2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

Application of Principal Components Analysis to Improve Fault Detection and Diagnosis on Semiconductor Manufacturing Equipment

Alexis Thieullen[1,2], Mustapha Ouladsine[1] and Jacques Pinaton[2]

Abstract— **With the evolutions in sensing technologies and the increasing use of advanced process control techniques, terabytes of data are recorded today during the manufacturing process of semiconductor devices. These large amount of data are then operated by Fault Detection and Classification (FDC) systems to assess the overall condition of production equipment. However, specific characteristics of semiconductor manufacturing such as highly correlated parameters, time-varying behaviors, or the large number of operating conditions tend to limit the efficiency of current indicators to detect and diagnose a failure occurence. There is therefore a significant requirement for the development and application of new methodologies to improve detection efficiency while reducing the complexity of condition monitoring, without losing detailed insight for efficient failure analysis. In this paper, we use data pretreatment algorithms from signal processing and time series analysis, and Multiway Principal Components Analysis (MPCA) methods to accurately represent equipment behavior and process dynamics and thus overcome issues inherent to semiconductor manufacturing context. A real-case application on a plasma etcher from STMicroelectronics Rousset 8' fab is proposed to highlight benefits of these methods.**

I. INTRODUCTION

Semiconductor devices manufacturing could be broadly described as a *batch multilayer process*. Products from multiple technologies are grouped into batches (containing up to 25 wafers) and manufactured according to a sequence of operations during which electronic circuits are gradually created on a wafer (Fig. 1). The manufacturing processes encompass several hundred operations at regular intervals, possibly on the same machinery, resulting in *reentrant flows*. Each operation made by a specific tool on a product is associated to fixed production *recipes* that describes the set of processing *steps* to be followed as closely as possible to achieve a satisfactory product quality.

As most of industrial batch processes, semiconductor manufacturing processes are highly nonlinear, time varying and subject to significant disturbances. Causes of these disturbances may arise from equipment (repairs, chamber cleaning, preventive maintenance, gradual build-up on chamber, machine aging, or sensor drift), product (different incoming wafer state, changes in materials), or process (first wafer effect, different preprocess chambers or steady states, warm-up). This generally result in gradual drift or abrupt shift, leading to variability wafer-to-wafer, within a batch, or batch-to-batch.

This work was supported by STMicroelectronics Rousset.

[1]Laboratoire des Sciences de l'Information et des Systemes, University of Aix-Marseille, 13397 Marseille, France {alexis.thieullen,mustapha.ouladsine}@lsis.org

[2]STMicroelectronics Rousset, 13106 Rousset, France {jacques.pinaton,alexis.thieullen}@st.com

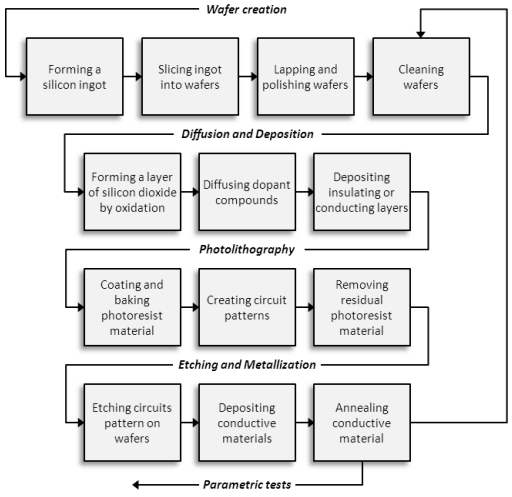

Fig. 1. Semiconductor Front End Manufacturing Process Flow Chart

To ensure the successful completion of a recipe and a consistent replicability of product quality, tens or even hundreds parameters are collected from tool sensors, representing a direct information on the tools current condition. As an extension of traditional statistical process control (SPC) approach, FDC is then used to detect and classify equipment failures for enhancement of tool effectiveness and product yield.

Nowadays, most of industrial FDC systems relies on the individual monitoring of sensor data. It consists in calculating several statistics of collected parameters (mean, standard deviation, maximum, minimum, range, *etc*.) on predefined time windows (recipe, step, *etc*.). These univariate indicators are then monitored throughout conventional SPC control charts to detect drifts or shifts. However, this approach presents several disavantages:

1. Its unidimensional nature leads to not considering the relationships, whether statics or dynamics, between collected parameters. This is a problem with multivariate processes where sensor measurements are highly correlated because of physical and chemical principles that drive the process operation.
2. The set of indicators used to monitor a specific recipe[1] is usually highly dependent on the production context. The large number of different equipment types, operations, products, and collected parameters prevents complete coverage. Experts analysis is therefore essential to define an

[1]Also known as "monitoring strategy".

TABLE I
PCA EXTENSIONS TO PROCESSES CHARACTERISTICS

Processes	Method	References	Method features
Batch processes	Multiway PCA (MPCA)	[1], [2]	Data unfolding
Nonlinear processes	Nonlinear PCA (NLPCA)	[3], [4], [5]	Nonlinearity is
	Kernel PCA (KPCA)	[6]	considered
Time varying processes	Recursive PCA (RPCA)	[7]	Model updating
	Adaptative PCA	[8]	
Dynamic processes	Dynamic PCA (DPCA)	[9]	Time dependencies
	EWMA Hybrid-wise Multiway PCA (E-HMPCA)	[10]	are considered

effective strategy. In return, only critical parameters (within the meaning of failure modes experts) are monitored, resulting on frequent non-detection problems.
3. Indicators developped through this approach are very sensitive to events on equipment(*e.g.* preventive or corrective maintenance) and changing operating conditions, making control charts limits difficult to set and maintain, resulting in false alarms problems.
4. Another major issue comes with the multiplicity of production contexts. Monitoring a fab like STMicroelectronics Rousset requires more than 60000 FDC control charts. It is difficult to ensure the effectiveness and relevance of such coverage, and to establish an efficient industrial diagnosis (*e.g.* physical parameter, recipe step, *etc*.) without long and exhaustive analysis.

Multivariate statistical process control such as principal component analysis (PCA) has found wide application in fault detection and diagnosis using collected sensor data. PCA is a projection method for mapping original high dimensional and correlated data onto a lower dimensional space with minimum loss of information. Originally applicable to a continuous, linear, and static processes [11], [12], [13], application of PCA was extended for handling industrial systems which exhibit batch data, nonlinear behaviour, changing process conditions, or time-correlated parameters. Some of these extensions are listed in Table I. In addition, some recent approaches combine several process characteristics, such as Batch Dynamic PCA [14] or Adaptative Kernel PCA [15].

In this paper, we use E-HMPCA to perform fault detection and diagnosis for a batch process on a plasma etcher. According [10], the main advantage of E-HMPCA is its ability to consider process dynamics, and both batch-to-batch and time-to-time (during recipe duration) correlations. Moreover, large data matrix calculations are avoided. E-HMPCA approach for fault detection is presented in the next section.

II. FAULT DETECTION VIA E-HMPCA

A. Data unfolding

Data from batch processes are usually stored in a three-dimensional matrix X, $(I \times J \times K)$, where I, J, and K are respectively the number of batches, variables, and observations (sampling times). Intended to handle batch data, E-HMPCA relies on hybrid-wise unfolding that combines the advantages of both batch-wise and variable-wise unfolding approaches [16]. To achieve this, X is unfolded according I, mean-centered, and rearranged in a variable-wise structure to obtain a $(IK \times J)$ matrix (Fig.2). However, correlations are still considered in a static way. Thus, the next step of the methodology will consist to consider time dependencies by employing an Exponentially Weighted Moving Average (EWMA) approach.

B. Process dynamics integration

After the unfolding step, the resulting matrix X, $(IK \times J)$ is decomposed by PCA:

$$X = TP^T \tag{1}$$

The *loading* matrix P gives the coefficients for each original variable when calculating the principal components. The *score* matrix T contained the projections of the original data onto the principal components. These matrices can be partitioned according the number l of the more significant principal components which are sufficient to explain the variability of the data:

$$T = [\hat{T}_l \mid \tilde{T}_{J-l}] \ , \ P = [\hat{P}_l \mid \tilde{P}_{J-l}] \tag{2}$$

An important number of selection criteria have been proposed in the literature to determine an optimal number of principal components. The most common are based on heuristics or statistical decision-making practices. A non exhaustive list can be found in [17].

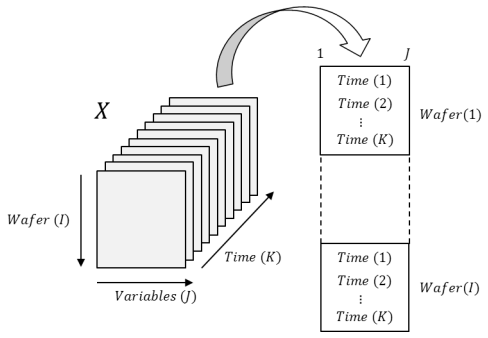

Fig. 2. Variable-wise unfolding

EWMA is then used to incorporate process dynamics, first into the score and then into the covariance matrices:

$$\begin{aligned} t_{E,k} &= \lambda t_k + (1-\lambda) t_{E,k-1} = \lambda \sum_{j=1}^{k} (1-\lambda)^{k-j} t_j \\ S_{E,k} &= \frac{t_{E,k}^T \times t_{E,k}}{I-1} \end{aligned}$$

where t_k is a $(I \times l)$ matrix corresponding to the score of the k^{th} observation of each wafer in the matrix $\hat{T}$, and $S_{E,k}$ is the EWMA filtered covariance matrix at time k (e.g. $t_{E,1}$ contained the EWMA filtered projection of the firsts data collected for each variable and each wafer). Thus batch dynamic is considered, taking into account all the previous observations from recipe start. The coefficient λ $(0 \leq \lambda \leq 1)$ represents the degree of weighting decrease, that determines the weight of older data. According [10], a small λ is helpful in detecting small variations from the model but increases the delay time of detection, while a large one has the opposite effect.

In the same way, EWMA is used to filter the residual subspaces projection to consider process dynamic:

$$\begin{aligned} e_{E,k} &= \lambda e_k + (1-\lambda) e_{E,k-1} \\ &= \lambda \sum_{j=1}^{k} (1-\lambda)^{k-j} e_j \end{aligned}$$

where

$$e_k = (\tilde{P}\tilde{P}^T) x_k$$

is a $(I \times J - l)$ matrix (also called error matrix) corresponding to the projection of the k^{th} observation of each batch on the residual subspace.

C. Fault detection indices

First, a new observation $x_{new,k}$ is projected on filtered principal components and residual subspaces through score and error matrices:

$$\begin{aligned} t_{E,new,k} &= \lambda t_{new,k} + (1-\lambda) t_{E,new,k-1} \\ &= \lambda \sum_{j=1}^{k} (1-\lambda)^{k-j} t_{new,j} \\ e_{E,new,k} &= \lambda e_{new,k} + (1-\lambda) e_{E,new,k-1} \\ &= \lambda \sum_{j=1}^{k} (1-\lambda)^{k-j} e_{new,j} \end{aligned}$$

In E-HMPCA, fault detection is ensured by classical PCA detection indices Hotelling's T^2 and Squared Prediction Error SPE:

- Hotelling's T^2 statistic measures variations in the principal components subspace, and is expressed by using the diagonal matrix $\Lambda_{E,k}$ whose elements are the eigenvalues of the filtered correlation matrix $S_{E,k}$ in the decreasing order:

$$T^2_{E,new,k} = t^T_{E,new,k} \Lambda^{-1}_{E,k} t_{E,new,k}$$

- SPE is the magnitude of a sample projection on the residual subspace. A change in variable correlation indicates an unusual situation because the variables do not conserve their normal relations. Under this situation, the sample increases its projection and the magnitude reaches unusual values compared to those obtained during normal conditions. SPE expression is given by:

$$\begin{aligned} SPE_{E,new,k} &= ((\tilde{P}\tilde{P}^T) x_{new,k})^T (\tilde{P}\tilde{P}^T) x_{new,k} \\ &= e^T_{E,new,k} e_{E,new,k} \end{aligned}$$

The process is considered reliable if the T^2 and the SPE statistics are under their upper control limit (UCL), which are respectively expressed as follows:

$$\begin{aligned} UCL_{T^2_{E,new,k}} &= \frac{l(I^2-1)}{I(I-l)} F_{l,I-l} \\ UCL_{SPE_{E,new,k}} &= \frac{v_{E,k}}{2m_{E,k}} \chi^2_{2m^2_{E,k}/v_{E,k}} \end{aligned} \quad (3)$$

where $m_{E,k}$ and $v_{E,k}$ are mean and variance of the new data at time k [10].

III. CASE STUDY

A. Description of the context

In order to demonstrate the effectiveness of E-MHPCA, we study in this paper a plasma etching tool of STMicroelectronics Rousset 8" fab. To reduce the complexity, we only focus here on one etch operation and one product. The related production recipe consist in a serie of twelve steps. Measured variables are sampled at 1 second intervals during the etch process, for approximately 320 measures. We have collected data from 37 sensor for one month of production, which represents more than 1100 wafers. These parameters can be classified according Table II. We attempt to detect a gas leak in the etch chamber. First effects were observed (but not considered as a fault) on FDC control charts around the 394^{th} production run in our data.

TABLE II
CLASSIFICATION OF SENSOR PARAMETERS

Main groups	Subgroups	Number of parameters
RADIO FREQ.	Top Match RF	8
	Bottom Match RF	5
	ESC Behavior	2
TEMPERATURE	Helium Cooling	2
	Bottom Temp.	2
	Top Temp.	2
GAS	Gas	8
	OES	4
PRESSURE	Pressure	4

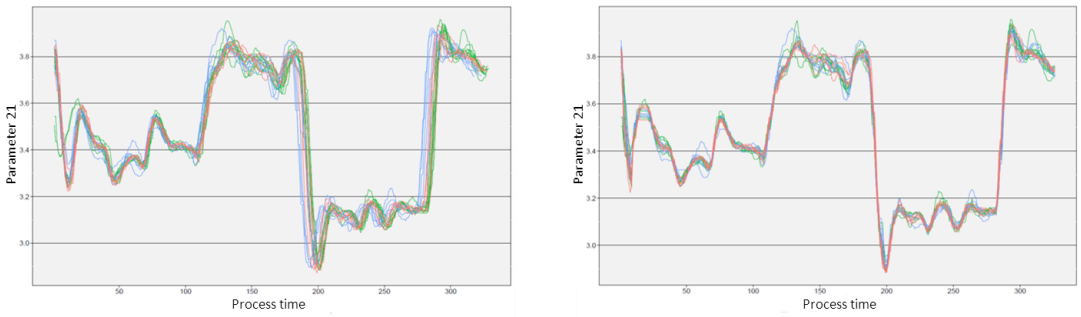

Fig. 3. Results of DTW synchronization on batch trajectories for one process parameter. Raw trajectories are on the left, the synchronized ones on the right. Each color is associated to a different lot.

B. Data preprocessing

Before applying PCA, several preprocessing steps are mandatory due to the industrial nature of the data. A common complication is that data from real industrial batch processes are often characterized by unsynchronized trajectories of variable duration. However, the synchronization of the trajectories to a common length is mandatory for the application of many statistical analysis approaches. One way to avoid this is to use statistic indicators on the data from each wafer over all available samples and work with only means, standard deviations, *etc*. Another approach would be to select a specified number of samples, corresponding to some critical process event. However, this type of approach leads to the loss of part of the information contained in the original data. In this work, Dynamic Time Warping (DTW) has been utilized to synchronize raw traces onto a reference trace, based on data from an equipment in proper operating mode. DTW was first developped in the domain of speech recognition. It was well utilized in the context of batch processing [10], [16], [18], [19]. The principle is to nonlinearly warps two trajectories in such a way that similar events are aligned and a minimum distance between them is obtained. We can see on Fig.3 an example of DTW on noisy data from a real manufacturing process. Readers interested in a further study of DTW algorithm and applications can refer to the works cited above.

A second step for data preprocessing consists in scaling into zero mean and unit variance each column of the data matrix X before the application of PCA. This is mandatory by the fact that original variables are not only heterogeneous in their mean (data are expressed in different measurement units), but also on their dispersion and nature (measurement units are not expressed in similar quantities). This is thus necessary to get each variable to a common framework of comparability prior further analysis.

A third step concerns normality of sensor data. In most of multivariate SPC practices data are supposed normally distributed. Although multivariate normality is not a very strict assumption when PCA is used for data reduction or exploratory purposes, violating the normality assumption when modeling with PCA and monitoring with Hotelling's T^2 and SPE could lead to several issues with fault detection [20]. Actually, control limits given in equation 3 are statistically relevant for normally distributed scores and residuals [2], [21]. However, this assumption is regularly violated in the context of real industrial data. If several alternative approaches for dealing with non-normal distributed data have been proposed, it may be sometimes sufficient to apply data transformations (*e.g.* using logarithm or square root functions) to obtain multivariate normaly distributed data.

To ensure normality of our data, the Doornik-Hansen test for multivariate normality [22] is applied on filtered scores and residual matrices $t_{E,k}$ and $e_{E,k}$ for each time k. Results show that we can retain the normality hypothesis, and therefore use T^2 and SPE control limits as defined in equation 3.

C. Results

E-HMPCA model is built from a sample of 50 wafers selected from wafers processed before the failure, representing an equipment in regular condition. Based on cross-validation [23]. The idea behind cross-validation broadly lies in partitioning the original data set and select the number of components which maximize the predictive properties of model, builded and tested on different parts of the data set. The selection criteria is then based on the sum of squares of prediction errors. According this approach, the first ten components are retained for more than 72% of cumulated variance intercepted (i.e. the amount of inertia, or variation in the data set, explained by retained components relative to the total explainable inertia). The "forgetting" factor λ was set to 0.1, this value being considered preferable in order to detect small variations of the process.

SPE and Hotelling's T^2 are used to detect abnormal behaviors on the whole production recipe for each of the 1051 remaining wafers. An exemple of SPE and T^2 with their 99% confidence thresholds is given in Fig. 4 for a process in regular operating conditions. However, both SPE and T^2 control charts shows a very sharp peak starts around time interval 309 for a wafer being part of the training sample, as we can see on Fig. 5. SPE contribution plot, which represents the contribution of each original variable in the index value at each time k, is presented in Fig. 6 for time interval 309 of the process. The interest of such

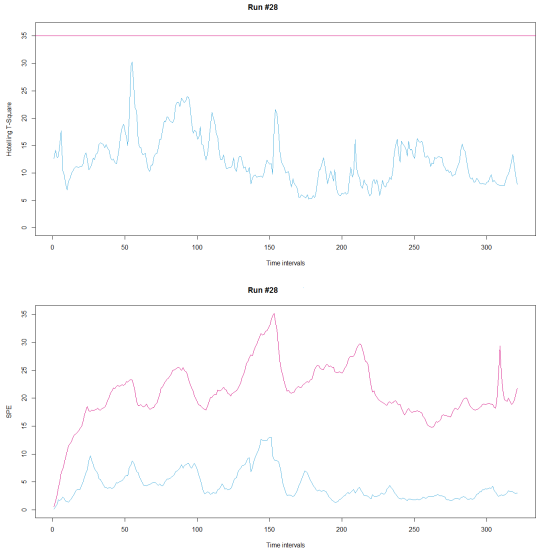

Fig. 4. T^2 and SPE control charts for a process under regular operating conditions. Red lines correspond to the 99% confidence thresholds for each indice at time k. Blue lines correspond to the indices value.

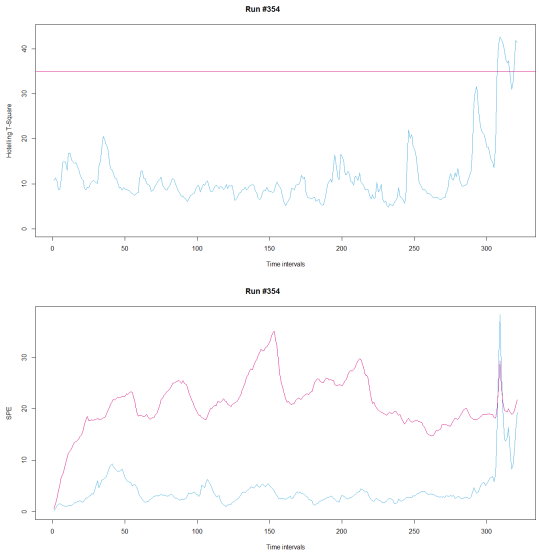

Fig. 5. T^2 and SPE control charts for a process with a fault occurence between time intervals 308 and 310.

a graph is in isolating the most representative parameters at identified failure times, and thus achieve first assumptions for the fault diagnosis. Based on subsequent tests on affected product, engineers have identified deficiencies related to a sporadic problem of very short duration occuring during the etch process, that affects chamber pressure (linked to parameter 21 on Fig. 6). According this analysis, we removed this wafer from the training sample, since its process is not representative of intended operating conditions. However this first analysis showed that we could provide relevant information (i.e. failure time and impacted parameters) for an efficient diagnosis, with relatively simple and well-known tools widely used in classical PCA methods.

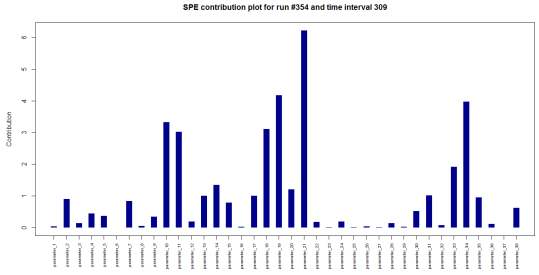

Fig. 6. SPE contribution plot for time interval 309.

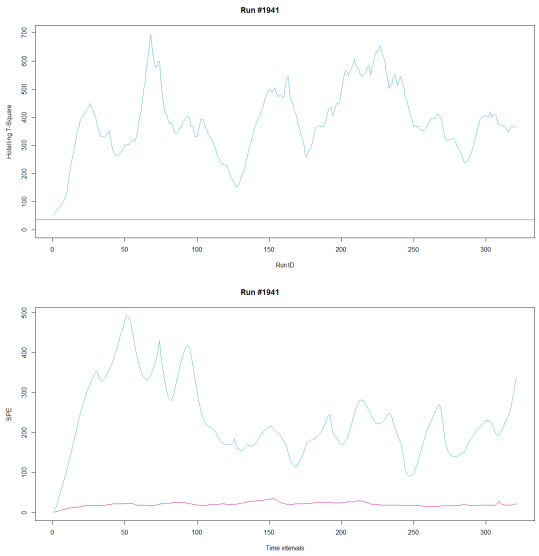

Fig. 7. T^2 and SPE control charts for run #1941. Respective control limits are transgressed throughout the duration of the recipe.

According T^2 and SPE results on test data, wafers processed after the 393^{th} run (which corresponds to the run ID 1940 in our data) seem representative of non-regular operating conditions, as shown in Fig. 7 for the following run. The corresponding contribution plot at time interval 50, which exhibits the greatest violation of the control limit, is given on Fig. 8. We can see that parameters 21, 22 and 29 present the most important contributions. According Table II classification, parameters 22 and 29 correspond to the subgroup "Gas". In addition, parameter 21 is linked to the subgroup "Pressure". It seems consistent with the description of the fault given in III-A. After an additional analysis conducted by process engineers, time interval 50 could be identified as a key moment regarding the impact of the failure on the recipe achievement.

IV. CONCLUSION AND PROSPECTS

Including process dynamics in score and error matrices, E-HMPCA allows to model accurately a manufacturing batch process. We have seen with an exemple on a real etch equipment from STMicroelectronics Rousset, that two control charts can effectively replace current monitoring strategies that require to monitor dozens of control charts. This approach has the additional advantage to address multivariate

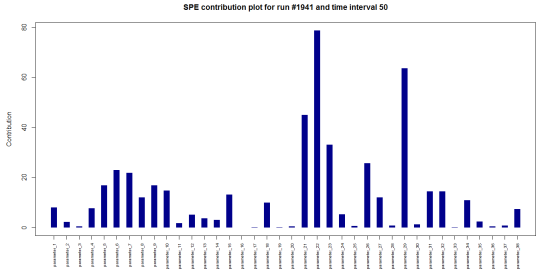

Fig. 8. SPE contribution plot for time interval 50.

data collected directly from sensor, and not univariate indicators that restrict the original information contained in the data. This explains the improvement of detection capability compared to actual FDC systems. This is evidenced by the accurate detection of both ponctual process variations on a specific time step of a process (Fig. 5), and a progressive change in the equipment behavior that impact the whole recipe. In addition conventional methods of diagnosis, such as contribution plots, can provide relevant information for a first diagnostic phase.

Future works can be divided in two parts. The first part concerns the improvement of the algorithms. Several methods have been developed for linear PCA in order to improve components selection, fault detection indices, or diagnostic methods. Adapting these methods to E-HMPCA may lead to a better detection capability and a reduction of false alarms rates. The second parts focus on the adaptation to semiconductor manufacturing issues, for real-time fault detection and diagnosis of production equipment in an industrial environment. This requires improvements to handle recipe changes, and any behavioral changes due to interventions on equipment (*e.g.* preventive or corrective maintenance).

REFERENCES

[1] P. Nomikos and J. F. MacGregor, "Monitoring batch processes using multi-way principal component analysis," *AIChE Journal*, vol. 40, no. 8, pp. 1361–1375, 1994.

[2] ——, "Multivariate spc charts for monitoring batch processes," *Technometrics*, vol. 37, no. 1, pp. 41–59, 1995.

[3] D. Dong and T. McAvoy, "Nonlinear principal component analysis-based on principal curves and neural networks," *Computers and Chemical Engineering*, vol. 20, no. 1, pp. 65–78, 1996.

[4] ——, "Batch tracking via nonlinear principal component analysis," *AIChE Journal*, vol. 42, no. 8, pp. 2199–2208, 1996.

[5] F. Jia, E. Martin, and A. Morris, "Non-linear principal components analysis for process fault detection," *Computers & Chemical Engineering*, vol. 22, no. 1, pp. S851–S854, Mar. 1998.

[6] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[7] W. Li, H. H. Yue, S. Valle-Cervantes, and S. J. Qin, "Recursive pca for adaptive process monitoring," *Journal of Process Control*, vol. 10, pp. 471–486, 2000.

[8] D. Lee and P. Vanrolleghem, "Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis," *Biotechnol. Bioeng.*, vol. 82, no. 4, pp. 489–497, May 2003.

[9] W. Ku, R. H. Storer, and C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 179–196, 1995.

[10] Y. Zhang, "Improved methods in statistical and first principles modeling for batch process control and monitoring," Ph.D. Dissertation, University of Texas, Austin, Aug. 2008.

[11] J. V. Kresta, J. F. MacGregor, and T. E. Marlin, "Multivariate statistical monitoring of process operating performance," *The Canadian Journal of Chemical Engineering*, vol. 69, no. 1, pp. 35–47, Feb. 1991.

[12] M. H. Kaspar and W. H. Ray, "Chemometric methods for process monitoring and high-performance controller design," *AIChE Journal*, vol. 38, no. 10, pp. 1593–1608, Oct. 1992.

[13] J. Zhang, E. B. Martin, and A. J. Morris, "Fault detection and diagnosis using multivariate statistical techniques : Process operations and control," *Chemical engineering research & design Journal*, vol. 74, no. 1, pp. 89–96, 1996.

[14] J. Chen and K.-C. Liu, "Non-linear principal components analysis for process fault detection," *Chemical Engineering Science*, vol. 57, pp. 63–75, 2002.

[15] M. Ding, Z. Tian, and H. Xu, "Adaptive kernel principal component analysis," *Chemical Engineering Science*, vol. 90, no. 5, pp. 1542–1553, May 2010.

[16] J.-M. Lee, "Statistical process monitoring based on independent component analysis and multivariate statistical methods," Ph.D. thesis, University of Science and Technology, Pohang, 2004.

[17] B. Mnassri, "Analyse de donnes multivariees et surveillance des processus industriels par analyse en composantes principales," Ph.D. thesis, Aix-Marseille University, 2012.

[18] A. Kassidas, J. F. MacGregor, and P. A. Taylor, "Synchronization of batch trajectories using dynamic time warping," *Chemical Engineering Science*, vol. 44, no. 4, pp. 864–875, Apr. 1998.

[19] G. Barna, "Procedures for implementing sensor-based fault detection and classification (fdc) for advanced process control (apc)," *SEMATECH Technology Transfer Document 97013235A-XFR*, 1997.

[20] G. A. Cherry, "Methods for improving the reliability of semiconductor fault detection and diagnosis with principal component analysis," Ph.D. Dissertation, University of Texas, Austin, Dec. 2006.

[21] H. H. Yue and S. J. Qin, "Reconstruction based fault detection using a combined index," *Ind. Eng. Chem. Res.*, vol. 40, pp. 4403–4414, 2001.

[22] J. A. Doornik and H. Hansen, "An omnibus test for univariate and multivariate normality," *Oxford Bulletin of Economics and Statistics*, vol. 70, pp. 927–939, 2008.

[23] S. Wold, "Cross-validatory estimation of the number of components in factor and principal components models," *Technometrics*, vol. 20, no. 4, pp. 397–406, 1978.