Proceedings of International Workshop on Data Mining and Artificial Intelligence (DMAI' 08)
24 December, 2008, Khulna, Bangladesh

Selecting Signature Optical Emission Spectroscopy Variables Using Sparse Principal Component Analysis

Beibei Ma, Seán McLoone, John Ringwood
Department of Electronic Engineering
National University of Ireland, Maynooth, Ireland
beibei.ma, sean.mcloone, john.ringwood@eeng.nuim.ie

Niall Macgearailt
Department of Electronic Engineering
Dublin City University, Ireland
niall.macgearailt2@mail.dcu.ie

Abstract

Principal component analysis (PCA) is a widely used technique in optical emission spectroscopy (OES) sensor data analysis for the low dimension representation of high dimensional datasets. While PCA produces a linear combination of all the variables in each loading, sparse principal component analysis (SPCA) focuses on using a subset of variables in each loading. Therefore, SPCA can be used as a key variable selection technique. This paper shows that, using SPCA to analyze 2046 variable OES data sets, the number of selected variables can be traded off against variance explained to identifying a subset of key wavelengths, with an acceptable level of variance explained. SPCA-related issues such as selection of the tuning parameter and the grouping effect are discussed.

1. Introduction

Principal component analysis (PCA) is widely known as a dimension reduction technique. Using PCA, a high dimensional data set can be decomposed into the sum of a small number of principal components (PCs). Given an $n \times m$ data matrix $\mathbf{X}$, n being the number of observations and m being the number of variables, there is [3]

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \ (\mathbf{T} \in \mathbb{R}^{n \times p}, \mathbf{P} \in \mathbb{R}^{m \times p}), \qquad (1)$$

where $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_p$. $\mathbf{T}$ and $\mathbf{P}$ are referred to as score and loading matrices, respectively. Since the loading of each PC is generally a linear combination of *all* the original variables, PCA cannot be used directly for variable selection.

Research on obtaining sparse components has been conducted for over two decades. The earliest method, proposed in 1958, is referred to as varimax [7]. Using varimax rotation, a number of the coefficients of the loading vectors can be adjusted to have greater values than the remaining coefficients. Such adjustment can help in the selection of key variables, but it is hard to quantify the distinction between small and large coefficients.

Jeffers [5] proposed a straight-forward method for achieving PCA sparsity. For each loading, any coefficients that are less than 70% of the greatest one are set to zero, regardless of their sign. This method can lead to a selection deficiency in two cases, one where the variables have small coefficients and the other where the variables have high mutual correlations [1].

In [10], the 'simple principal components' is proposed. This focuses on restricting the coefficients of the loadings to have integer values, such as -1,0 and 1, to help simplify variable selection.

The first true algorithmic method for achieving sparse loadings was proposed in 2003 by Jolliffe *et al.* [6] and is known as SCoTLASS (Simplified Component Technique for Least Absolute Shrinkage and Selection). This employs a penalty term referred to as the Least Absolute Shrinkage and Selection Operator (LASSO) [9] to force loadings to be sparse. Nevertheless, it is not practical due to the relatively high computational cost [12].

A recently proposed algorithm, known as semidefinite programming, is described in [2]. Using this method, the normal loadings are constrained by a cardinality condition, that is, a limit on the number of the nonzero elements in each loading. By relaxing this constraint, the problem is converted into a convex optimization problem and hence, can use semidefinite programming as a solution. The generated PCs are shown to be able to explain larger variance than competing algorithms, but the computational cost is high.

In 2004 Zou *et al.* [12] proposed an alternative approach to solving the SPCA problem, which they referred to as elastic net for SPCA (EN-SPCA). EN-SPCA can be implemented in two forms. One is similar to an approach used to solving the LASSO problem, details of which will be provided in the next section, and the other is the so called soft thresholding algorithm, designed for handling large data sets (thousands of variables). Both EN-SPCA

1-4244-2136-7/08/$20.00 ©2008 IEEE

implementations are computational alternatives to semidefinite programming, but the later implementation has the key advantage that it can scale to much larger problems that the semidefinite programming algorithm.

This paper explores the application of SPCA to key variable selection in optical emission spectroscopy (OES). OES is an optical emission detection technique, used for detecting the optical intensity of the chemicals in a plasma as a function of wavelength and time. Because the wavelengths are the 'fingerprints' of the corresponding chemical species, OES data can be used to trace the chemical reactions in a plasma chamber. Consequently, OES is increasingly being used by semiconductor manufacturing to assist with the monitoring and control of plasma etch processes. However, in practice direct use of OES spectra is limited due to the difficulties with handling and interpreting the large number of variables that are associated with such spectra.

The reminder of the paper is organized as follows. Section 2 gives a theoretical description of SPCA. Section 3 describes the methods used to estimate SPCA model accuracy. Section 4 shows experimental results on the application of SPCA to Optical Emission Spectroscopy (OES) data from a semiconductor manufacturing plasma etch process. Finally, the conclusions are presented in Section 5.

2 Theoretical Framework and Numerical Solution

EN-SPCA employs a penalised regression estimator to solve the SPCA problem. This exploits the fact that PCA can be formulated as a least squares regression problem, thereby facilitating the inclusion of the LASSO penalty, which is known from regression theory to yield sparse solutions. The theory of LASSO and EN-SPCA are quite involved, hence only a brief overview will be provided here. For a more complete development of the algorithms see [9],[12] and [11].

2.1 PCA as a LS problem

Given a data matrix $\mathbf{X}$ (as defined in Eq. (1)) and defining its transpose as the matrix $\mathbf{Z} = \mathbf{X}^{\mathrm{T}}$, the LS estimation of the PCA loading matrix $\mathbf{P} \in \mathbb{R}^{m \times p}$ can be expressed as

$$(\widehat{\mathbf{A}}, \widehat{\mathbf{P}}) = \arg\min_{\mathbf{A},\mathbf{P}} \{\sum_{i=1}^{n} |\mathbf{z}_i - \mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{z}_i|^2\}, \quad \text{s.t. } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{2}$$

Where $\mathbf{z}_i$ are the columns of $\mathbf{Z}$. The constraint on matrix $\mathbf{A}$ ensures that the orthogonality of $\mathbf{P}$ is guaranteed. While theoretically the LS estimate is the best unbiased estimate of $\mathbf{P}$ [9], in practice better mean square error performance can by obtained by biasing the regression coefficients towards zero. This is typically achieved by adding either an L_1 or an L_2 norm penalty to the LS cost function. The L_2 implementation, referred to as Ridge regression or regularisation, is frequently employed to address data ill-conditioning and singularity issues and benefits from a straightforward algebraic solution. In contrast, the L_1 implementation has traditionally been avoided due to the associated computational issues, but in recent years it has been receiving increasing attention as a variable selection method, due to its tendency to yield sparse solutions. In this context it is referred to as LASSO [9].

2.2 Ridge Estimation

The Ridge estimate of $\mathbf{A}$ and $\mathbf{P}$ (denoted by $\widehat{\mathbf{A}}^{\mathrm{R}}$ and $\widehat{\mathbf{P}}^{\mathrm{R}}$) is expressed as

$$(\widehat{\mathbf{A}}^{\mathrm{R}}, \widehat{\mathbf{P}}^{\mathrm{R}}) = \arg\min_{\mathbf{A},\mathbf{P}} \{\sum_{i=1}^{n} ||\mathbf{z}_i - \mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{z}_i||_2^2 + \gamma_2 \sum_{j=1}^{p} ||\mathbf{p}_j||_2^2\}, \quad \text{s.t. } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{3}$$

where $||\cdot||_2$ represents the L_2 norm, γ_2 is the Ridge tuning parameter ($\gamma_2 \geq 0$) and $\mathbf{p}_j$ is the jth column vector of $\mathbf{P}^{\mathrm{R}}$. The main benefit of employing the Ridge estimator is that it can handle data matrices in which the number of observations is less than the number of variables. Unlike regression this comes at no cost since the Ridge estimate of $\mathbf{P}$ is simply a scaled version of the LS estimate and scaling does not affect the PCA decomposition.

2.3 LASSO Estimation

Mathematically, the LASSO estimate of $\mathbf{A}$ and $\mathbf{P}$, denoted by $\widehat{\mathbf{A}}^{\mathrm{L}}$ and $\widehat{\mathbf{P}}^{\mathrm{L}}$ can be expressed as

$$(\widehat{\mathbf{A}}^{\mathrm{L}}, \widehat{\mathbf{P}}^{\mathrm{L}}) = \arg\min_{\mathbf{A},\mathbf{P}} \{\sum_{i=1}^{n} ||\mathbf{z}_i - \mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{z}_i||_2^2 + \gamma_1 \sum_{j=1}^{p} ||\mathbf{p}_j||_1\}, \quad \text{s.t. } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{4}$$

or equivalently as

$$(\widehat{\mathbf{A}}^{\mathrm{L}}, \widehat{\mathbf{P}}^{\mathrm{L}}) = \arg\min_{\mathbf{A},\mathbf{P}} \{\sum_{i=1}^{n} ||\mathbf{z}_i - \mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{z}_i||_2^2\}, \quad \text{s. t. } \sum_{j=1}^{p} ||\mathbf{p}_j||_1 \leq c_1 \text{ and } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{5}$$

where $||\mathbf{p}_j||_1 = \sum_{i=1}^{m} |b_{ij}|$, γ_1 is the LASSO tuning parameter and c_1 is a corresponding upper bound. For each γ_1,

there exists a c_1 that gives an equivalent constraint on the regression coefficients. The second formulation (Eq. (5)) has the attraction that it is in the form of a quadratic programming problem with linear inequality constraints making its solution mathematically tractable.

Although the Ridge and LASSO penalty both cause the regression coefficients to shrink towards zero a powerful feature of the LASSO is that it is more likely to drive coefficients to exactly zero, hence generating sparse solutions. This arises because of the distinctive shape of the LASSO penalty as illustrated in Fig.1 for the 2-D case. This shows the elliptical contours of a LS quadratic cost function with the feasible region corresponding to the L_1 (LASSO) constraint indicated in grey. The L_1 norm produces a diamond shaped region with corners aligned with the axes. In contrast, as illustrated in Fig.1(b), the L_2 norm generates a circular boundary. Increasing the constraint penalty causes the feasible regions to shrink towards the origin. As this happens there is a tendency for the optimum elliptical contour to intersect the diamond region at the corners and since these are aligned with the axes this leads to sparse solutions. This is not the case with the circular boundary of the Ridge penalty.

2.4 Naive Elastic Net and Elastic Net

The drawback of LASSO is that if the number of observations (n) is less than the number of the variables (m), LASSO will at most choose n variables. In contrast, Ridge can extend the selection to m variables. However, Ridge cannot provide sparse solutions. The naive elastic net method was developed to address these problems. Mathematically, the naive elastic net estimate is defined as:

$$(\widehat{\mathbf{A}}^{\mathrm{N}}, \widehat{\mathbf{P}}^{\mathrm{N}}) = \arg\min_{\mathbf{A},\mathbf{P}} \{ \sum_{i=1}^{n} |\mathbf{z}_i - \mathbf{A}\mathbf{P}^{\mathrm{T}}\mathbf{z}_i|^2 + \gamma_2 \sum_{j=1}^{k} ||\mathbf{p}_j||_2^2 + \sum_{j=1}^{k} \gamma_{1j} ||\mathbf{p}_j||_1 \}, \text{ s.t. } \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}, \tag{6}$$

where $\gamma_{1j}, j{=}1 \ldots k$ are scalars used to individually penalize the loadings.

In [11], Zou and Hastie argue that naive elastic net estimation introduces a double bias, but that the reduction in estimation variance is no greater than with LASSO or Ridge estimation alone. To correct for this double shrinkage, the authors propose re-scaling the naive elastic net estimate by a factor $(1 + \gamma_2)$, giving the final elastic net SPCA estimate as:

$$(\widehat{\mathbf{A}}^{\mathrm{EN}}, \widehat{\mathbf{P}}^{\mathrm{EN}}) = (1 + \gamma_2)(\widehat{\mathbf{A}}^{\mathrm{N}}, \widehat{\mathbf{P}}^{\mathrm{N}}). \tag{7}$$

Because the constraint function $\gamma_2 ||\mathbf{p}_j||_2^2 + \gamma_{1j} ||\mathbf{p}_j||_1$ is strictly convex, elastic net estimation retains a special property of least squares regression known as the grouping effect [12]. This means that highly correlated variables will be assigned similar regression coefficients. While this property can be beneficial in many applications, as will be demonstrated later, it is not ideal for variable selection.

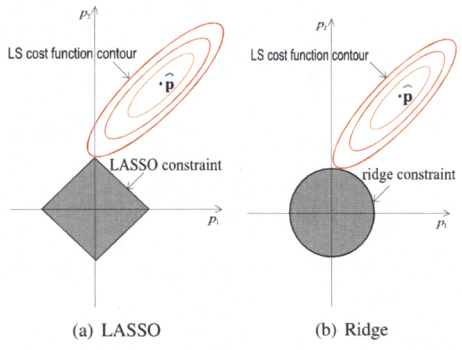

Figure 1. Solution for different estimators

2.5 Numerical Solution

Simple matrix transformations can be used to estimate the least squares or Ridge regression estimates of $\mathbf{P}$, but efficient computation of the LASSO estimate is much more challenging, and was only effectively addressed with the development of the least angle regression (LARS) algorithm in 2002 [8] (S refers to its close relation to LASSO and stagewise regression).

In LARS, all the coefficients are first set to zero and then revised successively until the least squares solution is reached. LARS-EN is a modified implementation of LARS designed to fit the elastic net framework [8]. In particular, it includes a special algorithm, called soft-thresholding, that is scalable to high dimensional data sets [12].

3 Estimating Model Accuracy

3.1 Variance Explained

SPCA employs a similar approach to PCA, known as adjusted variance [12], to measure the estimation accuracy. Like PCA, the sparse scores ($\widehat{\mathbf{T}}^{\mathrm{S}}$) are used to calculate the variance as:

$$\widehat{\mathbf{T}}^{\mathrm{S}} = \mathbf{X}\underline{\widehat{\mathbf{P}}}^{\mathrm{EN}}, \tag{8}$$

where $\underline{\widehat{\mathbf{P}}}^{\mathrm{EN}}$ denotes $\widehat{\mathbf{P}}^{\mathrm{EN}}$ normalized to have unit length columns. However, unlike PCA, $\widehat{\mathbf{P}}^{\mathrm{EN}}$ is not orthogonal, hence the variances explained by individual PCs are not independent of each other and hence not additive. To correct for this the sparse scores matrix must first be orthogonalised leading to the following variance estimation algorithm [4, 8].

- Orthogonalize $\widehat{\mathbf{t}}_j^{\,S}$ (the jth column vector of $\widehat{\mathbf{T}}^S$) by applying the recursion

$$\widehat{\mathbf{t}}_j^{\,S*} = \widehat{\mathbf{t}}_j^{\,S} - \widehat{\mathbf{T}}_{(j-1)}^S[(\widehat{\mathbf{T}}_{(j-1)}^S)^{\mathrm{T}}(\widehat{\mathbf{T}}_{(j-1)}^S)]^{-1}(\widehat{\mathbf{T}}_{(j-1)}^S)^{\mathrm{T}}\widehat{\mathbf{t}}_j^{\,S},$$

for $j = 1, \ldots, p$ (p is the number of sparse principal components), where $\widehat{\mathbf{T}}_{(j)}^S = [\widehat{\mathbf{t}}_{(1)}^S, \ldots, \widehat{\mathbf{t}}_{(j)}^S]$.

- Collect the orthogonalized vectors into a matrix $\widehat{\mathbf{T}}^{S*}$, *i.e.*

$$\widehat{\mathbf{T}}^{S*} = [\widehat{\mathbf{t}}_1^{\,S*}, \ldots, \widehat{\mathbf{t}}_j^{\,S*} \ldots, \widehat{\mathbf{t}}_p^{\,S*}]. \quad (9)$$

- Compute the variance explained (V_e) by the p sparse components as

$$V_e = \mathrm{trace}\{(\widehat{\mathbf{T}}^{S*})^{\mathrm{T}}\widehat{\mathbf{T}}^{S*}\}. \quad (10)$$

The jth diagonal entry of $(\widehat{\mathbf{T}}^{S*})^{\mathrm{T}}\widehat{\mathbf{T}}^{S*}$ corresponds to the variance explained by the jth sparse PC.

3.2 SPMSE

When using sparse PCs to reconstruct a data set, many columns of the reconstructed data are in fact zero, because of the zero elements in the sparse loadings. Therefore, a fairer assessment of the accuracy of reconstruction is to only compare the reconstruction against the original data over the regions where the reconstruction exists. Here, a sparse mean square error measure is proposed, denoted SPMSE, where S stands for the sparse, P for percentage and MSE for mean square error. This is given by

$$SPMSE = \frac{||\widehat{\mathbf{X}}_s - \mathbf{X}_s||_f^2}{||\mathbf{X}_s||_f^2} \times 100\%, \quad (11)$$

where $||\cdot||_f$ is the Frobenius norm, $\widehat{\mathbf{X}}_s$ consists of the nonzero columns of the reconstructed data matrix $\widehat{\mathbf{X}}$ and $\mathbf{X}_s$ is the corresponding subset of the original data matrix $\mathbf{X}$. Based on SPMSE, SV_e, variance explained by the sparse components, is proposed as

$$SV_e = 100\% - SPMSE. \quad (12)$$

In contrast to V_e (Eq. 10), SV_e only calculates the variance for the non-zero reconstructed-channels, so SV_e can more effectively reflect the reconstruction accuracy. Note, that since the sparse components are not orthogonal, the reconstruction of $\mathbf{X}$ is defined as

$$\widehat{\mathbf{X}} = \widehat{\mathbf{T}}^S(\widehat{\underline{\mathbf{P}}}^{EN})^{\mathrm{T}}[\widehat{\underline{\mathbf{P}}}^{EN}(\widehat{\underline{\mathbf{P}}}^{EN})^{\mathrm{T}}]^{-1}. \quad (13)$$

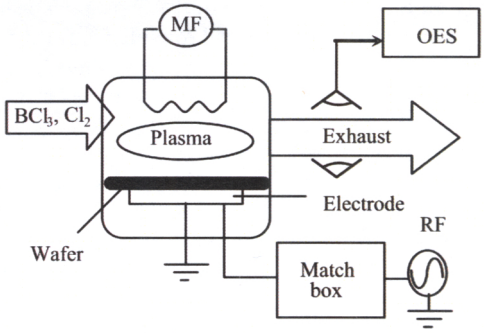

Figure 2. Diagram of OES data collection

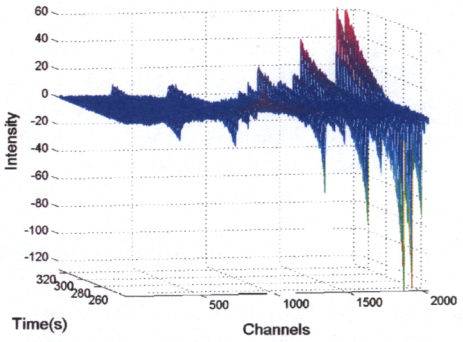

Figure 3. Sample OES data set

4 Results

4.1 Data Description

The OES data under consideration is from a plasma etch chamber used in the manufacture of semiconductor chips (Fig. 2). It is collected for the exhaust plasma leaving the chamber and consists of 2046 channels (each channel corresponds to one wavelength) recorded at a sampling interval of 0.76s. Fig. 3 shows a typical data set collected for a 77s etch step on a single wafer. The data, which has been mean-centred, is clearly highly redundant. A PCA analysis reveals that a single PC can capture 95.8% of the variance in the data.

4.2 Selecting the EN-SPCA Tuning Parameters

As noted previously when the number of variables is large (in this case 2046) the soft-thresholding EN-SPCA algorithm developed in [12] can be used to efficiently compute components. In this formulation the L_2 penalty tuning

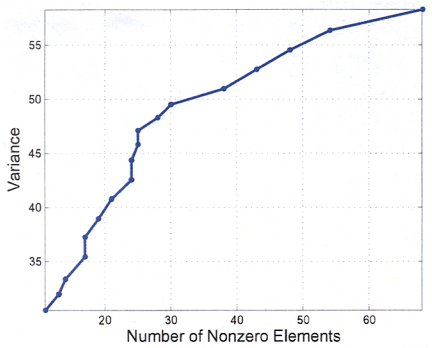

Figure 4. The relationship between N_{NE} and V_e

parameter (γ_2) is set to infinity leaving only the L_1 penalty parameters (γ_{1j} - one for each computed PC, j=1 ...p) to be determined experimentally. These parameters essentially determine the sparseness of the corresponding PCs.

For example, Table 4.2, shows how the number of nonzero elements (N_{NE}) varies as a function of γ_{11}, for the first sparse component of the OES data. The table also shows how the variance explained by the component (expressed as a percentage of the total data variance). As might be expected V_e decreases as more and more components are forced to zero. In contrast, the SV_e value remains large for all values of γ_{11} demonstrating that the sparse PC achieves good accuracy for those channels where a reconstruction exists.

Fig. 4 demonstrates the corresponding relationship between N_{NE} and V_e and provides a useful guide for making a judgment call on the trade-off between sparsity and variance explained. As can be seen, in this instance there is a 'knee' in the graph at $N_{NE} = 25$, beyond which there is a marked decrease in the rate of variance increase with included variables. This corresponds to choosing γ_{11} as 29000.

The variance explained by the sparse PC for $\gamma_{11} = 29000$ is 47% which compares to 95.8% for the unrestricted principal component. The corresponding SV_e value, which measures the reconstruction accuracy on the non-zero reconstructed-channels only, is more than 91%, by the 25 channels used in its computation.

4.3 The Grouping Effect

Figure 5 shows the distribution of the first sparse loading computed with $\gamma_{11} = 290000$. For comparison purposes the distribution of the first PCA component loading is included in Figure 6. The PCA loading elements are all

γ_{11}	N_{NE}	V_e (%)	SV_e (%)
0	2046	95.8057	95.8057
5000	1564	95.6068	96.129
10000	1373	95.0602	95.9379
50000	611	84.1086	89.6853
100000	189	65.0638	85.9841
150000	68	58.2831	90.3285
170000	54	56.3396	90.88
190000	48	54.5187	89.956
210000	43	52.7259	89.0393
230000	38	50.9482	88.6207
250000	30	49.4871	91.1976
270000	28	48.2702	90.5483
290000	**25**	**47.0887**	**91.127**
310000	25	45.8107	88.6538
330000	24	44.3460	86.9806
350000	24	42.5365	83.4315
370000	21	40.7546	84.3206
390000	19	38.9135	84.0687
410000	17	37.2303	84.522
430000	17	35.4131	80.3966
450000	14	33.3417	83.5658
470000	13	31.9432	83.0192
490000	11	30.4860	86.5150

Table 1. N_{NE}, V_e and SV_e, corresponding to different γ_{11} values for the first sparse PC

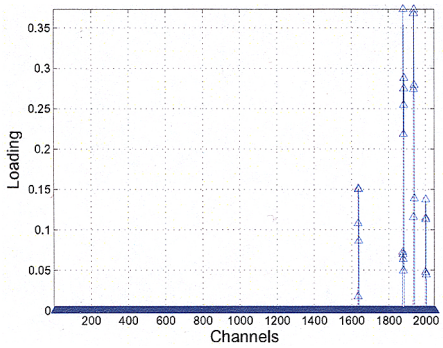

Figure 5. The loading of the first sparse principal component obtained for the sample OES spectra in Fig. 3 using SPCA with $\gamma_{11} = 29000$

non-zero with several clusters of large values centered on the active OES channels. These clusters arise because of the spectral bleed between adjacent channels. In contrast, the SPCA loading has only 25 non-zero entries and these are in four distinct clusters of points. Analysis of these four sets of points show that they are all highly correlated as can be seen in Fig. 7 (correlation coefficient > 0.99). This is a direct consequence of the grouping effect that is a feature of EN-SPCA, that is, EN-SPCA has a tendency to give equal weighting to strongly correlated variables and, as such, selects all the correlated variables as a group, rather than selecting a single representative example. This is useful when trying to identify groups of related variables, but is not ideal for a variable selection algorithm.

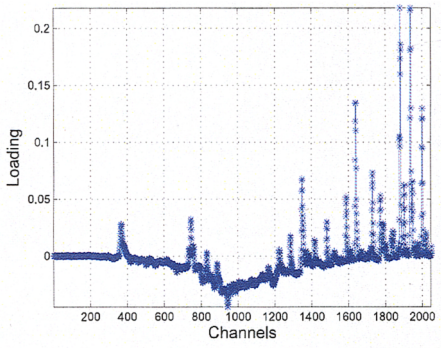

Figure 6. The loading of the first principal component obtained for the sample OES data in Fig. 3 using PCA

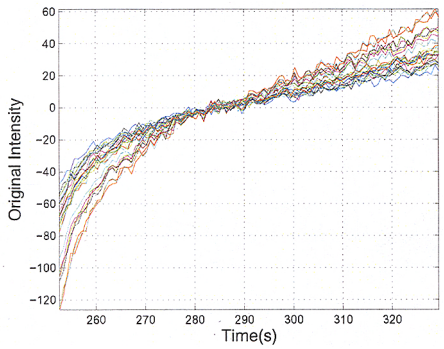

Figure 7. Intensity changes of the nonzero loadings (over time) for the first sparse PC of the OES data

5 Conclusions

This paper introduces SPCA as a variable selection tool for the identification of key variables in large data sets. SPMSE has been proposed as a measure that better reflects the estimation accuracy of SPCA, given the sparse structure of the model. Using analysis of OES data from a plasma etch chamber the main features of SPCA have been illustrated, particularly in relation to how it provides a trade-off between variance explained and a sparse representation. The existence of a grouping effect in the selection of variables has also been highlighted as a weakness of the method.

Acknowledgments.

The authors gratefully acknowledge the financial support of Enterprise Ireland (grant IP/2006/0325).

References

[1] J. Cadima and I. T. Jolliffe. Loadings and Correlations in the Interpretation of Principal Components. *Journal of Applied Statistics*, 22(2):203–214, 1995.

[2] A. D'Aspremont, L. Ghaoui, M. I. Jordan, and G. Lanckriet. A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, 49(3):434–448, 2007.

[3] P. Geladi and B. R. Kowalski. Partial Least-squares Regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[4] D. Gervini and V. Rousson. Criteria for Evaluating Dimension-reducing Components for Multivariate Data. *American Statistician*, 58(1):72–76, 2004.

[5] J. N. R. Jeffers. Two Case Studies in the Application of Principal Component Aanlysis. *Applied Statistics*, 16(3):225–236, 1967.

[6] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[7] F. H. Kaiser. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 23:187–200, 1958.

[8] K. Sjöstrand, M. B. Stegmann, and R. Larsen. Sparse Principal Component Analysis in Medical Shape Modeling. In *Proceedings of SPIE*, volume 6144, page 61444X, March 2006.

[9] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of Royal Statistical Society B*, 58(1):267–288, 1996.

[10] S. K. Vines. Simple Principal Components. *Applied Statistics*, 49(4):441–451, 2000.

[11] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of Royal Statistical Society B*, 67(2):301–320, 2005.

[12] H. Zou, T. Hastie, and R. Tibshirani. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.