataset $\mathbf{X}_3^{test}$ is used to simulate a sensor bias in X_1. Hence, the sensor readings is expressed by $X_1' = \delta_1' + 1.2\lambda X_8 + 0.8X_9 + \epsilon_1$, where $\delta_1' = 2$. The descendants of X_1 are not affected because the underlying structure remains the same (only the measurement was affected, not the true underlying state). Thus, the values of descendants are simulated based on X_1. The last dataset $\mathbf{X}_4^{test}$ applies the same setting as $\mathbf{X}^{train}$ and represents the process under NOC. In this study, only single faults were simulated. This is the most common situation under the reasonable assumption of independent process/operation/sensor faults, where the occurrence of a single event is much more likely than the simultaneous occurrence of multiple (rather rare) events.

General settings. In this study, we compare the proposed approach with the conventional PCA approach. Both monitoring schemes and their control charts are constructed based on the same data splitting procedure. The training set $\mathbf{X}^{train}$ is used to build the model. The control limits are obtained by 10-fold cross-validation of $\mathbf{X}^{train}$. Note that the control limits can also be established using an independent validation set if data is sufficient.

Evaluation: Individual mode. The effectiveness of fault detection is assessed based on the *sensitivity* expressed in terms of True Positive Rate (TPR), and *specificity* expressed in terms of True Negative Rate (TNR). The first three testing sets are used to assess the sensitivity of different monitoring schemes. The results are displayed in Table 2. The performances of T^2 and L^g are similar in detecting the process change in correlation and the step change, i.e. $\mathbf{X}_1^{test}$ and $\mathbf{X}_2^{test}$. The control charts for detecting the correlation change are shown in Fig. 10. For detecting the sensor bias in $\mathbf{X}_3^{test}$, both Q and L^g control charts show high sensitivity. The specificity is evaluated through $\mathbf{X}_4^{test}$, where it is possible to verify that the three control charts are very close to the theoretical value of the false alarm rate, which corresponds to the significance level established for the control limits (0.01).

Once an abnormal observation is detected, the next step is to isolate the faulty variables. Among the detected abnormal observations (or batch), the correct isolation rate is computed as the percentage of the observations, where the faulty variables are successfully isolated by different methods. As shown in Table 3, although the correct isolation rate of DC_k^Q in $\mathbf{X}_1^{test}$ outperforms the other two methods, the sample size is relatively small. The proposed approach shows a better performance in $\mathbf{X}_2^{test}$ in terms of the number of detected observations, being close to T^2. The results of $\mathbf{X}_3^{test}$ shows that DC_k^Q produces the highest correct isolation rate among the three approaches. It is not straightforward

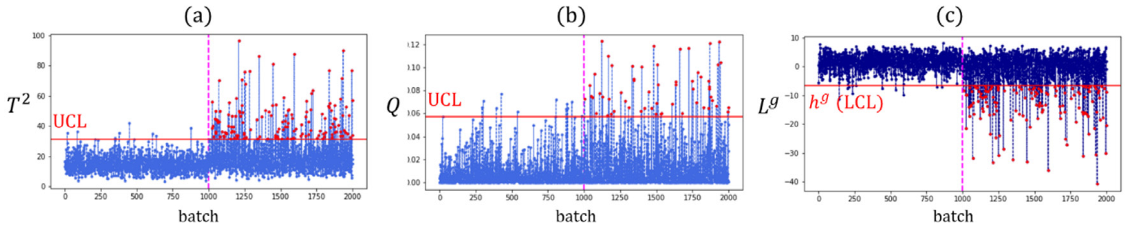


Fig. 10. Control charts for the process change in correlation. The two datasets, $\mathbf{X}^{train}$ and $\mathbf{X}_1^{test}$, are split by the dashed line, and the red points indicate the observations signaled as out-of-control: (a) T^2 control chart; (b) Q control chart; (c) L^g control chart.

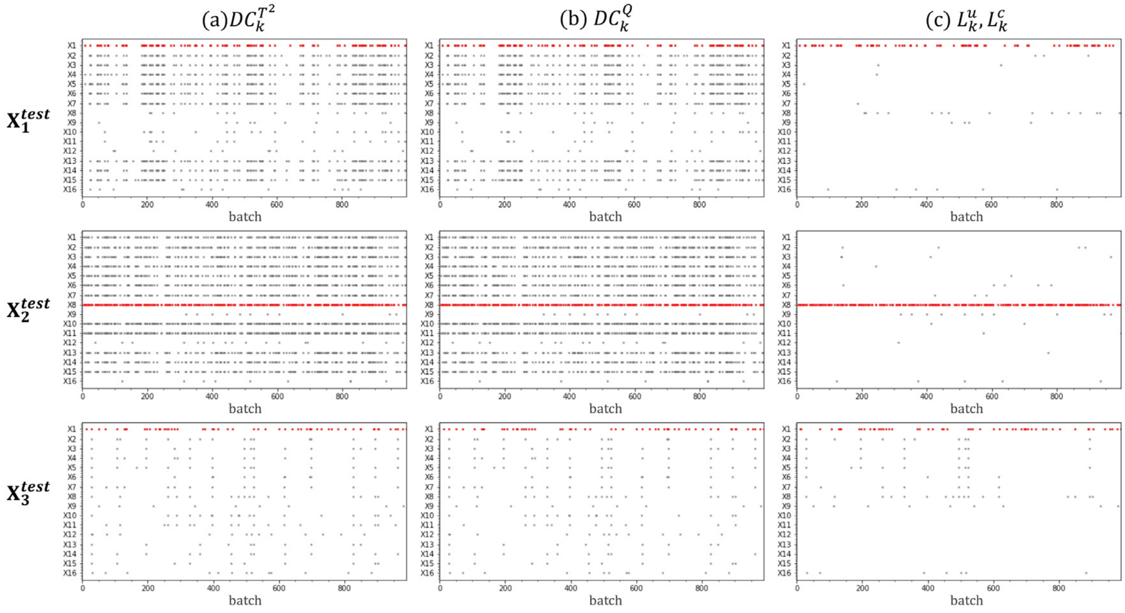


Fig. 11. Isolated faults for each observation based on: (a) Contributions of T^2, (b) Contributions of Q, (c) Diagnosis based on L_k^u and L_k^c. A red point represents a correctly identified faulty variable, and gray points represent other signaled variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
The performance of fault diagnosis: *Correct isolation rate (# of ooc observations)*.

Correct isolation rate (%) (# of ooc observations)	$DC_k^{T^2}$	DC_k^Q	L_k^u, L_k^c
$\mathbf{X}_1^{test}$	14.3 (147)	<u>**51.7**</u> (58)	45.4 (141)
$\mathbf{X}_2^{test}$	21.0 (62)	0.0 (8)	<u>**69.5**</u> (59)
$\mathbf{X}_3^{test}$	0.0 (110)	<u>**6.3**</u> (1)	5.6 (1)

Table 4
The performance of detection under batch mode.

Evaluation		L^g (%)
Sensitivity	$\mathbf{X}_1^{test}$	87.9
	$\mathbf{X}_2^{test}$	84.8
	$\mathbf{X}_3^{test}$	100.0
Specificity	$\mathbf{X}_4^{test}$	100.0

however to establish which approach exhibits the best performance. However, it is possible to conclude that the proposed method is very competitive when compared to the classic PCA approach.

The ideal monitoring should be capable of identifying the root cause, or at least providing a reduced set of variables that could be involved in the fault mechanism. As shown in Fig. 11, the proposed approach produces a significantly lower number of signaled variables. The induced process faults in $\mathbf{X}_1^{test}$ and the operation disturbances in $\mathbf{X}_2^{test}$ suffer from the smearing-out effect, propagating the anomaly to other relevant variables. In this context, the number of signaled variables based on the contribution plot is very large.

Evaluation: Batch mode. The proposed BN-based monitoring scheme can also monitor the process by batch, i.e., processing sets of observations, instead of individual observation. Batch mode monitoring can be more efficient for monitoring high sampling rate processes. Based on the same datasets, we assess the effectiveness of batch mode monitoring by setting the batch size to $n_b = 30$. Instead of calculating the likelihood metrics for each sample, 33 values are computed for both the training set and all the testing sets, representing the state of each batch. The control charts of L^g in batch mode are presented in Fig. 12. As shown in Table 4, batch mode performs well in terms of both sensitivity and specificity.

By visualizing the causal structure with the proposed indices, the state of each variable can be clearly presented. Fig. 13(a) presents the diagnosis results of an abnormal batch of $\mathbf{X}_1^{test}$. The correlation between X_1 and X_8 has been changed, which leads to changes in the distribution of X_1 and its descendants, i.e., $L_{k,b}^u < h_k^u$ (see those nodes marked in yellow). Based on the conditional local likelihood index, it is possible to check the distribution of residuals, excluding the effects caused by parents. Since the $L_{c,b}^c$ of descendants of X_1 are all under control, this implies that the changes in their distributions are caused by X_1. Similarly, a step-change in X_8 is propagated to its descendant distributions, as shown in Fig. 13(b). By looking at the conditional index, the variable X_8 can be identified as the root cause (see the red rim). Fig. 13(c) presents the results of a sensor bias in X_1. Since the underlying process remains the same as NOC, only X_1 with inaccurate

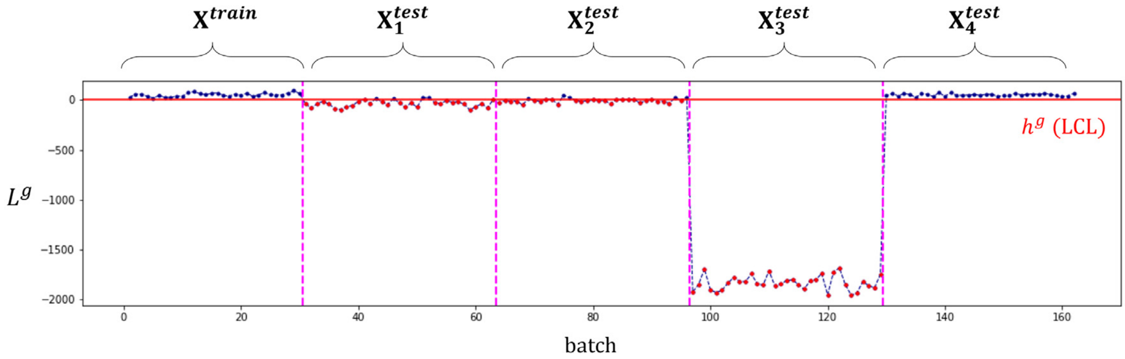
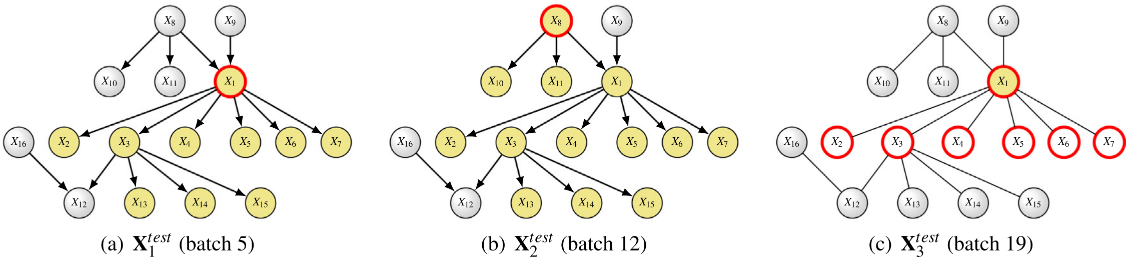


Fig. 12. Control chart of L^g in batch mode.



(a) X_1^{test} (batch 5) (b) X_2^{test} (batch 12) (c) X_3^{test} (batch 19)

Fig. 13. Diagnosis results based on the BN structure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

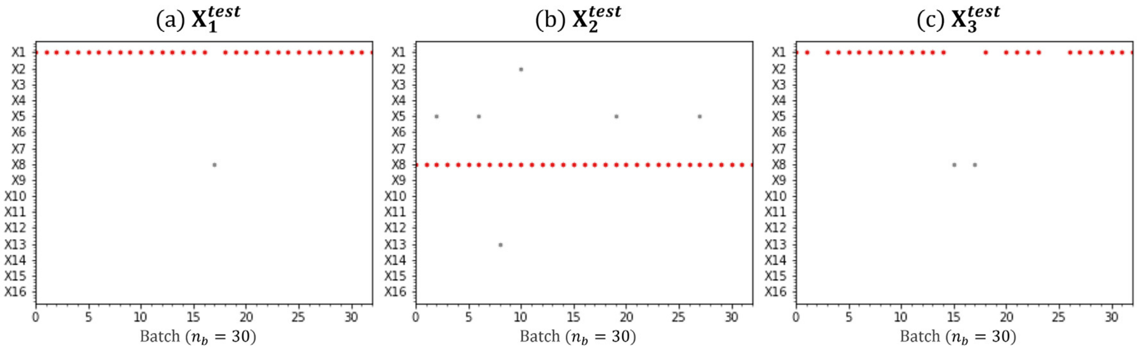


Fig. 14. Isolated faults of each batch. A red point represents a correctly identified faulty variable, and gray points represent other signaled variables.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
The performance of fault diagnosis: *Correct isolation rate*.

Dataset	Correct isolation rate (%) (#of ooc batch)
$\mathbf{X}_1^{test}$	96.6 (29)
$\mathbf{X}_2^{test}$	100.0 (28)
$\mathbf{X}_3^{test}$	78.8 (33)

sensor readings shows a change in its unconditional distribution. The incorrect sensor readings were taken into account for computing the $L_{k,c}^c$ of the descendants of X_1, which result in the anomalies in their conditional distributions (see nodes with red rim).

A causal structure with the information of likelihood indices provides an overview of the affected regions. By employing the labeling procedure described in Section 4.3, the suspicious root causes and the possible fault types can be identified. The correct isolation rates are summarized in Table 5. The performance of batch mode is better than that of individual mode (Table 3), as the metrics of batch mode are based on more evidence, i.e., sets of observations. The batch mode also illustrates its effectiveness in terms of the number of isolated variables, as shown in Fig. 14.

Furthermore, the labeling procedure can provide more information about possible types of faults by checking the children of a given faulty variable. As shown in Table 6, the proposed approach successfully labels the faulty variable with the corresponding root cause for most batches. This information can help reducing the investigation time.

5.2. Case study 2: Etching process in semiconductor manufacturing

In this section, a case study conducted on real data from a semiconductor manufacturing plant, was used to assess the capability of the proposed approach. The previous case study examined the effectiveness of BN-monitoring in both individual mode and batch mode, and in this case study the individual mode is adopted due to limitations in the number of samples available.

Data description. The dataset is collected from the LAM 9600 plasma etching process at Texas Instrument Inc (He & Wang, 2007; Wise, Gallagher, Butler, White, & Barna, 1999). The data consists of 107 normal wafers and 20 faulty wafers from three experiments, and 19 sensor reading variables (see Appendix D). The process consists of six steps, such as gas flow stabilization, etching on different layers (He & Wang, 2007). As the focus in this case study is fault analysis instead of

Table 6
Results of the labeling procedure.

Dataset	Induced fault	# of batches correctly labeled fault type
$\mathbf{X}_1^{test}$	Process: Correlation change between X_1 and X_8	32 labeled C in X_1
$\mathbf{X}_2^{test}$	Operation: Step change in X_8	32 labeled S in X_8
$\mathbf{X}_3^{test}$	Sensor: Reading bias of X_1	32 labeled B in X_1

Table 7
Description of experiments.

Experiment	# of normal wafers	# of faulty wafers	Fault description
Exp29	34	9	TCP (+50), RF (−12), RF (+10), Pr (+3), TCP (+10), BCl3 (+5), Pr (−2), Cl2 (−5), He Chuck (unknown)
Exp31	36	5	TCP (+30), Cl2 (+5), BCl3 (−5), Pr (+2), TCP (−20)
Exp33	37	6	TCP (−15), Cl2 (−10), RF (−12), BCl3 (+10), Pr (+1), TCP (+20)

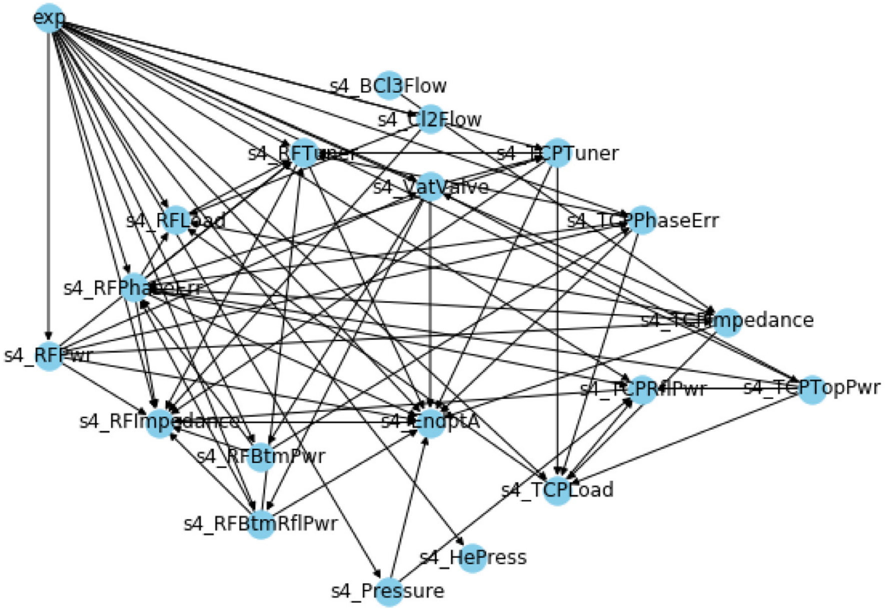


Fig. 15. The fitted conditional linear Gaussian Bayesian network. The categorical variable is labeled as exp.

selecting important features, we only consider the data points from one of the main steps, namely etch of the aluminum layer (i.e. step4). Other studies in the literature focus on the detection accuracy by considering more process steps.

Three experiments $\{29, 31, 33\}$ were performed and the faults are intentionally induced by changing the settings of different controllable variables (see Table 7). These experiments aim to simulate the scenario of sensor failure, which means that the sensors fail to detect the changes in the process. To generate a faulty wafer, the set-point of a controllable variable was changed during the experiment. After the experiment, the data collected from the controllable variable was manually reset to the same level as its normal baseline. Therefore, the values of controllable variables would appear normal, while other relevant variables might exhibit abnormal. More details about these experiments can be found in study of Wise et al. (1999).

General settings. As experiments were run at different time periods, February, March, and April, respectively, the process drift and changes on covariance among variables can be observed (He & Wang, 2007; Wise et al., 1999). To obtain the optimized models, both PCA and BN are constructed for each experiment.

The effectiveness of monitoring methods was evaluated by their sensitivity and specificity, i.e., True Positive Rate (TPR) and True Negative Rate (TNR). Thus, five normal wafers of each experiment are excluded from the learning process and are used for testing. Let $D^{train} = \{\mathbf{X}_{29}^{train}, \mathbf{X}_{31}^{train}, \mathbf{X}_{33}^{train}\}$ be a set of training data used for learning the structure of the Bayesian network, and $D^{test} = \{\mathbf{X}_{29}^{test}, \mathbf{X}_{31}^{test}, \mathbf{X}_{33}^{test}\}$ is a set of testing data consists of both normal wafers and faulty wafers.

BN modeling. Typically, the interactions among process variables do not change in a short period. Hence, we assume that the BNs of each experiment share the same structure. To learn a general BN structure under limited training data, the Conditional Linear Gaussian Bayesian Network (CLGBN) is considered (Lauritzen & Wermuth, 1989; Scutari, 2009) (see Appendix C). With a joint dataset $\mathbf{X}_{CLGBN} = [\mathbf{X}^{train}, Z]$, where $\mathbf{X}^{train} = [\mathbf{X}_{29}^{train}, \mathbf{X}_{31}^{train}, \mathbf{X}_{33}^{train}]$, and Z is a vector corresponding to the categorical variable, which indicates the experiment ID, an CLGBN model is obtained though the hill climbing algorithm (see Fig. 15) available in bnlearn (R package) (Scutari, 2009). The blacklist provided by SMEs is listed in Appendix D. The causal connections of CLGBN are kept as the general structure $G = (V, E)$, except the categorical variable and its arcs are removed. The parameters of the three BNs, $\theta_{29}^G, \theta_{31}^G$, and θ_{33}^G, are learned based on the corresponding experimental data given the general structure G. In this case study, the CLGBN model is considered as a workaround for learning the global structure given such limited historical dataset. Authors suggest that practitioners can determine the appropriate model depending on the available resources, such as type of variables or size of dataset.

Evaluation. Control limits for each monitoring statistic were obtained as described in Section 4.1. The results obtained were compared with those from PCA-based monitoring method, and are shown in Table 8. The Q statistic and the global likelihood index L^g show equally high sensitivities, while the detection rate of T^2 is relatively low. As shown in Table 8, T^2 outperforms the other metrics in terms of specificity.

The performance of a monitoring method is evaluated by checking if the root cause variables are successfully isolated. For example, in the

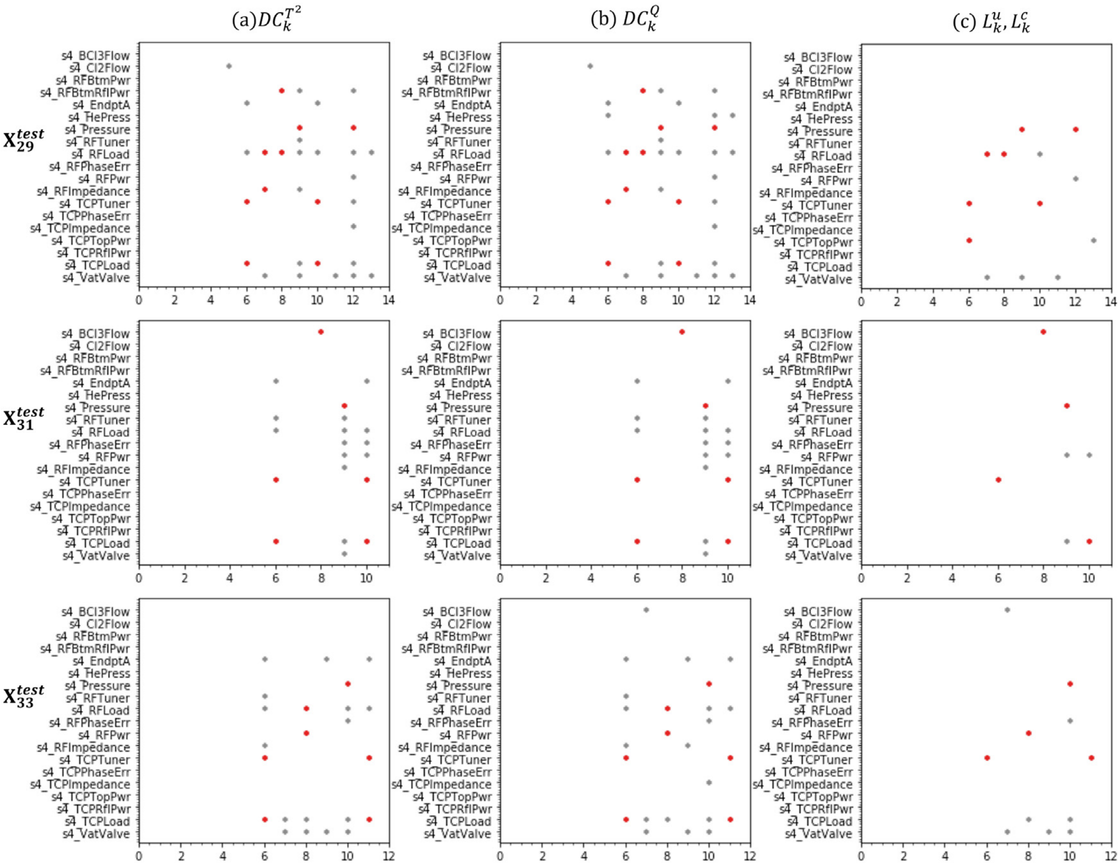


Fig. 16. Isolated faults for each wafer based on: (a) Contributions of T^2, (b) Contributions of Q statistics, (c) Diagnosis based on the L_k^u and L_k^c. In each experiment, the first five wafers are normal wafers. A red point represents a correctly identified faulty variable, and gray points represent other signaled variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

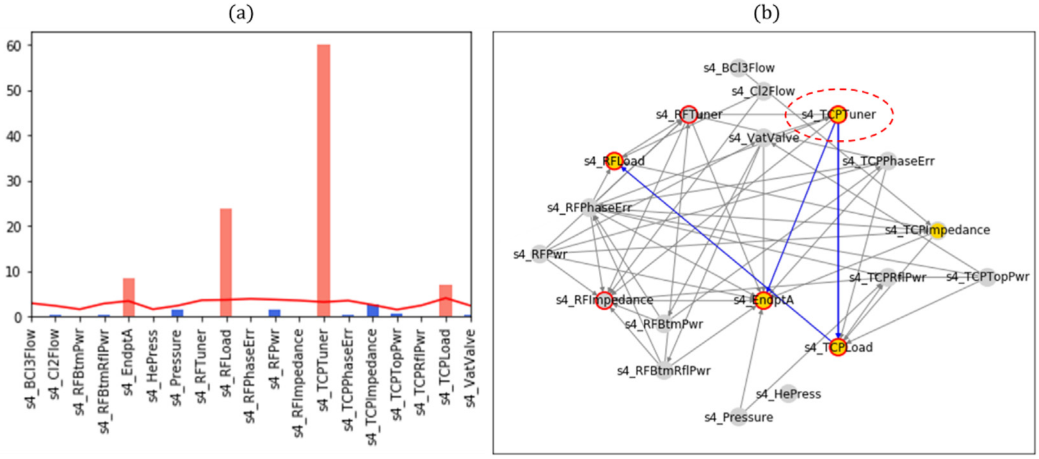


Fig. 17. Isolated faults of each wafer based on different methods: (a) Contributions of Q statistics; (b) L_k^u and L_k^c with the labeling procedure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

case of a faulty wafer with an induced fault, where the set-point of TCP power was changed by 50, we expect the variables related to TCP to be identified.

As shown in Table 9, all methods successfully target the root cause for 14 wafers among 20 faulty wafers. However, by looking at the number of identified variables (see Fig. 16), the variables identified based on likelihood metrics are much less than the variables identified by other two approaches.

The diagnosis results of the 11*th* wafer of $\mathbf{X}_{33}^{test}$ are shown in Fig. 17. Based on the contribution plot of Q statistics, four variables are identified (see Fig. 17(a)). As illustrated in Section 4.3, the proposed diagnosis procedure consists of two steps, where the set of abnormal variables is first extracted based on likelihood metrics L_k^u and L_k^c, and then a labeling process is conducted to identify the fault type. In this case, four abnormal variables are identified by likelihood metrics (see the node marked in yellow with a red rim). After the

Table 8
Fault detection ability.

Dataset	Sensitivity (%)			Specificity (%)		
	PCA-T^2	PCA-Q	BN-L^g	PCA-T^2	PCA-Q	BN-L^g
$\mathbf{X}_{29}^{test}$	4/9	**8/9**	**8/9**	**5/5**	3/5	4/5
$\mathbf{X}_{31}^{test}$	4/5	**4/5**	**4/5**	**5/5**	3/5	**5/5**
$\mathbf{X}_{33}^{test}$	5/6	**6/6**	**6/6**	**5/5**	4/5	4/5
Overall	13/20	**18/20**	**18/20**	**15/15**	10/15	13/15

Table 9
Fault diagnosis. The last three columns represent the number of wafers successfully diagnosed by different methods: T^2, Q, L_k^u and L_k^c.

Dataset	# of faulty wafers	$DC_k^{T^2}$	DC_k^Q	L_k^u, L_k^c
$\mathbf{X}_{29}^{test}$	9	6	6	6
$\mathbf{X}_{31}^{test}$	5	4	4	4
$\mathbf{X}_{33}^{test}$	6	4	4	4
Overall	20	14	14	14

labeling procedure, the size of the identified variables is reduced to 1 because the other three variables are descendants of TCP Tuner (see Fig. 17(b)). With a causal structure, the efficiency of troubleshooting can be improved by checking the information provided by neighbor nodes.

As referred in Section 4.3, we suggest to analyze the labeling results only when the data is sufficient, e.g. $n_b > 1$. In this case study, since the sample size is small, the individual mode has been carried out. Besides, the dataset is collected from designed experiments with manual induced faults, where this type of fault is not common. Thus, only the isolation of the variable is discussed.

5.3. Discussion

In this paper, we present a new process monitoring based on BN and a structured analysis of the information available in its nodes, and compare it with PCA-based statistical process control. Analyzing the results from the two case studies presented in the previous section, we can further analyze the differences between the two approaches. Normally, a PCA-based approach requires two indices for fault detection, the T^2 statistic and Q statistic, which are two complementary indices used to monitor the PCA subspace and residual subspace, respectively. The PCA subspace represents existing correlations among variables, while the residual subspace represents unstructured variation. Therefore, in the first case study, T^2 could capture a correlation change and a step change, and Q was able to detect the sensor bias (see Table 2). In the second case study, in the sensor fault case, the Q statistic could also catch these anomaly slightly better than T^2 (see Table 8). The idea of the proposed global likelihood index is similar to monitoring correlations and residuals, but in the local sense, i.e., looking at direct associations between variables. The index is computed based on a causal structure, which considers the direct correlations among variables. By adding up the conditional local likelihood, the summarized index can also reflect the anomalies which cannot be explained by their parents, i.e., detecting higher residual. In this case, we can use a single global likelihood index to provide similar performance as the two indices of PCA (see Tables 2 and 8).

After detecting the abnormal observations, the PCA-based approach uses Diagonal Contributions (DC) to spot the faulty variables, and the BN-based approach combines two local likelihood indices to find the faulty variables. Although the correct isolation rates were similar for both approaches (see Tables 3 and 9), the proposed approach produces a significantly lower number of signaled variables (false positives) as shown in Figs. 11 and 17. This is because many process faults suffer from the smearing-out effect, propagating the anomaly to other relevant variables (thus resulting in a high rate of false positives). The proposed conditional local likelihood eliminates the effects caused by parents and therefore it can discover the root cause more efficiently.

6. Conclusions and future perspectives

In this paper, we explore the use of Bayesian Networks to improve the diagnostic capabilities, while maintaining the detection accuracy of state-of-the-art methods. We propose a new process monitoring scheme based on an interpretable machine learning model and evaluate it on several simulated and real case studies. As a BN describes a process in a structured way, the interactions between process variables can be easily incorporated and illustrated. Given the flexibility of combining existing knowledge, the diagnosis result can better reflect the physical meaning. Three metrics are employed for online monitoring. A global likelihood index is adopted to monitor the overall system (level 1 - detection), and two local likelihood indices are used to check the changes in local distributions (level 2 - diagnosis). These statistics are instrumental for diagnosing the type of root cause and their location, through a new labeling procedure. The granularities of these metrics can be determined by practitioners depending on the sample frequency of a process, which is also a novel approach in process monitoring. A process with a high sampling rate can be monitored in a batch way so that the indices can provide more precise information with more evidence. Furthermore, unlike many fault diagnosis models that require historical dataset with fault labels to provide better interpretability, the proposed method only requires NOC data to build the structure. In this context, the scope of monitoring is not limited to or conditioned by the previous faulty patterns.

Through the simulated case study, we demonstrate the detection rate and correct isolation rate of a BN-based monitoring scheme in an individual mode are similar to the performance of the classic PCA approach. And the BN-based method can isolate a smaller number of variables in the diagnostic phase. The experiment of a batch mode further shows a significant improvement in both detection and diagnosis. Besides, through the labeling procedure, the identified variables are correctly labeled with the fault type, which provides more information for the subsequent troubleshooting and repairing stages. An industrial example was also considered to compare the two process monitoring approaches on a real dataset. The result shows that both approaches have similar performance in detection rate and correct isolation rate. By looking at the number of isolated variables, the BN-based approach again outperforms the PCA-based approach.

The effectiveness of the proposed BN-based process monitoring has been illustrated in both fault detection and fault diagnosis. Nevertheless, some factors should be taken into account before implementation. Since a Bayesian network is a directed acyclic graph (Pearl, 2014), it cannot be used to model a closed-loop control system, such as the classic well-known Tennessee Eastman Process (Downs & Vogel, 1993). Other possible network structures for a closed-loop system should be investigated in future study. Besides, an ideal BN should be compatible with the physical laws, so we highly recommend to include the SMEs during the development stage.

To demonstrate the capabilities of the proposed BN-based monitoring scheme, we have used a linear stationary system. There are several remaining aspects to be investigated in the future. For instance, the next logical step is to extend the proposed approach to a non-linear system or non-stationary system. The performance of a large network should be studied as well. Since the global likelihood is calculated by summing the local likelihood, a single index may not be sufficient in monitoring large systems. Decomposing a large network into several communities can be one of the possible solutions (Clauset, Newman, & Moore, 2004). It may improve the diagnosis efficiency as well as by early targeting faulty regions. Furthermore, only one induced fault is designed in the simulation cases and the industrial case. Future studies should also include the scenario of multiple faults and other types of faults, such as gradual drifts. In this paper, we demonstrate that a BN-based process monitoring approach is able to improve the interpretability of fault diagnosis compare to the traditional approach. Although some limitations and remaining issues needed further investigation, we believe this is an interesting direction towards eXplanatory Artificial Intelligence.

Table A.10
Notations.

Indexes and parameters:	
n	Number of samples
m	Number of variables
B	Number of batches in $\mathbf{X}$
n_b	Batch size
$\mathbf{X}$	$n \times m$ data matrix
Σ	Covariance matrix of $\mathbf{X}$
p	Number of retained principal components
$\mathbf{T}$	$n \times p$ matrix of PCA score
$\mathbf{P}$	$m \times p$ matrix of PCA loadings
$\mathbf{E}$	$n \times m$ residual matrix
T^2	T^2 statistic of Hotelling
Q	Squared Prediction Error (SPE), also known as Q statistics
DC_k^{index}	Contribution of variable k for $index$ statistic, $index \in \{T^2, Q\}$
$G(V, E)$	Graph defined by a set of nodes V and a set of arcs E
V	Set of nodes, where V_k corresponds to variables $X_k \in \mathbf{X}$
$e_{ij} \in E$	Arc from node V_i (i.e., variable X_i) to node V_j (i.e. variable X_j)
$\mathbb{X}_{pa(k)}$	Set parent nodes (i.e., variables) of variable X_k
$\mathbb{X}_{ch(k)}$	Set child nodes (i.e., variables) of variable X_k
θ_G	Parameters of graph G
T_G	Topological sorted permutation of graph G
$\mathbf{X}_b$	$n_b \times m$ data matrix of batch b
$L_{k,b}^u$	Unconditional local likelihood of variable X_k of batch b
$L_{k,b}^c$	Conditional local likelihood of variable X_k of batch b
L_b^g	Global likelihood of batch b
h_k^u	Control limit of L_k^u
h_k^u	Control limit of L_k^c
h^g	Control limit of L^g
$\mathbb{X}_{abn}$	Set of abnormal variables for which $L_{k,b}^u < h_k^u$ and $L_{k,b}^c < h_k^c$, $\forall X_k \in \mathbb{X}_{abn}$
$\mathbb{X}_{label}$	Set of abnormal variables after applying the labeling procedure

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is conducted in the framework of the project MADEin4, which has received funding from the ECSEL JU (Electronic Components and Systems for European Leadership Joint Undertaking) under grant agreement No 826589. The JU receives support from the European Union's Horizon 2020 research and innovation program and France, Germany, Austria, Italy, Sweden, Netherlands, Belgium, Hungary, Romania and Israel. The authors are also grateful to our local industrial, who provided precious process knowledge for validating the case study.

Appendix A. Notations

The mapping notations used in this paper are summarized in Table A.10.

Appendix B. Principal component analysis and its application in process monitoring

Principal Component Analysis (PCA) is a popular multivariate method that has been widely adopted for process monitoring. The objective of PCA is to reduce the number of variables by projecting data into a lower dimension space that explains most of the original information. Let $\mathbf{X}$ be a data matrix with n samples and m variables (usually preprocessed, namely centered and possibly also scaled so that all variables have zero mean and unit variance). The covariance matrix of $\mathbf{X}$ is denoted by $\boldsymbol{\Sigma}$. PCA transforms the original variables into new orthogonal variables, by decomposing $\mathbf{X}$ as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{B.1}$$

where $\mathbf{T}$ is an n by p matrix of scores, $\mathbf{P}$ is an m by p matrix of loadings, p is the number of retained principal components, and $\mathbf{E}$ is an n by m residual matrix. The columns of $\mathbf{P}$ are the eigenvectors of $\boldsymbol{\Sigma}$ associated with the p largest eigenvalues, and the remaining eigenvectors are denoted by $\tilde{\mathbf{P}} \in \mathbb{R}^{m\times(m-p)}$. By applying PCA, the original data is decomposed into two complementary spaces: the PCA subspace $\mathbf{T}\mathbf{P}^T$, and the residual subspace $\mathbf{E} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}\mathbf{P}^T$, where $\tilde{\mathbf{T}} = \tilde{\mathbf{P}}^T\mathbf{X}$. Two complementary statistics are often employed to monitor the variance in these two spaces, namely the Hotelling's T^2 statistic of the scores and the Squared Prediction Error (SPE).

The Hotelling's T^2 monitors the variance in the PCA subspace and is computed as:

$$T^2 = \mathbf{x}^T\mathbf{P}\boldsymbol{\Lambda}_p^{-1}\mathbf{P}^T\mathbf{x} \tag{B.2}$$

where $\mathbf{x} \in \mathbb{R}^m$ is a sample vector and $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with the first p eigenvalues in the main diagonal. The upper control limit (UCL) of T^2 is defined as follows:

$$UCL(T^2) = \frac{p(n-1)(n+1)}{n^2 - np}F_{\alpha,p,n-p} \tag{B.3}$$

where $F_{\alpha,p,n-p}$ is the upper α percentile of distribution F with the degree of freedom p and $n-p$.

The variation in the complementary residual space is monitored by SPE of residuals $\mathbf{e} \in \mathbb{R}^m$, which is also known as the Q statistic,

$$Q = \mathbf{e}^T\mathbf{e} = \mathbf{x}^T\tilde{\mathbf{P}}\tilde{\mathbf{P}}^T\mathbf{x} \tag{B.4}$$

The UCL of the Q statistic is defined as:

$$UCL(Q) = \theta_1\left(\frac{z_\alpha\sqrt{2\theta_2h_0^2}}{\theta_1} + 1 + \frac{\theta_2h_0(h_0-1)}{\theta_1^2}\right)^{\frac{1}{h_0}} \tag{B.5}$$

with $\theta_i = \sum_{j=p+1}^{m}\lambda_j^i$, $i \in \{1,2,3\}$ and $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, where λ_j is jth eigenvalue and z_α is the upper α percentile of the standard normal distribution.

Fault detection can be done by monitoring T^2 and Q statistics. Once a fault is detected, the contribution charts can be used to isolate the variables that may be connected to it. Many approaches have been developed to define T^2 contribution (Nomikos & MacGregor, 1995; Qin et al., 2001; Westerhuis, Gurden, & Smilde, 2000). In this paper, the Diagonal Contributions (DC) proposed by Qin is employed (Alcala & Qin, 2011). The general DC of variable k for the T^2 and Q statistics is calculated as:

$$DC_k^{index} = \mathbf{x}^T\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T\boldsymbol{x} \tag{B.6}$$

where $\mathbf{M}^{T^2} = \mathbf{P}\boldsymbol{\Lambda}_p^{-1}\mathbf{P}^T$ and $\mathbf{M}^Q = \tilde{\mathbf{P}}\tilde{\mathbf{P}}^T$. ξ_k is the kth column of the identity matrix, denoted by $\xi_k = [0,\dots,1,\dots,0]^T$.

Since DC_k^{index} has a quadratic form, its distribution can be approximated by $g_k\chi^2(h_k)$ distribution (Box 1954). The control limits of DC_k^{index} can be obtained for a given significance level α (Alcala & Qin, 2011). Parameters g and h are calculated as:

$$g_k = \frac{tr\{\mathbf{S}\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T\}^2}{tr\{\boldsymbol{S}\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T\}} = \boldsymbol{S}\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T \tag{B.7}$$

$$h_k = \frac{tr\{\mathbf{S}\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T\}^2}{tr\{\mathbf{S}\xi_k\xi_k^T\mathbf{M}^{index}\xi_k\xi_k^T\}} = 1 \tag{B.8}$$

Note that many researchers have pointed out that using control limits to identify the significant variables may mislead the diagnosis result because of the smearing out effect, that leads to an increased contribution of variables unrelated with the fault (Van den Kerkhof et al., 2013; Westerhuis et al., 2000). Nevertheless, a contribution plot with control limits is still one of the most common approaches for conducting PCA-based diagnosis. In this paper, the metrics introduced above are employed as a benchmark for comparison. More details and extensions of PCA-based monitoring schemes can be found in the literature (Qin, 2012).

Table D.11
Blacklist.

	BCl3Flow	Cl2Flow	RFBtmPwr	RFBtmRflPwr	EndptA	HePress	Pressure	RFTuner	RFLoad	RFPhaseErr	RFPwr	RFImpedance	TCPTuner	TCPPhaseErr	TCPImpedance	TCPTopPwr	TCPRflPwr	TCPLoad	VatValve
BCl3Flow		1	1	1		1	1				1					1			
Cl2Flow	1		1	1		1	1				1					1			
RFBtmPwr	1	1				1	1				1					1			
RFBtmRflPwr	1	1				1	1				1					1			
EndptA	1	1	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1
HePress	1	1		1	1		1				1					1			1
Pressure	1	1		1		1					1					1			
RFTuner	1	1				1	1				1					1			
RFLoad	1	1				1	1				1					1			
RFPhaseErr	1	1				1	1				1					1			
RFPwr	1	1				1	1									1			
RFImpedance	1	1				1	1				1					1			
TCPTuner	1	1				1	1				1					1			
TCPPhaseErr	1	1				1	1				1					1			
TCPImpedance	1	1				1	1				1					1			
TCPTopPwr	1	1				1	1				1								
TCPRflPwr	1	1				1	1				1					1			
TCPLoad	1	1				1	1				1					1			
VatValve	1	1				1	1				1					1			

Appendix C. Conditional linear Gaussian Bayesian network

Conditional Linear Gaussian Bayesian Network (CLGBN) is a hybrid BN which consists of discrete and continuous variables, where the continuous ones cannot be the parents of the discrete ones (Lauritzen & Wermuth, 1989). The local distribution of the continuous variables X_k given its parents $\mathbf{X}_{pa(k)} = \mathbf{X}_{pa(k),D} \cup \mathbf{X}_{pa(k),C}$ is defined as a conditional Gaussian distribution:

$$f(X_k|\mathbf{X}_{pa(k)}) = \mathcal{N}\left(\alpha(\mathbf{X}_{pa(k),D}) + \beta(\mathbf{X}_{pa(k),C})^T \sigma^2(\mathbf{X}_{pa(k),D})\right) \tag{C.1}$$

where $\mathbf{X}_{pa(k),C}$ is the matrix of continuous parent variables, $\mathbf{X}_{pa(k),D}$ is the matrix of discrete parent variables, and α and β are the coefficients of the linear regression model of X_k given its continuous parents. This model can be different depending on the values of its discrete parents $\mathbf{X}_{pa(k),D}$.

Appendix D. Blacklist

Existing domain knowledge can be included in BN structure learning by incorporating an association matrix $\mathbf{A}$ defined by SMEs. Each element $a_{i,j} \in \mathbf{A}$ specifies the infeasible causality between variables X_i and X_j. Hence, if X_i cannot cause X_j, then $a_{i,j} = 1$. Other elements not specified will be learned from data. The association matrix considered in the framework of Case Study 2 is provided in Table D.11.

References

Alcala, C. F., & Qin, S. J. (2011). Analysis and generalization of fault diagnosis methods for process monitoring. *Journal of Process Control*, *21*(3), 322–330.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bauer, M., & Thornhill, N. F. (2008). A practical method for identifying the propagation path of plant-wide disturbances. *Journal of Process Control*, *18*(7–8), 707–719.

Cai, B., Huang, L., & Xie, M. (2017). Bayesian networks in fault diagnosis. *IEEE Transactions on Industrial Informatics*, *13*(5), 2227–2240.

Campos, L. M. d. (2006). A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, *7*(Oct), 2149–2187.

Chiang, L. H., Jiang, B., Zhu, X., Huang, D., & Braatz, R. D. (2015). Diagnosis of multiple and unknown faults using the causal map and multivariate statistics. *Journal of Process Control*, *28*, 27–39.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*(Nov), 507–554.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), Article 066111.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

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

Elidan, G., Ninio, M., Friedman, N., & Schuurmans, D. (2002). *Data perturbation for escaping local maxima in learning*.

Ge, X., Wang, B., Yang, X., Pan, Y., Liu, B., & Liu, B. (2021). Fault detection and diagnosis for reactive distillation based on convolutional neural network. *Computers & Chemical Engineering*, *145*, Article 107172.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)* (pp. 80–89). IEEE.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

He, Q. P., & Wang, J. (2007). Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, *20*(4), 345–354.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

Hotelling, H. (1947). Multivariate quality control. *Techniques of Statistical Analysis*.

Jackson, J. E., & Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics*, *21*(3), 341–349.

Van den Kerkhof, P., Vanlaer, J., Gins, G., & Van Impe, J. F. (2013). Analysis of smearing-out in contribution plot based fault isolation for statistical process control. *Chemical Engineering Science*, *104*, 285–293.

Knuth, D., & Addison-Wesley (1997). *Addison-Wesley series in computer science and information processing*: *vol. 1*, *The art of computer programming: Fundamental algorithms*. Addison-Wesley.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 31–57.

Li, W., Li, H., Gu, S., & Chen, T. (2020). Process fault diagnosis with model- and knowledge-based approaches: Advances and opportunities. *Control Engineering Practice*, *105*, Article 104637.

Lou, C., Li, X., Atoui, M. A., & Jiang, J. (2020). Enhanced fault diagnosis method using conditional Gaussian network for dynamic processes. *Engineering Applications of Artificial Intelligence*, *93*, Article 103704.

Lv, F., Wen, C., Bao, Z., & Liu, M. (2016). Fault diagnosis based on deep learning. In *2016 American control conference (ACC)* (pp. 6851–6856). IEEE.

MacGregor, J. F., & Kourti, T. (1995). Statistical process control of multivariate processes. *Control Engineering Practice*, *3*(3), 403–414.

Mahadevan, S., & Shah, S. L. (2009). Fault detection and diagnosis in process data using one-class support vector machines. *Journal of Process Control*, *19*(10), 1627–1639.

Margaritis, D. (2003). *Learning Bayesian network model structure from data*: *Tech. rep.*, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.

Miller, P., Swanson, R. E., & Heckler, C. E. (1998). Contribution plots: a missing link in multivariate quality control. *Applied Mathematics and Computer Science*, *8*(4), 775–792.

Mori, J., Mahalec, V., & Yu, J. (2014). Identification of probabilistic graphical network model for root-cause diagnosis in industrial processes. *Computers & Chemical Engineering*, *71*, 171–209.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics*, *37*(1), 41–59.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, *36*(2), 220–234.

Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal*, *60*(9), 3092–3100.

Qin, S. J., Valle, S., & Piovoso, M. J. (2001). On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *15*(9), 715–742.

Raich, A., & Çinar, A. (1997). Diagnosis of process disturbances by statistical distance and angle measures. *Computers & Chemical Engineering*, *21*(6), 661–673.

Rato, T. J., Delgado, P., Martins, C., & Reis, M. S. (2020). First principles statistical process monitoring of high-dimensional industrial microelectronics assembly processes. *Processes*, *8*(11), 1520.

Rato, T. J., & Reis, M. S. (2015a). On-line process monitoring using local measures of association: Part I—Detection performance. *Chemometrics and Intelligent Laboratory Systems*, *142*, 255–264.

Rato, T. J., & Reis, M. S. (2015b). On-line process monitoring using local measures of association. Part II: Design issues and fault diagnosis. *Chemometrics and Intelligent Laboratory Systems*, *142*, 265–275.

Rato, T. J., & Reis, M. S. (2017). Markovian and non-Markovian sensitivity enhancing transformations for process monitoring. *Chemical Engineering Science*, *163*, 223–233.

Rato, T., Reis, M., Schmitt, E., Hubert, M., & De Ketelaere, B. (2016). A systematic comparison of PCA-based statistical process monitoring methods for high-dimensional, time-dependent processes. *AIChE Journal*, *62*(5), 1478–1493.

Reis, M. (2019). Multiscale and multi-granularity process analytics: A review. *Processes*, *7*, 1–21.

Reis, M. S., & Gins, G. (2017). Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis. *Processes*, *5*(3), 35.

Reis, M. S., Gins, G., & Rato, T. J. (2019). Incorporation of process-specific structure in statistical process monitoring: A review. *Journal of Quality Technology*, *51*(4), 407–421.

Reis, M. S., Rendall, R., Rato, T. J., Martins, C., & Delgado, P. (2021). Improving the sensitivity of statistical process monitoring of manifolds embedded in high-dimensional spaces: The truncated-Q statistic. *Chemometrics and Intelligent Laboratory Systems*, *215*, Article 104369.

Reis, M. S., & Saraiva, P. M. (2006). Heteroscedastic latent variable modelling with applications to multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems*, *80*(1), 57–66.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Samanta, B., Al-Balushi, K., & Al-Araimi, S. (2003). Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection. *Engineering Applications of Artificial Intelligence*, *16*(7–8), 657–665.

Scutari, M. (2009). Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

Siegel, J. E., Pratt, S., Sun, Y., & Sarma, S. E. (2018). Real-time deep neural networks for internet-enabled arc-fault detection. *Engineering Applications of Artificial Intelligence*, *74*, 35–42.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis, Vol. 26*. CRC Press.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press.

Sun, W., Paiva, A. R., Xu, P., Sundaram, A., & Braatz, R. D. (2020). Fault detection and identification using Bayesian recurrent neural networks. *Computers & Chemical Engineering*, *141*, Article 106991.

Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., et al. (2003). Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, *19*(suppl_2), 227–236.

Teyssier, M., & Koller, D. (2012). Ordering-based search: A simple and effective algorithm for learning Bayesian networks. arXiv preprint arXiv:1207.1429.

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, *5*(3), 246–255.

Vedam, H., & Venkatasubramanian, V. (1999). PCA-SDG based process monitoring and fault diagnosis. *Control Engineering Practice*, *7*(7), 903–917.

Venkatasubramanian, V., & Chan, K. (1989). A neural network methodology for process fault diagnosis. *AIChE Journal*, *35*(12), 1993–2002.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, *27*(3), 293–311.

Verma, T., & Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Uncertainty in artificial intelligence* (pp. 323–330). Elsevier.

Verron, S., Tiplica, T., & Kobi, A. (2010). Fault diagnosis of industrial systems by conditional Gaussian network including a distance rejection criterion. *Engineering Applications of Artificial Intelligence*, *23*(7), 1229–1235.

Westerhuis, J. A., Gurden, S. P., & Smilde, A. K. (2000). Generalized contribution plots in multivariate statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems*, *51*(1), 95–114.

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574.

Wise, B. M., Gallagher, N. B., Butler, S. W., White, D. D., & Barna, G. G. (1999). A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process. *Journal of Chemometrics*, *13*(3–4), 379–396.

Wu, H., & Zhao, J. (2018). Deep convolutional neural network model based chemical process fault diagnosis. *Computers & Chemical Engineering*, *115*, 185–197.

Yang, W.-T., Blue, J., Roussy, A., Pinaton, J., & Reis, M. S. (2020). A physics-informed run-to-run control framework for semiconductor manufacturing. *Expert Systems with Applications*, *155*, Article 113424.

Yang, L., & Lee, J. (2012). Bayesian belief network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. *Robotics and Computer-Integrated Manufacturing*, *28*(1), 66–74.

Yang, F., Shah, S. L., & Xiao, D. (2012). Signed directed graph based modeling and its validation from process knowledge and process data. *International Journal of Applied Mathematics and Computer Science*, *22*(1), 41–53.

Zhang, Z., Jiang, T., Li, S., & Yang, Y. (2018). Automated feature learning for nonlinear process monitoring–An approach using stacked denoising autoencoder and k-nearest neighbor rule. *Journal of Process Control*, *64*, 49–61.