Toward better QSAR/QSPR modeling: simultaneous outlier detection and variable selection using distribution of model features

Dongsheng Cao · Yizeng Liang · Qingsong Xu · Yifeng Yun · Hongdong Li

Received: 23 January 2010 / Accepted: 3 November 2010 / Published online: 13 November 2010

Abstract Building a robust and reliable QSAR/QSPR model should greatly consider two aspects: selecting the optimal variable subset from a large pool of molecular descriptors and detecting outliers from a pool of samples. The two problems have the specific similarity and complementarity to some extent. Given a particular learning algorithm on a particular data set, one should consider how the interaction could happen between variable selection and outlier detection. In this paper, we describe a consistent methodology for simultaneously performing variable subset selection and outlier detection using the idea of statistical distribution which can be simulated by the establishment of many cross-predictive linear models. The approach exploits the fact that the distribution of linear model coefficients provides a mechanism for ranking and interpreting the effects of variable, while the distribution of prediction errors provides a mechanism for differentiating the outliers from normal samples. The use of statistic of these distributions, namely mean value and standard deviation, inherently provides a feasible way to effectively describe the information contained by the original samples. Several examples are used to demonstrate the prediction ability of our proposed approach through the comparison of different approaches as well as their combinations.

Keywords QSAR/QSPR · Outlier detection · Variable selection · Monte Carlo · Statistical distribution

Introduction

High throughput screening has been a major recent technological improvement in drug development process. In conjunction with combinatorial chemistry, it can not only fast discover active compounds but also identify those that will be suitable as drugs in terms of absorption, distribution, metabolism, and excretion, as well as toxicology. With the need for knowledge-guided screening of compounds, the virtual filtering and screening have been recognized as techniques complementary to high-throughput screening. To large extent, these techniques rely on quantitative structure–activity/property relationship (QSAR/QSPR) analysis [1]. The QSAR/QSPR methodology aims at finding a model, which allows for correlating the activities to structures within a family of compounds. QSAR/QSPR has successfully been applied to modeling and prediction of many physicochemical and biological properties, such as boiling point, melting point, aqueous solubility, toxicity, retention index and the activities of drugs etc. [2–10].

Building a robust and reliable QSAR/QSPR model should simultaneously consider the following two aspects: (1) From the point of view of sample space, it is well known that the model built by some training set will be strongly dependent upon the structures defined by the training set. If there are some special chemicals called as

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9401-1) contains supplementary material, which is available to authorized users.

D. Cao · Y. Liang (✉) · Y. Yun · H. Li
Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, People's Republic of China
e-mail: yizeng_liang@263.net

Q. Xu (✉)
School of Mathematical Sciences and Computing Technology, Central South University, Changsha 410083, People's Republic of China
e-mail: dasongxu@gmail.com

outliers departing from the bulk of the data, the regression model will be distorted toward these points. Therefore, whilst minimizing the training error, the bulk of data is represented suboptimally. Robust regression approaches are usually used to remove influence of outliers [11–13]. Several well-known examples include M-estimators [14], least median of squares (LMS) [15], least trimmed squares (LTS) [16], robust principal component regression (RPCR) [17], robust partial least squares (RPLS) [18, 19] and robust principal components regression based on principal sensitivity vectors (RPPSV) [20]. The resultant regression models optimally account for the bulk of data points. When performed correctly, the removal of significant outliers will allow for the development of stronger models. (2) From the point of view of variable space, it is generally considered unacceptable for QSAR/QSPR models to incorporate large number of descriptors, for two reasons. Firstly, in most cases, only a small number of descriptors have substantial influence on a particular type of properties or biological activity. The use of over many descriptors likely generates noise in an established model and subsequently affects its prediction accuracy [21]. Secondly, it is extremely difficult to interpret a model with a large number of descriptors. Currently, there is no feasible way to select the optimal set of variables, resulting in the global maximum of prediction power estimated by validation approaches. An exhaustive enumeration for all possible 2^p-1 (p is the number of descriptors) combinations is computationally impractical for usual chemical datasets with large number of descriptors because of the combinatorial explosion of the number of subsets. Therefore, several elaborate approaches are usually employed to extract the set of descriptors used for modeling, including simulated annealing algorithm (SA) [22], genetic algorithm (GA) [23], genetic function approximation (GFA) [24], particle swarm optimization algorithm (PSO) [25], stepwise regression regarding different criteria [26], the lasso [27] and least angle regression (LARS) [28] etc.

Given a particular learning algorithm on a particular data set, it is worth noting that variable selection and outlier detection should be not treated separately [29]. Certainly, outlying of a molecule depends on the descriptors used for characterization. Molecules that are seen as outliers when investigated in a set of descriptors may be within the bulk of all molecules when described with a different set of descriptors. Thus, when dealing with the QSAR/QSPR datasets with redundant variables, we should fail to detect all outliers and even may regard normal molecules as outliers. Similarly, when the dataset is contaminated by the outliers, the descriptors selected by variable selection approaches may include some uninformative ones, which may be additionally selected to describe the influence of the outliers. Figure 1 shows such a picture, in which ten two-dimensional data points marked by circle are fitted to a plain A. This plain parallel to x_2 axis has a function form of $y = 5x_1 + $ noises (The true model function is $y = 5x_1$). However, due to some reason, the sample a records a wrong response value y and so is changed into one outlier b. In such a situation, a new plain marked by B is generated to interpret the relationship between **X** and **y**, which uses two features to establish the function form of $y = 2.96x_1 - 2.45x_2$. Thus, due to the influence of an outlier, a total different model is generated; what is more, a new variable x_2 is additionally introduced into the model to make up the large error caused by the outlier. Such a situation has been reported initially by Kirchner et al. [30] and Cronin et al. [31, 32] in the QSAR investigation of in vitro human skin permeability coefficients. After seven outlying molecules are removed in their study, a different picture is revealed and the model established by three descriptors has been changed into one by two descriptors. What is more, following the removal of these outliers, a much improved and mechanistically comprehensible QSAR was developed. So, considering the interaction between variable selection and outlier detection is imperative for building a robust and reliable model. Currently, some approaches based on different ideas are proposed to alleviate the problem to some extent [33–37].

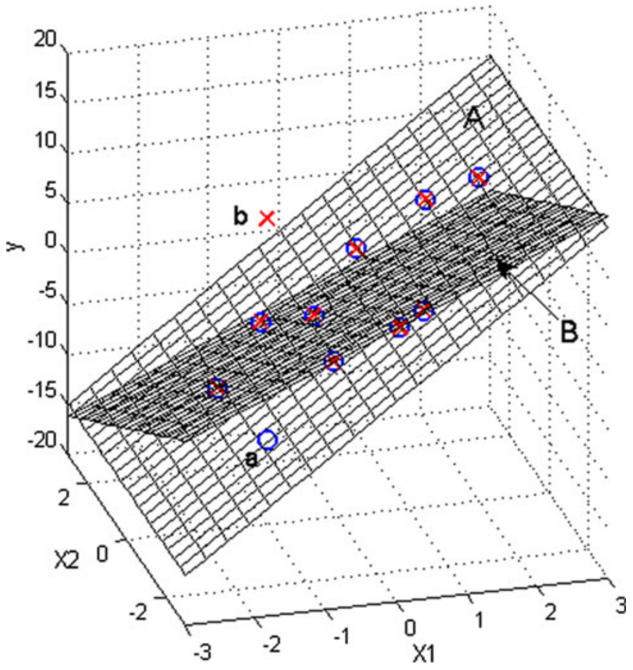


Fig. 1 An example of the interaction between outlier detection and variable selection. The true model being $y = 5x_1$ is a plain parallel to x_2 axis, marked by A. However, after one sample a becomes an outlier b, a different plain ($y = 2.96x_1 - 2.45x_2$) marked by B is used to describe the relationship between **X** and **y**. Thus, an additional variable only useful for the outlier is selected to make up the influence of the outlier

In the present study, we attempt to address these two problems simultaneously by constructing a consistent framework, based on the idea of the statistical distribution. Our approach exploits the fact that the distribution of linear model coefficients provides a mechanism for ranking and interpreting the effects of variable, while the distribution of the prediction errors provides a mechanism for differentiating the outliers from normal samples. Thus the statistic of these distributions, namely mean value and standard deviation, are then used to quantitatively describe the size of models features (i.e., prediction errors and variable coefficients). Finally, the clean training dataset itself can stepwise be obtained by a backward elimination strategy. Two real QSAR/QSPR datasets along with simulated examples are used to test our proposed approach through the comparison of different approaches as well as their combinations. The results obtained clearly indicated that our approach can largely reduce the risk of selecting a wrong model mainly caused by the interaction between outlier detection and variable selection and improve the prediction performance of models.

Theory and method

Constructing statistical distribution of model features

The statistical distribution can provide very abundant information about the random variables [38]. Most approaches of statistical inference are based on such a statistical distribution [39–44]. In this work, we just make use of such a strategy to construct the statistical distribution of model features, such as prediction errors and variable coefficients, and subsequently make statistical inference. Among these methods, the Monte-Carlo approach is usually employed to extract the information and used for statistical inference [45]. In general, the Monte-Carlo approach can be used to generate a distribution of some statistic of interest by repeatedly calculating that statistic randomly selected portions of the data because of its good asymptotic property [46–48]. In QSAR/QSPR study, if we model a given QSAR/QSPR data by a single training/validation set division, then we can obtain predictive errors of this validation set and all variable coefficients, characterizing the behavior of model features (i.e., prediction errors and variable coefficients) within these two sets. However, these model features highly depend on the way in which we divide the data into the training set and validation set. Different training/validation data division should yield different model features. Thus, by changing the training/validation data division by Monte Carlo, we can obtain a large number of QSAR/QSPR models and corresponding model features so as to get some insight into the data structure statistically. What kind of information about these model features can be obtained from their distribution? Generally speaking, some parameters of interest can be acquired as a function of the probability density function or of the empirical cumulative distribution function of a random variable (e.g., model features), which will make statistical inference about model features easier. Suppose that $z_1, \ldots, z_m$ is to be used to estimate a population parameter, θ. A function of a population distribution function, defining the parameter θ, can usually be expressed as:

$$\theta = \int \mathrm{g}(\mathbf{z}) d\mathrm{P}_m(\mathbf{z}) \tag{1}$$

Here $\mathrm{g}(\mathbf{z})$ is the statistic used to estimate θ, whose expectations we might be interested in. $\mathrm{P}_m(\mathbf{z})$ is the probability density of $\mathbf{z}$. Thus, by constructing different $\mathrm{g}(\mathbf{z})$, one can obtain different statistics θ describing the specific information (e.g., mean value or standard deviation) of a population distribution.

The working procedure of MCOVS method

The analysis of the QSAR/QSPR data by cross-prediction in the Monte-Carlo schemes can intensively probe large model population. The model features generated by individual model can not necessarily reflect their true behavior, since they are seriously affected by the random division of training set and validation set. However, the distribution of model features constructed by model population can provide more objective information. Based on such a distribution, we developed a new methodology to establish a high quality QSAR/QSPR model by studying the distribution of prediction errors of each sample and coefficients of each variable obtained from the training set. We refer to this approach as Monte-Carlo outlier and variable screening approach (MCOVS). Figure 2 shows a flow chart for the complete algorithm. Firstly, with the help of the Monte-Carlo approach, the whole training data is further randomly divided into the training set and the validation set, respectively. Generally, the training set should be sufficiently large (i.e., 70–90%) so that a reliable model can be developed. For the training set, five fold cross-validation is used to select the optimal number of PLS components. In our study, partial least squares (PLS) regression is employed instead of multiple linear regression (MLR) to establish the QSAR/QSPR models, since PLS could deal with colinearity problem in regression, allowing the construction of models based on datasets with more descriptors than compounds. Thus, the prediction error would be obtained for each validation sample, which has proved to be much more sensitive to diagnose the outliers [49]. This cycle was executed according to the predetermined number

T of times. Finally, the prediction errors for each sample and the coefficients for each variable are recorded, respectively. The distributions such obtained are plotted and then their statistic can be constructed to quantitatively describe the property of each sample or variable.

Selecting relevant variable subset using coefficient distribution

When selecting different data subsets by the working procedure of MCOVS described above, one can obtain a large number of coefficients for each variable. Thus, the distribution of coefficients can be constructed and studied. Could the distribution of coefficients of each variable provide more information about the variable? The answer is positive.

To observe the behavior of variables conveniently and directly, the absolute values of mean values of coefficient distribution are plotted versus standard deviations. Such a visualization of variables may be useful for understanding the role of useful variables. Figure 3 shows such a picture, in which the whole area is split into four parts, marked by A–D, expressing four types of variables. A large mean value for some variable indicates that it plays a very important role in the model population. However, a large standard deviation of some variable indicates that regression coefficient vary strongly during the individual iterations of cross-prediction. A probable reason is that such variables are responsible for the outlying molecules, since their values largely depend on the presence or absence of these outliers in the training subset. Ideally those important variables should be the ones that possess both large mean value and small standard deviation. We can therefore construct the following measure of variable importance [50]:

$$\mathbf{c}_i = \frac{mean(\mathbf{s})}{std(\mathbf{s})} \tag{2}$$

where **s** is the coefficient vector for the *i*th variable, generated by Monte-Carlo. *mean*(**s**) and *std*(**s**) represent the mean value and standard deviation, respectively. Thus, the variable with the largest $\mathbf{c}_i$ value should be the most important one in the pool of variables. These variables with the smaller $\mathbf{c}_i$ value should be removed due to their small contribution to models. However, to avoid the influence of outliers, we revise Eq. 2 as follows:

$$\mathbf{c}_i = \frac{median(\mathbf{s})}{rstd(\mathbf{s})} \tag{3}$$

where *median* (**s**) is the median of **s**, and *rstd* (**s**) is the robust version of standard deviation, which is given by:

$$rstd(\mathbf{s}) = \frac{quantile(\mathbf{s}, \alpha) - quantile(\mathbf{s}, 1 - \alpha)}{2z^{\alpha}} \tag{4}$$

where *quantile*(**s**, α) is α quantile of **s** for the *i*th variable, z^{α} is the α percentile point of standard normal distribution (i.e., $z^{\alpha} = 1.645$ when $\alpha = 0.95$). Thus, Eq. 4 is more robust to outliers than Eq. 2.

Detecting outliers using prediction error distribution

Similarly, the distribution of prediction errors generated by a large number of models can also contain more sample information (i.e., whether this sample is an outlier or not). Likewise, the mean value *mean* (*j*) and the standard deviation *std* (*j*) of the prediction error distribution for the *j*th sample are employed to describe this distribution.

$$mean(j) = \frac{1}{k}\sum_{i=1}^{k} error(i) \tag{5}$$

$$std(j) = \left(\frac{1}{k-1}\sum_{i=1}^{k}(error(i) - mean(j))^2\right)^{1/2} \tag{6}$$

where *k* is the total times of which the *j*th sample is found in the validation set. The *error* (*i*) is the prediction error of the *j*th sample in the *i*th cycle. Thus, a large mean value of prediction errors for some sample indicates that we can always obtain large prediction errors no matter how the training datasets are fluctuated. So we can conclude that such a sample should be **y** outlier. It should be noted that cross-prediction can provide information on potentially outlying molecules [51–53]. If for example only one

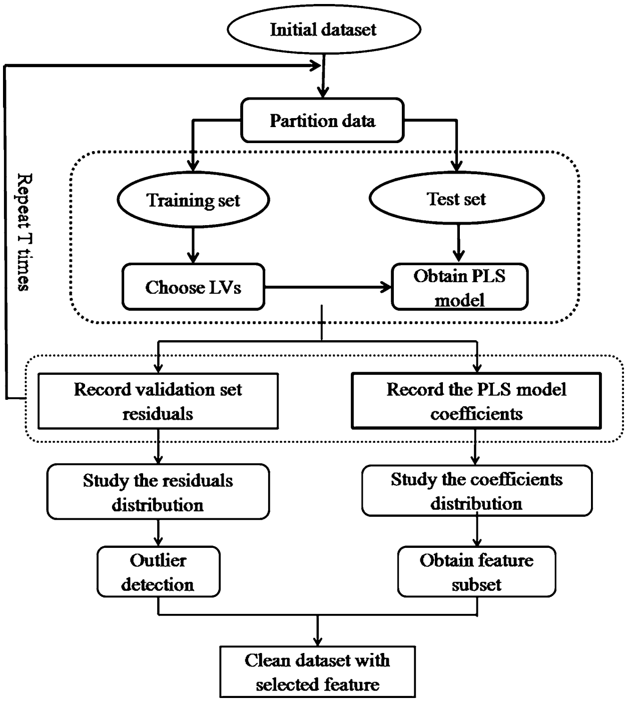
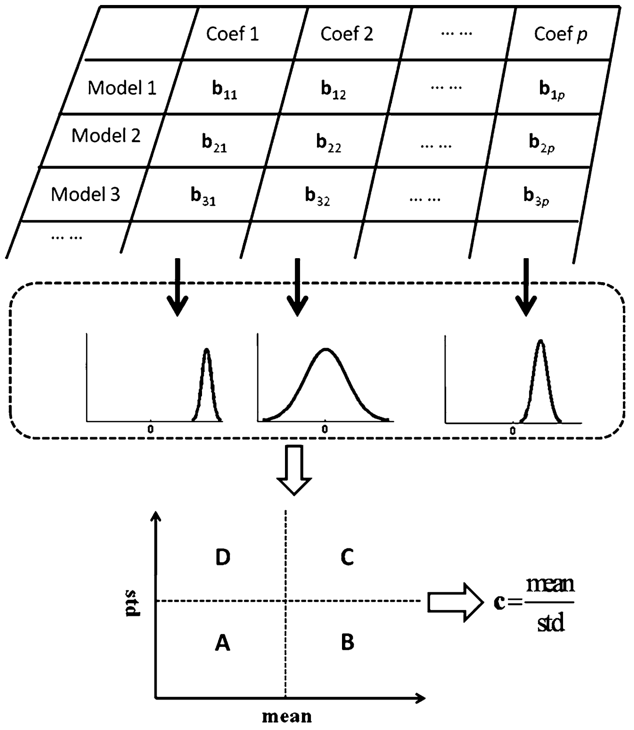

Initial dataset
Partition data
Repeat T times
Training set
Test set
Choose LVs
Obtain PLS model
Record validation set residuals
Record the PLS model coefficients
Study the residuals distribution
Study the coefficients distribution
Outlier detection
Obtain feature subset
Clean dataset with selected feature

Fig. 2 Flow chart showing the steps used in the MCOVS approach



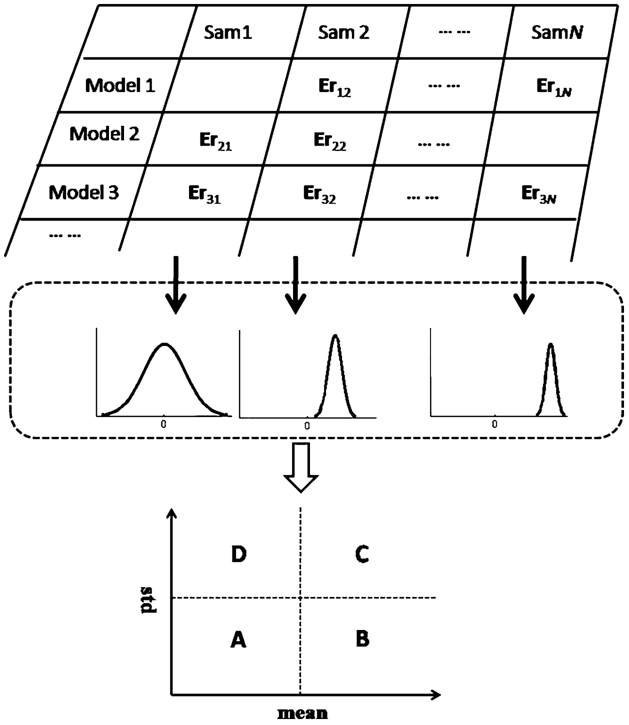
	Coef 1	Coef 2	……	Coef *p*
Model 1	b_{11}	b_{12}	……	b_{1p}
Model 2	b_{21}	b_{22}	……	b_{2p}
Model 3	b_{31}	b_{32}	……	b_{3p}
……				

Fig. 3 The process of obtaining variable importance by means of MCOVS. By studying the distribution of variable coefficients of each variable, we can classify all variables into four parts marked by *A–D*. Based on such observations, the measure of variable importance can be obtained to rank all variables

outlying molecule has many chlorine atoms and chlorine is an important variable, then the full data set may be able to calibrate the effect of chlorine and make good predictions, but the data set with the molecule excluded will likely lead to a large prediction residual on that molecule. So, the prediction errors obtained by cross-prediction allow us to easily detect these outlying molecules compared to the fitted residuals [49]. Moreover, in MLR, if an external data point $\mathbf{x}_i$ is predicted and has a leverage of $h = \mathbf{x}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_i$, its prediction error has variance $s^2\{e_i\} = MSE(1 + h)$. Herein, *MSE* represents mean square error. From this formula, we can see that the variability of the sampling distribution of e_i is affected by how far $\mathbf{x}_i$ is from the centroid $\overline{\mathbf{X}}$, through the term h. The further $\mathbf{x}_i$ from $\overline{\mathbf{X}}$ is, the greater is the quantity and the larger is the variance of e_i Hence, the variation of e_i obtained from different cases will be greater when $\mathbf{x}_i$ is far from the mean value than the ones near the mean value [54, 55]. We can therefore detect **X** outliers by standard deviation of prediction errors. Figure 4 shows the process of detecting the outliers. Similar to Fig. 3, the standard deviations of prediction errors versus mean values of prediction errors are plotted to directly describe the behavior of the samples. Thus, the whole area is split into four parts marked by A-D, which indicate four types of samples: normal samples marked by A, **y** outliers marked by B, abnormal samples marker by C, and **X** outliers marked by D. With the help of the distribution of cross-predictive errors such obtained, it seems to be able to simultaneously detect all types of outliers and reduce the risk caused by the masking effect [49].

Improving model performance using backward elimination strategy

The use of MCOVS allows us to rank the variables and prediction errors. We therefore can easily remove unimportant variables and detect the outliers within current variable space. However, it should be noted that the removal of redundant variables probably leads to the change of outliers. Furthermore, different outliers probably lead to the selection of different variables. In practice, this interaction is very common phenomenon, especially for QSAR/QSPR. One-step strategy is very insufficient to uncover the interaction. A probably effective approach is to take stepwise ways to reduce the risk that the interaction between outlier detection and variable selection brings about to some extent. The backward elimination strategy (BES) refers to a search that begins at the full set of variables. The main reason for this choice is that going backward from the full set of variables may easily capture interacting variables and the interaction between outlier



	Sam1	Sam 2	……	Sam*N*
Model 1		Er_{12}	……	Er_{1N}
Model 2	Er_{21}	Er_{22}	……	
Model 3	Er_{31}	Er_{32}	……	Er_{3N}
……				

Fig. 4 The process of detecting outliers by means of MCOVS. By studying the distribution of prediction errors of each sample, different types of outliers can easily be diagnosed. *A*: normal samples, *B*: **y** outliers, *C*: abnormal samples, *D*: **X** outliers

detection and variable selection. The MCOVS approach based on BES can be further executed in the following way:

(1) At each MCOVS, all variables are ranked according to Eq. 3 and the specific number of outliers can be detected according to Eqs. 5 and 6. Record the index of these outliers and compute root mean squared error for cross-validation set (RMSECV) with the dataset without outliers. RMSECV such obtained can effectively avoid the influence of outliers on the objective function and thereby help to generate an accurate and reliable variable subset.

(2) Remove the variable with the smallest $\mathbf{c}_i$ value. To improve computational efficiency, we can also remove a few variables with small $\mathbf{c}_i$ values once.

(3) Perform MCOVS with all training dataset on the remaining variables. Again, the remaining variables are ranked and another specific number of outliers can be detected. Because of the influence of variable subset, the outliers detected by this iteration may be different from the outliers detected previously. So we compute the frequency of all outliers detected previously and select the same number of outliers as the ones in this iteration.

(4) Repeat steps (1)–(3). After *B* iteration, *B* subset of descriptors and *B* RMSECV values are obtained. The RMSECV values of the resulting sequence of models are used to determine a learning curve which, typically, first decreases then reach a minimum and increases again. So, we can select the subset with the minimum RMSECV value as the optimal subset of descriptors. Meantime, the outliers with minimum RMSECV value are regarded as the final ones.

Datasets

In order to illustrate the performance of the proposed approach when dealing with the QSAR/QSPR datasets with outliers and redundant descriptors, two real datasets in the chemical literature together with additional datasets from one of our research projects are used to test this MCOVS approach.

Simulated data

For illustrative purposes, simulated data analogous to real QSAR/QSPR data was designed to test this approach ($\mathbf{X}$ (50 × 100) and $\mathbf{y}$ (50 × 1)). In these examples, matrix $\mathbf{X}$ contained an independent column ($\mathbf{X}$), which represented some molecular descriptors, and a dependent column ($\mathbf{y}$) being related to by $\mathbf{y} = f(\mathbf{X})$. Thus, the response variable y represented output values for each molecule, which, in a real QSAR/QSPR data set, would be the activity or property value of each chemical molecule. To better uncover the interaction between outlier detection and variable selection, we constructed two simulated QSAR/QSPR datasets ($\mathbf{y}$ values without and with noises), respectively. The simulated QSAR/QSPR datasets can be constructed in the following steps:

1. An $\mathbf{X}_\mathrm{C}$ matrix of the (50,100) size is constructed from random numbers from 0 to 1. Principal component analysis (PCA) on the centered matrix $\mathbf{X}_\mathrm{C}$ is performed, yielding scores $\mathbf{T}$ and loadings $\mathbf{V}$.
2. Multiplication of the first five score vectors $\mathbf{T}$(:,1:5) by the first five loading vectors $\mathbf{V}$(1:5,:) gives a simulated pure data matrix $\mathbf{X}$p (50,100), which does not contain any noise.
3. The $\mathbf{y}$ vectors of the responses can be calculated using the first five score vectors: $\mathbf{y} = 5\times\mathrm{PC1} + 4\times\mathrm{PC2} + 3\times\mathrm{PC3} + 2\times\mathrm{PC2} + 1\times\mathrm{PC1}$; where PCs are the vector of scores on PCs. For the QSAR/QSPR data with noises, the noises are generated from the normal distribution with mean value of 0 and standard deviation of about 5% of standard deviation of the normal y values.
4. About 25% variations of the normal $\mathbf{y}$ values are added to the response values of samples 46–50 so that samples 46–50 become outliers.
5. An $\mathbf{UNI}$ matrix of the (50,100) size is generated from random numbers from 0 to 1. The $\mathbf{UNI}$ matrix contains some noise features unrelated to the response values $\mathbf{y}$.
6. We construct four simulated QSAR/QSPR datasets, respectively. The dataset without noise features and outliers: $\mathbf{SIMU1} = \mathbf{X}$p(1:45,:); the dataset with outliers and without noise features: $\mathbf{SIMU2} = \mathbf{X}$p; the dataset with noise features and without outliers: $\mathbf{SIMU3} =$ [$\mathbf{X}$p(1:45,:),$\mathbf{UNI}$(1:45,:)]; the dataset with outliers and noise features: $\mathbf{SIMU4} =$ [$\mathbf{X}$p(1:50,:),$\mathbf{UNI}$(1:50,:)]. For the QSAR/QSPR data with noises, four similar simulated datasets are generated.

QSAR/QSPR data

Two real QSAR datasets are used for testing the proposed approach. The first data consists of a set of 114 angiotensin converting enzyme (ACE) inhibitors originally taken from the work of Depriest et al. [56], which describes their use for CoMFA modeling by 56 molecular descriptors. Activities are spread over a wide range, with pIC50 values ranging from 2.1 to 9.9. The second data is boiling point data for the alkanes with different branches. A total of 233 molecules were collected from different literature [57–59, 61]. 61 molecular descriptors which consist of connectivity indices and walk and path counts are used to characterize these alkanes. The range of experimental boiling points is between −42.10 and 525.04.

Results and discussion

Simulated data

For comparison of different approaches, robust partial least square (RPLS) and iterative variable elimination partial least square (IVE-PLS) are used to detect the outliers and select compact subset of variables, respectively. RPLS accounts for outlying molecules regarding descriptors as well as dependent variables [19]. It is based on a robust variation of the SIMPLS algorithm. Robust estimators are used at two occasions of the SIMPLS algorithm: (1) a robust PCA estimate of the center and the covariance matrix is used for deriving the PLS-components. (2) Robust linear regression based on the robust estimate of the center and the covariance matrix is employed to estimate the PLS regression coefficients. IVE-PLS, proposed by Polanski and Gieliciak [60, 61], is backward elimination approach to model selection. Its main concept is to iteratively remove variables according to the signal-to-noise ratio of their PLS regression coefficients by leave-one-out cross-validation. Furthermore, the combinations of RPLS and IVE-PLS are also used for new modeling approaches to further improve the prediction power. Two strategies were investigated: initial removal of outliers followed by variable elimination (RPLS + IVE-PLS) and initial variable elimination followed by the removal of outliers within this new representation of the dataset (IVE-PLS + RPLS).

Results for these approaches applied to simulated data are shown in Tables 1 and 2. For SIMU1 data without outliers and noise variables, it is conceivable that four approaches, namely PLS, IVE-PLS, RPLS and MCOVS-PLS, all obtain satisfactory prediction results. For SIMU2 data contaminated only by outliers, PLS obtains the poor prediction ability due to the existence of five outliers. However, RPLS and MCOVS-PLS have the ability to detect the five outliers (i.e., 46–50) and so obtain better prediction compared to PLS. For SIMU3 data contaminated only by noise variables, PLS model obtains the poor prediction results again due to the influence of noise variables. However, IVE-PLS and MCOVS-PLS can find informative variables and thereby improve the prediction results. For SIMU1, SIMU2 and SIMU3 datasets, we can clearly see that RPLS and IVE-PLS only lay stress on the influence of single direction. MCOVS-PLS not only has ability to detect the outliers, but also has ability to select compact subset of features.

SIMU4 data is a more complex example contaminated simultaneously by five outliers and 100 noise variables. From Table 2, we can see that single modeling approaches such as PLS, IVE-PLS and RPLS perform poorly in prediction. RPLS does not accurately detect five true outliers when the dataset contains noise variables. This indicates the influence of noise variables on the outliers. The data points that are seen as outliers within the space of 100 useful variables may not be the outliers within the space of 200 variables containing 100 noise ones. IVE-PLS improves the prediction ability of PLS to some extent by

Table 1 The prediction results of the first three simulated data by different modeling approaches

Methods	RMSECV[a]	Q^2	LVs[b]	Vars	Outliers
SIMU1					
PLS	1.62e−15	1.000	5	100	–
IVE-PLS	1.62e−15	1.000	5	100	–
RPLS	1.62e−15	1.000	5	100	–
MCOVS-PLS	1.62e−15	1.000	5	100	–
SIMU2					
PLS	0.289	0.996	3	100	–
RPLS	1.82e−15	1.000	5	100	46–50
MCOVS-PLS	1.71e−15	1.000	5	100	46–50
SIMU3					
PLS	0.574	0.986	13	100 + 100	–
IVE-PLS	1.68e−15	1.000	5	94 + 0	–
MCOVS-PLS	1.72e−15	1.000	5	94 + 0	–

[a] Root mean square error for cross validation

[b] The number of the latent variables for PLS

Table 2 The prediction results of SIMU4 by different modeling approaches

Methods	RMSECV	Q^2	LVs	Vars	Outliers
PLS	0.577	0.986	12	100 + 100	–
IVE-PLS1[a]	0.245	0.997	7	95 + 29	–
RPLS	0.636	0.981	13	100 + 100	2, 7, 8, 24, 30
IVE-PLS1 + RPLS	0.237	0.998	8	95 + 29	5, 8, 12, 24, 30
RPLS + IVE-PLS1	0.225	0.998	8	95 + 29	2, 7, 8, 24, 30
MCOVS-PLS	1.16e−15	1.000	5	14 + 0	46–50
IVE-PLS2[b]	0.301	0.996	5	88 + 0	–
IVE-PLS2 + RPLS	1.52e−15	1.000	5	88 + 0	46–50
RPLS + IVE-PLS2	0.315	0.996	5	75 + 0	2, 7, 8, 24, 30

[a] Iterative variable elimination partial least squares with the smallest RMSECV value

[b] Iterative variable elimination partial least squares without uninformative or noise variables

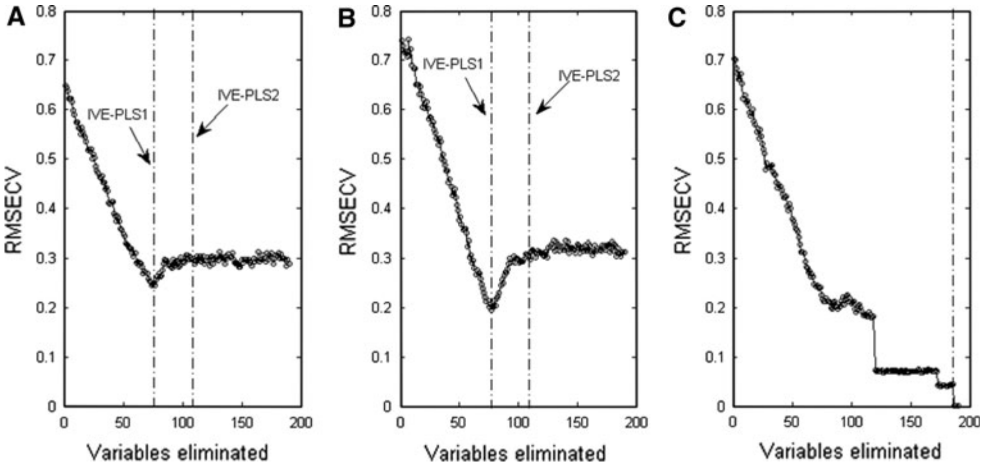


Fig. 5 The plots of RMSECV versus the number of variables eliminated using different modeling approaches for SIMU4 dataset. **A**: IVE-PLS1 and IVE-PLS2; **B**: RPLS + IVE-PLS1 and RPLS + IVE-PLS2; **C**: MCOVS-PLS

removal of noise variables. But it does not remove all uninformative variables and 29 noise variables are retained in the final model, which may indicates that the outliers influence the selection of variables. To achieve the better performance, 29 noise variables must be selected to make up the residuals caused by the five outliers. Compared to IVE-PLS1, IVE-PLS2 yields the worse prediction results when all noise variables are eliminated. Here, IVE-PLS2 is referred to as IVE-PLS without noise variables. In practice, we cannot control the endpoint where all noise variables are just eliminated (see Fig. 5A). The comparison between IVE-PLS1 and IVE-PLS2 sufficiently demonstrates that the process of selection of variables will be seriously influenced when the dataset contains the outliers. Additional variables probably useful only for these outliers may thereby be introduced to the final model. As for the combination approaches, IVE-PLS1 + RPLS and RPLS + IVE-PLS1 obtain the similar results. But they do not obtain clean dataset not containing both outliers and noise variables. There still exists to be 29 noise variables and all five outliers after they are carried out. The simple combinations of two approaches indeed improve the prediction ability of PLS, but it does not yield the best prediction results. In addition, we also list the results of the combination of IVE-PLS2 + RPLS and RPLS + IVE-PLS2. Because IVE-PLS2 removes all noise variables, RPLS successfully detected all five outliers. Figure 5B shows that the results of variables elimination by RPLS + IVE-PLS1 and RPLS + IVE-PLS2. However, MCOVS-PLS obtains the perfect prediction performance again. Figure 5C shows the process of variables elimination by MCOVS-PLS. we can see from this plot that with the increasing of iterations RMSECVs gradually decrease. When the number of iterations reaches 187, we obtained the best prediction, which gives the RMSECV value of approaching 0. From the simulated data, we can conclude that a good modeling approach simultaneously detecting outliers and selecting variables is of practical necessity for deeply improving prediction power. MCOVS-PLS has successfully done it!

For the simulated QSAR/QSPR data with noises, the results similar to the ones without noises were obtained. MCOVS-PLS again yields the best prediction accuracy among all modeling approaches. Discussed and analyzed fully in the Supplementary Material.

QSAR/QSPR data

ACE data

Results for these approaches applied to ACE dataset are listed in Table 3. To compare all the approaches impartially, 10 molecules in our experiment was detected as outliers for robust approaches including RPLS, RPLS + IVE-PLS, IVE-PLS + RPLS and MCOVS-PLS. From Table 3, we can see that PLS obtains the worst prediction performance among all approaches, which gives RMSECV value of 1.30 ± 0.10 and Q^2 value of 0.67. Such results are not surprising in that classical, nonrobust PLS regression model is built on the full dataset including all compounds and descriptors. IVE-PLS and RPLS yield the better prediction accuracy than classical PLS by the removal of redundant descriptors and outlying molecules, respectively. Such results indicate that PLS is not only strongly sensitive to outliers, but also seriously affected by uninformative variables. As for two combination approaches, IVE-PLS + RPLS and RPLS + IVE-PLS give quite similar prediction results and yield the better prediction accuracy than single modeling approaches. In addition, it is shown from Table 3 that the two approaches detect different molecules as outliers. The very possible reason is the strong interaction between variables and samples. Some outliers (e.g., 15 41 42 77 88) in the original chemical space of 56 descriptors have changed into normal samples in the reduced chemical space of 14 descriptors. Among all modeling approaches,

Table 3 The prediction results of the ACE dataset by different modeling approaches

Methods	RMSECV	Q^2	LVs	Vars	Outliers
PLS	1.30 ± 0.10	0.67	4	56	–
IVE-PLS	1.13 ± 0.08	0.76	4	19	–
RPLS	1.15 ± 0.08	0.75	4	56	15, 24, 36, 41, 42, 77, 88, 105, 106, 108
MCOVS-PLS[a]	0.86 ± 0.05	0.86	14	14	3, 6, 24, 36, 68, 77, 90, 105, 108, 114
IVE-PLS + RPLS	0.92 ± 0.07	0.83	14	14	3, 24, 35, 36, 77, 88, 92, 105, 108, 114
RPLS + IVE-PLS	0.91 ± 0.06	0.83	14	16	15, 24, 36, 41, 42, 77, 88, 105, 106, 108

[a] The results of MCOVS-PLS is considerably significant compared to other methods based on the threshold of $p = 0.05$

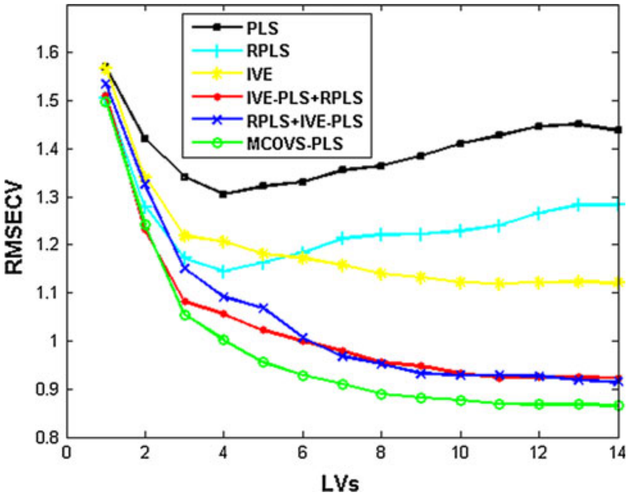
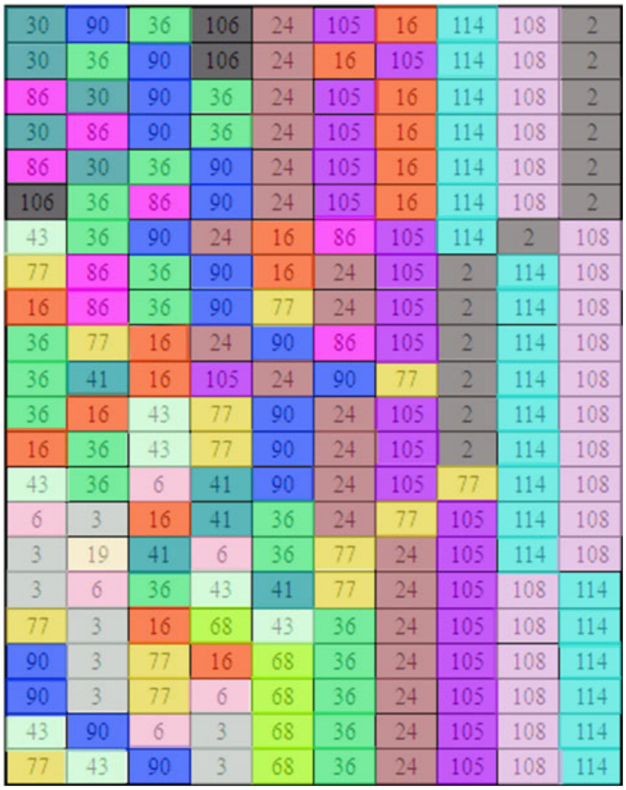


Fig. 6 ACE data: RMSECVs of 500 Monte Carlo cross-validations versus the PLS components in the final PLS model using the dataset obtained by each approach

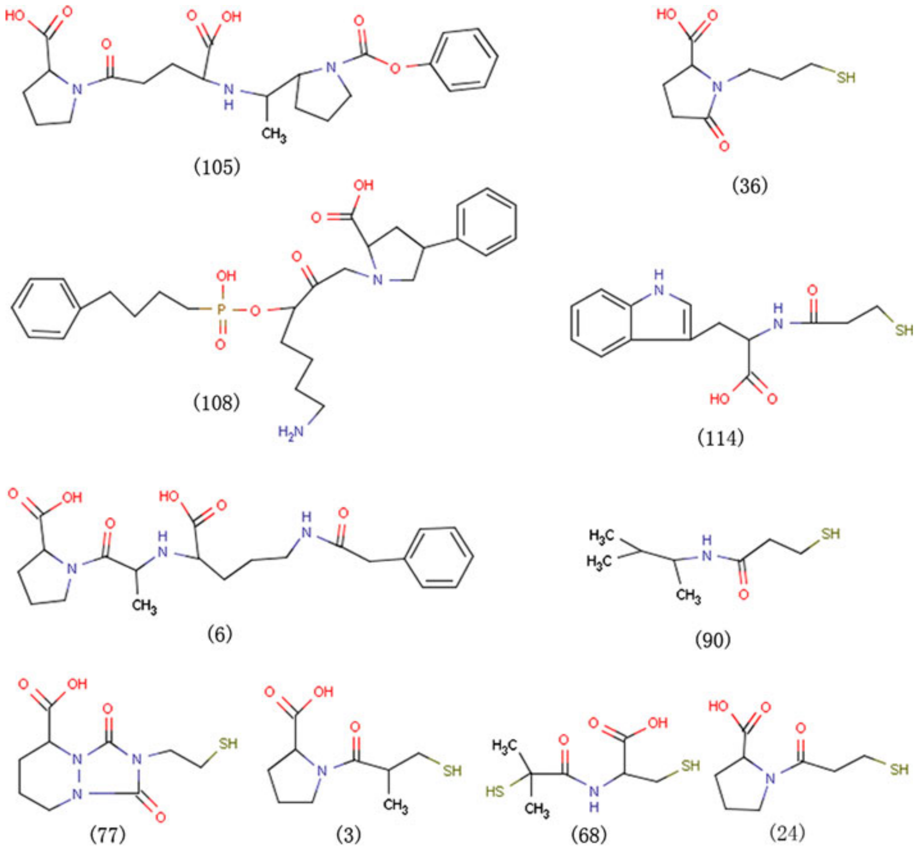


30	90	36	106	24	105	16	114	108	2
30	36	90	106	24	16	105	114	108	2
86	30	90	36	24	105	16	114	108	2
30	86	90	36	24	105	16	114	108	2
86	30	36	90	24	105	16	114	108	2
106	36	86	90	24	105	16	114	108	2
43	36	90	24	16	86	105	114	2	108
77	86	36	90	16	24	105	2	114	108
16	86	36	90	77	24	105	2	114	108
36	77	16	24	90	86	105	2	114	108
36	41	16	105	24	90	77	2	114	108
36	16	43	77	90	24	105	2	114	108
16	36	43	77	90	24	105	2	114	108
43	36	6	41	90	24	105	77	114	108
6	3	16	41	36	24	77	105	114	108
3	19	41	6	36	77	24	105	114	108
3	6	36	43	41	77	24	105	108	114
77	3	16	68	43	36	24	105	108	114
90	3	77	16	68	36	24	105	108	114
90	3	77	6	68	36	24	105	108	114
43	90	6	3	68	36	24	105	108	114
77	43	90	3	68	36	24	105	108	114

Fig. 7 The pathway analysis of outlying molecules detected with different chemical space of descriptors. Each row represents the outliers detected with current iteration and each iteration removes two most unimportant descriptors. The residuals of outliers in each row are ranked (from small residual to large residual)

MCOVS-PLS obtains the best prediction, which gives RMSECV value of 0.86 ± 0.05 and Q^2 value of 0.86. This may indicate that MCOVS-PLS yields a cleaner dataset than other modeling approaches since it simultaneously considers the influence of two directions (i.e., sample and variable space). To well compare the performance of every approach, Fig. 6 shows the RMSECV values of 500 Monte-Carlo cross-validations on the dataset obtained by each approach. From Fig. 6, we can clearly see that the prediction results obtained by single modeling approaches are significantly inferior to ones by two combination approaches. However, MCOVS-PLS outperforms all other approaches considerably. As mentioned above, outlying of a molecule depends on the descriptors used for characterization. So, analyzing the outlying molecules within different representation of the dataset will be of great importance for deeply understanding complex phenomenon, e.g., the mechanism of producing outliers and the interaction between outlier detection and variable selection etc. [62]. Thus, the pathway analysis of outlying molecules detected within different chemical space of descriptors is shown in Fig. 7. We can clearly see that with the increasing of the iteration the outlying of molecules has minor changes and exhibits certain regularity. For example, samples 24, 36, 105, 108 and 114, as the outliers, always coexist in the process of variable reduction and finally tend towards stability. Such observations illustrate that the molecules may not be described by current molecular descriptors or PLS model and so become outlying molecules. These outliers can easily be detected by current robust approaches. The conclusion can be illustrated by the fact that RPLS and IVE-PLS + RPLS also detect them as outliers. However, some molecules such as sample 20, 106

Fig. 8 The outlying molecules detected by MCOVS-PLS



and 2 initially become outliers in the first several iterations but become normal samples later, especially for sample 2. This may indicate that these molecules are seriously affected by redundant descriptors. The introduction of redundant descriptors makes these molecules produce large prediction errors. So, RPLS probably detect them as outliers since RPLS does not consider the influence of variables on the outliers. We can also see that some molecules such as samples 90, 16, 86, 77, 6, 43, 41, 3 etc. seriously depend on the certain process of iteration. A very probable reason is that these molecules depend on certain molecule descriptors that are not important for most of molecules. The existence of such outliers may make some unimportant variables for most of molecules enter into the model established. To obtain further analysis, we check the structure of sample 77 due to their unregularity in the pathway plot. We find that only sample 77 and 2 have similar structure, which exhibits the abnormal behavior in the PLS model. The reason for abnormal behavior is that they are far away from the bulk of the dataset. Likewise, we can also see that samples 3, 24 and 36 have very similar main body structure and therefore are detected as outliers. A possible reason is that this main body structure of these molecules is collectively pointing to another mechanism of action and thereby does not be well modeled by current PLS. A further demonstration may require professional knowledge from pharmaceutical chemist. Finally, the outlying molecules detected by MCOVS-PLS are shown in Fig. 8.

As a further comparison with PLS, we split 114 ACE molecules into the training set of 76 molecules and the test set of 38 molecules according to the original article. The training set is used for developing a model and the test set is then used for predicting. The prediction results of ACE dataset are listed in Table 4. From this table, one can see that the improvement between MCOVS-PLS and PLS is quite significant statistically. Especially, when 4 outliers are removed from the test set, the RMSET value of MCOVS-PLS becomes 0.85 ± 0.02, compared to PLS (whose RMSET is 1.09 ± 0.08), the improvement is also statistically significant. For cross validation and the test set, p values are all 0 ($< 1 \times 10^{-6}$), respectively. When 4 outliers are removed from the test set, p value becomes also 0. These p values sufficiently indicated that MCOVS-PLS indeed provides statistically significant improvements compared to the original PLS method.

A good QSAR model not only has good summary statistics such as RMSECV and Q^2, but also yields high stability of predictions for each sample. Several studies have indicated that the high stability of predictions of models correlates with the accuracy of predictions [63–67]. Thus, the standard deviation of prediction errors for each sample can be used as an additional metric characterizing the

Table 4 The prediction results of ACE dataset by training/test protocol

Methods	CV[a]		Test[b]		Test[c]	
	RMSECV	Q^2	RMSET	RT^2	RMSET	RT^2
PLS	1.26 ± 0.09	0.72	1.42 ± 0.07	0.53	1.09 ± 0.08	0.60
MCOVS-PLS	0.95 ± 0.03	0.83	1.17 ± 0.02	0.67	0.85 ± 0.02	0.82

[a] Only five outliers (2, 6, 24, 36, 68) are removed from the training set

[b] All test samples are used

[c] Four outliers (1, 29, 32, 38) are removed from the test set

predictability and stability of the established models. Figure 9 shows such pictures in which the standard deviation of prediction errors of each sample versus the mean value of prediction errors are plotted for PLS, IVE-PLS, RPLS and MCOVS-PLS. From Fig. 9A, we can see that the distribution for normal samples seems to be much more diverse from the STD direction, which may indicate that classical PLS yields an inaccurate and nonrobust QSAR model. Compared with Fig. 9A, B showing the results after IVE-PLS significantly obtains a more compact plot. It seems that the reduction of descriptors helps to stabilize the model and so allows for a more accurate and robust QSAR model than classical PLS. Similarly, Fig. 9C seems to give the better results than classical PLS by the removal of outliers. It is worth noting that IVE-PLS and RPLS consider different directions to build better models, respectively. IVE-PLS gives the better stability of prediction compared to RPLS, although they obtain the similar prediction accuracy. However, MCOVS-PLS indicated by Fig. 9D yields the most compact distribution of samples compared to other approaches. That is, MCOVS-PLS not only gives the best results among these approaches, but also produces the highest stability of prediction. This again indicates that MCOVS-PLS can give a very robust and reliable QSAR model.

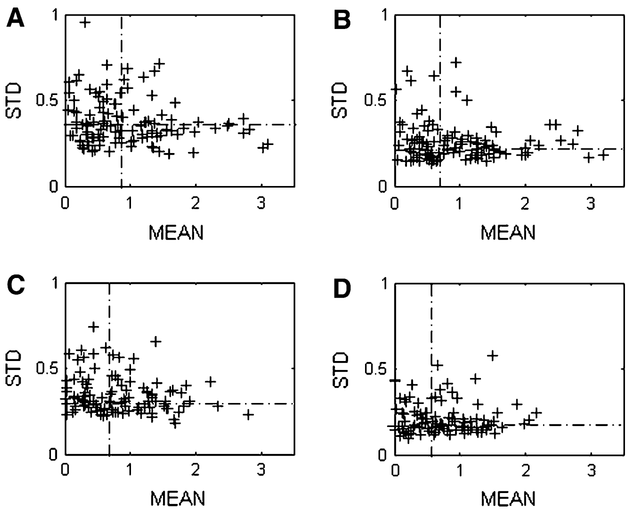


Fig. 9 ACE data: the plot of the standard deviations of prediction errors of each sample versus the mean values of prediction errors after different modeling approaches are carried out. The two *dash dot lines* in each plot represent the center of these data points, respectively. **A**: PLS, **B**: IVE-PLS, **C**: RPLS, **D**: MCOVS-PLS

Boiling point data

The normal boiling point is one of the major physicochemical properties used to characterize and identify an organic compound. Besides being an indicator for the physical state (e.g., liquid and gas) of an organic compound, the boiling point also provides an indication of its volatility. Moreover, boiling points can be used to predict or estimate other physical properties. The boiling point is often the first property measured for a new compound, so the prediction of boiling points for new chemicals is very important according to the QSPR model. Table 5 lists the prediction results of different approaches. For robust approaches, 20 outlying molecules were deleted and the remaining ones were used to establish the model. From Table 5, the PLS model built on the original dataset produces the poor prediction results considerably, which seems to indicate that there exist to be outliers and redundant descriptors in the dataset. IVE-PLS and RPLS again yield the better prediction accuracy than the classical PLS model by the removal of redundant descriptors and outlying molecules, respectively. It should be noted that RPLS remarkably outperforms IVE-PLS, which may indicate that the influence of outlying molecules is very strong. Furthermore, the combinations of two approaches indeed further improve the prediction accuracy of single approaches since they simply enjoy the advantages of two single

Table 5 The prediction results of the boiling point dataset by different modeling approaches

Methods[a]	RMSECV	Q^2	ME[b]	LVs	Vars
PLS	8.49 ± 0.29	0.9922	82.60	14	62
IVE-PLS	6.35 ± 0.32	0.9954	33.48	8	18
RPLS	3.32 ± 0.26	0.9987	12.88	10	62
MCOVS-PLS[c]	2.30 ± 0.06	0.9994	12.02	12	14
IVE-PLS + RPLS	3.19 ± 0.09	0.9985	14.48	9	18
RPLS + IVE-PLS	2.65 ± 0.06	0.9992	12.31	9	14

[a] 213 molecules are used for constructing different models except for PLS and IVE-PLS

[b] The maximum error generated by 2000 Monte Carlo cross-validation

[c] The results of MCOVS-PLS is considerably significant compared to other methods based on the threshold of $p = 0.05$

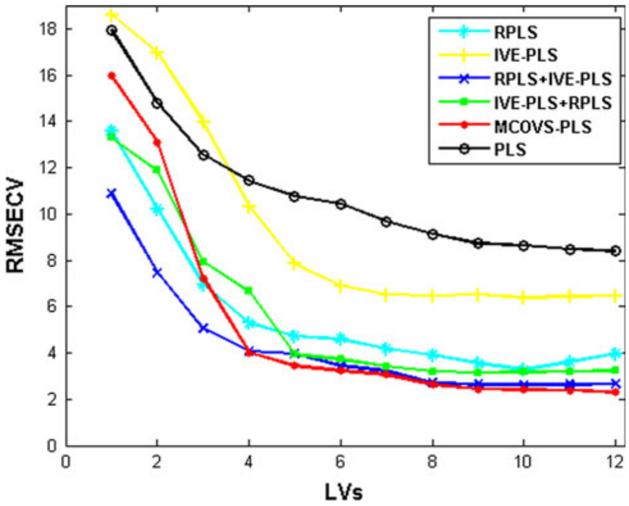


Fig. 10 Boiling point data: RMSECVs of 2000 Monte Carlo cross-validations versus the PLS components in the final PLS model using the dataset obtained by each approach

approaches. However, the results by RPLS + IVE-PLS are far better than ones obtained by IVE-PLS + RPLS. A main reason may be that the prediction performance is probably dominated by the influence of outlying molecules. In addition, it is worth noting that RPLS + IVE-PLS uses the fewer 14 descriptors to describe the information of molecules compared to IVE-PLS + RPLS. Such observations are conceivable since the inhomogeneity of molecules is largely reduced. It is shown from Table 5 that MCOVS-PLS outperforms all other modeling approaches, which gives RMSECV value of 2.30 ± 0.20 and Q^2 value of 0.9994. Such result is 13.3% and 26.9% relative improvement in RMSECV compared to RPLS + IVE-PLS and IVE-PLS + RPLS, respectively. A comparison of ME (maximum error produced by Monte Carlo cross-validation) for each approach shows that better results for each molecule were obtained from MCOVS-PLS and RPLS + IVE-PLS. The RMSECV values of 2,000 Monte Carlo cross-validations on the dataset obtained by each approach are shown in Fig. 10. Likewise, the prediction results of boiling point data by training/test protocol (154 molecules are used for training and other 79 ones are used for testing) are listed in Table 6, from which one can see that MCOVS-PLS again provides statistically significant improvements compared to PLS (For cross validation and test set, p values are 0 and 0.514, respectively. When 4 outliers are removed from the test set, p value becomes 0). Figure 11 shows the stability analysis of models obtained by each approach. We can clearly see that the PLS model built on the full dataset seems to produce a dispersive result (Fig. 11). Some potential outlying molecules should exist in the dataset. After variable selection by IVE-PLS, the results indicated by Fig. 11B obtain certain improvement, especially for STD direction. The reduction of molecular descriptors can improve the stability of models. However, Fig. 11C showing the results of RPLS obtains a more compact plot compared to IVE-PLS. This indicates that the influence of outlying molecules is very strong for the model established. In addition, compared to Fig. 11A, when the outlying molecules contaminate the dataset, they will also have a large influence on the normal samples, This time, MCOVS-PLS again yields the most compact plot among all modeling approaches (Fig. 11D).

Conclusions

The application of MCOVS-PLS allows one to simultaneously detect the outliers and select a compact subset of variables before a final modeling is carried out. The approach makes full use of the distribution of model features to construct a consistent framework simultaneously considering both sample space and variable space. The use of relevant distribution of model features inherently provides a feasible way to effectively describe the information contained by the original samples. It can be concluded that the MCOVS approach performs well for outlier detection and variable selection in regression models and that the regression model after discarding of the outliers and redundant variables has a better prediction performance. The MCOVS approach is a computer-intensive program and so enjoys the advantage of being completely automatic. The approach was applied to PLS, but it can be

Table 6 The prediction results of boiling points by training/test protocol

Methods	CV[a]		Test[b]		Test[c]	
	RMSECV	Q^2	RMSET	RT^2	RMSET	RT^2
PLS	9.07 ± 0.40	0.9930	5.38 ± 0.13	0.9930	3.76 ± 0.17	0.9952
MCOVS-PLS	2.63 ± 0.05	0.9993	5.02 ± 0.05	0.9993	1.81 ± 0.06	0.9985

[a] Only 15 outliers (2, 8, 13, 21, 34, 44, 45, 50, 68, 101, 116, 136, 148, 153, 154) are removed from the training set

[b] All test samples are used

[c] Four outliers (45, 51, 53, 55) are removed from the test set

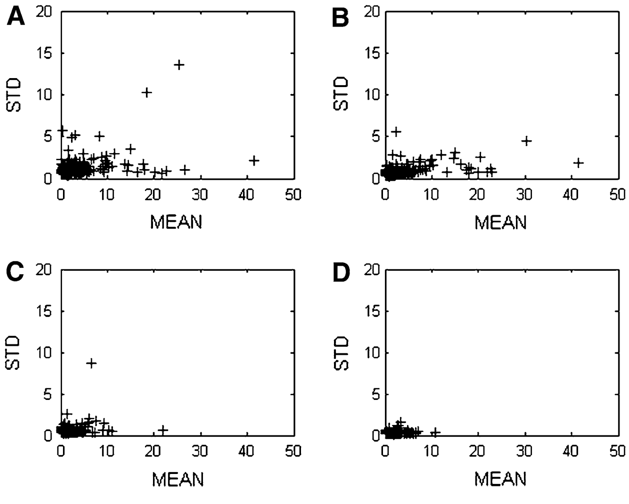


Fig. 11 Boiling point data: the plot of the standard deviations of prediction errors versus the mean values of prediction errors after different modeling approaches are carried out. For PLS, two samples beyond the scope of axis are not displayed. **A**: PLS, **B**: IVE-PLS, **C**: RPLS, **D**: MCOVS-PLS

considered to be equally useful for MLR or other related approaches.

The findings of this study clearly illustrate that the interaction between outliers and variables indeed exists. When the QSAR/QSPR datasets are contaminated by outliers and redundant variables, there is a large chance of selecting a wrong model that fails to reflect the true relationship. However, MCOVS-PLS has effectively reduced the risk that the interaction brings about to large extent and thereby yields the best prediction performance compared to other approaches. Moreover, the pathway analysis of outliers detected within different chemical space of descriptors should be of practical necessity since it helps the researchers to deeply understand the mechanism of producing outlying molecules and the relationship between outlying molecules and molecular descriptors.

Acknowledgments We would like to thank the reviewers for their useful discussions, comments and suggestions throughout this entire work. This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 20875104 and No. 10771217), the international cooperation project on traditional Chinese medicines of ministry of science and technology of China (Grant No. 2007DFA40680). The studies meet with the approval of the university's review board.

References

1. Dudek AZ, Arodz T, Galvez J (2006) Comb Chem High Throughput Screen 9:213
2. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO (2007) J Chem Inf Model 47:150
3. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) J Chem Inf Model 45:786
4. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ (2004) J Chem Inf Comput Sci 44:1497
5. Gunturi SB, Narayanan R (2007) QSAR Comb Sci 26:653
6. Konovalov DA, Coomans D, Deconinck E, Vander Heyden Y (2007) J Chem Inf Model 47:1648
7. Liang YZ, Yuan DL, Xu QS, Kvalheim OM (2008) J Chemometr 22:23
8. Rucker C, Meringer M, Kerber A (2005) J Chem Inf Model 45:74
9. Karthikeyan M, Glen RC, Bender A (2005) J Chem Inf Model 45:581
10. Cronin MTD, Livingstone DJ (2004) Predicting chemical toxicity and fate. CRC Press, Boca Raton
11. Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York, p 329
12. Liang Y-Z, Kvalheim OM (1996) Chemom Intell Lab Syst 32:1
13. Konovalov DA, Llewellyn LE, Vander Heyden Y, Coomans D (2008) J Chem Inf Model 48:2081
14. Huber PJ (2004) Robust statistics in Wiley Series in probability and statistics. Wiley, New York
15. Rousseeuw PJ (1984) J Am Stat Assoc 79:871
16. Agull J, Croux C, Van Aelst S (2008) J Multivar Anal 99:311
17. Walczak B, Massart DL (1995) Chemom Intell Lab Syst 27:41
18. Juan AG, Rosario R (1998) J Chemometr 12:365
19. Hubert M, Branden KV (2003) J Chemometr 17:537
20. Zhang MH, Xu QS, Massart DL (2003) Chemom Intell Lab Syst 67:175
21. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) J Chem Inf Comput Sci 44:1630
22. Sutter JM, Dixon SL, Jurs PC (2002) J Chem Inf Comput Sci 35:77
23. Clark DE, Westhead DR (1996) J Comput Aided Mol Des 10:337
24. Rogers D, Hopfinger AJ (2002) J Chem Inf Comput Sci 34:854
25. Shen Q, Jiang J-H, Jiao C-X, Shen G-l, Yu R-Q (2004) Eur J Pharm Sci 22:145
26. Xu L, Zhang W-J (2001) Anal Chim Acta 446:475
27. Tibshirani R (1996) J R Stat Soc B Methodol 58:267
28. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Ann Stat 32:407
29. Rainer G, Torsten S (2008) J Comput Chem 29:847
30. Kirchner H (2000) Altern Lab Anim 28:364
31. Cronin MTD, Dearden JC, Moss GP, Murray-Dickson G (1999) Eur J Pharm Sci 7:325
32. Cronin MTD, Schultz TW (2003) J Mol Struct THEOCHEM 622:39
33. Cavill R, Keun HC, Holmes E, Lindon JC, Nicholson JK, Ebbels TMD (2009) Bioinformatics 25:112
34. Tolvi J (2004) Soft Comput Fusion Found Methodol Appl 8:527
35. Wiegand P, Pell R, Comas E (2009) Chemom Intell Lab Syst 98:108
36. Menjoge RS, Welsch RE (2010) Comput Stat Data Anal 54:3181
37. Aksenova T, Volkovich V, Villa AEP (2005) Robust structural modeling and outlier detection with GMDH-type polynomial neural networks, in artificial neural networks: formal models and their applications. ICANN, p 881
38. Plomin R, Haworth CMA, Davis OSP (2009) Nat Rev Genet 10:872
39. Manly BFJ (1998) Randomization, bootstrap and Monte Carlo in biology, in texts in statistical science, 2nd edn. Chapman and Hall, London, p 399
40. Robert CP, Casella G (1999) Monte Carlo statistical methods in Springer texts in statistics. Springer, New York
41. Efron B, Tribshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall/CRC, New York, p 436
42. Efron B (1979) Ann Stat 7:1

43. Efron B, Gong G (1983) Am Stat 37:36
44. Efron B, Tibshirani R (1986) Stat Sci 1:54
45. Gentle JE (2006) Elements of computational statistics. Springer Science and Business Media, Inc., New York
46. Shao J (1993) J Am Stat Assoc 88:486
47. Xu Q-S, Liang Y-Z (2001) Chemom Intell Lab Syst 56:1
48. Xu Q-S, Liang Y-Z, Du Y-P (2004) J Chemometr 18:112
49. Cao D-S, Liang Y-Z, Xu Q-S, Li H-D, Chen X (2010) J Comput Chem 31:592
50. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C (1996) Anal Chem 68:3851
51. Riccardo L (1994) J Chemometr 8:65
52. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579
53. Bak A, Gieleciak R, Magdziarz T, Polanski J (2005) J Chem Inf Model 46:2310
54. Myers RH (2005) Classical and modern regression with applications. PWS-KENT, Boston
55. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear regression models. Irwin, Chicago
56. Sutherland JJ, O'Brien LA, Weaver DF (2004) J Med Chem 47:5541
57. Cao C, Liu S, Li Z (1999) J Chem Inf Comput Sci 39:1105
58. Rucker G, Rucker C (1999) J Chem Inf Comput Sci 39:788
59. Wessel MD, Jurs PC (1995) J Chem Inf Comput Sci 35:68
60. Polanski J, Gieleciak R (2003) J Chem Inf Comput Sci 43:656
61. Bak A, Polanski J (2007) J Chem Inf Model 47:1469
62. Kim K (2007) J Comput Aided Mol Des 21:63
63. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A (2008) J Chem Inf Model 48:1733
64. Beck B, Breindl A, Clark T (2000) J Chem Inf Comput Sci 40:1046
65. Chalk AJ, Beck B, Clark T (2001) J Chem Inf Comput Sci 41:457
66. Schwaighofer A, Schroeter T, Mika S, Laub J, ter Laak A, Sulzle D, Ganzer U, Heinrich N, Muller K-R (2007) J Chem Inf Model 47:407
67. Kolossov E, Stanforth R (2007) SAR QSAR Environ Res 18:89