Chemometrics and Intelligent Laboratory Systems 136 (2014) 147–154

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A novel variable reduction method adapted from space-filling designs

Davide Ballabio [a,*], Viviana Consonni [a], Andrea Mauri [a], Magalie Claeys-Bruno [b], Michelle Sergent [b], Roberto Todeschini [a]

[a] Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milano, Italy
[b] Aix Marseille Université, LISA EA4672, 13397, Marseille Cedex 20, France

ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 20 May 2014
Accepted 24 May 2014
Available online 2 June 2014

Keywords:
Unsupervised variable reduction
Wootton
Sergent
Phan-Tan-Luu's algorithm
Linear correlation

ABSTRACT

Unsupervised variable reduction methods are intended for reducing the presence of redundancy and multicollinearity in the data. These are common issues when dealing with multivariate analysis associated to large number of variables. With respect to supervised selection, unsupervised reduction aims at selecting subsets of variables able to preserve information, but eliminating redundancy, noise and linearly or near-linearly dependent variables, without considering any dependent response.
In this study, we propose the V-WSP algorithm for unsupervised variable reduction, which is a modification of the recently proposed WSP algorithm for design of experiments (DOE). Convergence, performances and comparison with several benchmark algorithms, as well as with other DOE strategies adapted to variable reduction, were evaluated on both simulated, benchmark and real QSAR datasets. The proposed algorithm demonstrated to converge to similar solutions with respect to other reduction strategies, with the advantage to be faster and simpler.

1. Introduction

The large number of variables and the associated presence of redundancy, multicollinearity, random noise and chance correlation are common problems when dealing with multivariate modelling [1–3]. The presence of irrelevant variables can change the underlying data patterns and consequently it can influence results of several multivariate methods.

The problem of data correlation is relevant in Quantitative Structure Activity/Property relationship (QSAR/QSPR) approaches, which analyse the relationships between molecular properties and suitable sets of molecular descriptors calculated using computational methods. This issue has proved difficult due to the amounts of redundancy and multicollinearity contained in QSAR data sets, since nowadays thousands of descriptors can be easily calculated. However, QSAR models should be parsimonious in order to give stable and reliable predictions and thus only relevant descriptors should be included in the model, while descriptors contributing to redundancy and multicollinearity of the data should be removed [4].

Therefore, a common strategy for overcoming the problem of data correlation is to decrease the number of variables. This can be carried out by means of both unsupervised (variable reduction) and supervised (variable selection) algorithms. When dealing with supervised selection, such as for Genetic Algorithms coupled with regression or classification methods, a response to be modelled is taken into account in order to achieve the selection. While supervised selection is moderately well known, this is not the case for unsupervised variable reduction, which refers to the procedure that aims at selecting a subset of variables able to preserve as much information of the original data as possible, but eliminating redundancy, noise and linearly or near-linearly dependent variables, without taking into account a dependent response. Moreover, unsupervised reduction can facilitate the subsequent supervised selection, which can suffer from the presence of highly correlated data and chance correlation, thus giving overfitted results [5].

The majority of unsupervised methods for variable reduction proposed in literature are based on linear correlation between variables [6,4], as well as eigenvalues obtained by singular value decomposition [2,7] and loadings of Principal Component Analysis [8].

In this study, we propose an adaptation of the WSP method, which has been developed for space-filling designs of experiments (SFD) to variable reduction (V-WSP). In fact, several DOE methods are related to the selection of representative sets of samples [9–13]. Here, we translated this purpose to the selection of a representative set of variables based on linear correlation. In the first part of the paper, theory of the

* Corresponding author at: Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza, 1-20126 Milano, Italy.
E-mail address: davide.ballabio@unimib.it (D. Ballabio).

http://dx.doi.org/10.1016/j.chemolab.2014.05.010
0169-7439/

proposed algorithm is introduced. Then, the performance of V-WSP is evaluated on both simulated, benchmark and real QSAR datasets and its effectiveness is discussed by comparison with results of other algorithms for variable reduction. Finally, results of supervised selections performed on both the original and reduced sets of variables were compared.

2. Materials and methods

2.1. WSP algorithm adapted for unsupervised variable reduction

Recently, a construction method of new space-filling designs for high dimensional spaces was proposed [10]. This was derived from the so called WSP designs based on Wootton, Sergent, Phan-Tan-Luu's algorithm. The construction of WSP designs is established on the selection of well distributed points in accordance with the algorithm proposed by Sergent et al. [14–16]. Points are chosen from a set of candidate points so as to be at a pre-fixed minimal Euclidean distance from every point in the defined multidimensional space, but WSP can also support adaptive corrections for specific problems [17].

In this study, the proposed WSP algorithm was adapted in order to select a representative set of variables instead of points. Variables are chosen in an unsupervised way so as to be at a fixed minimal correlation from every variable in the defined multidimensional space. Given a data matrix with n rows (samples) and p columns (variables), the algorithm for calculating the V-WSP method is the following:

- step 1: choose an initial variable (seed) j and a correlation threshold (thr);
- step 2: calculate the Pearson linear correlation coefficients ($\mathbf{c}$) between j and all other variables;
- step 3: eliminate variables d such as absolute value of $c_{dj} \geq thr$;
- step 4: variable j is selected and replaced by the variable with the highest absolute correlation value with j among the remaining variables;
- step 5: repeat steps 2, 3 and 4 until there are no more variables to select.

2.2. Parameters for variable reduction evaluation

Results and comparison between full and reduced sets of variables were analysed by means of two parameters. The amount of correlation and redundancy in the reduced set of variables was quantified by means of the K multivariate correlation index [7,18]. This is defined in terms of the distribution of eigenvalues obtained by the diagonalization of the correlation matrix of the data set and it is equal to 1 when all variables are perfectly correlated, while it is equal to 0 when variables are orthogonal.

The similarity between the structure information of the complete set of variables and the reduced subset was quantified with a Procrustes criterion. Procrustes analysis is a statistical method to match two data sets measured from the same samples with different sets of variables. It determines a linear transformation, based on translation, reflection, orthogonal rotation, and scaling, of the points in the first data set to best conform them to the points in the second data set [19–21]. The Procrustes goodness-of-fit criterion is the sum of squared errors; it is equal to 0 if two datasets coincide, while it is equal to 1 if data structures are completely dissimilar.

2.3. Benchmark algorithms for variable reduction

Performance of V-WSP was evaluated by comparison with the following methods for variable reduction. Originally proposed by Jolliffe [8], B2 and B4 methods are based on loadings of Principal Component Analysis (PCA). The B2 method consists in a sequential analysis of all the Principal Components (PC), starting from the last one (the less significant). For each PC, the first not already chosen variable with the highest absolute loading value is removed. In the non-iterative version this is made only once; in the iterative version, PCs are calculated every time a variable is removed from the dataset. The idea beyond this method is that last PCs bring the less relevant information (i.e. redundancy and noise), thus variables that most represent these PCs are those related to redundancy and noise in the dataset. The B4 method consists in a sequential analysis of all the principal components, starting from the first one. For each PC, the first not already chosen variable with the highest absolute loading value is selected. Since the first PCs have most of the information, variables which are most representative of those first PCs are retained in the dataset. In order to choose the number of variables to be retained, the number of significant PC must be selected. A simple method based on eigenvalues (Corrected Average Eigenvalue Criterion, CAEC) was adopted in this study: CAEC accepts as significant only the components with eigenvalue larger than the average eigenvalue multiplied by 0.7 [22]. Note that when data are autoscaled, the average eigenvalue is equal to 1.

The K Inflation factor (KIF) is a variable reduction method based on the K multivariate correlation index [7]. This method is based on the idea that data structure is mostly preserved by removing the variable q for which the remaining variables show the minimum multivariate correlation. This means that when variable q is excluded from the data, the remaining multivariate correlation derived from the remaining variables is maximally decreased. The KIF_j value associated to the j-th variable is an inflation factor obtained by considering the total multivariate correlation K_p and the multivariate correlation index calculated on the data by removing the j-th variable, $K_{p/j}$. It is suggested to retain all variable associated with a KIF index value not greater than a suggested threshold equal to 0.50 [7].

The Pairwise correlation method is based on a simple algorithm, which is included in some commercial QSAR softwares, such as Dragon 6 [23]. For each pair of correlated descriptors (variables) with a correlation coefficient equal to or larger than a defined correlation threshold, the one showing the largest pair correlation with all the other descriptors is removed in an iterative way. Similar strategies were proposed in literature. For example, the CORCHOP algorithm identifies variables whose correlation with one another is higher than a predefined threshold and suggests an appropriate member of the pair to remove [6].

The Canonical Measure of Correlation (CMC index) between sets of variables is a method for determining the subset of variables that reproduce as well as possible the main structural features of the complete data set [2,24]. The CMC index can be used following a stepwise procedure, which consists in comparing each variable in turn with the entire set of available variables and excluding the most correlated one. The procedure is repeated iteratively by using the remaining variables until only two variables remain. At the end of this elimination procedure, variables can be ranked on the basis of their CMC values and the subset of variables with the smallest CMC values can then be included in the reduced set of variables.

Auto-Associative Multivariate Regression Trees (AAMRT) were suggested as variable reduction strategy. They are based on Multivariate Regression Trees (MRT), but in AAMRT variables are not only used as explanatory variables, but also as response variables. In this way, AAMRT divide samples into groups with similar response values by using explanatory variables, and variables in the tree nodes are supposed to be the most responsible for the cluster structure in the data. Therefore, the set of variables selected in the tree nodes can be retained as the result of the unsupervised data reduction [3].

Unsupervised Forward Selection (UFS) is a data reduction algorithm that starts with the two descriptors with the smallest correlation and selects additional descriptors based on their multiple correlations with those already chosen. The reduction process stops when the correlation value of each remaining variable with those already selected exceeds a defined threshold. Thus, UFS selects a reduced subset of variables that is as close to orthogonality as possible [4].

Since V-WSP is based on the same principles as the WSP algorithm for the selection of a representative set of samples, two other DOE algorithms were also considered and modified to variable reduction purposes. One is

the Kennard–Stone (KS) algorithm that selects a representative set of samples on the basis of their inter-distances [11]; the other is the Distance-Based Optimal Design (DBOD) that partitions samples in two sets on the basis of a Distance-Based Optimal Design [9]. In this study, the original KS and DBOD strategies were readapted, to be applied on variables instead of sample. The absolute value of the correlation coefficient between variable pairs was used as the similarity criterion and, in particular, 1 minus the absolute value of correlation coefficient was considered as the distance measure between variables.

2.4. Datasets

The performance of V-WSP was evaluated on two benchmark datasets (Aphid and Coffee) and two large QSAR datasets (Biodegradation and LogP).

Aphid (Alate adelges) consists of 19 different variables measured on 40 winged aphids (samples) [25]. Coffee is another benchmark dataset for variable reduction and consists of 43 samples described by 13 variables [26].

The Biodegradation dataset is constituted by 1055 chemicals and was used to develop QSAR models for the study of the relationships between chemical structure and biodegradation of molecules [27]. In this study, 758 molecular descriptors, belonging to 5 different blocks (40 constitutional indices, 72 topological indices, 385 2D matrix-based descriptors, 181 2D autocorrelations, 80 Burden eigenvalues), were calculated for each molecule by means of DRAGON software [23]. Since the effect of unsupervised variable reduction was evaluated by means of supervised selection performed on both the original and reduced sets of variables, an additional external validation set of molecules was considered for the final validation of the supervised selection. This external set was constituted of 670 molecules (191 ready biodegradable and 479 not ready biodegradable).

The LogP dataset is constituted by 12,403 molecules described by 1265 molecular descriptors, belonging to 11 different blocks: 42 constitutional indices, 32 ring descriptors, 72 topological indices, 46 walk and path counts, 37 connectivity indices, 48 information indices, 334 2D matrix-based descriptors, 213 2D autocorrelations, 96 Burden eigenvalues, 21 Eta indices, and 324 Edge Adjacency indices. Molecules were retrieved from the PHYSPROP dataset [28], which was previously used to calibrate QSAR models for predicting the octanol–water partition coefficient of molecules (LogP) [29]. All chemicals were screened and molecular structures checked in order to cure the dataset. Molecules were randomly divided into two subsets: 8683 molecules were used to perform the unsupervised variable reduction and the subsequent supervised variable selection based on both reduced and original sets of variables, while 3720 molecules were just used to validate the models obtained in the supervised selection.

Data are available for downloading together with the code for calculating the V-WSP algorithm, as detailed in the software section.

2.5. Software

Molecular descriptors of the Biodegradation and LogP datasets were calculated by means of DRAGON [23]. Multivariate Regression Trees (AAMRT) were calculated with the GUIDE software [30]. Unsupervised Forward Selection (UFS) was calculated by means of the software distributed by the Centre for Molecular Design, University of Portsmouth [4]. The Kennard–Stone algorithm for variable reduction was adapted from the MATLAB routine released by Michal Daszykowski (Department of Chemometrics, Institute of Chemistry, The University of Silesia) and available at http://www.chemometria.us.edu.pl [31].

All other calculations were performed in MATLAB (MathWorks) by means of routines built by the authors. The MATLAB routine for calculating the V-WSP algorithm and all datasets used in this study are available at the Milano Chemometrics and QSAR Research Group website (http://michem.disat.unimib.it/chm/download/wspinfo.htm). Since methods were compared also on the basis of their computational performances, all calculations were performed on the same machine (HP Z620 Workstation, processor 2.00 GHz, 16GB RAM).

3. Results

3.1. Demonstration of V-WSP on simulated and benchmark datasets

The use of the proposed V-WSP method is initially demonstrated on a simple simulated dataset comprised of 1000 samples and 100 variables. The dataset was created so that variables were divided in 5 uncorrelated blocks, while variables of each block had correlation higher than 0.80.

With a correlation threshold *thr* equal to 0.80, V-WSP included just 5 variables in the reduced set, one for each uncorrelated block. The *K* correlation index calculated on the full set of variables (0.82) was reduced to 0.04 when only the 5 selected variables were considered. Procrustes analysis on the first five principal components for the reduced and original data gave goodness-of-fit equal to 0.14. Five components were retained since five blocks of uncorrelated variables were created and therefore five sources of information were expected. The first five PCs explained 86% of variance of the original data.

V-WSP demonstrated to be able to include significant variables in the reduced set. In fact, one variable for each uncorrelated block was selected, as expected, and this allowed the reduction of redundancy and multicollinearity (the *K* correlation index was considerably reduced from 0.82 to 0.04) while preserving information on the data structure, as confirmed by the low Procrustes goodness-of-fit (0.14). Looking at loadings of the first two principal components calculated on the original set of 100 variables (Fig. 1), it is possible to identify the five uncorrelated blocks of variables. Variables included in the reduced set correctly covered the multidimensional space represented by all the original variables.

The V-WSP procedure was also applied on the benchmark Aphid and Coffee datasets (19 and 13 variables, respectively) by selecting one variable at a time as the algorithm seed. In this case, a number of solutions equal to the number of variables were obtained and thus, the selection frequency of variables was calculated, which allowed the variables to be ranked. The V-WSP solution is the set of the top-ranked variables (Tables 1 and 2).

The V-WSP correlation thresholds used for Aphid and Coffee were 0.85 and 0.50, respectively. The same correlation thresholds were

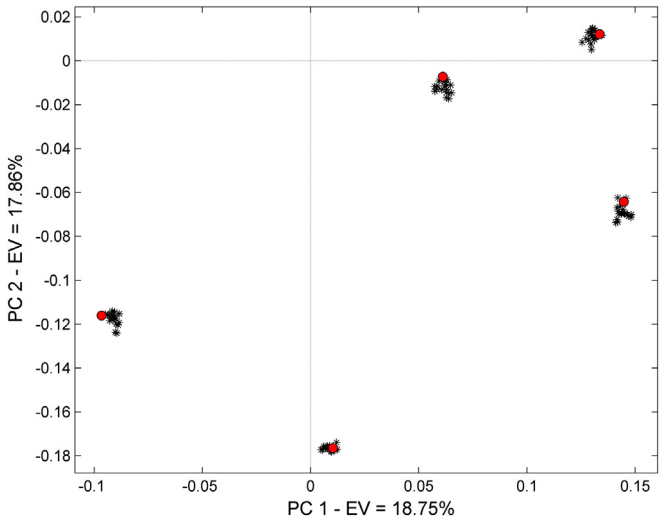

Fig. 1. Loadings of first and second principal components calculated on the full original simulated dataset. Variables selected by V-WSP are plotted as red circles. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

Table 1
Variable reduction results on the Aphid dataset.

Method	Size	Variables
V-WSP	5	5, 11, 17, 18, 19
KS	5	5, 11, 13, 17, 18
Pairwise correlation	5	5, 9, 11, 17, 19
DBOD	5	5, 10, 11, 17, 18
AAMRT	2	12, 13
CMC	5	5, 11, 17, 18, 19
UFS	5	5, 9, 11, 17, 19
KIF method	5	5, 9, 11, 18, 19
B2	4	2, 5, 11, 18
B2 iterative	4	5, 9, 11, 19
B4	4	5, 11, 13, 17

applied to UFS and Pairwise correlation methods. Thresholds were selected to include in the reduced set of variables approximately half of the initial set of variables. The threshold selected for the Coffee dataset was lower since it has a lower degree of correlation between variables, the average absolute correlation value being equal to 0.40, while Aphid has an average absolute correlation value equal to 0.69.

For Aphid dataset, 5 variables out of 19 (namely variables 5, 11, 17, 18 and 19) are present in all the 19 V-WSP solutions. This solution exactly corresponds to that obtained by CMC algorithm, while partly matches solutions of the other reduction algorithms, as shown in Table 1. In particular, KS and DBOD selected a common subset of 4 variables (5, 11, 17, 18), as well as UFS and Pairwise correlation (5, 11, 17, 19), while KIF included variables 5, 11, 18, and 19. In previous analyses of the Aphid data, four or five variables were supposed to be necessary in order to account for as much information as that of the original set of variables [2]. In particular, variables 5, 11 and 17 resulted on average not much correlated with all the other variables and thus they were retained in the reduced sets of variables by several algorithms. On the other hand, variables 1, 7, 10, and 15 were never retained. Therefore, V-WSP results appear to be consistent with those obtained by the majority of the considered methods.

Results of variable reduction obtained on the Coffee dataset are listed in Table 2. Even on this dataset, V-WSP gave consistent results when compared to other methods. The subset of variables selected by V-WSP was very similar to those selected by KS, DBOD, Pairwise correlation and KIF. Variables 1, 2, 3, 4 and 6 were always included in the 13 V-WSP solutions. These variables were included in the reduced sets by the majority of considered algorithms. Variables 10 and 11 were selected 8 and 6 times in the V-WSP solutions, respectively. Even this couple of variables is represented in subsets selected by other algorithms. Finally, variables 7, 8 and 13 were those with less frequencies of selection in the V-WSP solutions: variables 8 and 13 were never selected by other methods, while variable 7 was retained by just two algorithms (UFS and B4).

Table 2
Variable reduction results on the Coffee dataset.

Method	Size	Variables
V-WSP	7	1, 2, 3, 4, 6, 10, 11
KS	7	1, 2, 3, 4, 6, 9, 10
Pairwise correlation	8	1, 2, 3, 4, 6, 9, 10, 11
DBOD	7	1, 2, 3, 4, 6, 11, 12
AAMRT	2	3, 10
CMC	9	1, 2, 3, 4, 6, 9, 10, 11, 12
UFS	5	1, 2, 3, 4, 7
KIF	8	1, 2, 3, 4, 6, 9, 10, 11
B4	6	1, 2, 3, 4, 7, 11
B2	6	1, 2, 4, 10, 11, 12
B2 iterative	6	1, 2, 3, 4, 6, 10

3.2. Method sensitivity to the correlation threshold

The effect of the correlation threshold to be used in the V-WSP procedure was evaluated on the Biodegradation QSAR dataset by changing the correlation threshold from 0.50 to 0.99. The number of selected variables, the *K* multivariate correlation index and Procrustes goodness of fit calculated between the scores of the first 4 PCs of the original and reduced sets of variables were used as performance indicators. The first 4 PCs were able to explain the majority of the data information in the original dataset, each PC being able to explain more than 4% of information and the cumulative explained variance being equal to 74%. If the reduced set includes only relevant variables, then the number of significant PCs from this set is supposed to be equal to that obtained from the original set of variables. Data were always autoscaled when calculating PCA.

Increasing the correlation threshold led to the increasing of the number of variables included in the reduced set, as expected (Fig. 2). In fact, step 3 of the V-WSP procedure eliminates variables with absolute value of correlation higher than the fixed threshold, *thr*; therefore, the lower the threshold is, the higher the number of removed variables is. When looking at Procrustes analysis, increasing the correlation threshold led to the decreasing of Procrustes goodness of fit. When selecting a high correlation threshold (which corresponds to a high number of included variables), the reduced data can better fit the original one, thus giving a lower Procrustes goodness of fit. However, the goal of variable reduction is not the preservation of the exact original data structure, but the elimination of redundant information. With Procrustes analysis it is possible to check how much the data structure (expressed in terms of PC scores) is changed after the removal of correlated variables. With a correlation threshold equal to 0.80, the data structure of the original set did not change considerably in the reduced set of variables (i.e., Procrustes goodness of fit equal to 0.27). Finally, increasing the correlation threshold led to a linear increasing of *K* correlation index, as expected, since more correlated variables were retained in the reduced set of variables.

3.3. Method sensitivity to the seed selection

In presence of large datasets, such as those associated to QSAR modelling, one major issue can be the computational time. Since QSAR data are often characterised by thousands of variables (molecular descriptors), it is not feasible to follow the same procedure used on the Aphid and Coffee datasets, that is, repeating the V-WSP algorithm by selecting one variable at a time as the algorithm seed. For QSAR

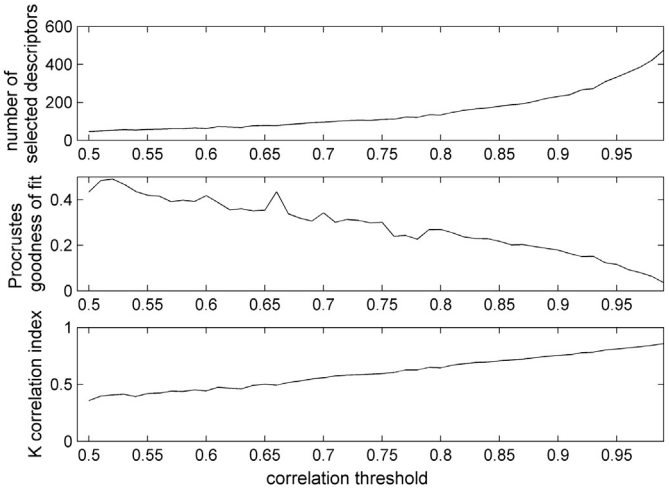

Fig. 2. Number of selected descriptors, Procrustes goodness of fit and *K* correlation index as a function of the correlation threshold in the V-WSP variable reduction of Biodegradation dataset.

datasets, the molecular weight (MW) can be used as a reasonable seed to carry out the V-WSP algorithm.

In order to evaluate the effect of the initialisation of the V-WSP algorithm, the following analysis was performed on the Biodegradation dataset, with correlation threshold of 0.80:

a) V-WSP was repeated by selecting one variable at the time as the seed, obtaining a number of reduced sets equal to the number of variables (758);
b) PCA was calculated on each reduced set and 4 PCs were always retained;
c) each reduced set was characterised in terms of the number of selected variables, the *K* multivariate correlation index and Procrustes goodness of fit (Fig. 3).

The 758 V-WSP solutions were generally composed of a comparable number of selected variables, in the range between 128 and 143, and with the *K* correlation index in the range between 0.63 and 0.66. Moreover, the maximum and average Procrustes goodness of fit were equal to 0.04 and 0.01, respectively. From these results, one can conclude that all the V-WSP solutions obtained by different initialisation had very similar data structures. Thus, these results can be considered as evidence of the convergence of the V-WSP algorithm to a consistent solution and low sensitivity to the algorithm seed.

3.4. Comparison with other variable reduction methods

The two QSAR datasets, LogP and Biodegradation, were used to evaluate the V-WSP algorithm performance in comparison with other variable reduction methods. These datasets were chosen for the comparison since they are real datasets characterised by a large number of variables and presence of multicollinearity.

The V-WSP procedure was carried out selecting the molecular weight (MW) as the seed of the algorithm. For both the datasets, the correlation threshold was set to 0.80, since this value gave reasonable results from early analyses.

The subset of variables obtained by means of V-WSP was compared with the subsets produced by the benchmark methods (B2, B4, Pairwise correlation, CMC, AAMRT, UFS, KIF) and the two DOE adapted algorithms (KS, DBOD). To implement these methods the following parameter values were selected: KIF threshold of 0.75; Pairwise correlation threshold equal to that used in the V-WSP algorithm (0.80); UFS correlation thresholds equal to 0.99 (i.e., default proposed by authors [4]) and 0.80 (i.e., the same correlation threshold adopted for V-WSP); for both KS and DBOD the a priori number of variables to be retained was set at the same number of variables retained by V-WSP; for CMC method, only variables with a CMC index lower than 0.2 were retained in the reduced set.

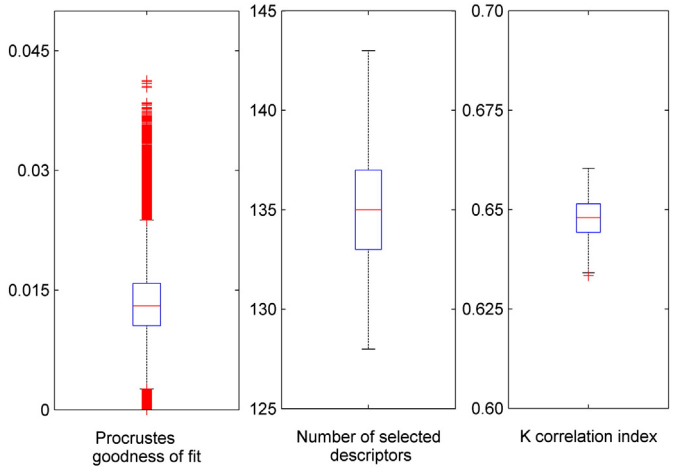

Fig. 3. Evaluation of V-WSP initialisation on the Biodegradation dataset: boxplot of Procrustes goodness of fit between 758 solutions, number of selected descriptors in each solution and *K* correlation index achieved on each solution. On each boxplot, the central mark is the median and the edges of the box are the 25th and 75th percentiles.

Strategies were compared in terms of the number of selected descriptors (i.e. variable subset size), Procrustes analysis goodness of fit considering the scores of the first 4 PCs of original and reduced data, *K* correlation index and computational time. Results achieved for the Biodegradation and LogP datasets are collected in Tables 3 and 4, respectively. For the Biodegradation dataset, the V-WSP algorithm led to a significant reduction of the original QSAR dataset providing with a subset of just 134 molecular descriptors out of 758 (corresponding to the 18% of the total number of variables). For this subset, the *K* correlation index was reduced from 0.91 (original data) to 0.65, indicating a significant reduction of the amount of correlation and redundancy. Finally, Procrustes analysis gave a goodness of fit equal to 0.27, indicating that the reduced set is able to reproduce the original data structure.

Also for the LogP dataset, V-WSP gave good results: a significantly reduced set of variables (corresponding to the 17% of the total number of original variables), which was able to maintain the original data structure (Procrustes goodness of fit equal to 0.18) and, at the same time, reduce the amount of correlation and redundancy (the *K* correlation index was lowered from 0.91 to 0.66).

The three strategies based on algorithms adapted from the design of experiments (V-WSP, Kennard–Stone KS, and Distance-Based Optimal Design DBOD), as well as the procedure based on the Pairwise correlation, gave similar results in terms of the number of retained descriptors, Procrustes goodness of fit and *K* correlation index. CMC gave similar results in terms of *K* correlation index and a lower degree of similarity with the initial dataset, the Procrustes goodness of fit being equal to 0.47 and 0.30 for the Biodegradation and LogP dataset, respectively. AAMRT produced the smallest set of descriptors, even if Regression Trees were not pruned during their calibration. Even B2, iterative B2 and B4 methods included a few descriptors. Variables selected by B2 gave a high *K* correlation index and thus this method did not reduce data multicollinearity. On the opposite, iterative B2 selected uncorrelated descriptors (low *K* correlation index), but the selected set did not reproduce well enough the original data structure, giving high Procrustes goodness of fit on both datasets. KIF and B4 had an intermediate result. UFS (correlation threshold equal to 0.80) gave a set of descriptors with similar characteristics to those selected by B4 and B2 methods. When the correlation threshold was increased to 0.99, UFS selected the subset associated with the best fitting of the original data structure (lowest Procrustes goodness of fit), but with a K correlation index slightly higher than V-WSP and a significantly higher computational time.

In conclusion, V-WSP demonstrated to be the fastest algorithm and required a significantly lower computational time with respect to strategies which gave comparable results (KS, Pairwise correlation, DBOD, UFS). Since the computational time is mainly related to the calculation of correlation values, the reduced computational time of V-WSP was due to the reduced number of correlation coefficients needed for the calculation. For example, V-WSP required the calculation of 17,109 correlation coefficients for the Biodegradation dataset, while KS required the calculation of a significantly higher number of correlation coefficients (92,661) and the Pairwise correlation approach had to be calculated on the full correlation matrix, thus 286,903 correlation coefficients were computed. Moreover, V-WSP requires the selection of a correlation threshold, but KS and DBOD do not select automatically the number of retained descriptors and the a priori choice of the number of descriptors to be included in the final set is probably more difficult and less intuitive than the correlation threshold required by the V-WSP strategy.

The loadings of the first and second principal components calculated on the original Biodegradation dataset are shown in Fig. 4, where the variables selected by each strategy are highlighted to better understand the final result of each variable reduction method. Descriptors selected by V-WSP covered the entire chemical space and a few descriptors were selected in the regions of clustered variables, that is, areas characterised by extremely correlated descriptors. The same consideration can be

Table 3
Variable reduction results on the Biodegradation dataset. The number of descriptors included in the reduced set, Procrustes goodness of fit calculated between the scores of the first 4 PCs of original and reduced data, *K* correlation index and computational time are reported for each method.

	Included descriptors	Procrustes goodness of fit	*K* correlation index	Computational time (seconds)
V-WSP	134	0.27	0.65	0.6
KS	134	0.26	0.63	6.2
Pairwise correlation	138	0.28	0.64	4.1
DBOD	134	0.25	0.63	117
AAMRT	12	0.26	0.76	9.9
CMC	132	0.47	0.62	3434
UFS (thr 0.99)	186	0.18	0.70	26
UFS (thr 0.80)	49	0.61	0.31	2.9
KIF	61	0.62	0.44	263
B4	49	0.65	0.43	2.4
B2	49	0.08	0.91	2.1
B2 iterative	19	0.81	0.13	59

extended to those methods which gave similar results, such as KS, DBOD, Pairwise correlation and UFS with a correlation threshold equal to 0.99. Other strategies produced unbalanced selections. For example, B2 selected the majority of descriptors with high loadings on the first principal component. On the opposite, B4 selected descriptors in the centre of the loading space.

3.5. QSAR modelling improvement by variable reduction

One advantage of variable reduction is to facilitate the subsequent QSAR modelling, since multicollinearity and redundancy in the data are reduced. Therefore, suitable supervised algorithms for variable selection (such as Genetic Algorithms, GA) can be used to further select descriptors on the basis of specific relationships with properties or activities. Working on reduced variable sets can greatly facilitate GA variable selection and improve QSAR modelling, since it is well-known that GA can suffer from the presence of highly correlated data and the high number of initial descriptors, which can often result in overfitted models [5].

In order to evaluate if and how a preliminary variable reduction can improve the supervised variable selection and the resulting QSAR models, the Biodegradation dataset was used as initial case study. The GA variable selection was undertaken both on the original dataset (758 variables) and the V-WSP subset (134 variables) early discussed. Classification models were developed in order to discriminate 356 ready biodegradable (RB) and 699 not ready biodegradable (NRB) molecules. Molecules were randomly divided into training and test sets, containing 80% (837) and 20% (218) of the total number of molecules (1055), respectively. The selection was performed maintaining the class proportions, that is, the number of test molecules of each class was proportional to the number of training molecules of that class. The training set was used to perform the supervised selection, while molecules of the test set were used just to evaluate the predictive ability of the trained models. However, since both training and test molecules were previously used to perform the unsupervised variable reduction, the supervised selection was further evaluated using an external validation set, constituted of 670 molecules (191 ready biodegradable and 479 not ready biodegradable).

Supervised selection was carried out by coupling Genetic Algorithms and Partial Least Square Discriminant Analysis (GA-PLSDA) [32–34]. Classification models were evaluated on the basis of specificity and sensitivity, which are the ability to correctly predict RB and NRB molecules, respectively. In particular, selection of Latent Variables (LVs) for PLSDA and optimisation in GA were performed by minimizing the classification error rate calculated in cross-validation with 5 cancellation groups divided in venetian blinds. The classification Error Rate (ER) was calculated as the complement of Non Error Rate (1-NER), where NER was calculated as the average of class sensitivities [35]. Being a two-class model, the sensitivity of one class corresponds to the specificity of the other class. These indices were used in order to better estimate classification performance in presence of a data set with unequal number of molecules in each class [35].

The performance parameters of the obtained QSAR classification models are collected in Table 5. GA selected 7 molecular descriptors from both the original and reduced set of variables. These two subsets had 3 common descriptors. The smaller number of descriptors included in the reduced set positively assisted the subsequent supervised selection based on GA. In fact, the PLSDA model obtained from the full set of variables required more latent variables (6), that is more complexity, than that achieved on the reduced set (4). Moreover, the PLSDA classification model associated with the supervised selection on the reduced set of variables gave better predictive performance, since the error rate was slightly lower (0.17) than that obtained from the selection on the full set (0.18), both for the test and external validation sets. Moreover, specificity of the NRB class (for both test and external

Table 4
Variable reduction results on the LogP dataset. The number of descriptors included in the reduced set, Procrustes goodness of fit calculated between the scores of the first 4 PCs of original and reduced data, *K* correlation index and computational time are reported for each method.

	Included descriptors	Procrustes goodness of fit	*K* correlation index	Computational time (seconds)
V-WSP	220	0.18	0.66	4
KS	220	0.18	0.65	109
Pairwise correlation	218	0.20	0.64	37
DBOD	220	0.16	0.64	800
AAMRT	9	0.17	0.81	178
CMC	257	0.30	0.67	127,391
UFS (thr 0.99)	333	0.07	0.72	1189
UFS (thr 0.80)	82	0.53	0.35	74
KIF	119	0.44	0.49	1442
B4	80	0.63	0.45	5
B2	80	0.09	0.89	5.71
B2 iterative	20	0.88	0.07	364

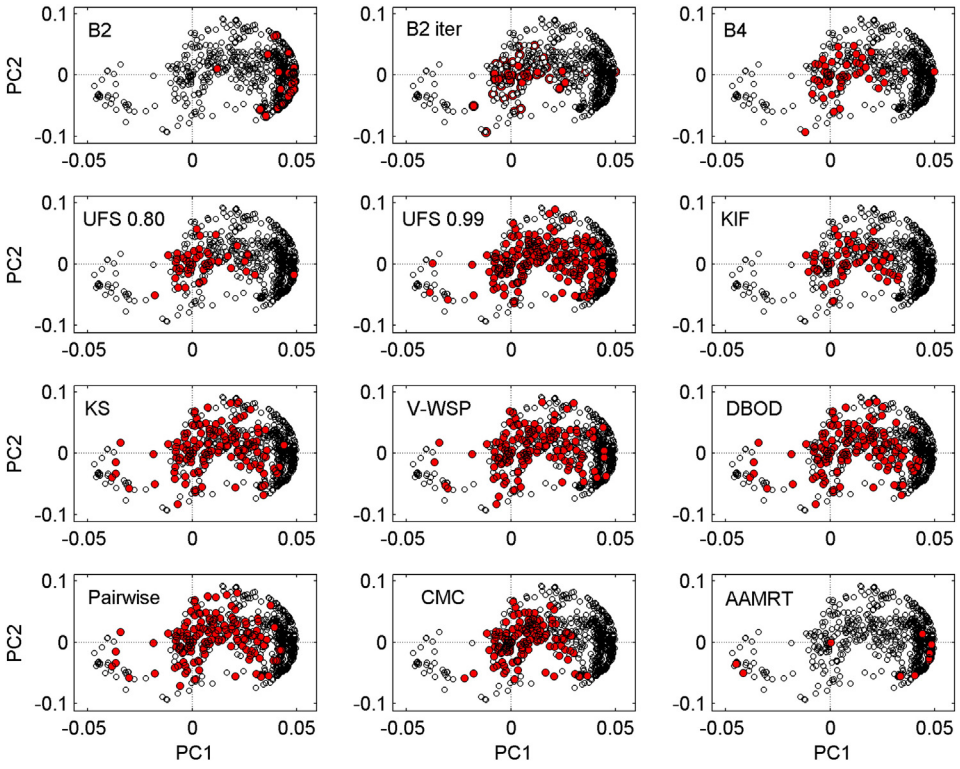

Fig. 4. Loadings of first and second principal components (cumulative explained variance equal to 62.66%) calculated on the full Biodegradation dataset. Variables selected by each variable reduction method are plotted as red circles. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

validation sets) was improved when the preliminary variable reduction was performed, while NRB sensitivity was improved on the external set of molecules. Finally, the use of the reduced set of descriptors as starting point for the supervised selection allowed in decreasing the Genetic Algorithm computational time from 3.7 h to 1.5 h.

Comparison of supervised selection based on reduced and original sets of variables was carried out on the LogP dataset too. Regression models were developed in order to predict octanol–water partition coefficient of molecules (LogP). As previously described in the data section, 8683 training molecules were used to perform the unsupervised variable reduction and the subsequent supervised variable selection, while 3720 test molecules were just used to validate the regression models. Supervised selection was carried out by coupling Genetic Algorithms and *k* Nearest Neighbours (GA-kNN) [36]. Similarities were calculated by means of Euclidean distances and predicted values were computed as weighted mean of experimental LogP values of *k* nearest neighbours. Regression models were evaluated on the basis of root means squared error (RMSE) and predictive squared correlation coefficient (Q^2) [37,38]. In particular, selection of optimal number of neighbours for *k*NN and optimisation in GA were performed by maximising Q^2 calculated in cross-validation with 2 cancellation groups divided in venetian blinds. Performance parameters of the obtained QSAR regression models are collected in Table 6. GA selected 19 molecular descriptors both on the original and reduced set of variables. Among these 19 variables, 8 descriptors were included in both solutions. As commented for the Biodegradation dataset, unsupervised reduction positively influenced the subsequent supervised selection. In fact, the *k*NN regression model calculated from the reduced set had better predictive performance both in cross validation (Q^2_{cv} equal to 0.84) and on the test set (Q^2_{ext} equal to 0.86) than the model obtained from the full set of variables (Q^2_{cv} equal to 0.80 and Q^2_{ext} equal to 0.83).

4. Conclusions

In this study, an adaptation of the WSP method, an existing algorithm for space-filling designs of experiments, to unsupervised variable reduction (V-WSP) is proposed. This method allows the selection of a representative set of variables based on linear correlation, so that multicollinearity and redundant information in the data can be reduced. V-WSP requires the selection of a correlation threshold and an initial variable (seed) to perform the reduction. The effect of changing the correlation threshold, which holds to the changing of number of variables included in the reduced set, was discussed. Moreover, the algorithm demonstrated to converge to similar solutions independently from the seed selection.

The performances of V-WSP were evaluated on simulated, benchmark and real QSAR datasets and compared with other methods for variable reduction. V-WSP gave similar results with respect to other methods. However, V-WSP demonstrated to converge to representative results with the benefit of being less time expensive in a computational point

Table 5
Biodegradation data: comparison of supervised variable selection (GA-PLSDA) based on original (758) and reduced (134) sets of descriptors. For each model, the model size and the number of Latent Variables selected in the PLSDA model (LV) are provided together with Error Rate (ER), specificity (Sp) and sensitivity (Sn) achieved in cross validation (with 5 cancellation groups) and on test and external validation sets. Sensitivity and specificity refer to NRB class.

Initial set of descriptors	Model size	LV	5 fold CV			Test set			External validation set		
			ER	Sn	Sp	ER	Sn	Sp	ER	Sn	Sp
134	7	4	0.18	0.80	0.84	0.17	0.86	0.81	0.17	0.88	0.79
758	7	6	0.17	0.82	0.84	0.18	0.86	0.79	0.18	0.87	0.77

Table 6
LogP data: comparison of supervised variable selection (GA-kNN) based on original (1265) and reduced (220) sets of descriptors. For each model, the model size and the number of neighbours (k) are provided together with root means squared error in regression (RMSE) and predictive squared correlation coefficient (Q^2) achieved in cross validation (with 2 cancellation groups) and on the test set.

Initial set of descriptors	Model size	k	2 fold CV		Test set	
			Q^2_{cv}	RMSECV	Q^2_{ext}	RMSEP
220	19	4	0.84	0.73	0.86	0.68
1265	19	3	0.80	0.81	0.83	0.74

of view. This can be an advantage when dealing with massive datasets such as those derived from databases and libraries related to QSAR and computational chemistry. V-WSP requires the selection of a correlation threshold, while other methods do not select automatically a number of descriptors by themselves or are based on pseudo thresholds. Thus, the final number of variables to be included in the reduced set must be user-defined. This is probably less intuitive to be decided with respect to the setting of a simple correlation threshold. Finally, results of supervised selections performed on both the original and reduced sets of variables demonstrated how variable reduction can improve the subsequent multivariate modelling.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

[1] S. Bagheri, N. Omidikia, M. Kompany-Zareh, Chemom. Intell. Lab. Syst. 128 (2013) 135–143.
[2] V. Consonni, D. Ballabio, A. Manganaro, A. Mauri, R. Todeschini, Anal. Chim. Acta. 648 (2009) 52–59.
[3] F. Questier, R. Put, D. Coomans, B. Walczak, Y. Vander Heyden, Chemom. Intell. Lab. Syst. 76 (2005) 45–54.
[4] D.C. Whitley, M.G. Ford, D.J. Livingstone, J. Chem. Inf. Comput. Sci. 40 (2000) 1160–1168.
[5] D.M. Hawkins, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[6] D.J. Livingstone, E. Rahr, Quant. Struct.-Act. Relat. 8 (1989) 103–108.
[7] R. Todeschini, V. Consonni, A. Maiocchi, Chemom. Intell. Lab. Syst. 46 (1999) 13–29.
[8] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[9] E. Marengo, R. Todeschini, Chemom. Intell. Lab. Syst. 16 (1992) 37–44.
[10] J. Santiago, M. Claeys-Bruno, M. Sergent, Chemom. Intell. Lab. Syst. 113 (2012) 26–31.
[11] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[12] A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K.H. Lee, J. Comput. Aided Mol. Des. 17 (2003) 241–253.
[13] A. Golbraikh, A. Tropsha, J. Comput. Aided Mol. Des. 16 (2002) 357–369.
[14] M. Sergent, R. Phan-Tan-Luu, J. Elguero, An. Quim. 93 (1997) 71–75.
[15] M. Sergent, R. Phan-Tan-Luu, J. Elguero, An. Quim. 93 (1997) 295–300.
[16] M. Sergent, Contribution de la Méthodologie de la Recherche Expérimentale à l'élaboration de matrices uniformes: application aux effets de solvants et de substituants, (PhD thesis) 1989.
[17] A. Beal, M. Claeys-Bruno, M. Sergent, Chemom. Intell. Lab. Syst. 133 (2014) 84–91.
[18] R. Todeschini, Anal. Chim. Acta. 348 (1997) 419–430.
[19] W.J. Krzanowski, Principles of Multivariate Analysis, Clarendon Press, Oxford, 2000.
[20] D.G. Kendall, Stat. Sci. 2 (1989) 87–99.
[21] J.C. Gower, Psychometrika 40 (1975) 33–51.
[22] H.F. Kaiser, Educ. Psychol. Meas. 20 (1960) 141–151.
[23] Talete srl, Dragon (Software for Molecular Descriptor Calculation), Version 6.0 — 2013, http://www.talete.mi.it/.
[24] R. Todeschini, D. Ballabio, V. Consonni, A. Manganaro, A. Mauri, Anal. Chim. Acta. 648 (2009) 45–51.
[25] J.N.R. Jeffers, Appl. Stat. 16 (1967) 225–236.
[26] H. Streuli, Lebensm. Technol. 20 (1987) 211.
[27] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, V. Consonni, J. Chem. Inf. Model. 53 (2013) 867–878.
[28] S.A. Rosenberg, A.E. Hueber, D. Aronson, S. Gouchie, P.H. Howard, W.M. Meylan, J.L. Tunkel, Sci. Technol. Lib. 23 (2004) 73–87.
[29] W.M. Meylan, P.H. Howard, J. Pharm. Sci. 84 (1995) 83–92.
[30] W.Y. Loh, Stat. Sin. 12 (2002) 361–386.
[31] M. Daszykowski, B. Walczak, D.L. Massart, Anal. Chim. Acta. 468 (2002) 91–103.
[32] M. Barker, W.S. Rayens, J. Chemom. 17 (2003) 166–173.
[33] R. Leardi, J. Chemom. 14 (2000) 643–655.
[34] R. Leardi, A. Lupianez, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.
[35] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC press, Boca Raton, FL, 2009.
[36] B.R. Kowalski, C.F. Bender, Anal. Chem. 44 (1972) 1405–1411.
[37] V. Consonni, D. Ballabio, R. Todeschini, J. Chem. Inf. Model. 49 (2009) 1669–1678.
[38] V. Consonni, D. Ballabio, R. Todeschini, J. Chemom. 24 (2010) 194–201.