Subset selection in dimension reduction methods

Luca Scrucca
Dipartimento di Economia, Finanza e Statistica
Università degli Studi di Perugia
06100 Perugia, Italy
luca@stat.unipg.it

Abstract

Dimension reduction methods play an important role in multivariate statistical analysis, in particular with high-dimensional data. Linear methods can be seen as a linear mapping from the original feature space to a dimension reduction subspace. The aim is to transform the data so that the essential structure is more easily understood. However, highly correlated variables provide redundant information, whereas some other feature may be irrelevant, and we would like to identify and then discard both of them while pursuing dimension reduction.

Here we propose a greedy search algorithm, which avoids the search over all possible subsets, for ranking subsets of variables based on their ability to explain variation in the dimension reduction variates.

keywords: Dimension reduction methods; Linear mapping; Subset selection; Greedy search.

Contents

1 Introduction 3

2 Linear mapping: a dimension reduction approach to multivariate data 4

2.1 Principal Components Analysis 5

2.2 Independent Component Analysis 6

2.3 Linear discriminant analysis 7

2.4 SIR and SAVE 7

3 Subset selection in dimension reduction methods 8

3.1 A criterion for variable selection 8

3.2 A greedy search algorithm for selecting the “best” subset 9

4 Data analysis examples 13

4.1 PCA: simulation data 13

4.2 PCA: Alate adelges data 14

4.3 ICA: simulation data 17

4.4 LDA: Iris data 20

4.5 SAVE: banknote data 22

4.6 Multivariate SIR: bleaching of cotton 25

5 Discussion 27

1 Introduction

A long-standing problem in statistics and related areas is how to find a suitable representation of high-dimensional multivariate data. Representation here means that we would like to transform the data so that the essential structure is more easily understood. Data may have dimension ranging from hundreds to perhaps thousands of features or variables, so a drastic reduction is sought. Problems of this type are found in pattern recognition and classification problems involving images (e.g. face recognition, character recognition) or speech (e.g. auditory models). In fields like social sciences, psychology, etc., the data have not severe high-dimensionality, so the reduction needed is not very drastic. However, interpretation may often be enhanced by a suitable choice of few components. In visualization problems, we need to reduce the dimension of the problem to two or three, at most four, dimensions in order to be able to graphically represent the data. A good representation is also a central goal of many techniques in data mining and exploratory data analysis. Furthermore, high-dimensional spaces are inherently sparse, a phenomenon responsible for the so-called *curse of dimensionality*. The latter refers to the fact that the sample size needed to estimate a function of several variables to a given degree of accuracy grows exponentially with the number of variables. Hence, a lower dimensional subspace may help to visualize patterns in the data that would otherwise go unnoticed.

For these reasons, dimension reduction techniques have played an important role in multivariate analysis. Dimensionality reduction is basically a mapping from a multidimensional feature space onto a space of fewer dimensions. However, dimension reduction without loss of information is only possible if the data fall exactly on a smooth, locally flat subspace; thus, the reduced dimensions are just coordinates in this subspace. More commonly, data are noisy and therefore does not exist an exact mapping.

Dimension reduction methods can be classified as linear or nonlinear methods. Linear methods attempt to find a globally flat subspace, while nonlinear methods attempt to find a locally flat subspace. As is the case with other techniques, linear methods are simpler and more completely understood, while nonlinear methods are more general but more difficult to analyze. In this paper we will focus on linear methods, such as principal component analysis and projection pursuit, which construct a system of q components obtained as linear combinations of the original

p variables ($q \leq p$). This process is often indicated as feature extraction (Webb, 2002).

In multivariate datasets it is often the case that variables are highly correlated and provide redundant information. If you have a large number of measurements from the same source it is possible that several of them may represent related characteristics. If this occurs then some of the extra measurements may lengthen the computation time by adding unnecessary information. When the number of variables is unnecessarily large, essentially the same information could be conveyed by fewer dimensions if the variables are wisely combined. In classification and pattern recognition problems, more features does not necessarily improve performance of a system and can lead to a reduction in accuracy (Ripley, 1996).

When the number of observed or measured variables, p, is large it is likely that a subset of k variables ($k < p$) contains virtually all the information available in the original variables. It is then useful to determine an appropriate value of k, and to decide which subset or subsets of k variables are best according to a given criterion. Variable selection methods are usually treated within each statistical procedure; see for example Jolliffe (2002, Chap. 6) for subset selection in the context of principal components analysis. As we noted above, linear dimension reduction methods may be viewed as a form of linear mapping and, therefore, a unified approach to variable selection might be pursued.

In Section 2 we state the problem as a linear mapping, and we briefly review some statistical techniques embodied in this view. In the following Section we discuss subset selection in the context of dimension reduction methods, introducing both a criterion and a greedy search algorithm for ranking variables subsets based on the chosen criterion. Section 4 reports some simulation studies and data analysis examples for a variety of dimension reduction techniques, which illustrate how the proposed approach can be used in practical applications. The final Section contains some final remarks and comments.

2 Linear mapping: a dimension reduction approach to multivariate data

Most dimension reduction methods can be expressed as a linear mapping from a random vector $X \in \mathbb{R}^p$, that without loss of generality will be assumed to have zero mean, to a lower dimension random vector $Z \in \mathbb{R}^q$ $(q \leq p)$. Such linear mapping $X \mapsto Z$ can be written as

$$Z = \boldsymbol{B}^\top X$$

for a $(p \times q)$ matrix $\boldsymbol{B}$ (with $\text{rank}(\boldsymbol{B}) = q$) of coefficients defining the set of q linear transformations. The vector Z defines a set of q projections onto the subspace spanned by the columns of $\boldsymbol{B}$, and we refer to such components as dimension reduction (DR) variables. Often, $\boldsymbol{B}$ is made-up of orthogonal column-vectors $\boldsymbol{b}_j$ $(j = 1, \ldots, q)$, hence $\boldsymbol{b}_j^\top \boldsymbol{b}_j \neq 0$ and $\boldsymbol{b}_j^\top \boldsymbol{b}_l = 0$ (for $j \neq l$), or equivalently $\boldsymbol{B}^\top \boldsymbol{B}$ is equal to a $(q \times q)$ diagonal matrix. If we further assume that each vector $\boldsymbol{b}_j^\top$ has unit length, i.e. $||\boldsymbol{b}_j|| = 1$ for all $j = 1, \ldots, q$, then $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}$ and $\boldsymbol{B}$ is said to be orthonormal.

Suppose a random sample of size n is available, so $\boldsymbol{X}$ is a $(n \times p)$ matrix of n observations on p variables or features. The $(n \times q)$ matrix of DR variables is thus computed as

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} \tag{1}$$

Some common multivariate statistical methods, both supervised and unsupervised, may be expressed in this framework, and some of these are briefly reviewed in the following.

2.1 Principal Components Analysis

Principal components analysis (PCA), also known as Karhunen-Loève transform in the machine learning field, is possibly the dimension reduction technique most widely used in practice, perhaps due to its theoretical appealing and efficient algorithms available. It was first introduced by Pearson (1901), and developed independently by Hotelling (1933). A comprehensive and up-to-date reference is Jolliffe (2002).

PCA estimates a system of components that are uncorrelated and have maximal variance. Since $\widehat{\boldsymbol{\Sigma}}$, the sample covariance matrix of X, is a non-negative definite matrix, it allows the eigen decomposition

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$$

where $\boldsymbol{D}$ is a diagonal matrix of (non-negative) eigenvalues in decreasing order and $\boldsymbol{V}$ is the $(p \times p)$ matrix of eigenvectors (also called loadings in PCA), for which $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$. Principal components (PC) are computed as in equation (1) with $\boldsymbol{B} = \boldsymbol{V}$, and it is easily to check that

$\mathrm{Var}(\boldsymbol{Z}) = \boldsymbol{V}^{\top}\widehat{\boldsymbol{\Sigma}}\boldsymbol{V} = \boldsymbol{D}$, so the PCs are uncorrelated and with variances equal to the corresponding eigenvalues. The cumulative proportion of eigenvalues associated with the first q PCs provides a measure of the variance explained.

The first q PCs, given by $\boldsymbol{Z}_q = \boldsymbol{X}\boldsymbol{V}_q$ where $\boldsymbol{V}_q = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_q]$ is formed by the first q eigenvectors, span a subspace containing the "best" q-dimensional view of the data. Here "best" means that PCA estimates those orthogonal directions which best approximates the original points in the sense of minimizing the sum of squared distances from the points to their projections. The first few principal components are often useful to reveal structure in the data.

Since the variances depend on the scale of the variables, it is customary to first standardize each variable to have mean zero and standard deviation one. It is easy to show that a PCA on the standardized variables is equivalent to apply the spectral decomposition to the correlation matrix.

A related technique is the so-called *Simple Component Analysis* (Vines, 2000). Since PCs are often difficult to interpret, the goal of Simple Component Analysis is to replace the optimal but non-interpretable PCs by suboptimal but interpretable simple components. Typically, the resulting loadings are not orthogonal.

2.2 Independent Component Analysis

Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate statistical data. Such components are assumed to be both statistically independent and nongaussian (Hyvarinen and Oja, 2000; Hyvarinen, Karhunen, Oja, 2001). In general, ICA allows to recover the mixing matrix $\boldsymbol{A}$ in $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$, where $\boldsymbol{X}$ is a $(n \times p)$ matrix containing n measures from p observed signals assumed to be generated from a mixture of q $(q \leq p)$ independent signals collected in the $(n \times q)$ matrix $\boldsymbol{S}$. Typical applications arise in signal processing, where there are a number of signals emitted by some physical objects or sources, but we actually records only a mixture of the original source signals. This is also known as the blind source separation problem. Since ICA looks for maximally nongaussian directions/projections in multi-dimensional datasets, there exists a close connection with *projection pursuit* (Friedman, 1987).

ICA algorithms estimate the mixing matrix $\boldsymbol{A}$ based on a pre-whitening of the data, i.e. $\boldsymbol{X}$ is transformed in such a way it has zero mean and identity covariance matrix. This sphering step is usually performed through a PCA on the original variables. Then, the independent signals are obtained as $\widehat{\boldsymbol{S}} = \boldsymbol{X}\widehat{\boldsymbol{A}}^{-1}$. Therefore, this is essentially the same as in equation (1) with $\boldsymbol{B} = \widehat{\boldsymbol{A}}^{-1}$. If $q < p$, only the first q PCs are retained in the pre-whitening step.

2.3 Linear discriminant analysis

Suppose we have a set of g groups or classes, and for each case we know the class membership. The g centroids in the p-dimensional input space span at most a $(g-1)$ dimensional subspace, and if $p >> g$, projecting the data onto this subspace will provide a considerable drop in dimension.

Canonical variates, also known as CRIMCORDS (Gnanadesikan, 1977), are obtained through a projection along the orthogonal directions of maximal ratio of group means to within-group variance, i.e. onto the subspace spanned by the eigenvectors obtained from the decomposition $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{S}_W$ denotes the pooled within-class covariance matrix, and $\boldsymbol{S}_B$ denotes the between-classes covariance matrix. There will be at most $\min(p, g-1)$ positive eigenvalues, and each eigenvalue expresses the proportion of the between-classes variance explained by the corresponding linear combination. This may help to choose how many components to use.

Canonical variates, computed as in equation (1) with $\boldsymbol{B} = \boldsymbol{V}$, are used to obtain a graphical representation of the data such that class-centroids are maximally spread out. Since canonical variates are directly related to Gaussian linear discriminant analysis (LDA), they are also called linear discriminants (LD) (Mardia et al. 1979).

2.4 SIR and SAVE

Consider a regression with a response variable Y and a vector X of p predictors. The main goal of a regression analysis is to understand how the conditional distribution of the response Y given X depends on the value assumed by X. However, the attention is often restricted to the mean function $\mathrm{E}(Y|X)$ and perhaps the variance function $\mathrm{Var}(Y|X)$. Dimension reduction in the context of regression analysis aims at finding the smallest number of linear combinations $(q \leq p)$ of X such that

$$Y \perp\!\!\!\perp X | \boldsymbol{B}^\top X$$

where $\perp\!\!\!\perp$ indicates independence and $\boldsymbol{B}^\top X = (\boldsymbol{b}_1^\top X, \boldsymbol{b}_2^\top X, \ldots, \boldsymbol{b}_q^\top X)$. The structural dimension of a regression is defined as the smallest number of linear combinations for which the above conditional independence statement holds (Cook & Weisberg, 1999).

Thus, dimension reduction methods in regression aim at reducing the dimension of X without losing information on $Y|X$, and without requiring a pre-specified parametric model for $Y|X$. The columns of $\boldsymbol{B}$ span the central dimension reduction subspace $\mathcal{S}_{Y|X}$ for the regression of Y on X (Cook, 1998). This leads to the pursuit of sufficient summary plots which contain all the information on the regression problem that is available from the sample.

Several methods are available for estimating the central subspace, including *Sliced Inverse Regression* (SIR) (Li, 1991) and *Sliced Average Variance Estimation* (SAVE) (Cook and Weisberg, 1991). SIR gains information on $\mathcal{S}_{Y|X}$ from the inverse mean function, whereas SAVE uses both the inverse mean and variance functions. SAVE appears to be more comprehensive, but it requires the estimation of more parameters, and the resulting summary plot may not be as informative as that provided by SIR when most of the statistical information comes from the inverse mean function. Both methods require the use of a sliced version of the response variable for computing an estimate of $\boldsymbol{B}$.

3 Subset selection in dimension reduction methods

The linear mapping methods discussed in the previous section represent a form of feature extraction, where the components are reduced through a set of linear combinations of the original variables. In this context, variable selection aims at finding a subset of the original variables X which best linearly explain the DR variables Z.

3.1 A criterion for variable selection

A suitable statistic for evaluating the amount of variation explained by a subset of variables is provided by a modified version of the squared correlation coefficient for a multivariate linear regression model (Mardia, Kent and Bibby (1979) pp. 170–171).

Let $\mathscr{S}$ be the set of $\dim(\mathscr{S}) = k$ containing one of the possible subset of k variables from the original p ($k \leq p$). The statistic proposed can be defined as follows:

$$R^2(\mathscr{S}) = 1 - \operatorname{tr}\{(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{W} \boldsymbol{E}^\top \boldsymbol{E}\} \qquad (2)$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1 \ \boldsymbol{z}_2 \ \cdots \ \boldsymbol{z}_q]$ is the $(n \times q)$ matrix of column-centered DR variates, $\boldsymbol{E} = \boldsymbol{Z} - \boldsymbol{X}_k(\boldsymbol{X}_k^\top \boldsymbol{X}_k)^{-1}\boldsymbol{X}_k^\top \boldsymbol{Z}$ is the matrix of residuals for the regression of $\boldsymbol{Z}$ on $\boldsymbol{X}_k$, with the latter being the $(n \times (k+1))$ matrix containing the subset of k variables in $\mathscr{S}$ plus a column of 1s. The $(q \times q)$ diagonal matrix $\boldsymbol{W}$ allows to weight differently each DR variable, a common requirement in several methods. For example, in PCA components have associated eigenvalues expressing the importance of each direction; in this case we may set $\boldsymbol{W} = \operatorname{diag}(l_j / \sum_{h=1}^q l_h)$ for $j = 1, \ldots, q$, where l_j is the eigenvalue corresponding to the j-th component.

Some simplifications may occur in some circumstances:

- If there exists a single DR variable, i.e. $q = 1$, the statistic in equation (2) reduces to the usual coefficient of determination for the regression of the DR variate on the subset of k variables.
- If the DR variables are orthogonal, then $\boldsymbol{Z}^\top \boldsymbol{Z} = \operatorname{diag}(\boldsymbol{z}_j^\top \boldsymbol{z}_j)$, and since $\operatorname{Var}(Z_j) = \boldsymbol{z}_j^\top \boldsymbol{z}_j / n$, we have $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} = \frac{1}{n}\operatorname{diag}(1/\operatorname{Var}(Z_j))$ for any $j = 1, \ldots, q$.
- If the directions are principal components, then $\operatorname{Var}(Z_j) = l_j$ and $\boldsymbol{W} = \operatorname{diag}(l_j / \sum l_h)$, so $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{W} = \frac{1}{n}\operatorname{diag}(1/\sum l_h)$, i.e. the matrix is diagonal with constant element for any $j = 1, \ldots, q$.

3.2 A greedy search algorithm for selecting the "best" subset

The statistic in equation (2) can be used as a criterion to rank candidate subsets based on the maximization of the multivariate squared correlation coefficient, eventually computed taking into account the importance of each estimated direction. This amounts to find those features which best linearly explain the DR variables.

However, the space of all possible subsets of size k, with k ranging from 1 to p, has number of elements equal to $2^p - 1$. An exhaustive search become soon unfeasible, even for moderate values of p. To alleviate this problem, we propose a greedy search algorithm. At each stage it searches for the variable to add that best linearly explain the variation in DR variates not explained by the variables already selected, and then it assess whether one of the current selected features could be dropped once the new variable is entered in the current subset. These steps are iterated until all variables have been included or some other stopping rule has been satisfied.

A complete description of each steps of the proposed algorithm follows:

1. Select the first variable to be the one which maximizes the R^2 criterion in equation (2). Let $\mathscr{S}_0 = \emptyset$ be the set of included variables, which is of course empty at the beginning, and $\mathscr{S}'_0 = \{1, 2, \ldots, p\}$ be the set containing indices of all p variables. We choose the "best" variable, X_{i_1}, such that

   $$i_1 = \arg \max_{i \in \mathscr{S}'_0} R^2_1(\{i\})$$

   where $R^2_j(\{i\})$ is the statistic in equation (2) computed at step $j = 1$ for any subset of size $k = 1$.

   Then, define with $\mathscr{S}_1 = \{i_1\}$ the set of included variables, and with $\mathscr{S}'_1 = \mathscr{S}'_0 \setminus \{i_1\}$ the set of variables currently not included. Set $j = 2$ and go to the next step.

2. Select a variable to add, among those not already included, to be the one which maximizes the R^2 criterion. Formally, we choose the "best" variable, X_{i_j}, such that

   $$i_j = \arg \max_{i \in \mathscr{S}'_{j-1}} R^2_j(\mathscr{S}_{j-1} \cup \{i\})$$

   Then, update the subsets of currently included and excluded variables, which are, respectively, given by $\mathscr{S}_j = \mathscr{S}_{j-1} \cup \{i_j\}$, and $\mathscr{S}'_j = \mathscr{S}'_{j-1} \setminus \{i_j\}$.

3. Remove one of the variables in the current subset if not needed once a new variable is included. Let $R^2_{j-1}(\mathscr{S}_{j-1})$ be the maximum value calculated for the best subset of size $\dim(\mathscr{S}_{j-1})$ at the previous step, i.e. before the inclusion of variable X_{i_j}, and $R^2_j(\mathscr{S}_j \setminus \{i'_j\})$ be the maximum value computed omitting in turn each variable from $\mathscr{S}_j$, i.e.

   $$i'_j = \arg \max_{i \in \mathscr{S}_j} R^2_j(\mathscr{S}_j \setminus \{i\})$$

If $R^2_j(\mathscr{S}_j \setminus \{i'_j\}) > R^2_{j-1}(\mathscr{S}_{j-1})$ then the corresponding variable $X_{i'_j}$ may be dropped. This because the subset with X_{i_j} included and $X_{i'_j}$ removed provides a better explanation of variation in DR variates for a given subset size. Of course, the removed variable might be considered for inclusion at successive steps when the subset size increases.

If a variable has been dropped, update the subsets as follows: $\mathscr{S}_j = \mathscr{S}_j \setminus \{i'_j\}$, $\mathscr{S}'_j = \mathscr{S}'_j \cup \{i'_j\}$.

4. Set $j = j + 1$ and iterate steps 2 to 4 until a stopping rule is meet. The algorithm naturally terminates when all variables are included, but it might be stopped earlier when, for example, a certain number of variables have been included or a given proportion of variance has been explained.

The proposed greedy search is a forward-backward algorithm type. However, if p is very large we may want to skip the backward step (n. 3 above) to improve computationally efficiency, hence reducing the algorithm to a forward search.

To assess the above algorithm we conducted a small Monte Carlo study. We compared the proposed algorithm against an exhaustive search for different sample sizes $n = (50, 100, 500, 1000)$, and number of variables $p = (5, 10, 15)$. For each combination of design variables (n, p), we generated 100 sample from a multivariate normal, then we conducted a PCA on the generated data. These were simulated such that only the first 3 variables were important for PCA estimation, while the remaining $p-3$ variables were redundant. In Table 1 we reported the averages (and standard deviations) of computing time required by each type of search, followed by the percentage of correct subsets chosen by the greedy-search algorithm. The time needed by the greedy-algorithm is always a fraction of that needed by the exhaustive search, except for the case with the smallest p and the largest n. This difference in computing time grow very fast as p increases; this was expected since for each p the exhaustive search need to evaluate, respectively, $31, 1023, 32767$ subsets. To judge about accuracy of the greedy search, we compared the subsets with the largest R^2 value at each subset size $k = 1, \dots, p$ chosen by the all-subsets search with those identified by the proposed algorithm. In all cases we obtained a 100% accuracy. Although this may not hold in other cases, it indicates the good performance of the proposed greedy search, a fact also confirmed by further analyses discussed in the following section. Overall, the computing time required by the greedy-search algorithm is much smaller than that required by a full search; furthermore, it seems to be able to accurately select the "best" subsets for each size.

Table 1. *Results from a Monte Carlo study comparing the greedy search algorithm against an exhaustive search. For each combination of sample size (n) and number of variables (p) the table reports the average system time (seconds) and standard deviation (in parenthesis) for the proposed search algorithm and the exhaustive search, respectively, based on 100 simulations.*

n		p	
	5	10	15
50	0.089 (0.005)	0.581 (0.156)	0.759 (0.044)
	0.117 (0.001)	6.828 (1.844)	275.2 (0.958)
100	0.091 (0.005)	0.566 (0.166)	0.767 (0.029)
	0.118 (0.001)	6.743 (1.989)	283.5 (0.361)
500	0.131 (0.022)	0.821 (0.168)	0.976 (0.039)
	0.137 (0.001)	9.102 (1.787)	306.8 (1.963)
1000	0.564 (0.181)	1.498 (0.377)	1.422 (0.032)
	0.212 (0.069)	10.559 (2.458)	333.9 (1.319)

4 Data analysis examples

4.1 PCA: simulation data

Jolliffe (1972) presented a Monte Carlo study where each dataset was generated according to a predetermined model. Each model was constructed in such a way that certain variables were redundant since they were obtained, except for a random disturbance, as a linear combinations of other variables. Four models were considered and, for each, he labelled the different possible subset selections as "Bad", "Moderate", "Good" and "Best", according with the presence or absence of redundant variables. The interested reader may refer to the detailed description of the simulation schemes contained in Tab. 2 and 3 of Jolliffe (1972).

We replicated this simulation study applying our greedy search algorithm. One thousand samples of size 100 were generated according to one of the predefined models, then for each we selected the best d-dimensional subset of variables which maximally explain the variation in the set of (i) relevant principal components chosen according to a modified version of Kaiser's rule, and (ii) all the principal components. The true subset dimension d was set by design equal to $d = 3$ for all models except for model IV where it was equal to 4. The rule used in method (i) amounts to retain those PC whose eigenvalues are larger than 0.7 times the average of all eigenvalues (Jolliffe, 2002, p. 115). The results are shown in Table 2. For models I and IV the proposed algorithm always selected one of the "Best" subsets, while "Good" subset were always selected in the case of model II. For model III approximately 2/3 of the times "Good" models were selected and for the remaining cases "Best" models were selected. In all cases, there were no or little differences whether or not PCs selection was applied. This is a direct consequence of the fact that the first PCs account for most of the variability and we used the corresponding eigenvalues to weight PCs, as discussed in Section 3.1. Comparing our results with those of Jolliffe (1972, Tab. 4) we note that the overall performance of our selection algorithm is comparable with the procedures considered by Jolliffe. In particular, it appears to perform equally or better than those in Jolliffe (1972), except for his method B4 which has better results in the case of model II. One remarkable aspect is that our greedy search procedure always selected at least "Good" subsets, so it never selected "Bad" or even "Moderate" subsets.

Table 2. *Percentage of times based on 1000 simulations the greedy search algorithm selects the different types of subset.*

Model	"True" dimensionality (d)	PCs selection method	Type of subset			
			Bad	Moderate	Good	Best
I	3	Kaiser's rule	0	-	-	100
I	3	none	0	-	-	100
II	3	Kaiser's rule	0	-	100	0
II	3	none	0	-	100	0
III	3	Kaiser's rule	0	0	63.3	36.7
III	3	none	0	0	62.7	37.3
IV	4	Kaiser's rule	0	-	-	100
IV	4	none	0	-	-	100

4.2 PCA: Alate adelges data

These data were analyzed originally by Jeffers (1967) and later by various authors, including Jolliffe (1973). The dataset consists of 19 variables measuring body parts on 40 alate adelges. PCA based on the correlation matrix provides a first component which accounts for a large proportion (73.0%) of the total variation, a second component accounting for 12.5% of total variation, and the third component with 3.9%. Two components are surely needed, peraphs with some evidence for the third one. Jolliffe (2002) discussed results from applying several subset selection methods proposed in literature.

We applied the proposed greedy search for subset selection to the first three PCs and we obtained the results shown in Table 3. These results are also reported graphically in Figure 1. Only two variables are needed to achieve a 90% of total variation of the selected PCA components. The best 3-variables subset $\{13, 17, 11\}$, which accounts for a 95% of total variation, is also selected by two out of four selection methods reported by Jolliffe (2002, Table 6.4). The best 4-variables subset $\{13, 11, 5, 18\}$ is equal to one of those reported by Jolliffe (2002), and it differs from another subset only by the use of variable 17 in place of 18, but, as it can be seen in Table 3, they appears to provide almost the same information, so they can be used exchangeably. The marginal contribution of each term rapidly decreases as the number of variables are included in the subset, becoming almost null after the first five or six variables are considered (see bottom of Figure 1).

It is interesting to note that if variable selection is performed on just

Table 3. *Subset selection results from greedy search algorithm for PC directions on the alate adelges data.*

Step	Included	Excluded	Size	SS	R^2
1	tibia (13)		1	366.53	0.78586
2	ovispi (17)		2	423.11	0.90716
3	antspi (11)		3	446.38	0.95707
4	numspi (5)		4	454.49	0.97446
5	anal (18)	ovispi (17)	4	456.24	0.97821
6	ovispi (17)		5	459.89	0.98602
7	numhooks (19)		6	461.80	0.99012
8	antseg2 (7)		7	463.14	0.99300
9	ovipos (16)		8	463.85	0.99453
10	fwing (3)		9	464.55	0.99603
11	rostrum (15)		10	464.97	0.99691
12	antseg4 (9)		11	465.26	0.99754
13	antseg5 (10)		12	465.52	0.99810
14	hwing (4)	antseg2 (7)	12	465.60	0.99827
15	antseg1 (6)	fwing (3)	12	465.60	0.99827
16	length (1)	hwing (4)	12	465.65	0.99839
17	hwing (4)		13	465.92	0.99895
18	antseg3 (8)		14	466.15	0.99945
19	fwing (3)		15	466.25	0.99966
20	width (2)		16	466.30	0.99977
21	antseg2 (7)		17	466.34	0.99985
22	tarsus (12)		18	466.39	0.99996
23	femur (14)	tibia (13)	18	466.39	0.99997
24	tibia (13)		19	466.41	1.00000

the first two PCA components, the best 4-variable subset $\{13, 18, 5, 17\}$ does not contain the previously included variable 11 ("number of antennal spines"). In fact, this variable dominates the third PC with a coefficient whose size is five times as large as any other variable. Thus, the selection procedure correctly discard such variable whose contribution is not needed for explaining variation on the first two components.

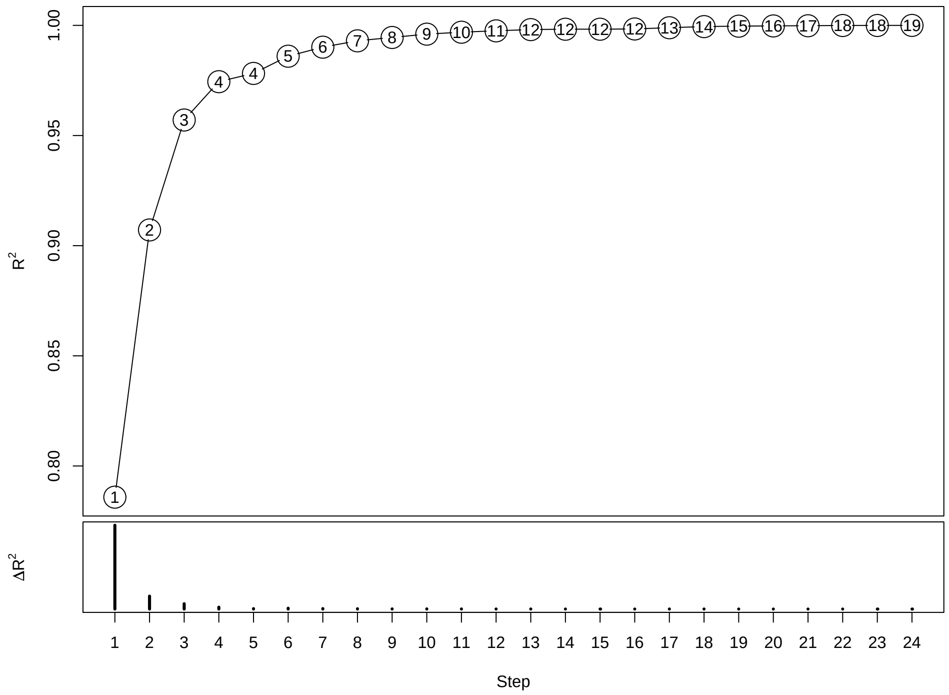

Figure 1. *R^2 values obtained at each step of the greedy search algorithm. For any point a number shows the size of the subset selected at each step. The graph at the bottom is a barplot of first differences in the R^2 criterion.*

4.3 ICA: simulation data

Consider the artificial signals shown at the top panel of Figure 2 and the observed mixed signals shown in Figure 3. The latter were generated using the mixing coefficients matrix $\boldsymbol{A} = \left[\begin{smallmatrix} 1 & -1 & 0.5 \\ 1 & 1 & 0.5 \end{smallmatrix}\right]$, and with the last feature generated from an independent gaussian random variable with mean zero and standard deviation 0.1. Therefore, the first two features contain all the information needed to recover the original source signals, the third feature being redundant once the first two have already been taken into account, and with the last feature which is irrelevant being simply noise.

ICA aims at recovering the source signals from the observed signals in Figure 3. Estimates are obtained using the FastICA algorithm (Hyvarinen and Oja, 2000) and they are shown in the bottom panel of Figure 2. Except for a change of sign, the estimated signals are almost identical to the source signals. However, not all the observed signals are required to obtain such estimates and we would like to identify only those features really needed.

We applied the proposed procedure in order to select a subset of the observed mixture signals which maximally explain the estimated ICA components. From Table 4 we can see that the first two mixed signals are correctly identified and they provide an almost perfect representation of the estimated ICA components, while the remaining observed signals can be quietly ignored. Applying the FastICA algorithm to the subset containing only the first two observed signals, we obtained components indistinguishable from those obtained using all the observed signals.

Table 4. *Subset selection results from the greedy search algorithm applied to estimated ICA components on artificial signals.*

Step	Included	Excluded	Size	SS	R^2
1	X_2		1	250.00	0.50000
2	X_1		2	499.81	0.99962
3	X_3		3	500.00	1.00000
4	X_4		4	500.00	1.00000

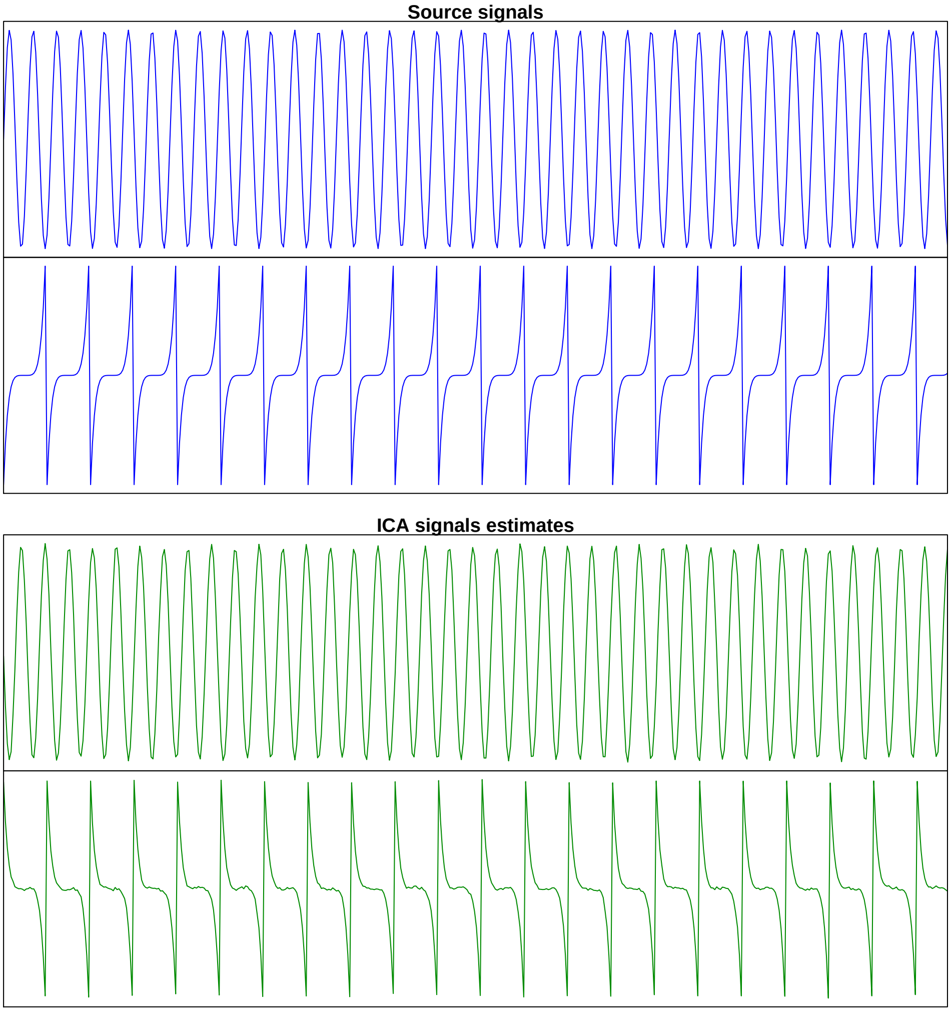

Figure 2. *The source signals (top panel) and ICA estimates of the original source signals (bottom panel).*

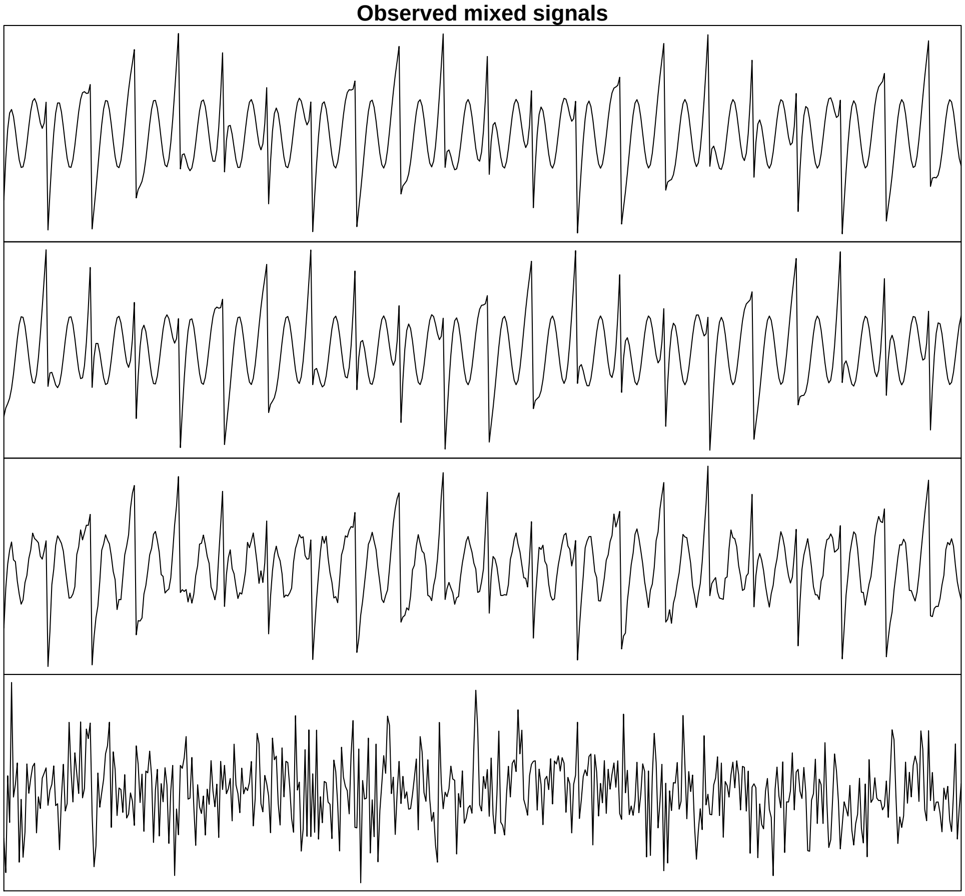

Figure 3. *The observed signals generated from a mixture of the underlying source signals shown in the top panel of Figure 2.*

4.4 LDA: Iris data

The well-known Iris dataset provides measurements on 4 characteristics (sepal length and sepal width, petal length and petal width) for 150 samples of either Iris Setosa, Versicolor or Virginica (Fisher, 1936). Prior to any formal discriminant analysis, it is often useful to graphically evaluate the existence of a natural grouping of cases.

The plot of canonical variates for the Iris dataset is shown on the top left panel of Figure 4. Since we can estimate at most $\min(p, g-1)$ directions, where g is the number of groups or classes, such two-dimensional graph contains all the information available from variation in group-means.

We applied our subset selection procedure and we obtained the results reported in Table 5. It is evident that Petal length accounts for a large amount of variations (96.16%), while Sepal length provides a negligible net contribution (about 0.2%).

Plots of canonical variates estimated using the subsets of best-2 and best-3 variables show the primary groups structure (see bottom panels of Figure 4). Using the subset {Pental length, Sepal width} the first LD direction is largely recovered ($R^2 = 0.9844$), but there are some differences in the second LD ($R^2 = 0.7379$). This can be further improved using Petal width, leading to essential the same LD directions obtained using all the features (in fact, R^2 is equal to 0.998 and 1.00 for the first and second LD, respectively).

Table 5. *Subset selection results from greedy search algorithm for canonical variates on the Iris data.*

Step	Included	Excluded	Size	SS	R^2
1	Petal length		1	4652.3	0.96162
2	Sepal width		2	4753.3	0.98250
3	Petal width		3	4828.2	0.99798
4	Sepal length		4	4838.0	1.00000

Given the small number of variables, we also conducted an exhaustive search over all possible subsets, obtaining the R^2 values shown in the top-right panel of Figure 4. In this graph, values reported by the greedy search algorithm are connected by a line: for any subset size the greedy search correctly identified the subset with the largest value of R^2.

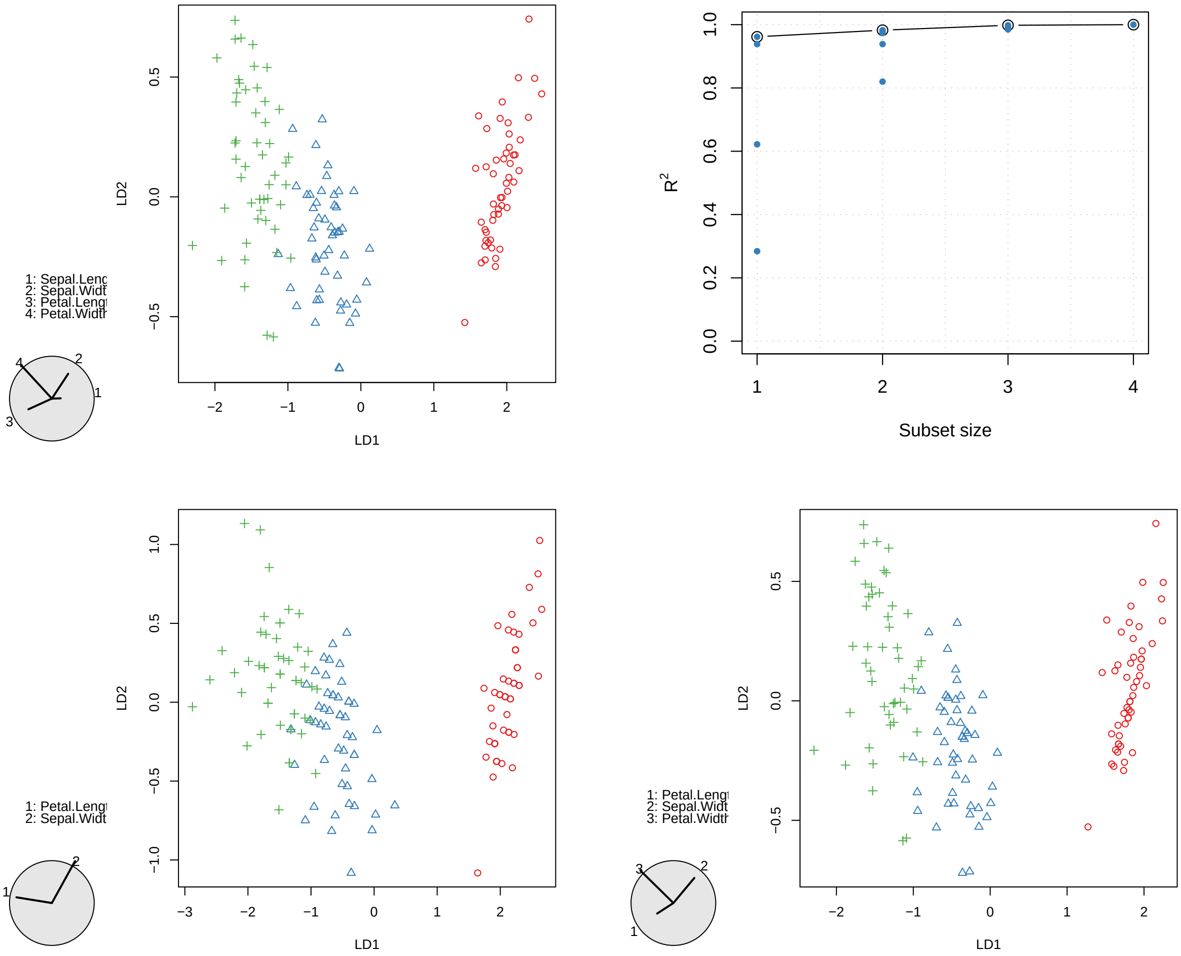

Figure 4. *Plot of canonical variates obtained using the full set of variables (top-left panel), the best 2-variables subset (bottom-left panel) and the best 3-variables subset (bottom-right panel). Points are marked according to the Iris species: Setosa=*○, *Versicolor=*△, *Virginica=*+. *The top-right panel shows the R^2 values obtained from an exhaustive search over all possible subsets; points connected by a line indicate the values for the subsets selected at each step by the greedy search algorithm.*

4.5 SAVE: banknote data

SAVE was applied by Cook (2000) to the Swiss bank notes data (Flury and Riedwyl, 1988). These give measurements, made on 100 genuine and 100 counterfeit notes, regarding different aspects of the size of a note (length at the top, bottom, left and right edges, and along the diagonal and center). Based on the summary plot obtained using the first two estimated SAVE directions (see the left panel of Figure 5), Cook argued that genuine notes could be accurately discriminate from counterfeit notes based on this summary plot, but he also noted the presence of a bimodal distribution among counterfeit notes and an outlying authentic note. We applied the proposed feature selection algorithm to such directions with weights given by the corresponding eigenvalues $(0.8715, 0.4314)$. Results are shown in Table 6: some features may clearly be dropped, since two or three of them explain a large amounts of variation in the estimated SAVE directions. Selecting the three variables which provides the largest R^2, we re-estimated the SAVE directions and obtained the summary plot shown in the right panel of Figure 5: this appears to be a very close approximation to the graph obtained from the full set of predictors. In particular, the above mentioned characteristics (separation between type of notes, bimodal distribution of counterfeit notes, and the presence of an outlier) are still visually evident.

The above feature selection analysis was based on the greedy search algorithm discussed in Section 3.2. However, given the small number of predictors, it is feasible to fully evaluate all the $2^6 - 1 = 63$ subset of size k, with k ranging from 1 up to 6. Figure 6 shows the R^2 values obtained for all possible features subsets, with those selected in the search path from the greedy search connected by a line. As it can be seen, the proposed algorithm always selected the subset with the largest R^2 value for any subset size.

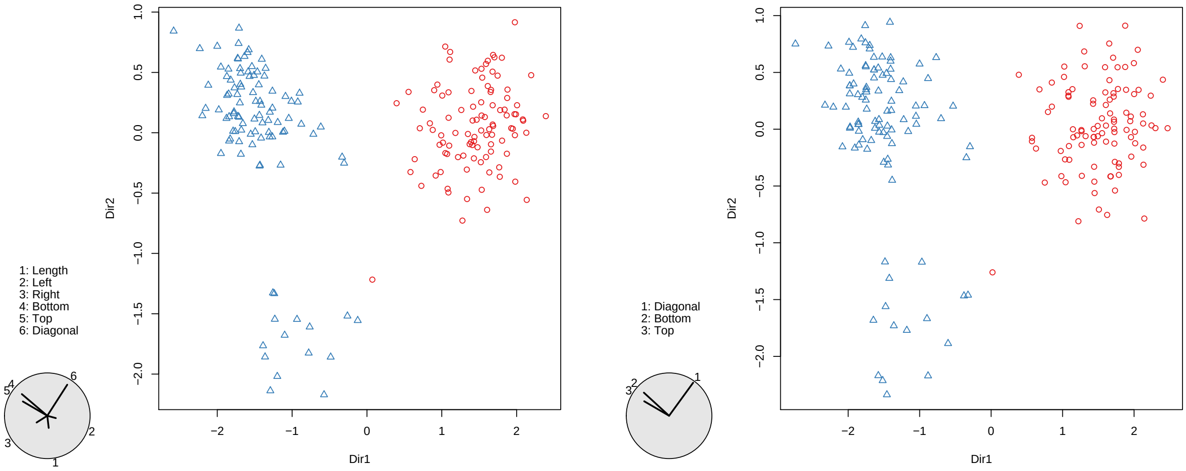

Figure 5. *Summary plots from SAVE for the bank note data based on all $p = 6$ predictors (left panel) and the selected predictors subset (right panel). The symbol* ○ *denotes genuine notes,* △ *counterfeit notes.*

Table 6. *Subset selection results from greedy search algorithm for SAVE directions on Swiss bank note data.*

Step	Included	Excluded	Size	SS	R^2
1	Diagonal		1	196.12	0.59848
2	Bottom		2	298.89	0.91208
3	Top		3	321.10	0.97988
4	Length		4	326.25	0.99558
5	Right		5	327.45	0.99925
6	Left		6	327.70	1.00000

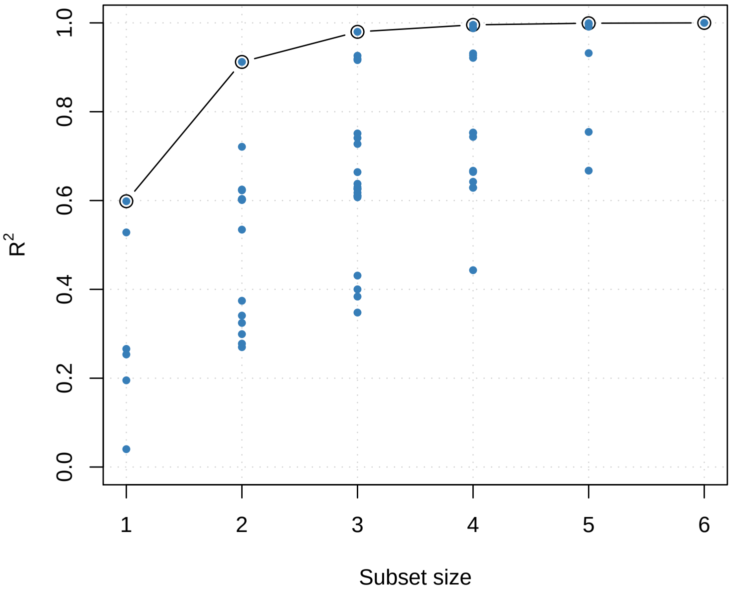

Figure 6. *R^2 values obtained from an exhaustive search over all possible subsets; points connected by a line indicate the values for the subsets selected at each step by the greedy search algorithm.*

4.6 Multivariate SIR: bleaching of cotton

Recently the SIR method for dimension reduction in regressions has been extended for dealing more efficiently with multivariate responses. Setodji and Cook (2004) proposed a new way of performing the slicing based on the k-means algorithm. The basic idea is to use the clusters, obtained through a slightly modified k-means algorithm to ensure a minimal cluster size, as a discrete response variable for slicing.

The proposed procedure was applied to a dataset used for studying the effects of four predictors in the pressure-kier bleaching of cotton measured by three response variables (for further details see Setodji and Cook (2004), Box and Draper (1987), p. 397). They claimed that only one SIR variate $\boldsymbol{z} = \boldsymbol{b}^\top \boldsymbol{X}$ is needed, where $\boldsymbol{b} = (.257, .916, .055, .304)^\top$. Based on the magnitude of the third coefficient, confirmed also by inspection of the coefficients for marginally standardized predictors, they declare X_3 the least important predictor since its coefficient is the smallest.

We applied the proposed greedy search to this dataset and we obtained the results shown in Table 7. The conclusion about the importance of X_3 is also supported by our subset selection analysis, with X_3 being the last predictor to enter the subset and with a contribution of less than 3%. A full search among all possible subsets was also conducted (see Figure 7). This indicates two aspects: (i) the greedy-search always selected the subset with the largest R^2 criterion for each subset size k ($k = 1, \ldots, 4$); (ii) the "best" $k = 3$ subset, namely $\{X_1, X_2, X_4\}$, is closely followed by the subset $\{X_2, X_3, X_4\}$ with $R^2 = 0.93213$, so the first and the third predictor provides almost the same information, with a slight prevalence for the former subset.

Finally, the coefficients estimated on the subset with X_3 removed are equal to $(.254, .913, .321)^\top$, very close to those obtained on the full set of predictors, and, consequently, the plots of each response variable versus the corresponding SIR variate (not shown) are basically identical to those reported by Setodji and Cook (2004, Fig. 1).

Table 7. *Subset selection results from greedy search algorithm for the estimated SIR variate.*

Step	Included	Excluded	Size	SS	R^2
1	X_4		1	8.4068	0.51306
2	X_2		2	14.7150	0.89806
3	X_1		3	15.9180	0.97145
4	X_3		4	16.3860	1.00000

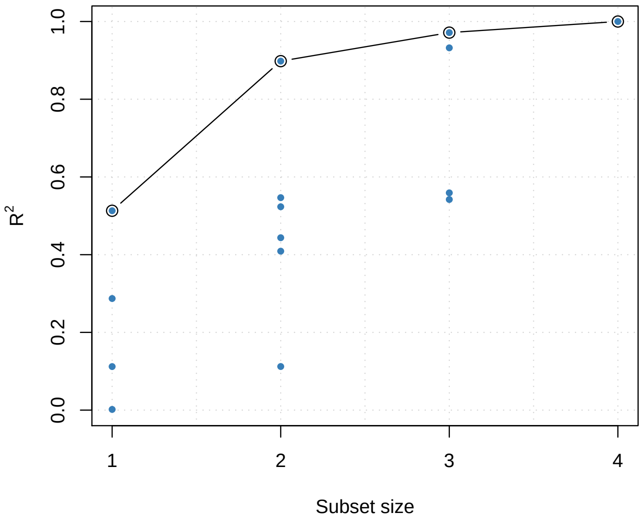

Figure 7. R^2 *values obtained from an exhaustive search over all possible subsets; points connected by a line indicate the values for the subsets selected at each step by the feature selection algorithm.*

5 Discussion

Dimension reduction methods play an important role in multivariate statistical analysis. Some of them are linear methods and they can be seen as a linear mapping from the original feature space to a dimension reduction subspace which hopefully will retain most of the relevant statistical information available in the data. However, highly correlated variables provide redundant information, whereas some other features may be irrelevant, and we would like to identify and then discard both of them while pursuing dimension reduction.

In this paper we proposed a greedy search algorithm for ranking subsets of variables based on their ability to explain variation in the dimension reduction variates. This greedy algorithm allows to avoid the search over all possible subsets, a number which soon becomes unfeasible even for moderates number of variables, say $p > 10$. The proposed greedy search is a forward-backward algorithm type which selects the "best" variable to be included among those not already selected, and then it assesses if any of the previous selected variables has became redundant and it could be dropped. If p is very large, as for instance in case of microarray data, the backward step may be skipped to improve computationally efficiency (Scrucca, 2006).

The proposed methodology has been applied to several simulated and real datasets, using different statistical techniques: projection pursuit tasks, from principal components analysis (PCA) to independent component analysis (ICA), classification settings, based on linear discriminant coordinates (LDA), and regression problems, using sliced inverse regression (SIR) and sliced average variance estimator (SAVE). In all cases we were able to find a reduced subset of variables while preserving the information contained in the dimension reduction subspace estimated from the full set of original variables.

No formal assessment on the best subset, i.e. how many variables are needed, is provided. We argue that the decision on how many variables to use should depend on the aim of the analysis. For example, in classification problems the ranked subsets could be evaluated on the basis of their misclassification error based on a test set or on a cross-validated set; in this case the subset with the smallest misclassification error should be selected. In visualization problems, graphical inspection of the results for increasing subset size compared to the configuration obtained from the full set of variables may lead to a final decision.

Finally, the greedy search algorithm has been implemented in R, a language and environment for statistical computing, freely available under GPL license (R Development Core Team, 2006). Source code is freely available upon request from the author.

References

Box, G.E.P. and Draper, N. (1987). *Empirical Model-Building and Response Surfaces.* New York: Wiley.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123–144.

Cook, D., Buja, A. and Cabrera, J. (1993). Projection Pursuit Indexes Based on Orthonormal Function Expansions. *Journal of Computational and Graphical Statistics*, **2** (3), 225–250.

Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, **4** (3), 155–172.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics.* New York: Wiley.

Cook, R.D. and Weisberg, S. (1991). Comment on “Sliced inverse regression for dimension reduction” by K.C. Li. *Journal of the American Statistical Association*, **86**, 328–332.

Cook, R.D. and Weisberg, S. (1999). *Applied regression including computing and graphics.* New York: Wiley.

Cook, R. D. and Yin, X. (2001). Dimension-reduction and visualization in discriminant analysis. *Australia and New Zealand Journal of Statistics*, **43**, 147–200.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics. A practical approach.* New York: Chapman and Hall.

Friedman, J. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association*, **82**, 249–266.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York: Wiley.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, 417–444.

Hyvarinen, A., Oja, E., (2000). Independent component analysis: algorithms and applications. *Neural Networks*, **13**, 411–430.

Hyvarinen, A., Karhunen, J., Oja, E., (2001). *Independent component analysis.* John Wiley & Sons: Toronto.

Jeffers, J.N.R. (1967). Two case studies in the application of principal component analysis, *Applied Statistics*, **16**, 225–236.

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis II - Real data, *Applied Statistics*, **22**, 21–31.

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

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–342.

Li K.C. (2000). High dimensional data analysis via the SIR/PHD approach. *Unpublished manuscript.*

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press: London.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **6**, 2, 559–572.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ripley, B.D. (1996). *Pattern recognition and neural networks.* Cambridge University Press: Cambridge, UK.

Scrucca, L. (2006). Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. Submitted to *Computational Statistics & Data Analysis.*

Setodji, M.C. and Cook, R.D. (2004). K-means inverse regression. *Technometrics*, **46**, 421–429.

Velilla, S. (1993). A note on the multivariate Box-Cox transformations to normality. *Statistics and Probability Letters*, **17**, 315–322.

Vines, S.K. (2000). Simple principal components, *Applied Statistic*, **49**, 441–451.

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

QUADERNI DEL DIPARTIMENTO DI ECONOMIA, FINANZA E STATISTICA
Università degli Studi di Perugia

1	Gennaio 2005	Giuseppe CALZONI Valentina BACCHETTINI	Il concetto di competitività tra approccio classico e teorie evolutive. Caratteristiche e aspetti della sua determinazione
2	Marzo 2005	Fabrizio LUCIANI Marilena MIRONIUC	Ambiental policies in Romania. Tendencies and perspectives
3	Aprile 2005	Mirella DAMIANI	Costi di agenzia e diritti di proprietà: una premessa al problema del governo societario
4	Aprile 2005	Mirella DAMIANI	Proprietà, accesso e controllo: nuovi sviluppi nella teoria dell'impresa ed implicazioni di corporate governance
5	Aprile 2005	Marcello SIGNORELLI	Employment and policies in Europe: a regional perspective
6	Maggio 2005	Cristiano PERUGINI Paolo POLINORI Marcello SIGNORELLI	An empirical analysis of employment and growth dynamics in the italian and polish regions
7	Maggio 2005	Cristiano PERUGINI Marcello SIGNORELLI	Employment differences, convergences and similarities in italian provinces
8	Maggio 2005	Marcello SIGNORELLI	Growth and employment: comparative performance, convergences and co-movements
9	Maggio 2005	Flavio ANGELINI Stefano HERZEL	Implied volatilities of caps: a gaussian approach
10	Giugno 2005	Slawomir BUKOWSKI	EMU – Fiscal challenges: conclusions for the new EU members
11	Giugno 2005	Luca PIERONI Matteo RICCIARELLI	Modelling dynamic storage function in commodity markets: theory and evidence
12	Giugno 2005	Luca PIERONI Fabrizio POMPEI	Innovations and labour market institutions: an empirical analysis of the Italian case in the middle 90's
13	Giugno 2005	David ARISTEI Luca PIERONI	Estimating the role of government expenditure in long-run consumption
14	Giugno 2005	Luca PIERONI Fabrizio POMPEI	Investimenti diretti esteri e innovazione in Umbria
15	Giugno 2005	Carlo Andrea BOLLINO Paolo POLINORI	Il valore aggiunto su scala comunale: la Regione Umbria 2001-2003

16	Giugno 2005	Carlo Andrea BOLLINO Paolo POLINORI	Gli incentivi agli investimenti: un'analisi dell'efficienza industriale su scala geografica regionale e sub regionale
17	Giugno 2005	Antonella FINIZIA Riccardo MAGNANI Federico PERALI Paolo POLINORI Cristina SALVIONI	Construction and simulation of the general economic equilibrium model Meg-Ismea for the italian economy
18	Agosto 2005	Elżbieta KOMOSA	Problems of financing small and medium-sized enterprises. Selected methods of financing innovative ventures
19	Settembre 2005	Barbara MROCZKOWSKA	Regional policy of supporting small and medium-sized businesses
20	Ottobre 2005	Luca SCRUCCA	Clustering multivariate spatial data based on local measures of spatial autocorrelation
21	Febbraio 2006	Marco BOCCACCIO	Crisi del welfare e nuove proposte: il caso dell'unconditional basic income
22	Giugno 2006	Mirko ABBRITTI Andrea BOITANI Mirella DAMIANI	Unemployment, inflation and monetary policy in a dynamic new keynesian model
23	Settembre 2006	Luca SCRUCCA	Subset selection in dimension reduction methods

ISSN 1722-618X

I QUADERNI DEL DIPARTIMENTO DI ECONOMIA
Università degli Studi di Perugia

1	Dicembre 2002	Luca PIERONI:	Further evidence of dynamic demand systems in three european countries
2	Dicembre 2002	Luca PIERONI Paolo POLINORI:	Il valore economico del paesaggio: un'indagine microeconomica
3	Dicembre 2002	Luca PIERONI Paolo POLINORI:	A note on internal rate of return
4	Marzo 2004	Sara BIAGINI:	A new class of strategies and application to utility maximization for unbounded processes
5	Aprile 2004	Cristiano PERUGINI:	La dipendenza dell'agricoltura italiana dal sostegno pubblico: un'analisi a livello regionale
6	Maggio 2004	Mirella DAMIANI:	Nuova macroeconomia keynesiana e quasi razionalità
7	Maggio 2004	Mauro VISAGGIO:	Dimensione e persistenza degli aggiustamenti fiscali in presenza di debito pubblico elevato
8	Maggio 2004	Mauro VISAGGIO:	Does the growth stability pact provide an adequate and consistent fiscal rule?
9	Giugno 2004	Elisabetta CROCI ANGELINI Francesco FARINA:	Redistribution and labour market institutions in OECD countries
10	Giugno 2004	Marco BOCCACCIO:	Tra regolamentazione settoriale e antitrust: il caso delle telecomunicazioni
11	Giugno 2004	Cristiano PERUGINI Marcello SIGNORELLI:	Labour market performance in central european countries
12	Luglio 2004	Cristiano PERUGINI Marcello SIGNORELLI:	Labour market structure in the italian provinces: a cluster analysis
13	Luglio 2004	Cristiano PERUGINI Marcello SIGNORELLI:	I flussi in entrata nei mercati del lavoro umbri: un’analisi di cluster
14	Ottobre 2004	Cristiano PERUGINI:	Una valutazione a livello microeconomico del sostegno pubblico di breve periodo all’agricoltura. Il caso dell’Umbria attraverso i dati RICA-INEA
15	Novembre 2004	Gaetano MARTINO Cristiano PERUGINI	Economic inequality and rural systems: empirical evidence and interpretative attempts
16	Dicembre 2004	Federico PERALI Paolo POLINORI Cristina SALVIONI Nicola TOMMASI Marcella VERONESI	Bilancio ambientale delle imprese agricole italiane: stima dell’inquinamento effettivo