Identifying key process characteristics and predicting etch rate from High-Dimension Datasets

E. Ragnoli[1a], S. McLoone[1], S. Lynn[1] ,J. Ringwood[1], N. Macgearailt[2]

[1]Department of Electronic Engineering, NUI Maynooth,, Ireland, [2]Dublin City University, Ireland

[a] Corresponding author, Email: emanuele.ragnoli@nuim.ie, Tel: +353 -01 7086907

Abstract - **In semiconductor manufacturing advanced process control (APC) refers to a range of techniques that can be used to improve process capability. As the dimensions of electronic devices have decreased, the application of APC has become more and more important for the critical stages of production processes. However, the economic disadvantage of employing APC is that it requires feedback information in the form of downstream metrology data, which is both time consuming and costly to obtain.**

I. INTRODUCTION

Virtual metrology (VM) is a cost-effective approach proposed to replace some of the actual metrology. In VM modelling and simulation are used to predict the key metrology characteristics from more accessible in-line measurements available on modern semiconductor manufacturing processes. This includes both traditional process variables such as flow rates, pressures and microwave power and advanced sensor data such as plasma impedance monitors (PIM) and Optical Emissions Spectroscopy (OES) signals [1]. In addition to the usual challenges presented by highly nonlinear and complex processes, a major issue when applying VM to semiconductor manufacturing processes is handling the large numbers of highly correlated variables that are typically recorded in modern facilities. For example, in the plasma etch process considered in this paper an OES derived dataset consisting of over 12000 input variables is available. The challenge is to extract from this data the key process characteristics and a model that reliably predicts the evolution of the process etch rate, a metrology variable that has a significant impact on manufacturing yield [2,3].

In this paper four different VM algorithms, Principal Component Regression (PCR), Partial Least Squares (PLS), Forward Selection Component Analysis (FSCA) and Forward Selection Regression (FSR) are investigated for this problem and practical measures for achieving realisable implementations when faced with computational and memory constraints are explored.

The paper is structured as follows: Section II explains the four VM algorithms. Section III describes the experimental setup, the dataset collected and issues such as training and test data selection. Results and analysis are presented in Section IV and finally the conclusions of are given in Section V.

II. FOUR DIFFERENT VM ALGORITHMS

A. PCR

Principal Component Regression (PCR) is a common regression technique based on Principal Component Analysis (PCA). PCA is a matrix factorisation technique that selects variables based on the directions of largest variance in the data. Specifically, using standard PCA notation, if $\mathbf{X}$ is an $n \times m$ data matrix then it can be decomposed as

$$\mathbf{X} = \sum_{i=1}^{r} \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \qquad (1)$$

where $\mathbf{t}_i$ and $\mathbf{p}_i$ are $n \times 1$ and $m \times 1$ vectors, respectively, and scalar r is the rank of $\mathbf{X}$. The vectors $\mathbf{t}_i$ and $\mathbf{p}_i$, referred to as scores and loadings, respectively are the principal components of matrix $\mathbf{X}$. The decomposition can also be expressed in matrix form with vectors $\mathbf{t}_i$ and $\mathbf{p}_i$ forming the columns of the corresponding matrices $\mathbf{T}$ ($n \times r$) and $\mathbf{P}$ ($m \times r$).

Statistically, the ordered principal components can be interpreted as the directions of largest variance in the data. In general, when there is a significant level of redundancy only a small number of components are needed to capture the information in the data. Hence, a selected number of principal components based on variance explained can constitute a new set of variables.

There are several algorithms which can be used to compute the principal components, for example nonlinear iterative partial least square (NIPALS), singular value decomposition (SVD), the power method (POWER) and eigen-value decomposition (EVD) [4], NIPALS [5], the one adopted in this work, is an iterative algorithm that computes the eigenvectors of the matrix X, one at the time in order of significance.

For a set of inputs $\mathbf{X}$ and a set of outputs $\mathbf{Y}$ Multiple Linear Regression (MLR) is a modelling technique that attempts to establish a linear relationship between them. MLR can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{B}$$

where $\mathbf{B}$ is the matrix of the linear coefficients. In PCR the input matrix is substituted by the selected scores of $\mathbf{X}$. Therefore, the MLR formula becomes:

$$\mathbf{Y} = \mathbf{T}\mathbf{B}$$

PCR is thus a two step process, first variable selection and then regression. The advantages are that it yields parsimonious models, it solves the collinearity problem and, by eliminating the less significant principal components, it reduces the impact of noise. The main disadvantage is that useful features for predicting **Y** may be in the discarded principal components.

B. PLS

Partial Least Squares (PLS) is an extension of the NIPALS algorithm for PCA that attempts to address the deficiency of PCR by taking into account the prediction of output **Y** in determining the decomposition of inputs into a set of loadings and scores. The main idea of PLS is to successively select directions of variation in the input data **X** that maximize the output variation that can be predicted. Thus the PLS decomposition consists of outer relations (**X** and **Y** individually)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \quad \text{and} \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}}$$

and an inner relation linking **X** and **Y**

$$\mathbf{U} = \mathbf{T}\mathbf{B}$$

with the final regression model given by

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q}^{\mathrm{T}}$$

If **Y** is one dimensional **U**=**Y** and **Q** is the identity matrix. In the NIPALS implementation of PLS the recursive computation of the columns of **T** and **U** is done simultaneously with information exchanged between them at each step to ensure optimal alignment [2].

C. FSR

Forward Selection Regression (FSR), Backward Selection Regression (BSR) and Stepwise Regression (SR) are well established variable selection/model building techniques in classical linear regression [6]. In FSR variables are added to the model one at a time. At each step the selected variable is the one that yields the best improvement in the model prediction. In contrast, BSR starts with a model containing all the candidate variables and the variables that have the least impact on performance are successively removed. SR techniques use a combination of both approaches in attempt to obtain more optimal results. All three methods are approximate solutions to the problem of finding the best subset of predictor variables given a set of *m* candidates. This is a NP-hard problem and is computationally intractable as the number of candidate variables increases. Consequently, optimality has to be sacrificed for computational efficiency. In this work FSR was chosen as it has the lowest computational complexity and has been found, in practice, to give comparable performance to other methods [7]. The specific implementation of FSR adopted is as follows:

1. Given regressor matrix **X** and output **y**, set $\widetilde{\mathbf{X}} = \mathbf{X}$ and $\hat{\mathbf{y}} = \widetilde{\mathbf{y}}$.
2. Select $\widetilde{\mathbf{x}} \in \widetilde{\mathbf{X}}$ such that

$$\widetilde{\mathbf{x}}^{*} = \arg\min_{\widetilde{\mathbf{x}} \in \widetilde{\mathbf{X}}} \{ \left\| \widetilde{\mathbf{y}} - \hat{\mathbf{y}}(\widetilde{\mathbf{x}}) \right\|_2^2 \} \quad \text{where} \quad \hat{\mathbf{y}}(\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}} \frac{\widetilde{\mathbf{x}}^{\mathrm{T}} \widetilde{\mathbf{y}}}{\widetilde{\mathbf{x}}^{\mathrm{T}} \widetilde{\mathbf{x}}}$$

3. Remove contribution of $\widetilde{\mathbf{x}}^{*}$ from X and y

$$\widetilde{\mathbf{X}} = (1 - \frac{\widetilde{\mathbf{x}}^{*} \widetilde{\mathbf{x}}^{*\mathrm{T}}}{\widetilde{\mathbf{x}}^{*\mathrm{T}} \widetilde{\mathbf{x}}^{*}}) \widetilde{\mathbf{X}}$$

$$\widetilde{\mathbf{y}} = (1 - \frac{\widetilde{\mathbf{x}}^{*} \widetilde{\mathbf{x}}^{*\mathrm{T}}}{\widetilde{\mathbf{x}}^{*\mathrm{T}} \widetilde{\mathbf{x}}^{*}}) \widetilde{\mathbf{y}}$$

4. Repeat from 2 until a specified stopping criterion has been satisfied.

The stopping criterion can be based on statistical significance testing, cross-validation on test data or a specified maximum model dimension. Here cross-validation on test data

D. FSCA

Forward Selection Component Analysis is an extension of FSR for selecting a subset of variables that best represent a set of variables **X**.

1. Given data matrix **X**, set $\widetilde{\mathbf{X}} = \mathbf{X}$.
2. Select $\widetilde{\mathbf{x}} \in \widetilde{\mathbf{X}}$ such that

$$\widetilde{\mathbf{x}}^{*} = \arg\min_{\widetilde{\mathbf{x}} \in \widetilde{\mathbf{X}}} \{ \left\| \widetilde{\mathbf{X}} - \hat{\mathbf{X}}(\widetilde{\mathbf{x}}) \right\|_F^2 \} \text{ where } \hat{\mathbf{X}}(\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}} \frac{\widetilde{\mathbf{x}}^{\mathrm{T}} \widetilde{\mathbf{X}}}{\widetilde{\mathbf{x}}^{\mathrm{T}} \widetilde{\mathbf{x}}}$$

3. Remove contribution of $\widetilde{\mathbf{x}}^{*}$ from **X**

$$\widetilde{\mathbf{X}} = (1 - \frac{\widetilde{\mathbf{x}}^{*} \widetilde{\mathbf{x}}^{*\mathrm{T}}}{\widetilde{\mathbf{x}}^{*\mathrm{T}} \widetilde{\mathbf{x}}^{*}}) \widetilde{\mathbf{X}}$$

4. Repeat from 1 until a specified stopping criterion has been satisfied.

The set of vectors $\widetilde{\mathbf{x}}^{*}$ returned by FSCA are orthogonal components that summarize the information contained in **X**. The corresponding columns of **X** are the selected variables (**x**). When the level of redundancy in the data is high the performance of FSCA in terms of explaining the variance in **X** with a small number of components can approach that of PCA. Having identified the FSCA decomposition the set of components $\widetilde{\mathbf{x}}^{*}$ or corresponding selected variables **x** can be used to build a regression model in a similar fashion to PCR.

While FSR is relatively efficient, FSCA is computationally demanding with $O(m^3)$ complexity compared to $O(m^2)$ for FSR. This makes direct application of FSCA intractable for large data sets. In these circumstances an approximation to FSCA can be obtained by partitioning the data set into sections and applying FSCA to each section to obtain a reduced set of variables. FSCA can then be applied to this reduced set of variables to obtain an aggregate set of components that are representative of the original data set.

III. EXPERIMENTAL SETUP

The four VM algorithms described above are compared on a Plasma Etching Optical Emission Spectroscopy (OES) dataset for etch rate prediction. The data collected consist of 2000 etch rate samples (Fig. 1) and for each of the samples a full OES spectra (2048 channels) is measured. In order to reduce the huge amount of OES data each of the OES spectra is summarised by their 6 statistical moments (Kurtosis, Mean, Skewness, Variance, Maximum and Minimum).

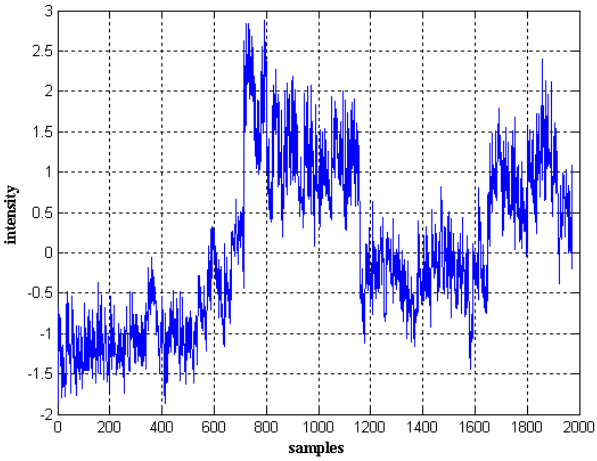

Figure 1: Measured etch rate for the 2000 samples.

In order to assess the potential for predicting etch rate the available data is split into training and test data sets with the training data used to build prediction models and the test data used to evaluate their performance. Two different partitions of the data set are considered as follows:

- Dataset 1: The first 1000 samples are used as training data and the remaining samples are used as test data.
- Dataset 2: The data is divided into odd and even sample numbers with the odd samples used as training data and the even samples used as test data.

Dataset 2 provides training data representative of the full range of etch rate variations and will be useful for establishing the extent to which the OES signals can be used to predict etch rate. In contrast, in dataset 1 provides less complete coverage of the operating space and will provide an indication of the stationarity of the relationship between the OES signals and etch rate.

The performance of the different VM models developed will be measured in terms of the Normalised Mean Square Prediction Error (NMSE) defined as:

$$\mathbf{NMSE} = \frac{\mathrm{var}(\mathbf{y} - \hat{\mathbf{y}})}{\mathrm{var}(\mathbf{y})} * 100 \,.$$

Here var() is the sample variance, and $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the measured and predicted output data, respectively.

IV. RESULTS AND COMPARISON

A.. PCA and PLS variable selection

Applying PCA to Datasets 1 and 2 (Figure 2) shows that there is little difference between the structure of the variance in each data set as a function of the number of principal components (PCs) It is also clear that there is huge redundancy in the data with 20 principal components able to explain more than 95% of the variation contained in data (12288 variables).

A similar pattern is observed with PLS, except that now the selected components explain much less of the variance in the input data. (more than 100 components are needed to capture 90% of the data variation). This is as expected since PLS focuses on explaining the output variation rather than the input variation.

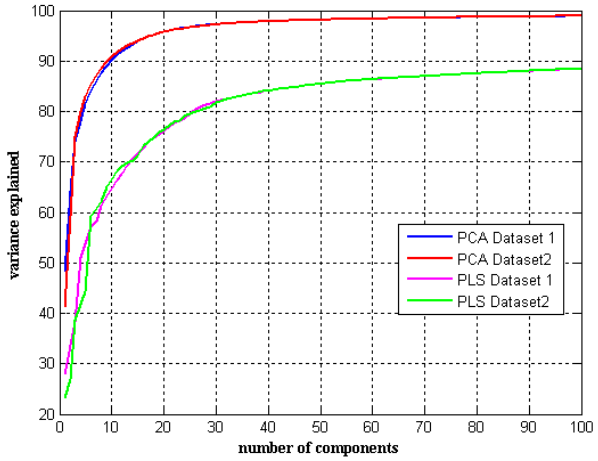

Figure 2: Plot of input data variation explained by PCA and PLS components explained as a function of the number of PCs

B.. Variable selection with FSR and FSCA

Tables 1 to 4 show the variable selected by FSR and FSCA for Dataset 1 and Dataset 2. The order of the variables is shown in the first column, the OES signal statistic and the channel number corresponding the selected variable are given in columns two and three respectively and the NMSE performance of the resulting linear regression models on the Training and data Test sets are given in the last two columns. The linear models are regressions on the currently and all previously selected variables.

In addition, a plot of the input variance explained as a function of the number of selected components is given Figure 3 for FSR and FSCA for each of the data sets.

The FSR results show that Mean of the OES signals is the most useful predictor of etch rate (especially on Dataset 1) . At the same time, the Minimum statistic appears to have little value as it has not been selected as a regressor for either dataset. A similar pattern is observed when with FSCA

selected components. As expected the NMSE performance of FSCA is substantially inferior to FSR.

Note that the FSCA components were computed using a 2 step process. First, FSCA is used to generate the top 30 FSCs for each of the six signal statistics. This results is an reduced data set of 180 selected variables. FSCA is then applied again to select the first 30 FSCs of the new dataset.

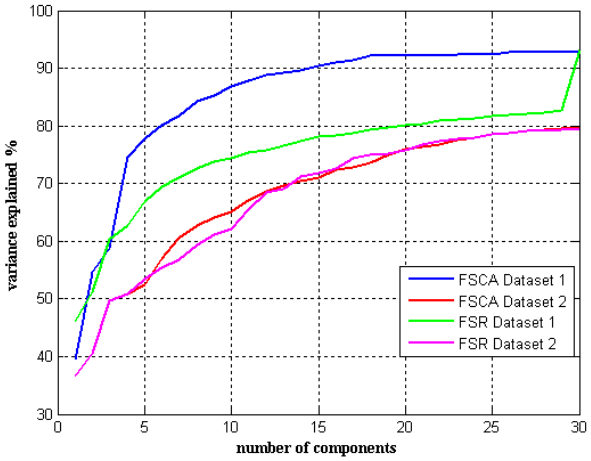

Figure 3: Input variance explained by FSCA/FSR components as a function of the number of FSCA/FSR components.

C. Prediction with PCR and PLS

Figure 4 shows the NMSE obtained using PCR on Dataset 1 and 2, in relation to the number of principal component selected. Figure 5 shows the same relationship for PLS.

As expected PLS outperforms PCR on both Datasets, when the number of principal components involved becomes big enough. This is due to the variable selection process, that chooses the principal components most correlated to the output.

Prediction results on Dataset 2 are better than results on Dataset 1. Obviously this is consequence of the different choice of training sets. In Dataset 1 the choice of training points spans the first half of the data, while in Dataset 2 it sampled through all the data. Hence, Dataset 2 is better suited for a detection of the global trend of etch rate. The significant drop in performance from training to test data in Dataset 1 suggests that the linear trend changes from the first to the second half of the process. This is likely due to process drift over time due to chamber seasoning and clouding of the OES window.

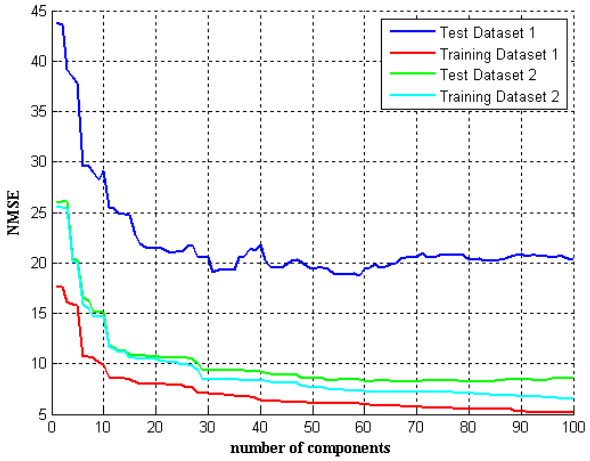

Figure 4: Prediction error for etch rate using PCR on Dataset 1 and 2

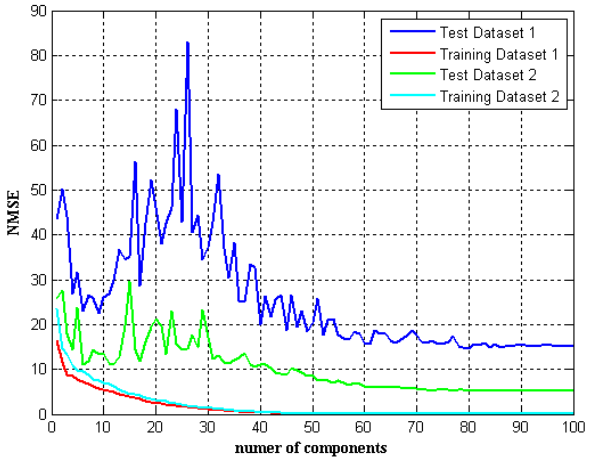

Figure 5: Prediction error for etch rate using PLS on Dataset 1 and 2

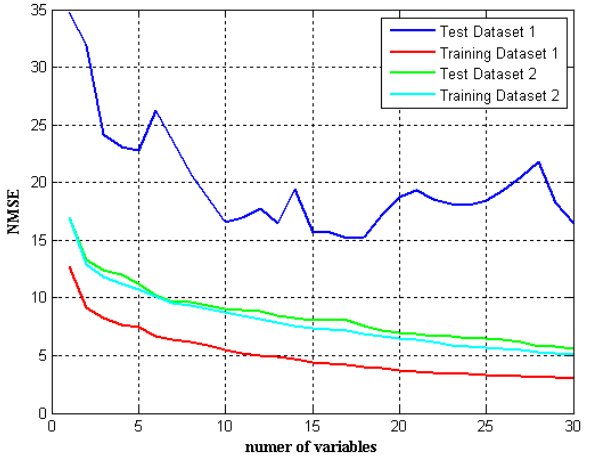

Figure 7: Prediction error for etch rate using FSR on Dataset 1 and 2

D. Prediction with FSR and FSCA

Prediction results of etch rate using FSR are shown in Figure 7. Consistently with PCR and PLS, FSR performs better on Dataset 2 than 1.

Moreover, on both Datasets, with a comparable number of components/variables involved, FSR gives a better performance than PCR and PLS. A linear fit of predicted etch rate and measured etch rate is shown in Figure 7 and Figure 8, for Dataset 1 and 2, respectively. The R-squared value of the linear fit in Figure 7 is 0.87 for test and 0.97 for training. For Figure 8 that is 0.94 for test and 0.95 for training.

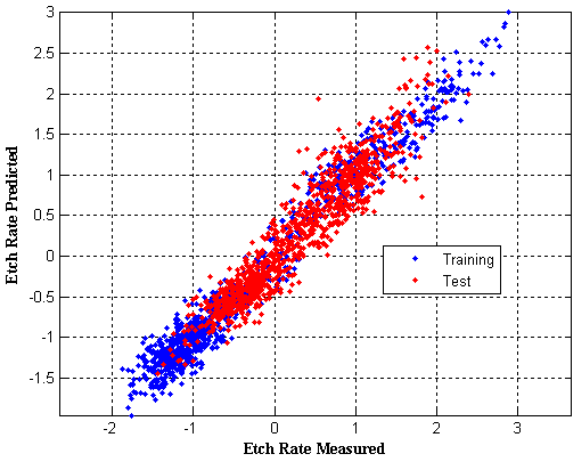

Figure 7: Linear fit of etch rate predicted with etch rate measured in Dataset 1

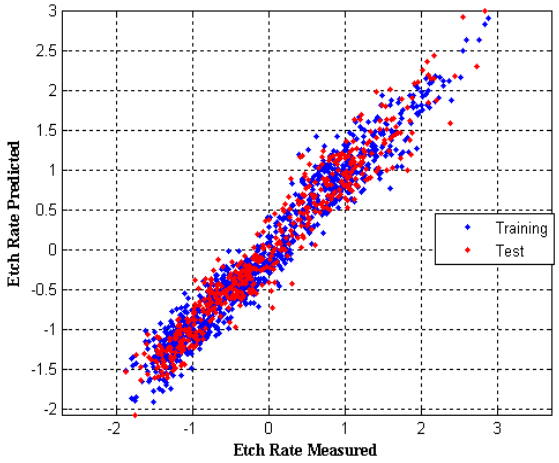

Figure 8: Linear fit of etch rate predicted with etch rate measured in Dataset 2

Figure 9 shows the prediction results of etch rate using FSCA. The performance is better again on Dataset 2 than 1, but overall it is worse than FSR.

A summary of the prediction results of the 4 VM algorithms, each of them computed with 30 principal components or 30 selected variables is shown in Table 5.

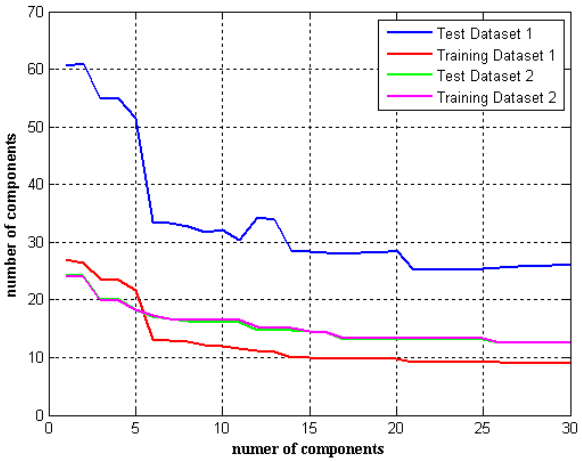

Figure 9: Prediction error for etch rate using FSCA on Dataset 1 and 2.

V. CONCLUSIONS

FSCA/FSR can be viewed as operating at the opposite end of the spectrum to their counterparts PCR/PLS. Whereas PCR/PLS generate a compact representation of the variance in the data by employing linear combinations of *all* input variables, FSCA/FSR seek to employ only a *few* variables to capture the observed variation. Thus, while PCR/PLS are optimal in the sense of maximising the variance explained, FSCA/FSR can often achieve comparable performance when there is a large degree of correlation among the input variables.

This is the case with OES data. Channels around peaks are generally strongly correlated and peaks arising from the same underlying chemistry are also strongly correlated. In PCA/PLS highly correlated variables are given equal weighting in computation of a PC, (the so-called grouping effect) whereas FSCA/FSR choose only one representative variable from each group. Therefore FSCA/FSR are much more effective that PCA/PLS for feature selection.

REFERENCES

[1] Y. J. Chang, Y. Kang, C. L. Hsu, C.T. Chang, and T. Y. Chan. Virtual metrology technique for semiconductor manufacturing. In *Proceedings of IJCNN '06. International Joint Conference on Neural Networks*, 2006.

[2] D. White, B. Goodlin, B. Gower, D. Boning, H. Chen, H. Sawin, and T. Dalton. Low open-area endpoint detection using a pca-based t2 statistic and q statistic on optical emission spectroscopy measurements. *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 2, 2000.

[3] B.M. Wise, N. B. Gallagher, D.D. Butler, S. W White, and G. G. Barna. 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*, vol. 13, no. 3-4, 1999.

[4] E.R. Malinowski. *Factor analysis in chemistry*. John Wiley and Sons, 1991.

[5] P. Geladi and B.R. Kowalski. Partial least squares regression: a tutorial,. *Analytica Chimica Acta*, 1986.
[6] N. Draper and H Smith. *Applied Regression Analysis, 2d Edition*. John Wiley & Sons, 1981.
[7] Leland Wilkinson and Gerard E. Dallal. Tests of significance in forward selection regression with an f-to-enter stopping rule. *Technometrics*, Vol. 23, No. 4, 1981.

Order	Statistics	Channel	Training NMSE	Test NMSE
1	Mean	1805	12.742	34.757
2	Mean	248	9.124	31.922
3	Mean	1418	8.239	24.154
4	Max	1386	7.69	23.043
5	Mean	668	7.452	22.752
6	Mean	658	6.626	26.244
7	Mean	1671	6.339	23.628
8	Skewness	1408	6.19	20.682
9	Skewness	678	5.866	18.681
10	Kurtosis	334	5.492	16.596
15	Mean	789	4.389	15.632
20	Mean	1510	3.743	18.74
25	Kurtosis	1720	3.305	18.415
30	Skewness	530	3.031	16.424

Table 1: FSR variable selection on Dataset 1

Order	Statistics	Channel	Training NMSE	Test NMSE
1	Mean	1766	16.938	16.935
2	Mean	68	12.869	13.306
3	Var	798	11.772	12.407
4	Var	678	11.211	11.994
5	Max	1405	10.73	11.234
6	Var	1915	10.168	10.263
7	Var	691	9.578	9.623
8	Kurtosis	346	9.306	9.644
9	Mean	1926	9.02	9.363
10	Kurtosis	334	8.729	9.089
15	Var	1886	7.391	8.078
20	Var	1733	6.451	6.965
25	Skewness	1879	5.659	6.492
30	Var	299	5.074	5.612

Table 2: FSR variable selection on Dataset 2

Order	Statistics	Channel	Training NMSE	Test NMSE
1	Min	40	27.001	60.583
2	Mean	18	26.496	60.935
3	Max	67	23.537	54.883
4	Mean	3	23.433	54.879
5	Var	526	21.561	51.424
6	Min	20	13.105	33.589
7	Min	33	12.994	33.327
8	Skewness	902	12.731	32.813
9	Min	57	12.253	31.723
10	Mean	698	11.992	32.08
15	Min	46	9.89	8.347
20	Kurtosis	533	9.714	8.536
25	Max	1511	9.144	5.435
30	Skewness	33	9.065	6.085

Table 3: FSCA variable selection on Dataset 1

Order	Statistics	Channel	Training NMSE	Test NMSE
1	Min	45	24.153	23.963
2	Kurtosis	784	24.153	23.962
3	Mean	18	20.16	19.879
4	Max	835	20.145	19.945
5	Max	877	18.238	18.249
6	Max	806	17.062	17.397
7	Min	40	16.671	16.642
8	Mean	1842	16.402	16.414
9	Kurtosis	445	16.157	16.54
10	Max	1824	16.156	16.527
15	Max	90	14.462	14.604
20	Min	2	13.209	13.319
25	Kurtosis	590	13.18	13.31
30	Skewness	6	12.494	12.648

Table 4: FSCA variable selection on Dataset 2

Regression Method	NMSE Training Dataset1	NMSE Training Dataset2	NMSE Test Dataset1	NMSE Test Dataset2
PCR	7.05	8.40	20.57	9.33
PLS	1.07	1.37	37.20	17.12
FSCA	9.06	12.49	25.91	12.65
FSR	3.03	5.07	16.42	5.61

Table 5: prediction error for the 4 VM algorithms (30 principal components or 30 selected variables).