Bayesian statistical monitoring of complex semiconductor manufacturing batch processes

Zhiqiang Ge\(^1\), Junhua Zheng\(^2\), Zhihuan Song\(^1\), Huaping Dai\(^1\)

\(^1\)State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, Zhejiang, China
\(^2\)Zhejiang Titan Design & Engineering Co., LTD, Hangzhou, Zhejiang, China

e-mail: zgge@iipc.zju.edu.cn (Ge Z), zhsong@iipc.zju.edu.cn (Song Z)

**Abstract:** Most current batch process monitoring methods have been implemented under a single operation mode, thus the batch uncertainty is assumed to be caused by batch-to-batch variations. However, due to the change of market requirements, the batch operation mode should also be changed frequently, especially in the semiconductor manufacturing process. This paper proposes an efficient method for monitoring those batch processes. First, the data is partitioned into multiple clusters, which correspond to different operation modes. Second, a sub-statistical model is built for each operation mode. Then the Bayesian inference strategy is introduced for result combination in different operation modes. The monitoring performance of the proposed method is evaluated by a real semiconductor process application case study.

### 1. INTRODUCTION

Batch and semi-batch processes play more and more important roles in the production of low volume, high-value products such as polymers, pharmaceuticals, and biological products [Nomikos, and MacGregor, 1994]. In contrast to the continuous process, some typical characteristics of the batch process complicate the implementation of many on-line monitoring techniques that have been successfully applied in continuous processes. These characteristics include finite duration, non-steady state behavior, batch-to-batch variations, and etc [Nomikos, and MacGregor, 1995]. By extending the conventional principal component analysis (PCA) method to the batch process, multiway PCA (MPCA) was proposed and successfully applied for batch process monitoring [Chen, and Chen, 2008; Zhao et. al., 2008; Yao, and Gao, 2008; Choi, et. al., 2008; Chen, and Zhang, 2010]. To our best knowledge, most of the developed technologies have assumed that the batch process ran under a single operation mode, process uncertainties are only caused by batch-to-batch variations.

However, due to the quick changes of modern market demands, different kinds of products should be produced, and thus frequent changes of process operation conditions are needed to meet this requirement. Along last decades, the semiconductor industry has maintained a high growth every year. While the production cost was reduced in the past by several methods, it has recently been recognized that the product quality should also be improved simultaneously. Recent efforts to meet the desired quality of semiconductor processes are mainly focused on advance process control and monitoring technologies [Wise, et. al., 1999].

The object of the present paper is a kind of semiconductor processes which run in multiple operation conditions. For those batch processes, products generated under different operation conditions always have different characteristics from each other. This paper aims to develop an efficient method for monitoring multimode batch processes. Different from existing methods, the new method first divides the whole batch dataset into several sub-datasets, which can be implemented by the data clustering method. Then a local sub-model is built for the process under each operation condition. For monitoring a new data sample, local monitoring results are generated in corresponding sub-regions of the batch process. Then a soft probabilistic strategy is proposed to combine monitoring results obtained by different sub-models, which is based on the Bayesian inference method. Simultaneously, the mode information of the monitored data sample can be easily obtained with the utilization of the Bayesian posterior probability, which can be calculated under different operation conditions. The rest of this paper is organized as follows. First, the basic PCA method is briefly described. In section 3, a detailed description of our proposed method is provided, which is followed by a real semiconductor process application case study in the next section. Finally, we make our conclusions.

### 2. PRINCIPAL COMPONENT ANALYSIS

Assuming the collected process dataset as \( \mathbf{X} \in \mathbb{R}^{n \times m} \), where \( n \) is the number of process variables, and \( m \) is the sample number for each variable. PCA is carried out upon the covariance matrix of \( \mathbf{X} \) (\( \sum = \frac{\mathbf{X}^T \cdot \mathbf{X}}{n-1} \)). After the Selection of the first \( k \) principal components, the PCA model can be built as [Nomikos, and MacGregor, 1994]

\[
\mathbf{X} = \mathbf{T}
\end{equation}

\( \mathbf{T} \in \mathbb{R}^{n \times k} \) and \( \mathbf{P} \in \mathbb{R}^{m \times (m-k)} \) are score matrices in the principal component subspace (PCS) and the residual
subspace (RS), \( \mathbf{P} \in \mathbb{R}^{m \times n} \) and \( \mathbf{P} \in \mathbb{R}^{m \times (n-1)} \) correspond to loading matrices in PCS and RS. \( \mathbf{E} \in \mathbb{R}^{m \times n} \) is the residual matrix after the PCA decomposition, which is always related to process noises. As a popular multivariate statistical method, PCA has been used for feature extraction, data compression, face recognition, process monitoring, and etc.

### 3. Proposed Bayesian Statistical Monitoring Approach

As we know, the dataset of the batch process is collected in a three-way manner \( \mathbf{X}(I \times J \times K) \), where \( I \) is the batch number, \( J \) is the variable number, and \( K \) is the total number of data samples during each batch running. In the present paper, the duration of each batch \( K \) is assumed to be constant. Traditionally, MPCA first unfolded the three-way dataset into a two-dimensional dataset, and then implemented PCA upon the two-dimensional dataset. The most two widely used unfolding methods are batch-wise unfolding and variable-wise unfolding. According to Sprang et. al. [2008], the unfolding method through the batch direction is considered to be the most appropriate method for MPCA monitoring, because it can catch most process characteristics along the batch trajectory. However, if the number of available batches for modeling is limited, the monitoring performance may be degraded. Besides, when MPCA is used for online monitoring, the future values of the monitored batch should be estimated, which may also deteriorate the monitoring performance of this method. In contrast, the unfolding method through the variable direction does not need any estimation when it is used for online monitoring. However, the trajectory characteristics may be lost by this method. Therefore, we intend to combine these two unfolding methods together, thus the process dataset is first unfolded through the batch direction for data auto-scaling, and then it is re-unfolded through the variable direction. Another advantage of this combined unfolding method is that the number of training samples becomes much more plentiful than that of the traditional method.

#### 3.1. Operation mode clustering

Before modeling of the multimode batch process, another important issue that should be considered is how to obtain local sub-datasets, which correspond to different operation conditions. A number of data clustering methods have already been developed in the data-mining area, including K-means clustering, fuzzy clustering, and etc. Besides, many multivariate statistical methods such as PCA, and fisher discriminant analysis (FDA) can also be employed for data clustering and pattern classification. In the present paper, the simple K-means clustering method is employed.

#### 3.2. Process monitoring based on Bayesian inference

After the original dataset has been partitioned into several sub-datasets, a sub-PCA model can be built for each operation mode. First, the sub-datasets divided from the original dataset can be represented as

\[
\mathbf{X}(I \times J \times K) = [\mathbf{X}_1(I_1 \times J \times K), \mathbf{X}_2(I_2 \times J \times K), \ldots, \mathbf{X}_Q(I_Q \times J \times K)] \tag{2}
\]

where \( Q \) is the number of operation modes in the process, and \( I_q(q = 1,2,\ldots,Q) \) is the batch number of the process dataset under operation mode \( q \), which obeys \( \sum_{q=1}^{Q} I_q = I \).

After the data scale step, these sub-datasets can be re-arranged as \( \mathbf{X}_q(KI_q \times J), q = 1,2,\ldots,Q \). Then, the sub-PCA model for each operation mode can be constructed as follows

\[
\mathbf{X}_q(KI_q \times J) = \mathbf{T}_q(KI_q \times J)\mathbf{P}_q^T(J \times R_q) + \mathbf{E}_q \tag{3}
\]

where \( q = 1,2,\ldots,Q \), \( R_q \) is the selected number of principal components.

For monitoring a new data sample \( \mathbf{x}_q \), the traditional \( T^2 \) and SPE statistics can be built, which are calculated as [Nomikos, and MacGregor, 1994]

\[
T^2_{q,t} = \sum_{i=1}^{K} \frac{\mathbf{T}^T_{q,t}}{\lambda_{q,i}} \tag{4}
\]

\[
SPE_{q,t} = \mathbf{e}^T_{q,t} \mathbf{e}_{q,t} \tag{5}
\]

where \( q = 1,2,\ldots,Q \), \( \lambda_{q,i} \) is the variance of the corresponding principal component in the \( q \)-th sub-PCA model, \( \mathbf{e}_{q,t} \) is the residual vector of the monitor data sample, the confidence limits of the \( T^2 \) and \( SPE \) statistics under different operation modes can be established as [Nomikos, and MacGregor, 1994]

\[
T^2_{q,lim} = \frac{R_q(KI_q - 1)}{KI_q - R_q} F_{R_q,(KI_q - R_q),\alpha} \tag{6}
\]

\[
SPE_{q,lim} = g_x x^2 \tag{7}
\]

where \( q = 1,2,\ldots,Q \), \( \alpha \) is the selected significance level, \( m_q \) and \( v_x \) are the mean and variance values of the \( SPE \) statistic within operation mode \( q \).

However, it is noted that depending on those sub-PCA models, we cannot get any mode information of the monitored data sample. Hence, which sub-PCA model should be employed for monitoring this new data sample is inexplicit. When an inappropriate model is used for monitoring, a false alarm may be triggered. The same problem will also arise when the monitored data sample is assigned to a single operation region by a hard assignment approach. This is because the new data sample may belong to different operation regions in probabilities. To address this issue, a soft assignment approach is proposed here. Thus, the new monitored data sample can be softly assigned to different operation modes with their corresponding probabilities. Therefore, although we have no mode information of the new data sample, the final monitoring result can be made through the combination of all subspace
monitoring results in different operation modes. In other words, the new method is robust to mode information of the process data. To build the new monitoring method in the probabilistic framework, the Bayesian inference strategy is used. Since each sub-PCA model has its own monitoring statistics and confidence limits, two following two transformations are introduced upon the traditional $T^2$ and $SPE$ statistics

$$T^2_q(x) \rightarrow \exp\left(-\frac{T^2_q(x)}{T_{q,\text{lim}}^2}\right)$$

(8)

$$SPE_q(x) \rightarrow \exp\left(-\frac{SPE_q(x)}{SPE_{q,\text{lim}}^2}\right)$$

(9)

where $q = 1, 2, \cdots, Q$. In the Bayesian framework, these two transferred probabilities can be considered as the probability distribution of the data sample under specific operation mode, which can represented as

$$P_{T^2}(x | q) = \exp\left(-\frac{T^2_q(x)}{T_{q,\text{lim}}^2}\right)$$

(10)

$$P_{SPE}(x | q) = \exp\left(-\frac{SPE_q(x)}{SPE_{q,\text{lim}}^2}\right)$$

(11)

Through the Bayesian inference, the posterior probabilities of the monitored sample corresponding to different operation modes are calculated as

$$P_{T^2}(q | x) = \frac{P_{T^2}(q, x)}{P_{T^2}(x)} = \frac{P_{T^2}(x | q)P(q)}{\sum_{q=1}^{Q} P_{T^2}(x | q)P(q)}$$

(12)

$$P_{SPE}(q | x) = \frac{P_{SPE}(q, x)}{P_{SPE}(x)} = \frac{P_{SPE}(x | q)P(q)}{\sum_{q=1}^{Q} P_{SPE}(x | q)P(q)}$$

(13)

where $P(q)$ is the prior probability of the $q$-th operation mode, which can be simply determined as

$$P(q) = \frac{K_I}{I} = \frac{I_q}{I}$$

(14)

After the posterior probability of the monitored data sample $x$ has been obtained, we are in the position to determine whether this sample is normal or not. Therefore, two new statistical-based indices related to each operation mode are defined as follows

$$P_{T^2,\text{f}}(x) = \Pr\{T^2_q(x_{\text{f},q}) \leq T^2_q(x)\}$$

(15)

$$P_{SPE,\text{f}}(x) = \Pr\{SPE_q(x_{\text{f},q}) \leq SPE_q(x)\}$$

(16)

where $q = 1, 2, \cdots, Q$. $x_{\text{f},q}$ is the training samples in the $q$-th operation mode. The values of these two probabilities can be calculated by integrating the $\chi^2$ probability density function with appropriate degree of freedom [Yu, and Qin, 2008]. Precisely, these two values can be considered as fault probabilities of the monitored data sample in the latent and residual planes, respectively. Therefore, these two probabilities are indications of whether the new monitored data sample is normal or not. If $P_{T^2,\text{f}}(x)$ and $P_{SPE,\text{f}}(x)$ exceed the confidence level $1 - \alpha$, some fault can be judged under the $q$-th operation region. Otherwise, this operation region is normal, and thus should be kept on monitoring.

Under the consideration that the new data sample may belong to different operation modes, two new Bayesian-based monitoring statistics can be constructed upon the calculated posterior and fault probabilities, which are presented in eqs. (12)-(13) and eqs. (15)-(16), respectively. The new Bayesian monitoring statistics are given as follows

$$BIF_{T^2}(x) = \sum_{q=1}^{Q} P_{T^2}(q | x)p_{T^2}(x)$$

(17)

$$BIF_{SPE}(x) = \sum_{q=1}^{Q} P_{SPE}(q | x)p_{SPE}(x)$$

(18)

Since the values of $P_{T^2,\text{f}}(x)$ and $P_{SPE,\text{f}}(x)$ are both ranged from zero to one, and the posterior probabilities $P_{T^2}(q | x)$ and $P_{SPE}(q | x)$ are restricted to $\sum_{q=1}^{Q} P_{T^2}(q | x) = 1$ and $\sum_{q=1}^{Q} P_{SPE}(q | x) = 1$, thus, the values of both two Bayesian monitoring statistics should also be ranged from zero to one. Hence, under a pre-specified significance level $\alpha$, the new monitored data sample $x$ is determined to be normal if both of the two statistic values are not larger than $1 - \alpha$. Otherwise, this data sample should be treated as abnormal, thus some fault is considered to be detected.

4. ILLUSTRATIONS AND RESULTS

To evaluate the monitoring performance of the proposed method, an industrial example is demonstrated. A series of three experiments were carried out to generate a total of 129 wafers, among which there were 21 fault wafers. The 21 faults were intentionally induced by changing the TCP power, RF power, pressure, Cl$_2$ or BC$_1$ flow rate and He chuck pressure. The dataset used in the present paper was collected from an AI stack etch process, which was performed on a commercial scale Lam 9600 plasma etch tool at Texas Instrument [Wise, et. al., 1999]. Several monitoring method has been tested in this process, including MPCA, k-Nearest neighbor rule, and etc [Cherry, and Qin, 2006; He, and Wang, 2007; ]. More detailed description of the process and induced faults can be found in reference [Wise, et. al., 1999]. The original dataset contains 40 variables, including process set-points, measured variables, and controlled variables such as gas flow rates, chamber pressure, and RF power. Among these 40 process
variables, 19 variables were chosen for fault detection previously. However, we have found that two variables (RFB reflected power and TCP reflected power) remain zero during all of the batch time. Therefore, the number of monitoring variables used in the present paper is reduced to 17, which are tabulated in Table 1. Besides, due to the large amount of missing data in two batches, only 107 normal batches and 20 fault batches are used. For simplicity, the duration of each batch is made as 85 sample points. Therefore, the normal and fault datasets can be represented as \( X_{\text{normal}} \) and \( X_{\text{fault}} \), respectively.

### Table 1: List of monitoring variables of the semiconductor manufacturing process

<table>
<thead>
<tr>
<th>No.</th>
<th>Variables No.</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BCl (_3), flow</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Cl (_2), flow</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>RF bottom power</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Endpoint A detector</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Helium pressure</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>Chamber pressure</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>RF tuner</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>RF load</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>Phase error</td>
<td></td>
</tr>
</tbody>
</table>

To examine the data characteristic of this batch process, a PCA model is first constructed based on the normal dataset. Choosing the first two principal components for data visualization, the result is shown in Figure 1. Obviously, there are three clusters, which correspond to the three experiments that generated this normal dataset. Hence, this batch process exhibits multiple operation modes behavior. To model and test the process separately, the normal process dataset is first partitioned into two parts: training dataset and testing dataset. 32 normal batches of each operation mode are randomly selected, thus a total number of 96 batches are selected for model development. Therefore, the training dataset and the testing dataset become as \( X_{\text{normal}}^\prime \) and \( X_{\text{test}} \), respectively.

To evaluate the feasibility and efficiency of the proposed method, the normal testing datasets and the fault datasets are both used. First, three normal batches are selected for testing. Monitoring results of the proposed method and MPCA for the first normal batch are given in Figure 2. It can be seen that both monitoring charts indicate that the process ran under the normal condition. Therefore, it can be inferred that the combination of monitoring results in different operation modes does not lose the efficiency for normal process monitoring. Figure 3 gives monitoring results of the sub-PCA model in each operation mode. The normal behavior of the process is only obtained by the first sub-PCA model, which indicates that this batch has a big probability in the first operation mode. To be clear, we can check posterior probability values of this batch in different operation modes, which are shown in Figure 4. As can be seen, the biggest
posterior probability value of this batch is found in the first operation mode. Similarly, monitoring results of the second and third normal batches can also be obtained. Due to the limited space, these similar results are not presented.

Figure 3: Monitoring results of the sub-PCA models for the first normal batch, (a) the first sub-PCA model; (b) the second sub-PCA model; (c) the third sub-PCA model

Figure 4: Posterior probabilities of the first normal batch in different operation modes

Figure 5: Monitoring results of the fault 6, (a) Proposed method; (b) MPCA

To test the fault detection capability of the proposed method, 20 fault batches are introduced for testing. For comparison, the fault detection results of both $T^2$ and $SPE$ statistics of MPCA are also given. Two faults are selected for detailed demonstration. First, monitoring results of fault 6 are given in Figure 5. It can be clearly seen from Figure 5 (a) that this fault can be detected by both $BIF_T$ and $BIF_{SPE}$ statistics during sample 65 and sample 75, which are highlighted by two ellipses. In contrast, neither $T^2$ and $SPE$ statistics of
MPCA can detect this fault. Similarly, monitoring results of the proposed method and MPCA for fault 8 are shown in Figure 6. Although this fault can be detected by the SPE statistic of MPCA, only several statistic values exceed the confidence limit around sample 30. For comparison, the monitoring performance can be greatly improved by the proposed method, because the values of both $BIF_T$ and $BIF_{SPE}$ statistics have continuously exceeded their corresponding confidence limits during samples 10-40, and samples 65-85.

![Figure 6: Monitoring results of fault 8, (a) Proposed method; (b) MPCA](image)

**5. CONCLUSIONS**

In the present paper, a novel Bayesian statistical method has been proposed for monitoring multimode batch processes. In this method, the Bayesian inference strategy was used for results combination in different operation modes, which render a soft assignment of the proposed method. According to the application study of the semiconductor process, the monitoring performance has been greatly improved by the proposed method. Besides, the current operation mode of the monitored data sample can be successfully identified. It is also straightforward that the proposed method can be used for fault identification, as long as the fault dataset is available from the process. By examining the posterior probability, the mode information of the process can be periodically updated. However, as mentioned above, we should pay more attentions to the identified new operation cases. In order to differentiate the new operation mode from the fault case, some process knowledge and experiences of the operation engineer would be very helpful.

**ACKNOWLEDGEMENT**

This work was supported by the National Natural Science Foundation of China. (61004134) and the National 863 High Technology Research and Development Program of China (2009AA04Z154).

**REFERENCES**


