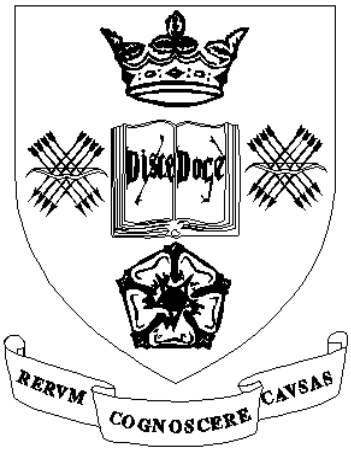
A Multiple Sequential Orthogonal Least Squares Algorithm for Feature Ranking and Subset Selection

S. A. Billings and H. L. Wei



Research Report No. 908

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street, Sheffield,
S1 3JD, UK

October 2005

A Multiple Sequential Orthogonal Least Squares Algorithm for Feature Ranking and Subset Selection

Stephen A. Billings and Hua-Liang Wei

Department of Automatic Control and Systems Engineering, University of Sheffield
Mappin Street, Sheffield, S1 3JD, UK

S.Billings@Sheffield.ac.uk, W.Hualiang@Sheffield.ac.uk

Abstract: High-dimensional data analysis involving a large number of variables or features is commonly encountered in multiple regression and multivariate pattern recognition. It has been noted that in many cases not all the original variables are necessary for characterizing the overall features. More often only a subset of a small number of significant variables is required. The detection of significant variables from a library consisting of all the original variables is therefore a key and challenging step for dimensionality reduction. Principal component analysis is a useful tool for dimensionality reduction. Principal components, however, suffer from two main deficiencies: Principal components always involve all the original variables and are usually difficult to physically interpret. This study introduces a new multiple sequential orthogonal least squares algorithm for feature ranking and subset selection. The new method detects in a stepwise way the capability of each candidate feature to recover the first few principal components. At each step, only the significant variable with the strongest capability to represent the first few principal components is selected. Unlike principal components, which carry no clear physical meanings, features selected by the new method preserve the original measurement meanings.

Keywords: Dimensionality reduction, high-dimensional data analysis, subset selection, principal component analysis, orthogonal least squares.

1. Introduction

Multivariate data analysis such as multiple regression and high-dimensional pattern classification, often involves a large number of variables or features. Quite often the sample features are unknown a priori and measurements are obtained with respect to more variables than is strictly necessary for conveying the main features. There is, therefore, the potential to greatly reduce the dimensionality without distorting the overall features. As one commonly used dimensionality reduction approach, subset selection aims to find a small number of significant variables or features from a library consisting of all the original variables. The remaining insignificant variables are, in a sense, irrelative or redundant, and can be ignored. In fact, the inclusion of these insignificant variables may often complicate data inspection without providing any extra information [Jolliffe 1972].

A large amount of work has been done on dimensionality reduction, see for example, [Oja 1983, Jain et al. 2000, Carreira-Perpinan 2001, Fodor 2002, Webb 2002]. Principal component analysis (PCA) and its variants [Jolliffe 2002] belong to the class of the most commonly used methods for dimensionality reduction. Note, however, that principal components (PCs) suffer from two main drawbacks: PCs are transformed variables that always involve all the original variables, and PCs are usually difficult to physically interpret. In many cases there may be some redundancy, linear correlation or linear dependency among the original variables. The redundant or irrelevant variables need not to be included in PCs and the exclusion of these insignificant variables could sake

to save a great amount of time and cost spent on obtaining and measuring these insignificant variables in future experiments. Bearing this fact in mind, efforts have been made to develop algorithms to select significant variables or eliminate insignificant variables from the full set of original variables, to form an effective subset that can be used to sufficiently recover the main information conveyed by the full data set [Jolliffe 1972, 1973, Krzanowski 1987, Pudil et al. 1994, Kohavi and John 1997, Mao 2005]. In fact, in many cases it is desirable to reduce not only the dimensionality in the transformed space, but also the number of variables that need to be considered or measured in the future in the measurement space [McCabe 1984].

This study introduces a new method for ranking significant variables and selecting a subset from a library consisting of all the original variables. In the new method, a general variable detection and subset selection problem is initially converted into a multivariate regression problem by treating principal components as the dependent variables (responses) and the original variables as the independent variables (predictors or explanatory variables). A new multiple sequential orthogonal least squares (MSOLS) algorithm is derived to detect the significant variables for multivariate regression. The main idea behind the new method is to detect, in a stepwise way, the significance of each candidate variable to represent the first few PCs. At each step only the variable with the strongest capability to present the first few PCs is selected and included in the subset.

The paper is organized as follows. In Section 2 the multiple sequential orthogonal least squares method is proposed in detail. In Section 3, the variable detection problem from principal components is converted into a subset selection problem for multiple regression, for which significant variables can easily be detected using the MSOLS algorithm. In Section 4, three examples relating to both artificial and real data sets are provided to illustrate the application of the new method for feature ranking and subset selection. The work is briefly concluded in Section 5.

2. The MSOLS algorithm

This section presents an orthogonal least squares type algorithm for variable detection and subset selection for a multiple linear regression. The models that are considered throughout the paper will be restricted to the linear in the regression coefficients form. The learning algorithm is multiple-regression oriented and will be implemented using a sequential orthogonal least squares method. This new learning scheme will thus be referred to as the multiple sequential orthogonal least squares (MSOLS) algorithm.

2.1 The multiple regression model

Denote the multiple response variables by $y_{1,}y_{2,}\cdots y_m$ and the set of n predictor (or independent) variables by $x_1, x_2, \cdots, x_n$. Assume that the relationship between y_i (i=1,2, …,m) and x_j (j=1,2, …,n) can be approximated by the linear-in-the-parameters regression model

$$y_i(k) = \beta_{i,0} + \sum_{j=1}^{n} \beta_{i,j} x_j(k) + e_i(k) \tag{1}$$

where $y_i(k)$ denotes the kth observation of the ith response variable and $x_j(k)$ denotes the kth observation of the jth predictor variable (i=1,2, …, m; j=1,2, …, n; k=1,2, …, N); $e_i(k)$ is assumed to be an unobservable error representing the discrepancy in the approximation; $\beta_{i,0}, \beta_{i,1}, \cdots, \beta_{i,n}$, called the regression parameters or

coefficients, are unknown constants that need to be estimated from the data. The linear model (1) can be expressed using a compact form

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{e}_i \tag{2}$$

where $\mathbf{y}_i = [y_i(1), y_i(2), \cdots, y_i(N)]^T$, $\boldsymbol{\beta}_i = [\beta_{i,0}, \beta_{i,1}, \cdots, \beta_{i,n}]^T$, $\mathbf{e}_i = [e_i(1), e_i(2), \cdots, e_i(N)]^T$ and $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n]$ with $\mathbf{x}_0 = [1,1,\cdots,1]^T$ and $\mathbf{x}_j = [x_j(1), x_j(2), \cdots, x_j(N)]^T$ for j=1,2,…, n.

The m matrix equations given by (2) can be put together to form an mth block-structured matrix equation below

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_m \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} & \mathrm{O} & \cdots & \mathrm{O} \\ \mathrm{O} & \mathbf{X} & \cdots & \mathrm{O} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \mathbf{X} \end{pmatrix} = \mathbf{I}_m \otimes \mathbf{X}$$

where $\mathbf{O}$ is a $N \times m(n+1)$ matrix whose all entries are zeros, the notation $\mathbf{I}_m$ denotes the mth order unit matrix, and the symbol ' $\otimes$ ' denotes the Kronecker product, which is defined for two matrix $\mathbf{A} = (a_{i,j})_{r\times s}$ and $\mathbf{B} = (b_{i,j})_{u\times v}$ as

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \cdots & a_{1,s}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \cdots & a_{2,s}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{r,1}\mathbf{B} & a_{r,2}\mathbf{B} & \cdots & a_{r,s}\mathbf{B} \end{bmatrix}$$

2.2 Subset selection for the multiple regression model

In many cases, the full set of the original predictor variables $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n$ may be redundant for representing the response variables, $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_m$, because of linear correlation and dependency. Furthermore, it is not necessary that all the n predictor variables are of the same importance for representing the response variables. The objective of subset selection is to find a subset $S_d = \{\mathbf{z}_1, \cdots, \mathbf{z}_d\} = \{\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_d}\} \subseteq S = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n\}$, where $\mathbf{z}_k = \mathbf{x}_{i_k}$, $i_k \in \{0,1,\cdots,n\}$ (k=1,2, …, d), so that $\mathbf{y}_i$ (i=1,2,…,m) can be satisfactorily approximated using a linear combination of $\mathbf{z}_1, \cdots, \mathbf{z}_d$ as below

$$\mathbf{y}_i = \theta_{i,1}\mathbf{z}_1 + \cdots + \theta_{i,d}\mathbf{z}_d + \mathrm{e}_i \tag{4}$$

or, in a compact matrix form

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\theta}_i + \mathrm{e}_i \tag{5}$$

where the matrix $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_d]$ is of full column rank, $\boldsymbol{\theta}_i = [\theta_{i,1}, \cdots, \theta_{i,d}]^T$ is a parameter vector, and $\mathbf{e}_i$ is an approximation error vector.

To select a subset for the multiple regression model, the significance of each variable needs to be detected initially by inspecting the capability of each predictor variable to represent all the response variables. A subset can then be determined by selecting the significant predictor variables or eliminating the insignificant predictor variables. In the following, a new multiple sequential orthogonal least squares (MSOLS) method is proposed for significant variable detection and subset selection.

2.3 The MSOLS algorithm for subset selection

The MSOLS algorithm starts from the *m*th order block-structured matrix equation (3). Let $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_{m(n+1)}$ be the column vectors of the matrix $\tilde{\mathbf{X}}$. Define

$$\mathrm{ERR}^{(1)}[j] = \frac{(\mathbf{y}^T \boldsymbol{\gamma}_j)^2}{(\mathbf{y}^T \mathbf{y})(\boldsymbol{\gamma}_j^T \boldsymbol{\gamma}_j)}, \quad j=1,2,\ldots, m(n+1), \tag{6}$$

$$\ell_1 = \arg \max_{1 \le j \le m(n+1)} \{\mathrm{ERR}^{(1)}[j]\} \tag{7}$$

$$\ell_1^0 = \mathrm{mod}(\ell_1, n+1) \tag{8}$$

where the function '$\mathrm{mod}(\cdot,\cdot)$' is defined as the modulus after division and thus $0 \le \ell_1 \le n$. The meaning of $\mathrm{ERR}^{(1)}[j]$ will be explained later. The first significant variable can then be selected as $\mathbf{z}_1 = \mathbf{x}_{\ell_1^0}$, and the associated orthogonal variable can be chosen as $\mathbf{q}_1 = \boldsymbol{\gamma}_{\ell_1}$.

Assume that a subset S_{r-1}, consisting of (*r*-1) significant variables, $\mathbf{z}_1, \cdots, \mathbf{z}_{r-1}$, has been determined at step (*r*-1), and the (*r*-1) selected variables have been transformed into a new group of orthogonalized variables $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{r-1}$ via some orthogonal transformation. To select the *r*th significant variable $\mathbf{z}_r$, let $\boldsymbol{\alpha}_{j,r} = \boldsymbol{\gamma}_j$ for j=1,2,…,$m(n$+1) and $j \notin \Gamma_r = \{\ell : \ell \text{ is multiple of } \ell_k^0 \text{ for } k = 1,2,\cdots, r-1\}$. Orthogonalize $\boldsymbol{\alpha}_{j,r}$ with $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{r-1}$ as below

$$\mathbf{q}_{j,r} = \boldsymbol{\alpha}_{j,r} - \sum_{k=1}^{r-1} \frac{\boldsymbol{\alpha}_{j,k}^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k \tag{9}$$

Following [Korenberg et al. 1988, Billings et al. 1989, Chen et al. 1989], the *error reduction ratio*, ERR[*j*;*r*], produced by adding the *j*th data vector $\boldsymbol{\alpha}_{j,r} = \boldsymbol{\gamma}_j$ into S_{r-1} for representing the overall response $\mathbf{y}$ is defined as

$$\mathrm{ERR}^{(r)}[j] = \frac{(\mathbf{y}^T \mathbf{q}_{j,r})^2}{(\mathbf{y}^T \mathbf{y})(\mathbf{q}_{j,r}^T \mathbf{q}_{j,r})} \tag{10}$$

Let

$$\ell_r = \arg \max_{1 \le j \le m(n+1)} \{\mathrm{ERR}^{(r)}[j]\} \tag{11}$$

$$\ell_r^0 = \mathrm{mod}(\ell_r, n+1) \tag{12}$$

The *r*th significant variable can then be chosen as $\mathbf{z}_r = \mathbf{x}_{\ell_r^0}$, and the associated orthogonal variable can be chosen as $\mathbf{q}_r = \mathbf{q}_{\ell_r, r}$. Subsequent significant variables can be detected in the same way step by step. At each step, the 'best' variable that accounts for the variation of the overall response $\mathbf{y}$ with the highest percentage is selected.

The above variable detection procedure involves sequential orthogonal transformations. For example, at step *r*, where a subset S_r of *r* significant variables, $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_r$, has been obtained. These *r* variables can be used to approximate the overall response $\mathbf{y}$ with a linear form

$$\mathbf{y} = \mathbf{P}^{(r)}\boldsymbol{\theta}^{(r)} + \mathbf{e}^{(r)} \tag{13}$$

where $\mathbf{P}^{(r)} = [\boldsymbol{\gamma}_{\ell_1}, \boldsymbol{\gamma}_{\ell_2}, \cdots, \boldsymbol{\gamma}_{\ell_r}]$, $\boldsymbol{\theta}^{(r)}$ is the regression parameter vector, $\mathbf{e}^{(r)}$ is the approximation error vector. From the above variable selection procedure, the full rank matrix $\mathbf{P}^{(r)}$ can be orthogonally decomposed as

$$\mathbf{P}^{(r)} = \mathbf{Q}^{(r)}\mathbf{R}^{(r)} \tag{14}$$

where $\mathbf{R}^{(r)}$ is a $r \times r$ unit upper triangular matrix and $\mathbf{Q}^{(r)}$ is an $r \times r$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_r$. Substituting (14) into (13), yields

$$\mathbf{y} = [\mathbf{P}^{(r)}(\mathbf{R}^{(r)})^{-1}][\mathbf{R}^{(r)}\boldsymbol{\theta}^{(r)}] + \mathbf{e}^{(r)} = \mathbf{Q}^{(r)}\mathbf{g}^{(r)} + \mathbf{e}^{(r)} \tag{15}$$

where $\mathbf{g}^{(r)} = [g_1^{(r)}, \cdots, g_r^{(r)}]^T = \mathbf{R}^{(r)}\boldsymbol{\theta}^{(r)}$ is an auxiliary parameter vector. Using the orthogonal property of $\mathbf{Q}^{(r)}$, $g_k^{(r)}$ can be directly calculated from $\mathbf{y}$ and $\mathbf{Q}^{(r)}$ as $g_k^{(r)} = (\mathbf{y}^T\mathbf{q}_k)/(\mathbf{q}_k^T\mathbf{q}_k)$ for *k*=1,2, …,*r*. The unknown parameter vector $\boldsymbol{\theta}^{(r)}$ can then be easily calculated from $\mathbf{g}^{(r)}$ and $\mathbf{R}^{(r)}$ by substitution using the special structure of $\mathbf{R}^{(r)}$.

The total sum of squares of the overall response $\mathbf{y}$ from the origin can then be expressed as

$$\mathbf{y}^T\mathbf{y} = \sum_{k=1}^{r}(g_k^{(r)})^2\mathbf{q}_k^T\mathbf{q}_k + (\mathbf{e}^{(r)})^T\mathbf{e}^{(r)} \tag{16}$$

Note that the total sum of squares $\mathbf{y}^T\mathbf{y}$ consists of two parts, the desired output $\sum_{k=1}^{r}(g_k^{(r)})^2\mathbf{q}_k^T\mathbf{q}_k$, which can be explained by the selected variables, and the part $(\mathbf{e}^{(r)})^T\mathbf{e}^{(r)}$, which represents the residual sum of squares. Thus, $(g_k^{(r)})^2\mathbf{q}_k^T\mathbf{q}_k$ is the increment to the desired total sum of squares of the output brought by $\mathbf{q}_k$. The *k*th error reduction ratio (ERR) introduced by $\mathbf{q}_k$ (or equivalently by $\mathbf{z}_k$), is defined as [Korenberg et al. 1988, Billings et al. 1989, Chen et al. 1989]

$$\mathrm{ERR}^{(r)}[k] = \frac{(\mathbf{y}^T\mathbf{q}_k)^2}{(\mathbf{y}^T\mathbf{y})(\mathbf{q}_k^T\mathbf{q}_k)}, \quad k=1,2, \ldots, r, \tag{17}$$

This ratio provides a simple but an effective index to indicate the significance of adding the *k*th variable into the model. The orthogonalization procedure for variable selection is usually implemented in a stepwise way, one variable at a time. The *sum of error reduction ratio* (SERR) [Wei et al. 2004] due to $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_r$ is defined as

$$\mathrm{SERR}[r] = \sum_{k=1}^{r}\mathrm{ERR}^{(r)}[k] \tag{18}$$

The selection procedure can be terminated at any time when SERR[*r*] satisfies some specified conditions, say when SERR[*r*] is greater than a given threshold. Otherwise, the variable detection procedure needs to be

continued and new significant variables need to be added to the subset, until the specified conditions are met. The pseudo-code of the MSOLS algorithm developed on the basis of the forward Gram-Schmidt transformation is given in Appendix A.

3. Detecting significant variables from PCs using MSOLS

Feature selection and feature extraction are two commonly encountered problems in statistical pattern recognition. Unlike regression analysis, where the objective is to achieve an approximation of the relationship between the response variables and the predictor variables and which can be solved using some supervised learning schemes, a typical dimensionality reduction problem, for instance feature extraction and feature selection in statistical pattern recognition, often involves a group of input data with a target to find significant patterns or features but without a clearly defined external supervisor (the desired response). Thus, in many cases the tasks of dimensionality reduction (subset selection) are fulfilled using some unsupervised learning algorithms [Haykin 1999, Webb 2002, Jain et al. 2000].

As will be seen, the feature detection and subset selection problem for a general non-regression high-dimensional data analysis can be directly converted into a multiple regression problem by treating the PCs as the response variables and the original measurement variables as the predictors; the MOLS algorithms can then be applied to detect significant variables and hence to select a subset.

3.1 PCA

Principal component analysis is a matrix based subspace decomposition method [Oja 1982, Jolliffe 2002], where a covariance (or correlation) matrix is initially constructed from collected data, and associated eigenvalues and eigenvectors (called direction vectors) of the covariance (or correlation) matrix are then calculated. Input data vectors in the original measurement space can be orthogonally projected onto the subspace (the new feature space) spanned by a few eigenvectors with maximum eigenvalues. The resulting projections are referred to as principal components (PCs), the significance of PCs is evaluated by the corresponding eigenvalues. The basic idea of PCA can briefly be summarized below.

Step 1. Data collection.

Assume that a total of N observations (patterns) are available and let $\mathbf{x}(k) = [x_1(k), x_2(k), \cdots, x_n(k)]^T$ be the kth feature vector in the measurement space. The data matrix can then be represented as $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \cdots, \mathbf{x}(N)]^T$. Note that the collected data are often centralized or standardized.

Step 2. Form the covariance (or *correlation*) *matrix* $\mathbf{\Sigma} = (1/(N-1))\mathbf{X}^T\mathbf{X} \approx (1/N)\mathbf{X}^T\mathbf{X}$.

Step 3. Calculate the eigenvalues and eigenvectors of the matrix $\mathbf{\Sigma}$.

Step 4. Sort the eigenvalues and the eigenvectors. Rearrange the eigenvalues in decreasing order such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Rearrange the eigenvectors accordingly and denote the rearranged eigenvectors by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$.

Principal component analysis aims to find a well-defined transform that maps the feature vectors $\mathbf{x}(k)$ in the n-dimensional measurement space to a new m-dimensional feature space without losing much information, where in general $m << n$. Dimensionality can often be greatly reduced by introducing an orthogonal transform involving only the first m eigenvectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m$, such that

$$\mathbf{y}_i = \boldsymbol{\alpha}_i^T \mathbf{x} = \sum_{j=1}^{n} \alpha_{i,j} x_j \tag{19}$$

where $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n}]^T$, $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ and $\mathbf{y}_i = [y_i(1), y_i(2), \cdots, y_i(N)]^T$ is referred to as the *i*th *principal component* (PC) (*i*=1,2, …, *m*). PCA is perhaps the most commonly used transform for feature extraction. Note that each new variable in the new feature space is a linear combination of all the original variables, this often makes it difficult to physically interpret the principal components in the new space.

3.2 Detecting significant variables of PCs

It has been noted that although the dimensionality may be greatly reduced, the number of variables need to be measured in the original measurement space is kept the same. In cases where there exists some linear dependency and linear correlation among the original variables, the inclusion of redundant variables in PCs is not necessary. Subset selection is thus desired.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ be the collected data matrix, where $\mathbf{x}_j = [x_j(1), x_j(2), \cdots, x_j(N)]^T$ for *j*=1,2, …, *n*. The vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ may be linearly dependent or correlated in the *n*-dimensional measurement space. The objective of subset selection is to find a subset $S_d = \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \cdots, \mathbf{x}_{i_d}\}$, which constitutes a basis for the original measurement space, where $i_s \in \{1,2,\cdots,n\}$, *s*=1,2,…,*d* with $d \le n$ (generally $d << n$ if the measurement space is of large dimension). This means that any data vector $\mathbf{x}_j$ in the measurement space can be well approximated using S_d. From (19), the *i*th principal component $\mathbf{y}_i$ (*i*=1,2, …, *m*) should also be well approximated using the selected subset S_d as below

$$y_i(k) = \sum_{j=1}^{d} \beta_{i,j} x_{k_j}(k) + e_i(k) \tag{20}$$

It is known from the definition of PCs that variables which are significant in the original measurement space are also significant when representing the transformed variables. In other words, variables that are significant for representing PCs must be significant to characterize the overall features in the original measurement space. Specifically, variables that are important to account for the variations in the first few PCs should also be important to account for the variations in the original feature space.

Motivated by the above observations, feature detection and subset selection can be achieved by detecting significant variables using the first few PCs. From the definition of PCs, the task of the significant variable detection problem for subset selection from the first *m* PCs can be viewed as a special case of detecting significant variables from a multiple linear regression by treating $y_1(k), y_2(k), \cdots, y_m(k)$ as the response variables and $x_1(k), x_2(k), \cdots, x_n(k)$ as the predictor variables.

In most PCA based methods, the number of PCs, denoted by *m*, has to be known before the calculation of the first *m* PCs. Although there exist several rules of thumb for choosing the number of PCs [Jeffers 1967, Jolliffe 1972, Jolliffe 2002], no standard criteria are available. In the present study, the number of PCs will be determined as below. Let $\mathbf{y}_i = \boldsymbol{\alpha}_i^T \mathbf{x}$ be the *i*th PC (*i*=1,2, …) and $\mathbf{y}^{(m)} = [\mathbf{y}_1^T, \mathbf{y}_2^T, \cdots, \mathbf{y}_m^T]^T$ (*m*=1,2, …) be the overall response formed by the first *m* PCs. For each *m*, an *m*th order block-structured matrix equation can be obtained, which is similar to (3). By performing the MSOLS algorithm on the *m*th order block-structured matrix

equation, the given *n* features can then be ranked in order of the capability to represent the overall response $\mathbf{y}^{(m)}$. As will be seen in the examples given in the next section, the order of the ranked features will become unchanged when *m* becomes large enough. The choice of the number of PCs will thus not be of importance when applying MSOLS to select a feature subset.

3.3 Determining the number of significant variables

Assume that the first *d* selected significant variables are $\mathbf{z}_1, \cdots, \mathbf{z}_d$, which compose a subset S_d. In the linear case, each data vector $\mathbf{x}_j$ (*j*=1,2, … , *n*) in the measurement space can be approximated using a linear combination of $\mathbf{z}_1, \cdots, \mathbf{z}_d$ as below

$$\mathbf{x}_j = \sum_{m=1}^{d} \theta_{j,m} \mathbf{z}_m + \mathbf{e}_j \tag{21}$$

or in a compact matrix form

$$\mathbf{x}_j = \mathbf{P}\boldsymbol{\theta}_j + \mathrm{e}_j \tag{22}$$

where the matrix $\mathbf{P} = [\mathbf{z}_1, \cdots, \mathbf{z}_d]$ is of full column rank, $\boldsymbol{\theta}_j = [\theta_{j,1}, \cdots, \theta_{j,d}]^T$ is a parameter vector, and $\mathbf{e}_j$ is an approximation error. Using the same orthogonalization procedure given in Section 2.3, the full rank matrix $\mathbf{P}$ can be orthogonally decomposed as

$$\mathbf{P} = \mathbf{QR} \tag{23}$$

where $\mathbf{R}$ is an $d \times d$ unit upper triangular matrix and $\mathbf{Q}$ is an $d \times d$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_d$. Similar to (17), the total sum of squares of the independent variable $\mathbf{x}_j$ from the origin can then be expressed as

$$\mathbf{x}_j^T \mathbf{x}_j = \sum_{k=1}^{d} g_{j,k}^2 \mathbf{q}_k^T \mathbf{q}_k + \mathbf{e}_j^T \mathbf{e}_j \tag{24}$$

The *k*th error reduction ratio (ERR) introduced by $\mathbf{q}_k$ (or equally by including $\mathbf{z}_k$), is given by

$$\mathrm{ERR}[j,k] = \frac{(\mathbf{x}_j^T \mathbf{q}_k)^2}{(\mathbf{x}_j^T \mathbf{x}_j)(\mathbf{q}_k^T \mathbf{q}_k)}, \quad k=1,2, \ldots, d, \tag{25}$$

The sum of error reduction ratio (SERR) due to $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_d$ (or equally due to $\mathbf{z}_1, \cdots, \mathbf{z}_d$) is

$$\mathrm{SERR}[j;d] = \sum_{k=1}^{d} \mathrm{ERR}[j;k] \tag{26}$$

The *total error reduction ratio*, TERR[*d*], which indicates what percentage of the overall variation in all the variables can be accounted for by the subset S_d, can be defined as

$$\mathrm{TERR}[d] = \frac{1}{n} \sum_{j=1}^{n} \mathrm{SERR}[j;d] \tag{27}$$

The criterion TERR[*d*] can be used to measure the performance of the selected subset S_d. If TERR[*d*]is larger than a given threshold, the associated subset S_d can then be considered to be sufficient to represent the overall features; otherwise, more significant variables need to be included to form a new subset.

The procedure to detect significant features (variables) for subset selection using the new MSOLS algorithm can briefly be summarized below, where the first m PCs $y_1(k), y_2(k), \cdots, y_m(k)$ are treated as the response variables, and the features $x_1(k), x_2(k), \cdots, x_n(k)$ are treated as the predictor variables.

- Data collection and pre-processing (centralization and standardization).
- Calculate eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and eigenvectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$ from the associated covariance (or correlation) matrix $\boldsymbol{\Sigma}$.
- Calculate principal components $y_i(k) = \sum_{j=1}^{n} \alpha_{i,j} x_j(k)$, i=1,2 …, m.
- Perform the MSOLS algorithm to detect significant features by treating $y_1(k), y_2(k), \cdots, y_m(k)$ as the response variable and $x_1(k), x_2(k), \cdots, x_n(k)$ as predictor variables.
- Evaluate the performance of the selected subset.

4. Experiments

In this section, two examples, one for artificial and another for real data sets, are provided to illustrate the application of the new method for significant variable detection.

4.1 Example 1– An artificial data set

Consider the model below

$$x_1(t) = c_1 + \sin(2\pi f_1 t) + \varepsilon_1(t) \tag{28a}$$

$$x_2(t) = c_2 + \sin(2\pi f_2 t) + \varepsilon_2(t) \tag{28b}$$

$$x_3(t) = c_3 + \sin(2\pi f_3 t) + \varepsilon_3(t) \tag{28c}$$

$$x_4(t) = x_1(t) + 2x_2(t) + \varepsilon_4(t) \tag{28d}$$

$$x_5(t) = x_2(t) + 2x_3(t) + \varepsilon_5(t) \tag{28e}$$

$$x_6(t) = 2x_1(t) + x_3(t) + \varepsilon_6(t) \tag{28f}$$

$$x_7(t) = x_1(t) + x_2(t) + x_3(t) + \varepsilon_7(t) \tag{28g}$$

$$x_8(t) = x_1(t) - 2x_2(t) + x_3(t) + \varepsilon_8(t) \tag{28h}$$

$$x_9(t) = x_1(t) + x_2(t) - 2x_3(t) + \varepsilon_9(t) \tag{28i}$$

$$x_{10}(t) = -2x_1(t) + x_2(t) + x_3(t) + \varepsilon_{10}(t) \tag{28j}$$

where c_1 =1, c_2 =2, c_3 =3, f_1 =1, f_2 =1.5, f_3 =3.5, and $\varepsilon_i \sim N(0, 0.05^2)$ for i=1,2, …,10. This model was simulated by setting the sampling period as $h = 0.01$ and 200 observations have been recorded to form a 200×10 data set. Although this data set involves 10 variables, that is, the measurement space is of 10 dimensions, 7 of the 10 variables are redundant and only 3 variables are required to depict the underlying system characteristics. The object here is to identify and rank the 10 variables, and then to select a subset without using any a priori information on either the data set or the simulated model.

The data set was centralised before analysis. By performing PCA on the centralised data, 10 eigenvalues, 6.6198, 6.5390, 6.4927, 0.0030, 0.0025, 0.0023, 0.0003, 0.0002, 0.0001, and the related eigenvectors

$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{10}$, of the associated covariance matrix have been obtained. A total number of 10 experiments have been done. In the *p*th experiment, the MSOLS algorithm was applied to the first *p* PCs: $\mathbf{y}_i = \boldsymbol{\alpha}_i^T \mathbf{x}$, *i*=1,…, *p*, and the 10 variables were ranked in order of their significance. The results for the 10 experiments are shown in Table I, where it is clear that the order of the ranked variables becomes stable from *p*=3, the number of PCs could thus be chosen as *m*=3.

TABLE I
RANKED VARIABLES OBTAINED BY CONSIDERING DIFFERENT NUMBER OF PCS FOR THE SIMULATED DATA FROM MODEL (28)

Index	Ranked variables for different number of PCs									
	1	2	**3**	4	5	6	7	8	9	10
1	3	10	**10**	10	10	10	10	10	10	10
2	8	3	**4**	4	4	4	4	4	4	4
3	9	8	**3**	3	3	3	3	3	3	3
4	6	9	**8**	8	8	8	8	8	8	8
5	5	2	**6**	6	6	6	6	6	6	6
6	4	4	**9**	9	9	9	9	9	9	9
7	7	6	**2**	2	2	2	2	2	2	2
8	10	5	**7**	7	7	7	7	7	7	7
9	2	1	**1**	1	1	1	1	1	1	1
10	1	7	**5**	5	5	5	5	5	5	5

TABLE II
THE TWO CRITERIA SERR AND TERR FOR THE SIMULATED DATA FROM MODEL (28)

	SERR[*j*;*d*] (%)										TERR[*d*] (%)
Index	Index *j*										
d	1	2	3	4	5	6	7	8	9	10	
1	66.64	17.03	16.49	0.03	29.78	29.93	0.14	25.13	24.57	100	30.96
2	86.69	96.57	16.49	100	45.57	45.81	59.03	54.71	54.27	100	65.91
3	**99.89**	**99.89**	**100**	**100**	**99.88**	**99.82**	**99.77**	**99.81**	**99.88**	**100**	**99.89**
4	99.89	99.95	100	100	99.88	99.83	99.77	100	99.89	100	99.92
5	99.93	99.95	100	100	99.88	100	99.79	100	99.90	100	99.95
6	99.94	99.96	100	100	99.88	100	99.81	100	100	100	99.96
7	99.95	100	100	100	99.91	100	99.82	100	100	100	99.97
8	99.95	100	100	100	99.91	100	100	100	100	100	99.99
9	100	100	100	100	99.91	100	100	100	100	100	99.99
10	100	100	100	100	100	100	100	100	100	100	100

Using the information given by the number of the 3rd column in Table I, the first variable to be included in the final subset should be x_{10}, followed by the variables x_4, x_3, etc. The remaining problem for subset selection is to determine the number of variables to be included in the subset. The two criteria, SERR[*j*;*d*] and TERR[*d*]defined in Section 3.3 were used to measure the performance of the selected subset consisting of *d* significant variables. The values of SERR[*j*;*d*] and TERR[*d*]for *j*=1,2, …, 10 and *d*=1,2, …,10 are shown in Table II, where the element of SERR in the *d*th row and *j*th column indicates what percentage of the variation in the *j*th variable x_j can be accounted for by the first *d* variables in column 3 of Table I. The total error reduction ratio, TERR[*d*], which refers to what percentage of the overall variation in all the 10 variables can be accounted for by the first *d* variables in column 3 of Table I, is also given in Table II. It is clear from Table II that there is a steep change in the total error reduction ratio, TERR[*d*], from *d*=2 to *d*=3, and from *d*=3 TERR[*d*] becomes stable. Variations in each of the 10 variables can be accounted for with a very high percentage using only the first three variables, x_{10}, x_4 and x_3, listed in column 3 of Table I. The final subset for the simulated data set was thus chosen to be $S_3 = \{x_{10}, x_4, x_3\}$.

4.2 Example 2– The Alate Adelges data

The alate adelges (winged aphids) related data set which has been studied by Jeffers [1967] using a PCA method, concerns an investigation into the variation in 40 individual winged aphids. This data has been studied as a benchmark by several other authors to test new variants of PCA methods or variable selection algorithms, see the detailed discussions in [Jolliffe 2002]. The data comprise 19 variables measured on each of 40 winged aphids (alate adelges) that had been caught in a light trap. The full 40×19 data matrix is available in [Krzanowski 1987], and a description of the variables is given in [Jeffers 1967].

Denote the 19 variables (features) by $x_1, x_2, \cdots, x_{19}$ and let $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$. The data set has been standardized before analysis. By performing PCA on the standardized Alate Adelges data, 19 eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{19}$, and related eigenvectors, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_{19}$, of the associated correlation matrix (standardized covariance matrix) were obtained. A total number of 10 experiments have been done. In the *p*th experiment, the MSOLS algorithm was applied to the first *p* PCs: $\mathbf{y}_i = \boldsymbol{\alpha}_i^T \mathbf{x}$, *i*=1,…, *p*, and the 19 variables were ranked in order of their significance. The results for the 10 experiments are shown in Table III, where it is clear that the order of the ranked variables becomes stable from *p*=6. In fact, the order of most ranked variables becomes stable from *p*=4. The number of PCs could thus be chosen as *m*=4, rather than *m*=6. The total error reduction ratio, TERR[*d*], which refers to what percentage of the overall variation in all the 19 variables can be accounted for by the first *d* variables in column 4 of Table III, is plotted in Fig. 1. Fig. 1 shows that about 70% overall variation in all the 19 variables can be accounted for by $S_1 = \{x_{13}\}$, 81% by $S_2 = \{x_{13}, x_{17}\}$, 85% by $S_3 = \{x_{13}, x_{17}, x_{11}\}$, 87% by $S_4 = \{x_{13}, x_{17}, x_{11}, x_8\}$, and 90% by $S_5 = \{x_{13}, x_{17}, x_{11}, x_8, x_5\}$, etc.

Note that for the same alate adelges data set, different criteria and different algorithms may lead to different subsets [Jolliffe 2000]. The subset $S_3 = \{x_{13}, x_{17}, x_{11}\}$, selected by MSOLS algorithm here is identical to that selected by the method B4 (a PCA based algorithm) [Jolliffe 1973]. Compared with other methods, the subset $S_3 = \{x_{13}, x_{17}, x_{11}\}$ selected by B4 is considered to be the best in all subsets of three variables [Jolliffe 2000].

TABLE III
RANKED VARIABLES OBTAINED BY CONSIDERING DIFFERENT NUMBER OF PCS FOR THE ALATE ADELGES DATA

Index	Ranked variables for different number of PCs									
	1	2	3	**4**	5	**6**	7	8	9	10
1	13	13	13	**13**	13	**13**	13	13	13	13
2	5	17	17	**17**	17	**17**	17	17	17	17
3	8	8	11	**11**	11	**11**	11	11	11	11
4	2	5	8	**8**	8	**8**	8	8	8	8
5	12	18	5	**5**	5	**5**	5	5	5	5
6	3	16	4	**4**	4	**4**	4	4	4	4
7	15	11	18	**18**	18	**18**	18	18	18	18
8	10	7	16	**16**	16	**16**	16	16	16	16
9	9	3	7	**19**	19	**19**	19	19	19	19
10	11	19	3	**14**	14	**14**	14	14	14	14
11	14	12	12	**7**	7	**7**	7	7	7	7
12	6	15	6	**3**	3	**3**	3	3	3	3
13	19	10	15	**6**	6	**6**	6	6	6	6
14	1	6	10	**12**	12	**12**	12	12	12	12
15	7	14	19	**15**	15	**9**	9	9	9	9
16	18	9	14	**10**	10	**10**	10	10	10	10
17	16	4	2	2	2	**15**	15	15	15	15
18	4	1	9	9	9	**2**	2	2	2	2
19	17	2	1	1	1	**1**	1	1	1	1

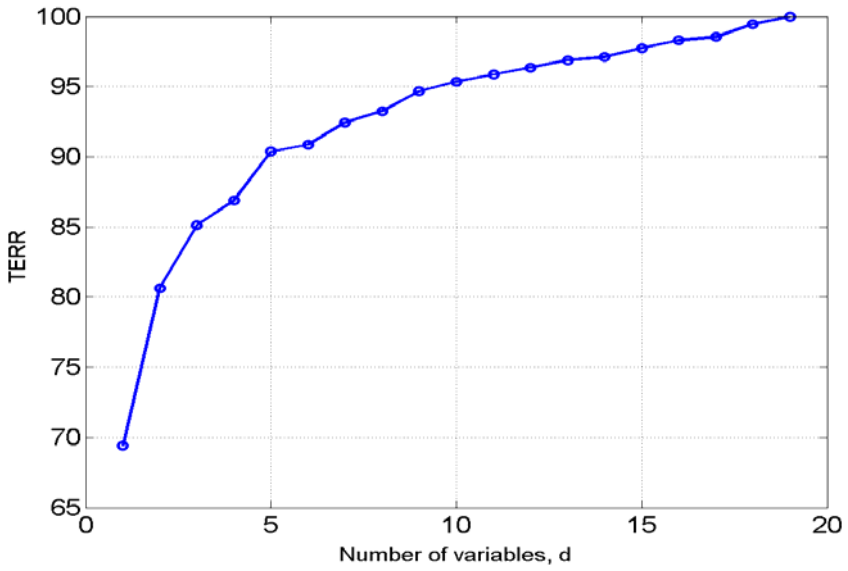


Fig. 1 The total error reduction ration, TERR [*d*], versus the number of variables in the subset for the Alate Adelges data

TABLE IV
RANKED VARIABLES OBTAINED BY CONSIDERING DIFFERENT NUMBER OF PCS FOR THE PIMA INDIANS DIABETES DATA

Index	Ranked variables for different number of PCs							
	1	**2**	3	4	5	6	7	8
1	2	**2**	2	2	2	2	2	2
2	5	**5**	5	5	5	5	5	5
3	3	**8**	8	8	8	8	8	8
4	6	**4**	4	4	4	4	4	4
5	8	**3**	3	3	3	3	3	3
6	4	**6**	6	6	6	6	6	6
7	1	**7**	7	7	7	7	7	7
8	0	**1**	1	1	1	1	1	1

TABLE V
THE TWO CRITERIA SERR AND TERR FOR THE PIMA INDIANS DIABETES TEST DATA SET

	SERR[*j*;*d*] (%)								TERR[*d*] (%)	
Index	Index *j*								Test data	Training data
d	1	2	3	4	5	6	7	8		
1	53.65	100	90.05	59.90	40.54	91.11	66.05	85.66	73.37	73.39
2	54.88	100	90.08	65.04	100	91.11	66.22	86.32	81.71	82.21
3	71.13	100	92.25	65.35	100	92.07	67.09	100	85.99	85.26
4	71.25	100	92.96	100	100	93.58	68.98	100	90.85	90.13
5	71.26	100	100	100	100	94.85	69.61	100	91.97	91.33
6	71.33	100	100	100	100	100	70.28	100	92.71	92.09
7	71.34	100	100	100	100	100	100	100	96.42	96.09
8	100	100	100	100	100	100	100	100	100	100

4.3 Example 3– The Pima Indians Diabetes data

The pima indians diabetes (PID) data is taken from UCI Machine Learning Repository [Newman et. al 1998]. The PID data set comprises 8 individual features, measured with 768 samples, where the first 500 samples were used for training and the remaining 268 samples were used for testing. The objective here is to rank the 8 features used in the 500×8 data set. Denote the 8 features by $x_1, x_2, \cdots, x_8$.

Similar to Example 1, the significance of the 8 features were detected and ranked by performing the MSOLS algorithm on a different number of PCs, and the associated results are shown in Table IV, where the number zero in the 2nd column means that only 7 features are required to represent the first PC. It is clear from Table IV that the number of PCs to be considered for subset selection is 2. The most significant feature select by MSOLS is x_2, referring to the plasma glucose concentration at 2 hours in an oral glucose tolerance test. This feature has been employed as one of the main criteria by World Health Organization (WHO) for diabetes diagnosis. The result here just reflects the fact that x_2 is significant for diabetes diagnosis.

The values of SERR[*j*;*d*] with respect to the PID test data set are shown in Table V, where the element of SERR in the *d*th row and *j*th column indicates what percentage of the variation in the *j*th variable x_j can be accounted for by the first *d* variables in column 2 of Table V. The total error reduction ratio, TERR[*d*], for both the training data set and the test data set are also listed in Table V.

5. Conclusions

A new learning scheme has been proposed for feature ranking and subset selection. By treating the first principal components as the response variables and all the candidate features as the predictor variables, the problem for feature ranking and subset selection can be viewed as a special case of variable detection and subset selection in multiple linear regression analysis. The new MSOLS algorithm can thus be applied to detect and rank the features according to the capability to represent the first few PCs. The selected feature subset can be evaluated by the total error reduction ratio (TERR), which refers to what percentage of the overall variation in all the features can be accounted for by a selected subset, and which can be used as a criterion to indicate how many features (variables) need to be included in the subset. Unlike principal components, which carry no clear physical meaning, variables (features) selected using MSOLS are physically interpretable. The applicability and usefulness of the new MSOLS algorithm and associated learning scheme for feature (variable) ranking and subset selection have been demonstrated using results from several examples for both artificial and real data sets, including the three examples given in Section 4.

Appendix A –The MSOLS algorithm

In the following, $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_n$ are the data vectors associated with the (*n*+1) candidate variables, $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_{m(n+1)}$ are the columns of the matrix $\tilde{\mathbf{X}}$ defined by (3); $\mathbf{z}_r$ is the significant variable selected at the *r*th step (*r*=1,2 …) .

The MSOLS algorithm:

Step 1: Set $I_1 = \{1, \cdots, L\}$; $L = m(n+1)$;

for *j*=1 to L

$$\boldsymbol{\alpha}_j = \boldsymbol{\gamma}_j;$$

$$\text{err}^{(1)}[j] = \frac{(\mathbf{y}^T \boldsymbol{\alpha}_j)^2}{(\mathbf{y}^T \mathbf{y})(\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j)}; \ \{\text{if } \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j < \delta, \text{set } \text{err}^{(1)}[j] = 0\};$$

end for

$$\ell_1 = \arg\max_{k \in I_1} \{\text{err}^{(1)}[k]\}; \ \ \ell_1^0 = \text{mod}(\ell_1, n+1);$$

$$\text{err}[1] = \text{err}^{(1)}[\ell_1]; \ \ \text{serr}[1] = \text{err}[1];$$

$$\mathbf{q}_1 = \boldsymbol{\alpha}_{\ell_1}; \ \mathbf{z}_1 = \mathbf{x}_{\ell_1^0}; \ \ \Gamma_1 = \{\ell : \ell \text{ is multiple of } \ell_1^0 \text{ and } \ell \in I_1\};$$

Step *r*, $r \geq 2$ **:**

for *r*=2 to *n*+1

$$I_r = I_{r-1} \setminus \Gamma_{r-1};$$

for $j \in I_r$

$$\boldsymbol{\alpha}_j = \boldsymbol{\gamma}_j - \sum_{k=1}^{r-1} \frac{\boldsymbol{\gamma}_j^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k;$$

$$\text{err}^{(r)}[j] = \frac{(\mathbf{y}^T \boldsymbol{\alpha}_j)^2}{(\mathbf{y}^T \mathbf{y})(\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j)}; \quad \{\text{if } \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j < \delta, \text{set } \text{err}^{(r)}[j] = 0\}; \tag{29}$$

end for (end loop for j)

$$J_r = \{\arg_{j \in I_r}(\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j < \delta)\}; \quad I_r = I_r \setminus J_r; \tag{30}$$

$$\ell_r = \arg\max_{k \in I_r}\{\text{err}^{(r)}[k]\}; \ \ell_r^0 = \text{mod}(\ell_r, n+1) \ ;$$

$$\text{err}[r] = \text{err}^{(r)}[\ell_r]; \ \ \text{serr}[r] = \sum_{k=1}^{r} \text{err}[k];$$

$$\mathbf{q}_r = \boldsymbol{\alpha}_{\ell_r}; \ \mathbf{z}_r = \mathbf{x}_{\ell_r^0}; \ \Gamma_r = \{\ell : \ell \text{ is multiple of } \ell_r^0 \text{ and } \ell \in I_r\};$$

end for (end loop for r)

The MSOLS algorithm provides an effective tool for selecting significant variables for a multiple regression with an iterative stepwise way. Variables are selected step by step, one variable at a time. Most numerical ill conditioning can be avoided by eliminating the candidate variables for which $\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j$ are less than a predetermined threshold δ, say $\delta = 10^{-\tau}$ with $\tau \geq 10$ (Eqs. (29), (30)). If required, the selection procedure can be terminated at any step when SERR[r] satisfies some specified conditions. This algorithm can be applied to not only the normal cases, where the number of features is smaller than the number of training samples ($n < N$), but also the cases where the number of features is much larger than that of training samples ($n \geq N$). The applications of MSOLS algorithm is not restricted to variable selection in linear regression, it can be used to select significant model terms for any linear-in-the-parameters models.

Acknowledgements

The authors gratefully acknowledge that this work was supported by EPSRC (UK).

References

S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems suing a forward regression orthogonal estimator," *Int. J. Control*, vol. 49, pp.2157-2189, June 1989.

M. A. Carreira-Perpinan, "Continuous latent variable models for dimensionality reduction and sequential data reconstruction," Ph.D. dissertation, Dept Computer Science, Univ. Sheffield, Sheffield, U.K., 2001.

S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, pp.1873-1896, Nov.1989

I. K. Fodor, "A survey of dimension reduction techniques," http://www.llnl.gov/CASC/sapphire/pubs , 2002.

S. Haykin, *Neural networks: A Comprehensive Foundation* (2nd ed.). New York: Prentice Hall, 1999.

A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4-37, Jan. 2000.

J. N. R. Jeffers, "Two case studies in the application of principal component analysis," *Appl. Statist.*, vol. 16, no. 3, pp. 225-236, 1967.

I. T. Jolliffe, "Discarding variables in a principal component analysis. I: artificial data." *Appl. Statist.*, vol. 21, no. 2, pp. 160-173, 1972.

I. T. Jolliffe, "Discarding variables in a principal component analysis. II: real data." *Appl. Statist.*, vol. 22, no. 1, pp. 21-31, 1973.

I. T. Jolliffe, *Principal Component Analysis* (2nd ed.). New York: Springer, 2002.

R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no.1-2, pp. 273-324, Dec 1997.

M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, vol. 48, no. 1, pp. 193-210, July 1988,

W. J. Krzanowski, "Selection of variables to preserve multivariate data structure using principal components," *Appl. Statist.*, vol. 36, no. 1, pp. 22-33, 1987.

K. Z. Mao, "Identifying critical variables of principal components for unsupervised feature selection," *IEEE Trans. Syst., Man, Cybern., B, Cybern*, vol. 35, no. 2, pp. 339-344, April 2005.

G. P. McCabe, "Principal variables," *Technometrics*, vol. 26, no. 2, pp. 137-144, May 1984.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, UCI Repository of Machine Learning Databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: Univ. California, Dept. Inform. Comput. Sci., 1998.

E. Oja, *Subspace methods of Pattern Recognition*. Letchworth, England: Research Studies Press, 1983.

P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Reconit. Lett.*, vol. 15, no. 11, pp. 1119-1125, Nov 1994.

A. R. Webb, *Statistical Pattern Recognition* (2nd ed.). New York: Wiley, 2002.

H. L. Wei, S. A. Billings, and J. Liu, "Term and variable selection for nonlinear system identification," *Int. J. Control*, vol. 77, no. 1, pp.86-110, Jan. 2004.