Article

Industrial Process Monitoring Based on Parallel Global-Local Preserving Projection with Mutual Information

Tianshu Wu 1,2, Hongpeng Yin 1,*, Zhimin Yang 3, Jie Yao 3, Yan Qin 1 and Peng Wu 1,3

1 College of Automation, Chongqing University, Chongqing 400044, China; tswu@cqu.edu.cn (T.W.); yan.qin@cqu.edu.cn (Y.Q.)
2 Cloudwalk, Chongqing 401320, China
3 Chongqing Chuanyi Automation Co., Ltd., Chongqing 401121, China
* Correspondence: yinhongpeng@cqu.edu.cn

Abstract: This paper proposes a parallel monitoring method for plant-wide processes by integrating mutual information and Bayesian inference into a global-local preserving projections (GLPP)-based multi-block framework. Unlike traditional multivariate statistic process monitoring (MSPM) methods, the proposed MI-PGLPP method transforms plant-wide monitoring into several sub-block monitoring tasks by fully taking advantage of a parallel distributed framework. First, the original datasets of the process are divided into a group of data blocks by quantifying the mutual information of process variables. The block indexes of new data are generated automatically. Second, each data block is modeled by the GLPP method. The variable information and local structure are well preserved during the whole projection. Third, Bayesian inference is introduced to generate final statistics of the process by the probability framework. To illustrate the algorithm performance, a detailed case study is performed on the Tennessee Eastman process. Compared with the principle component analysis and GLPP-based method, the proposed MI-PGLPP provides higher FDRs and superior performance for plant-wide process monitoring.

Keywords: global-local preserving projection; mutual information; Bayesian inference; plant-wide process monitoring

1. Introduction

With the rapid development of information technology, intelligent sensors are widely applied to industrial manufacturing processes. For this reason, advanced monitoring systems such as SCADA and IIoT are adopted to retrieve real-time data. Process monitoring has become an emerging concern for researchers and industrial engineers [1,2]. Statistical methods have been intensively researched and applied to multivariate statistical process monitoring (MSPM) [3]. The most common practice is to extract process features with dimension reduction methods, then construct a latent variable model for dynamic industrial process monitoring [4]. These methods are based on partial least-squares (PLS) [5], PCA [6], canonical correlation analysis (CCA) [7], Fisher discriminant analysis (FDA) [8], independent component analysis (ICA) [9] and subspace-aided approach [10] etc. These methods and their variants are used to monitor linear, nonlinear, time-variant and multimode processes. To further investigate the performance, Shen et al. presented a review [11] to compare the performance of the aforementioned methods on the same industrial process.

Most of the traditional methods focus on solving the MSPM problem by generating latent variables from either global information or local structure, i.e., PCA-based methods use the variance of the dataset to generate projection directions. Meanwhile, a trend of preserving both global and local structures was observed in recent works [12]. The global-local structure analysis (GLSA) [13] model is first applied to monitoring plant-wide processes, which shows promising application prospects. On the basis of the GLSA model, Luo L et al. presented GLPP [4] for industrial process monitoring and kernel GLPP
(KGLPP) [14] for nonlinear process monitoring. Their research further analyzes the intrinsic relationship between PCA and LPP, revealing its performance theoretically. Similarly, the FDA is introduced into GLPP in [3] to enhance the fault-identification performance by maximizing scattering between classes and minimizing scattering within classes. Huang et al. [15] proposed the KGLPP-Cam model with a new weighted distance named Cam weighted distance. This approach is used to reselect the neighbors and consequently overcome the non-uniform distributed characteristic of data. For the dynamic process monitoring problems, Tang et al. [16] proposed a hybrid framework by integrating canonical variate analysis into GLPP.

Recently, plant-wide process monitoring is becoming a hot issue and has attracted a lot of researchers’ interests. The multi-block methods are presented to generate sub-blocks of the original dataset, which contributes to significant performance improvement. In general, the multi-block methods consist of three major steps:

- Block division. The industrial process data are divided into several blocks by certain strategies, i.e., mutual information [17] or knowledge-based methods [18].
- Block modeling. Each block is modeled by related methods, i.e., PLS, PCA, ICA, etc.
- Statistics fusion. For an online dataset, several groups of statistics are generated by a well-trained monitoring model. In order to obtain consistent monitoring results, a flexible fusion strategy is required to generate the final statistic pair. According to the literature review, Bayesian inference and voting methods are the common strategies for plant-wide process monitoring.

Based on these steps, researchers have presented several application results. A distributed PCA (DPCA) method [19] is presented, which evolves into the distributed framework presented in paper [20]. This idea is also extended to big data applications [21] and quality monitoring [1]. For the purpose of utilizing process knowledge, a hierarchical hybrid DPCA framework is presented by introducing a knowledge-based strategy and mutual information [18]. For non-Gaussian industrial process monitoring, a multi-block ICA [22] is developed and applied to industrial processes. These data-driven methods have the capability to automatically divide the process into several sub-blocks. However, they are based on covariance and are subject to linear constraints. Process decomposition, as a crucial step in multi-block methods, holds significant importance for practical nonlinear fault detection and diagnosis.

Furthermore, a significant portion of research has been focused on the application of Bayesian methods for industrial process monitoring. For instance, Huang and Yan [23] proposed a dynamic process monitoring method based on DPCA, DICA and Bayesian inference. Jiang et al. [24] established a Bayesian diagnostic system based on optimal principal components (OPCs), enabling probabilistic fault identification. Zou et al. [25] employed Bayesian inference criteria for online performance assessment. Bayesian networks are commonly employed in industrial process monitoring to model dependencies and interactions among process variables, sensors and equipment [26]. Bayesian networks provide a systematic approach to handle uncertainty, identify causal relationships and propagate the effects of changes or faults throughout the entire system [27]. For most modern industrial processes, comprehending and evaluating multiple monitoring outcomes from different units or components using Bayesian techniques holds significant importance. However, Bayesian methods necessitate sufficient data for effective inference and updating, and their successful application relies on appropriate model selection and parameter configuration.

Inspired by the distributed PCA framework [20] and mutual information-based block division [17], a mutual information-based parallel GLPP method (MI-PGLPP) is proposed for plant-wide process monitoring in this paper. With the integration of mutual information, GLPP and Bayesian fusion in a multi-block framework, the advantages of MI-PGLPP are as follows:

- MI-PGLPP utilizes mutual information of the variables and divides data blocks automatically, which does not require prior knowledge of the process.
• MI-PGLPP naturally meets the independent condition of Bayesian inference since variables in each block are divided by the independence of mutual information.
• MI-DGLPP utilizes GLPP to obtain the latent matrix and transformation matrix of each data block. The intrinsic features of global and local structures are well preserved during the projection.

This paper consists of five sections. The rest is organized as follows: Section 2 briefly reviews the fundamental theory of the GLPP method and mutual information in MI-PGLPP. Section 3 presents the detail of MI-PGLPP for process monitoring. Section 4 gives a case study to evaluate the performance of the proposed method. Finally, Section 5 concludes the paper.

2. Preliminaries

This section presents a brief theoretical review of the GLPP method and mutual information. These two methods are integrated into the proposed MI-PGLPP method. The detailed information can be seen in the following parts.

2.1. Global-Local Preserving Projection

Benefiting from the dimensionality-reduction and feature-extraction capabilities of PCA and LPP, these two methods are often used in industrial process inspection to project original datasets into a low-dimensional space. During the data projection procedure, PCA preserves the variance information while LPP keeps the local neighborhood structure information [4]. Therefore, these two methods and their variants may lead to information loss when performing data projection. Unlike these methods, the GLSA-based [13] method aims to generate the spatial correlations of original datasets by preserving the variance and local structure information [3]. For a given dataset $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m \times n}$ with $n$ samples and $m$ variables, GLPP formulates a projection matrix $A \in \mathbb{R}^{m \times k}$ and projects $X$ into lower-dimensional data $Y = [y_1, y_2, \ldots, y_k] \in \mathbb{R}^{k \times n} (k \ll m)$,

$$y_i = A^T x_i$$

$Y = A^T X$ (1)

In order to retain the global and adjacent structures of dataset $X$, the objective function is formulated as follows [4,13]

$$J_{GLPP}(a) = \min_a \{ J_{Local}(a), J_{Global}(a) \}$$

$$J_{Local}(a) = \frac{1}{2} \sum_{ij} (y_i - y_j)^2 W_{ij}$$

$$J_{Global}(a) = -\frac{1}{2} \sum_{ij} (y_i - y_j)^2 \overline{W}_{ij}$$

where $a$ denotes the transformation vector and $y_i^T = a^T X$, adjacent matrix $W_{ij}$ and non-adjacent weighting matrix $\overline{W}_{ij}$ are defined as follows [4,13,28].

$$W_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma_1^2}}, & \text{if } x_j \in \Omega_k(x_i) \text{ or } x_i \in \Omega_k(x_j) \\ 0, & \text{otherwise} \end{cases}$$

$$\overline{W}_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{\sigma_2^2}}, & \text{if } x_j \notin \Omega_k(x_i) \text{ or } x_i \notin \Omega_k(x_j) \\ 0, & \text{otherwise} \end{cases}$$
where $\sigma_1$ and $\sigma_2$ are constant parameters and $\Omega_k(x_i)$ denotes the $k$ nearest neighbors of data point $x_i$. It is easy to find out that $\sigma_1$ and $\sigma_2$ are both in the interval $[0, 1]$. Equation (2) can be reformulated as $[3, 4]$

$$J_{GLPP}(a) = \min_a \left\{ \eta J_{Global}(a) - (1 - \eta) J_{Local}(a) \right\}$$

$$= \min_a \frac{1}{2} \begin{cases} \eta \sum_{ij} (y_i - y_j)^2 W_{ij} \\ -(1 - \eta) \sum_{ij} (y_i - y_j)^2 W_{ij} \end{cases}$$

(5)

where $\eta$ is the tradeoff between global and local information in projection.

With some basic arithmetic operations, Equation (5) can be written as,

$$J_{GLPP} = \min_a \left\{ \sum_{ij} (y_i - y_j)^2 R_{ij} \right\}$$

$$= \min_a a^T X(H - R)X^T a$$

$$= \min_a a^T XMX^T a$$

(6)

where $R_{ij} = \eta W_{ij} - (1 - \eta) W_{ij}, H$ is a diagonal matrix with $H_{ij} = \sum_j R_{ij}$ and $M = H - R$ denotes the Laplacian matrix $[4]$.

Furthermore, a common constraint is introduced to transform the above objective function into a constrained optimization problem $[15, 16],$

$$\min_a a^T XMX^T a$$

s.t. $a^T N a = 1$

(7)

where $N = \eta XHX^T + (1 - \eta) I$ and $I$ denotes identity matrix.

It is easy to transform Equation (7) into a generalized eigenvalue problem,

$$XMX^T a = \lambda XNX^T a$$

(8)

By solving Equation (8), the transformation matrix $A$ is finally obtained by the $k$-th eigenvectors $a_1, a_2, \ldots, a_k$ which correspond to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$.

2.2. Mutual Information

Mutual information is defined as a non-parametric index $[17]$ and used to quantify the mutual dependence of variables from the perspective of entropy $[29]$. For given vectors $x_1$ and $x_2$, the mutual information $MI$ is defined with the aid of the joint probability density function $[30],$

$$MI(x_1, x_2) = \int \int p_j(x_1, x_2) \log \frac{p_j(x_1, x_2)}{p_m(x_1), p_m(x_2)} dx_1 dx_2$$

(9)

where $p_m(x_1)$ and $p_m(x_2)$ denote marginal probability of $x_1$ and $x_2$, respectively. $p_j(x_1, x_2)$ represents the joint probability density function of variable $x_1$ and variable $x_2$.

By introducing Shannon entropy, Equation (9) can be clarified as follows $[31]:$

$$MI(x_1, x_2) = H_j(x_1) + H_j(x_2) - H_m(x_1, x_2)$$

(10)

where $H_j(x_1, x_2)$ is the joint entropy of $x_1$ and $x_2$, the Shannon entropy of $x$ can be defined as $[32]:$

$$H_j(x) = - \int_p p_j(x) \log p_j(x) dx$$

(11)
Similarly, the joint entropy $H_m(x_1, x_2)$ can be defined as [31]:

$$H_m(x_1, x_2) = -\int_{x_1, x_2} p_m(x_1, x_2) \log(p_m(x_1, x_2)) \, dx_1 \, dx_2$$  \hspace{1cm} (12)

Since mutual information describes the dependence of variables in high-dimensional space, it can be used in both linear and nonlinear analysis. Furthermore, it is obvious that the stronger relevance of two variables leads to larger mutual information.

3. Fault Diagnosis Based on MI-PGLPP

This section explains the detail of the proposed MI-PGLPP method. Its framework is shown in Figure 1, which consists of three major steps.

- Step 1: Block selection. The original dataset is divided into several subblocks.
- Step 2: Offline model training. The block model is obtained by the subblocks of the dataset.
- Step 3: Online monitoring. The statistics of each subblock is generated by the online data and the final statistics pair is calculated by fusion strategies.

Figure 1. The flowchart of MI-PGLPP.

3.1. Mutual Information-Based Variable Block Division

A large-scale process such as the plant-wide process takes a number of resources to train a universal monitoring model. However, the model does not always perform as well as the initial purpose. Thus, some distributed methods are proposed aiming at a plant-wide monitoring problem [20,21,33]. These methods divide the dataset into several blocks via physical structure analysis, empirical analysis and quantitative calculation. Mutual information is a typical technique of quantitative methods and used in various methods to deal with variable selection problems.

As in papers [17,29], mutual information is introduced to divide the original process data into several data blocks. In each block, the variables have stronger dependence, contributing to fewer model efforts and better data features. For dataset $X \in \mathbb{R}^{m \times n}$, the mutual information $MI_{x_i, x_j}$ is calculated as Equation (10),

$$MI_{x_i, x_j} = MI(x_i, x_j)$$  \hspace{1cm} (13)

where $i, j \in 1, 2, \ldots, m$. A numerical example for MI calculation is shown in Figure 2.
The division principle depends on the MI value. For example if $MI_{x_i,x_j} > MI_{x_i,x_k}$, variable $x_j$ will be divided into the block with variable $x_i$. The dataset $X$ can be divided into the following blocks,

$$X = [X_1, X_2, \ldots, X_N]$$

where $N$ denotes the block number, for $i \in [1,2,\ldots,N]$, $X_i \in \mathbb{R}^{m_i \times n}$ and $m_i$ denotes dimension of data block $X_i$.

This step aims to partition the original dataset of the process into a set of data blocks by quantifying the mutual information between process variables. It automatically generates block indices for the new data. Based on the calculation principles of mutual information, a higher MI value indicates stronger dependency between variables, which contributes to reduced modeling effort and improved data characterization.

3.2. GLPP-Based Block Data Modeling

For the $i$-th data block $X_i^T \in \mathbb{R}^{n \times m_i}$, the offline normal operation model can be trained by GLPP

$$X_i = Y_i A_i^T + E_i$$

$$Y_i = X_i A_i (A_i^T A_i)^{-1}$$

where $A_i = [a_{1i}, a_{2i}, \ldots, a_{mi}] \in \mathbb{R}^{m_i \times m_{ik}}$ denotes the transform matrix, $Y_i = [y_{1i}, y_{2i}, \ldots, y_{mi}] \in \mathbb{R}^{n \times m_{ik}}$ denotes the latent variable matrix, and matrix $E_i$ represents the residual.

For process monitoring of each block, $T^2$ and SPE statistics are applied to GLPP model and their control limits are defined as follows [34],

$$T_{i,limits}^2 = \frac{m_{ik}(n^2 - 1)}{n(n - m_{ik})} \sim F_a(m_{ik}, n - m_{ik})$$

$$SPE_{i,limits} = \theta_1 \left[ 1 - \theta_2^2 h_0  \frac{(1 - h_0)}{\theta_1^2} + \frac{z_a(2\theta_2^2 h_0^2)}{\theta_1} \right]^{\frac{1}{2}}$$

where $m_{ik}$ is the dimension of latent variables in block $i$, $\sim F_a(m_{ik}, n - m_{ik})$ denotes $\sim F$-distribution, $\alpha$ is the significance level. $h_0 = 1 - [(2\theta_1\theta_3)/(3\theta_2)]$, $\theta_2 = tr(cov(E))^2$, $\theta_3 = tr(cov(E_i))^3$ and $z_a$ denotes the standardized normal variable with confidence $1 - \alpha$.

In each block, the monitoring performance depends on the selection of parameter $\eta$. In order to meet the requirements of tradeoff between global information and local information in the latent matrix, several feasible principles have been applied according to
the dataset. For industrial time series analysis, a spectral radius based approach [4,16] is presented to select parameter $\eta$,

$$\eta = \frac{\delta(L)}{\delta(L) + \delta(\bar{L})}$$ (17)

where $\delta(L)$ and $\delta(\bar{L})$ denote the spectral radius of local and global Laplacian matrices, respectively.

For block $i$, the new data sample $x_{i}^{\text{new}} \in \mathbb{R}^{1 \times m_i}$ is projected onto the $i$-th well-trained offline model,

$$x_{i}^{\text{new}} = A_{i}y_{i}^{\text{new}} + e_{i,\text{new}}$$
$$y_{i}^{\text{new}} = (A_{i}^{T}A_{i})^{-1}A_{i}^{T}x_{i}^{\text{new}}$$
$$e_{i,\text{new}} = x_{i}^{\text{new}} - A_{i}y_{i}^{\text{new}}$$ (18)

The $T_{i,\text{new}}^2$ and $SPE_{i,\text{new}}$ statistics of the $i$-th block are calculated as follows,

$$T_{i,\text{new}}^2 = y_{i}^{\text{new}}(Y_{i}^{T}Y_{i}/n - 1)^{-1}y_{i}^{\text{new}}$$
$$SPE_{i,\text{new}} = e_{i,\text{new}}^{T}e_{i,\text{new}}$$ (19)

where $Y_{i}$ denotes the $i$-th latent variable matrix.

For each block, $\{T_{i,\text{new}}^2, SPE_{i,\text{new}}\}_{i=1,2,...,N}$ are calculated online with new monitoring data. The monitoring results of each block are simply generated by comparing statistics of new sample data with their control limits $\{T_{i,\text{limits}}^2, SPE_{i,\text{limits}}\}_{i=1,2,...,N}$. For example, if $T_{i,\text{new}}^2 > T_{i,\text{limits}}^2$, a process fault is triggered by the $T^2$ statistics of the $i$-th block model.

Since there are several blocks, the final monitoring results can be further calculated by some fusion strategies. According to the literature review, the probabilistic framework is the common solution to this problem. The following section introduces a Bayesian inference strategy to generate final monitoring results.

### 3.3. Bayesian Inference-Based Monitoring Result Fusion

Bayesian inference is prevalent in distributed monitoring such as distributed PCA methods [35], distributed ICA [36] and so on. This section presents Bayesian inference based monitoring result fusion.

Firstly, the normal operation and faulty case are tagged with $\mathcal{N}$ and $\mathcal{F}$, respectively. $p_{T^2}(\mathcal{N})$ and $p_{T^2}(\mathcal{F})$, $p_{SPE}(\mathcal{N})$ and $p_{SPE}(\mathcal{F})$ are the prior probabilities of $T^2$ and $SPE$ statistics of the $i$-th block corresponding to operation modes $\mathcal{N}$ and $\mathcal{F}$, which can be calculated by significance level $\alpha$ and $1 - \alpha$. The probabilities of $T^2$ and $SPE$ statistics are determined by

$$p_{T^2}(x_{i}^{\text{new}} | \mathcal{N}) = \exp - \frac{T_{i,\text{new}}^2}{T_{i,\text{limits}}^2}$$
$$p_{T^2}(x_{i}^{\text{new}} | \mathcal{F}) = \exp - \frac{T_{i,\text{limits}}^2}{T_{i,\text{new}}^2}$$

$$p_{SPE}(x_{i}^{\text{new}} | \mathcal{N}) = \exp - \frac{SPE_{i,\text{new}}}{SPE_{i,\text{limits}}}$$
$$p_{SPE}(x_{i}^{\text{new}} | \mathcal{F}) = \exp - \frac{SPE_{i,\text{limits}}}{SPE_{i,\text{new}}}$$ (21)
Based on Equations (20) and (21) and prior probabilities, the fault probabilities of data point $x_i^{\text{new}}$ can be calculated as,

$$p_{T^2}^i(F | x_i^{\text{new}}) = \frac{p_{T^2}^i(x_i^{\text{new}} | F)p_{T^2}^i(F)}{p_{T^2}^i(x_i^{\text{new}} | N)p_{T^2}^i(N) + p_{T^2}^i(x_i^{\text{new}} | F)p_{T^2}^i(F)}$$  

$$p_{SPE}^i(F | x_i^{\text{new}}) = \frac{p_{SPE}^i(x_i^{\text{new}} | F)p_{SPE}^i(F)}{p_{SPE}^i(x_i^{\text{new}} | N)p_{SPE}^i(N) + p_{SPE}^i(x_i^{\text{new}} | F)p_{SPE}^i(F)}$$  

(22)

where $p_{T^2}^i(x_i^{\text{new}})$ and $p_{SPE}^i(x_i^{\text{new}})$ are calculated as

$$p_{T^2}^i(x_i^{\text{new}}) = p_{T^2}^i(x_i^{\text{new}} | N)p_{T^2}^i(N) + p_{T^2}^i(x_i^{\text{new}} | F)p_{T^2}^i(F)$$  

$$p_{SPE}^i(x_i^{\text{new}}) = p_{SPE}^i(x_i^{\text{new}} | N)p_{SPE}^i(N) + p_{SPE}^i(x_i^{\text{new}} | F)p_{SPE}^i(F)$$  

(23)

With Equation (22), fault probabilities of each block can be calculated corresponding to $T^2$ and $SPE$ statistics. The probabilities set $\{p_{T^2}^i(F | x_i^{\text{new}})\}_{i=1,2,...,N}$ and $\{p_{SPE}^i(F | x_i^{\text{new}})\}_{i=1,2,...,N}$ are further integrated by a weighted probability approach [20],

$$p_{T^2}(F | x^{\text{new}}) = \sum_{i=1}^{N} \left\{ \frac{p_{T^2}^i(F | x_i^{\text{new}})p_{T^2}^i(x_i^{\text{new}} | F)}{\sum_{i=1}^{N} p_{T^2}^i(F | x_i^{\text{new}})} \right\}$$  

$$p_{SPE}(F | x^{\text{new}}) = \sum_{i=1}^{N} \left\{ \frac{p_{SPE}^i(F | x_i^{\text{new}})p_{SPE}^i(x_i^{\text{new}} | F)}{\sum_{i=1}^{N} p_{SPE}^i(F | x_i^{\text{new}})} \right\}$$  

(24)

In a Bayesian inference framework, the assumption of independence is essential and requires lower correlation between variables $x_i$ and $x_j$. Because sub-blocks are divided by mutual information with low dependence, variables in each sub-block are relatively independent of each other. As a result, fault-monitoring performance is guaranteed to the maximum extent.

Based on the final statistics $p_{T^2}(F | x^{\text{new}})$ and $p_{SPE}(F | x^{\text{new}})$, the monitoring result is determined by the significance level $\alpha$ as

$$NO: p_{T^2}(F | x^{\text{new}}) \leq \alpha \& p_{SPE}(F | x^{\text{new}}) \leq \alpha$$  

$$FO: p_{T^2}(F | x^{\text{new}}) > \alpha \text{ or } p_{SPE}(F | x^{\text{new}}) > \alpha$$  

(25)

where $NO$ and $FO$ denote normal operation and faulty operation, respectively.

### 3.4. Monitoring Procedure

The proposed MI-PGLPP based process monitoring can be applied as the procedures in Figure 3.
Figure 3. Procedures of the proposed method.

Step 1: Steps of offline modeling:
1. Normalize each data block by Z-score standardization and generate data mean and variance.
2. Analyze the variable dependence of training dataset by Equation (10), and then generate the data blocks and block index.
3. Construct adjacent weighting matrix \( W_{ij} \) and nonadjacent weighting matrix \( \bar{W}_{ij} \) by Equation (4).
4. Calculate the tradeoff parameter \( \eta \) by Equation (17) for each data block.
5. Perform GLPP on each data block \( X_i \) by Equation (7) and generate transformation matrices \( A_i \), latent matrices \( Y_i \) and \( NO \) model.
6. Define significance level \( \alpha \) and calculate control limits for each data block by Equation (16).

Step 2: Steps of online monitoring:
1. Acquire the new sample data \( x_{\text{new}} \) and perform normalization with the mean and variance of training samples.
2. Divide new data \( x_{\text{new}} \) into \( N \) blocks with the variable index generated by the training data.
3. Project each block data point on the \( NO \) model and generate residual \( e_{i,\text{new}} \) of each block by Equation (18).
4. Calculate the statistics of each block by Equation (19).
5. Calculate the prior probabilities, conditional probabilities and fault probabilities by Equations (20)–(23).
6. Perform statistics fusion with Bayesian inference by Equation (24).
7. Monitor the process by Equation (25).

4. Experiments and Results

In this section, the Tennessee Eastman (TE) process [37] is adopted to evaluate the performance of the proposed method. The TE process is a benchmark for data-driven
fault diagnosis strategies and is widely applied to illustrating process monitoring performance [11]. It consists of five major process unit [38], including a vapor–liquid flash separator, an exothermic two-phase reactor, a reboiled product stripper, a recycling compressor and a product condenser (mixer) [3]. As shown in Figure 4, the TE process produces two products and a byproduct from four gaseous materials [39].

The Tennessee Eastman process is simulated using mathematical models that capture the behavior of a chemical plant. The simulation involves solving equations based on principles of physics and chemistry to calculate the evolution of process variables over time. The simulation replicates the dynamics of a real chemical plant, considering factors such as reaction kinetics, heat transfer and fluid flow. It allows researchers to analyze the process behavior, evaluate control strategies and test techniques before implementing them in real-world applications.

The TE process has 53 monitored variables, including 22 process variables, 12 manipulated variables and 19 component measurements. In this study, 11 manipulated variables and 22 continuously monitored variables are selected as experimental datasets. A total of 21 faults were obtained under 21 different types of disturbances.

Faults in the TE process are classified into five categories: group 1 is step changes corresponding to faults 1–7, group 2 is random changes corresponding to faults 8–12, group 3 is a slow shift of reaction kinetics corresponding to fault 13, group 4 is unknown faults corresponding to faults 16–20 and the last group is valve sticking, corresponding to faults 14, 15 and 21. More information on faults in the TE process is presented in paper [3,17,40]. The training datasets are acquired under normal operating conditions, which consist of 960 samples, and all variables are collected every three minutes. Then introducing each fault into process at sample 160 to collect the fault datasets.

Figure 4. Flowchart of TE process [37].

The normal operation process is utilized as modeling input, and 21 faults are introduced to verify the monitoring performance. For further illustration, the feasibility of MI-PGLPP, PCA and GLPP-based methods are introduced to monitor the TE process. The significance level of three methods is set as $\alpha = 0.05$, and the principle contribution rate is set as $\sigma_{PCA} = 0.85$. For the GLPP method, the parameters are referred to paper [3] and selected as neighborhood number $k = 10$, local weight $\sigma_1 = 48$, global weight $\sigma_2 = 132$. The proposed MI-PGLPP also utilizes these parameter settings. Furthermore, the neighborhood number is set as $k = 2$ due to the group characteristics. The mutual-information-based
multiblock principal component analysis (MI–MBPCA) [17] and the global DISSIM (GDIS-SIM) model [41] are also constructed for comparison. During the construction, the window length is set to be 30 by trial and error. A total of 371 moving windows are generated in each test dataset and the fault is introduced from the 172th the moving window. The confidence level is defined as 0.99.

The fault detection rate (FDR), which is widely used for monitoring results evaluation, was also introduced in this study for performance comparison. Since there are two statistical results of the experiment, this study uses a voting strategy to determine the FDR.

\[
FDR = \max \{ FDR_{T^2}, FDR_{SPE} \}
\]

where \( FDR \) of certain statistics is defined as in Ref. [11],

\[
FDR_{\text{statis}} = \frac{\text{Count}( \{ |J| > J_{th} \mid FO \} )}{\text{Count}(FO)}
\]

In order to construct a multi-block model, the mutual information between each pair of variables was computed. In all, 33 continuous variables are used for process monitoring, represented as \( x = [x_1, x_2, \cdots, x_{33}] \), which was partitioned into 7 sub-blocks, consisting of 6 typical blocks and 1 unknown block. The variables within each sub-block are listed in Table 1.

According to FDRs in Table 2, the proposed method performs much better for 10 faults. They are fault 1, fault 2, fault 6, fault 7, fault 8, fault 12, fault 13, fault 14, fault 17 and fault 18. It is capable of handling monitor step, slow drift, random variation and unknown fault. The result shows that the proposed MI-PGLPP can cover all types of fault in the TE process. It is also seen that the results seem bad only for Fault 4. The reason may be that the monitoring variables in the Fault 4 are related-fault, which makes it difficult for the original data to divide the block through the mutual information in the proposed method. Moreover, Fault 4 is characterized by unique patterns that are more challenging to detect and classify compared to other faults. The \( T^2 \) and \( SPE \) statistics can generate alarms for most faults around sample 160 when faults are introduced. Figures 5 and 6 show the performance of MI-DLPP for monitoring fault 1, fault 2 and fault 7. All faults are introduced at sample 160. Both \( T^2 \) and \( SPE \) statistics can alarm fault occurrence after about four samples.

Table 1. Block Division Result.

<table>
<thead>
<tr>
<th>Block Number</th>
<th>Variables</th>
<th>Block Number</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{11}, x_{14}, x_{21}, x_{22}, x_{23}, x_{24}, x_{26}, x_{32} )</td>
<td>5</td>
<td>( x_1, x_{25} )</td>
</tr>
<tr>
<td>2</td>
<td>( x_7, x_{13}, x_{16}, x_{20}, x_{27} )</td>
<td>6</td>
<td>( x_{12}, x_{29} )</td>
</tr>
<tr>
<td>3</td>
<td>( x_{10}, x_{17}, x_{28}, x_{33} )</td>
<td>7</td>
<td>( x_{15}, x_{30} )</td>
</tr>
<tr>
<td>4</td>
<td>( x_{18}, x_{19}, x_{31} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of fault detection rates (%).

<table>
<thead>
<tr>
<th>Fault No.</th>
<th>PCA</th>
<th>GLPP</th>
<th>MI–MBPCA</th>
<th>GDISSIM</th>
<th>MI–PGLPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97.88</td>
<td>99.25</td>
<td>65</td>
<td>99.15</td>
<td>99.75</td>
</tr>
<tr>
<td>2</td>
<td>96.5</td>
<td>98.25</td>
<td>29.74</td>
<td>98.43</td>
<td>98.75</td>
</tr>
<tr>
<td>3</td>
<td>2.63</td>
<td>7.64</td>
<td>4.28</td>
<td>5.75</td>
<td>11.88</td>
</tr>
<tr>
<td>4</td>
<td>20.88</td>
<td>81.75</td>
<td>39.33</td>
<td>20.94</td>
<td>7.88</td>
</tr>
<tr>
<td>5</td>
<td>24.13</td>
<td>43.23</td>
<td>44.83</td>
<td>24.15</td>
<td>31</td>
</tr>
</tbody>
</table>
Table 2. Cont.

<table>
<thead>
<tr>
<th>Fault No.</th>
<th>PCA</th>
<th>GLPP</th>
<th>MI-MBPCA</th>
<th>GDISSIM</th>
<th>MI-PGLPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>99.13</td>
<td>100</td>
<td>93.45</td>
<td>99.19</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>99.75</td>
<td>99.37</td>
<td>58.45</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>96.88</td>
<td>96.63</td>
<td>92.83</td>
<td>96.97</td>
<td>98.25</td>
</tr>
<tr>
<td>9</td>
<td>1.75</td>
<td>7.5</td>
<td>7.95</td>
<td>3.25</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>29.63</td>
<td>44.88</td>
<td>82.75</td>
<td>29.64</td>
<td>53.13</td>
</tr>
<tr>
<td>11</td>
<td>74.88</td>
<td>67.25</td>
<td>94.13</td>
<td>40.67</td>
<td>52.88</td>
</tr>
<tr>
<td>12</td>
<td>96.38</td>
<td>98.63</td>
<td>98.83</td>
<td>98.45</td>
<td>99.50</td>
</tr>
<tr>
<td>13</td>
<td>93.63</td>
<td>94.5</td>
<td>90</td>
<td>93.62</td>
<td>95.13</td>
</tr>
<tr>
<td>14</td>
<td>99.25</td>
<td>99.62</td>
<td>57.75</td>
<td>99.38</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>12.25</td>
<td>13.75</td>
<td>10.64</td>
<td>14.38</td>
</tr>
<tr>
<td>16</td>
<td>27.38</td>
<td>27</td>
<td>96.32</td>
<td>13.58</td>
<td>40</td>
</tr>
<tr>
<td>17</td>
<td>76.25</td>
<td>90.5</td>
<td>93.35</td>
<td>76.37</td>
<td>95.38</td>
</tr>
<tr>
<td>18</td>
<td>90.13</td>
<td>88.75</td>
<td>86.33</td>
<td>89.35</td>
<td>91</td>
</tr>
<tr>
<td>19</td>
<td>12.5</td>
<td>25.5</td>
<td>37.6</td>
<td>11</td>
<td>42</td>
</tr>
<tr>
<td>20</td>
<td>49.75</td>
<td>50.5</td>
<td>84.45</td>
<td>31.83</td>
<td>55.75</td>
</tr>
<tr>
<td>21</td>
<td>47.25</td>
<td>51.63</td>
<td>42.32</td>
<td>39.35</td>
<td>50.75</td>
</tr>
<tr>
<td>Average</td>
<td>59.03</td>
<td>61.22</td>
<td>62.30</td>
<td>56.27</td>
<td>64.21</td>
</tr>
</tbody>
</table>

Figure 5. SPE monitoring results of the proposed method: (a) Fault 1; (b) Fault 2; (c) Fault 7 (Red dashed line indicates the control limits, and the blue solid line is the online monitoring statistics).
Figure 6. $T^2$ monitoring results of the proposed method: (a) Fault 1; (b) Fault 2; (c) Fault 7 (Red dashed line indicates the control limits, and the blue solid line is the online monitoring statistics).

For the purpose of further exploring the performance of MI-PGLPP, more experimental results are presented below, in Figures 7–10. Figures 7 and 8 are the monitoring results of fault 8 whereas Figures 9 and 10 are the detection performance of Fault 12.

Figure 7. $T^2$ monitoring results of Fault 8: (a) PCA; (b) GLPP; (c) MI-PGLPP (Red dashed line indicates the control limits, and the blue solid line is the online monitoring statistics).
Figure 8. $SPE$ monitoring results of Fault 8: (a) PCA; (b) GLPP; (c) MI-PGLPP (Red dashed line indicates the control limits, and the blue solid line is the online monitoring statistics).

Figure 9. $T^2$ monitoring results of Fault 12: (a) PCA; (b) GLPP; (c) MI-PGLPP (Red dashed line indicates the control limits, and the blue solid line is the online monitoring statistics).
Therefore, a higher FDR implies better monitoring performance. The FDRs of three methods are shown in Table 2, and bold numbers indicate optimum performance with certain failures. It is not difficult to see that MI-PGLPP contributes most of the highest FDR even for fault 3, fault 9 and fault 15, which are generally recognized as the most difficult monitoring faults. Moreover, the average FDR of MI-PGLPP is also higher than those of the other two methods.

All methods show acceptable performance on Fault 8, while MI-PGLPP outperforms the other two methods on Fault 12 in this study. These results fully demonstrate the effectiveness of MI-PGLPP for the TE process, which is recognized as a typical benchmark of plant-wide process monitoring applications.

5. Conclusions

In this paper, the MI-PGLPP method is proposed and applied to monitoring plant-wide processes. Different from PCA and GLPP, MI-PGLPP adopts a parallel distributed framework. To uncover the nonlinear characteristics between data blocks, the proposed method combines mutual information with classical GLPP, partitioning high-dimensional datasets into blocks and projecting the local structure and global information of each block into a latent variable space. For process monitoring applications, the new dataset is projected onto the transformation matrices to generate latent matrices and residuals. Leveraging the fully explored data, Bayesian inference is employed to integrate the statistical information of each data block. Ultimately, experimental monitoring of the TE process demonstrates that MI-PGLPP outperforms the other methods in terms of FDR. Further research is needed to reveal the fundamental reasons behind the performance degradation of MI-PGLPP under certain faults, which could contribute to improving the monitoring outcomes for such faults in the future.

Author Contributions: Conceptualization, T.W.; methodology, T.W.; software, T.W.; validation, Z.Y., J.Y. and P.W.; writing—original draft preparation, T.W.; writing—review and editing, Y.Q.; supervision, H.Y. All authors have read and agreed to the published version of the manuscript.
Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in [37].

Conflicts of Interest: The authors declare no conflict of interest.

References


3. Tang, Q.; Chai, Y.; Qu, J.; Fang, X. Industrial process monitoring based on Fisher discriminant global-local preserving projection. J. Process Control 2019, 81, 76–86. [CrossRef]


12. Huang, J.; Yan, X. Related and independent variable fault detection based on KPCA and SVDD. J. Process Control 2016, 38, 99–110. [CrossRef]


15. Huang, C.; Chai, Y.; Liu, B.; Tang, Q.; Qi, F. Industrial process fault detection based on KGLPP model with Cam weighted distance. J. Process Control 2021, 106, 110–121. [CrossRef]


33. Qin, Y.; Arunan, A.; Yuen, C. Digital twin for real-time Li-ion battery state of health estimation with partially discharged cycling data. *IEEE Trans. Ind. Inform.* 2023, *early access*. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.