A New Strategy of Outlier Detection for QSAR/QSPR

DONG-SHENG CAO,[1] YI-ZENG LIANG,[1] QING-SONG XU,[2] HONG-DONG LI,[1] XIAN CHEN[1]
[1]*Research Center of Modernization of Traditional Chinese Medicines, Central South University, Changsha 410083, People's Republic of China*
[2]*School of Mathematical Sciences and Computing Technology, Central South University, Changsha 410083, People's Republic of China*

Received 13 January 2009; Revised 5 April 2009; Accepted 11 May 2009
DOI 10.1002/jcc.21351
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: The crucial step of building a high performance QSAR/QSPR model is the detection of outliers in the model. Detecting outliers in a multivariate point cloud is not trivial, especially when several outliers coexist in the model. The classical identification methods do not always identify them, because they are based on the sample mean and covariance matrix influenced by the outliers. Moreover, existing methods only lay stress on some type of outliers but not all the outliers. To avoid these problems and detect all kinds of outliers simultaneously, we provide a new strategy based on Monte-Carlo cross-validation, which was termed as the MC method. The MC method inherently provides a feasible way to detect different kinds of outliers by establishment of many cross-predictive models. With the help of the distribution of predictive residuals such obtained, it seems to be able to reduce the risk caused by the masking effect. In addition, a new display is proposed, in which the absolute values of mean value of predictive residuals are plotted versus standard deviations of predictive residuals. The plot divides the data into normal samples, ***y*** direction outliers and ***X*** direction outliers. Several examples are used to demonstrate the detection ability of MC method through the comparison of different diagnostic methods.

 J Comput Chem 00: 000–000, 2009

Key words: QSAR/QSPR; outliers; Monte-Carlo cross-validation; robust regression; regression diagnostics

Introduction

Quantitative structure-activity/property relationship (QSAR/QSPR) is an important method which has been applied to modeling and prediction of many physicochemical and biological properties successfully, such as boiling point, melting point, aqueous solubility, toxicity, retention index, and the activities of many drugs, etc.[1–15] The basis of such relationships is based on an assumption that compounds of similar structure will exhibit similar properties. That is to say, a data set with very similar chemical structures should give accurate prediction of analogous molecular property when used to establish a QSAR/QSPR model. In addition, a crucial issue in QSAR/QSPR is the predictive ability of the models, good predictive ability is essential for practical use of the model (e.g., directing the synthesis of more active or less toxic chemicals, contributing to the toxicological profiling of chemicals for regulatory purposes). Whereas, there are tens of thousands of chemical molecules spanning the whole chemical space.[16] Thus, the diversity of chemicals will bring about some difficulties for the establishment of QSAR/QSPR model undoubtedly.[17] Among these difficulties we may list: (1) it is well known that the model built by some training set will be strongly dependent upon the structures defined by the training set. Moreover, if there are some special chemicals called as outliers departing from the bulk of the data set, they will destroy the similarity of the chemicals and influence the fitting and the subsequent prediction ability of the QSAR/QSPR model. A number of detection methods were reported in the QSAR/QSPR literature.[18–20] (2) For multiple linear regression (MLR) models, which are in general recommended for QSAR/QSPR modeling, the criterion to decide whether a model generated is good or not is commonly defined by the square coefficient of fitting model (R^2) and the square coefficient of cross validated (Q^2). Generally speaking, the values of R^2 and Q^2 should approximate mutually for a good model.[21] However, as a result of the diversity of

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Y.-Z. Liang; e-mail: yizeng_liang@263.net

Contract/grant sponsor: National Nature Foundation Committee of People's Republic of China; contract/grant numbers: 20875104, 10771217

Contract/grant sponsor: The international cooperation project on traditional Chinese medicines of ministry of science and technology of China; contract/grant numbers: 2006DFA41090, 2007DFA40680

chemicals, R^2 is sometimes quite different from Q^2. Finding out the reason is very important for the establishment of a good model. (3) Since any linear QSAR/QSPR model based on MLR is constructed on a particular dataset, it is important to build a model by using a dataset, in which the structures of the chemicals could be well defined, and then to test the model by using the other different dataset of chemicals with similar structures. So, how to select a training set and a test set without outliers to establish a QSAR/QSPR model is worthy of discussing and studying.[22,23]

On the basis of the above difficulties, the present investigation aims to detect the outliers in the model in that the aforementioned problems are related to the existence of the outliers to a great extent. Outliers are observations that appear to break the pattern shown by the main body of the data. There are many reasons for the presence of outliers, from recording errors to a non-representative sampling design. So important is outlier detection in multivariate regression calibration that Kutner and Neter have a comprehensive discussion in their book.[24]

There are two approaches in current statistics to cope with outliers: diagnostics and robust estimators. As discussed by Rousseeuw and Leroy, diagnostics and robust regressions have the same goal although they proceed in opposite order.[25] Diagnostic approach starts by identifying the outliers and then fits the rest of the data by many regression methods. The commonly used diagnostic methods,[26–30] such as the mean standard deviation and the hat matrix leverage, depend on testing whether a statistic exceeds some critical value derived from a special distribution, such as the normal or χ^2 distribution. These methods have a good ability when used solely for the identification of the prediction outliers and single calibration outlier. But, they will give inaccurate diagnostic results when multiple outliers coexist in the model established.[31] Multiple outliers will distort the measures of the mean value and the covariance matrix to such an extent that these observations may not be recognized when analyzed by hat matrix leverage. This phenomenon is termed the masking effect. Moreover, many spiffy diagnostic methods, such as minimum volume ellipsoid (MVE),[32,33] ellipsoidal multivariate trimming (MVT),[34,35] minimum covariance determinant (MCD)[36,37] resampling by half-means (RHM) and smallest half-volume (SHV),[38] have been used to detect outliers. The RHM and SHV methods proposed by Egan and Morgan have a high computational effectiveness than the other methods. Although the aforementioned methods can handle the masking effect to some extent, they just emphasized the outliers in samples direction. They are also based on the idea of classification and attempt to find the main body of the data, and then outliers are looked upon as the other sort which are different from the majority of the data and are therefore removed. In the process, neither of them takes the dependent variables into account. So, these methods are not sufficient to discover all outliers in regression analysis when used solely to detect the outliers.

In a robust regression model, the capital goal is to construct estimator which fits the majority of the data and examine the residuals from this fit to detect the outliers. So, many robust regression methods such as M-estimators, least median of squares (LMS), least trimmed squares (LTS), robust principal component regression (RPCR), robust partial least squares (RPLS), and robust principal components regression based on principal sensitivity vectors (RPPSV), have been used to detect the outliers.[39–50] These methods use the robust estimators instead of the minimum of error sum of squares, for this reason, they are not susceptible to the outliers. In all methods, LMS provides a breakdown point of 50%, which is the maximum that can be achieved by a robust method. The RPCR method developed by Massart[45] was made up of two steps which were also stabilized with the help of the combination of MVT and LMS. A more thorough discussion of these methods can be found in Rousseeuw and Leroy.[32] These robust methods have a good performance to detect the outliers in dependent variable direction, but they become somewhat incompetent to detect the sample outliers (except for RPCR and RPLS).

The discarding of the outliers is based on the model established; the different models may generate different outliers. So, taking the outliers into account solely without the model is not suitable. Moreover, the coexistence of the outliers in sample and dependent variable direction may also influence the quality of the model established. On the basis of the aforementioned, we offer a different view of the multivariate outlier detection problem, which is based on the Monte-Carlo cross-validation method and should handle all the outliers simultaneously.

Theory and Method

Notation

The data matrix $\boldsymbol{X}$ has m observations in the rows and n variables in the columns. Vectors are shown a bold lowercase, while scalar variables are shown in lowercase.

Outliers in QSAR/QSPR

In fact, there may be three types of outliers influencing the quality of the model in QSPR/QSAR study. Figure 1A illustrates different outliers in an example of simple regression. The first one is the outliers in the dependent variable $\boldsymbol{y}$ direction which break away from the normal distribution of $\boldsymbol{y}$ and will cause a large error sum of squares. In the example, point 1 is an outlier in $\boldsymbol{y}$ direction. Robust regression methods should cope with this problem easily, if there are a few outliers without masking effect. The second one is the outliers in the predictor or independent variable $\boldsymbol{X}$ direction. This sort of outliers is far away from the main body of the samples. In the example, point 2 and 3 are $\boldsymbol{X}$ outliers or leverage points, because their $\boldsymbol{x}$ value is outlying. But, point 2 is a good leverage point which does not cause a large error sum of squares and point 3 is a bad leverage point. When the QSAR/QSPR data contaminated by the leverage points are used to establish the model, a negligible variation may cause a large fluctuation for this model. A third type of outliers, so called outliers towards the model, can be found only after building the regression model. They represent a different relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. Model outliers are a special sort of outliers which exist in the QSAR/QSPR data set extensively due to diverse molecular structures in QSAR/QSPR study. In Figure 1A, points marked by 4 may be an example of the outliers toward the model which are not only an outlier in $\boldsymbol{y}$, also

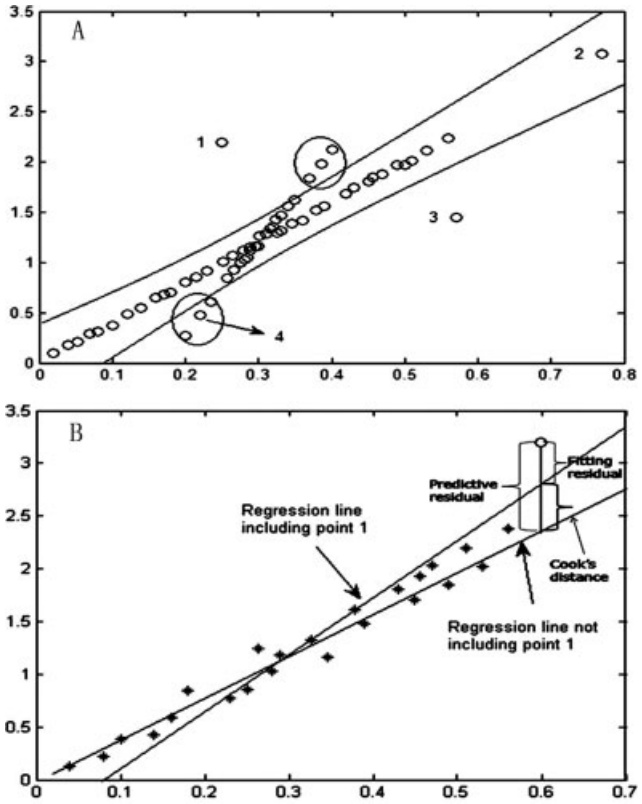

Figure 1. (A) Simple regression example with (1) y outlier; (2) good leverage point; (3) bad leverage point; (4) model outlier (the two curves represent the margin of 3Δ). (B) A data set with a single influential data point. Cook's distance measures the difference between PRESS residual and ordinary residual. So, PRESS residual is more likely to disclose the y outlying samples.

an outlier in $\boldsymbol{X}$. In addition, these three types of outliers mentioned commonly coexist in a model. A good outlier detection method should identify them exhaustively and simultaneously.

Cook's Distance

Cook's distance considers the influence of each case on all m fitted value,[22] which could be usually regarded as a good measure for outlier detecting. Cook's distance measure, denoted by D_i, is an average influence measure, showing the effect of the ith case on all m fitted values and model parameters:

$$D_i = \frac{\sum_{j=i}^{m}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\text{MSE}} = \frac{(\beta - \beta_{(i)})^T(\mathbf{X}^T\mathbf{X})(\beta - \beta_{(i)})}{p\text{MSE}} \quad (1)$$

where $\hat{y}_j$ is the fitted value for the jth case when all m cases are used in fitting the regression function and $\hat{y}_{j(i)}$ is the predicted value for the jth case obtained when the ith case is omitted in fitting the regression function. p and MSE are the freedom degree of this regression model and the mean squares of error, respectively. Cook's distance measure D_i can also be calculated without fitting a new regression function each time a different case is deleted. An algebraically equivalent expression is:

$$D_i = \frac{e_i^2}{p\text{MSE}}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \quad (2)$$

where e_i and h_{ii} is the fitting residual value and the leverage value for the ith case. Note from the above formula that D_i depends on two factors: (1) the size of the residual e_i and (2) the leverage value h_{ii}. So, the Cook's distance measure considers the effect of both the predictor variables and the dependent variables. Cook's distance can also be extended to diagnostics of multiple outliers by measuring the joint effect of deleting more than one case.

$$D_i(I) = \frac{\sum_{j=i,j\notin I}^{m}(\hat{y}_j - \hat{y}_{j(I)})^2}{p\text{MSE(I)}} = \frac{(\beta - \beta_{(I)})^T(\mathbf{X}^T\mathbf{X})(\beta - \beta_{(I)})}{p\text{MSE(I)}} \quad (3)$$

Here, I represent the indices corresponding to a subset of cases. The other symbols have the same meanings as in Cook's distance for the case of a single observation. The quantity $D_i(I)$ can be interpreted in an analogous way to D_i. However, the selection of cases to be included in I is not at all obvious. Therefore, the computation for all pairs, triplets and so on, lead to C_n^m runs, where $m = 1, 2, \ldots, n/2$. This is such a difficult task that the computers are not able to finish it.[25]

Fitting Residual and Predictive Residual

It may be worth noting the difference between the fitting residual and the predictive residual. If y_i is far outlying, both from $\boldsymbol{y}$ direction and $\boldsymbol{X}$ direction, the fitted least squares regression function based on all cases including the ith one may be influenced to come close to y_i, yielding a fitted $\hat{y}_i$ near y_i. In that event, the fitting residual, say e_i, will be small and will not disclose that y_i is outlying. On the other hand, if the ith case is excluded before the regression function is fitted, the least squares fitted value $\hat{y}_{i(i)}$ is not be influenced by the outlying y_i observation, and the residual for the ith case will then tend to be larger, and therefore, more likely to disclose the outlying y observation. A simple figure is useful to illustrate ordinary residual, predictive residual, and the notion of influence (Cook's distance). Figure 1B shows such a single regressor. Clearly, the slope and intercept of the regression change considerably if the single observation, say point ○, is set aside. It can be seen that the predictive residual is considerably larger in magnitude than the fitting residual. Compared with the Cook's distance, which measures the average difference between the fitted value $\hat{y}_i$ and the estimated expected value $\hat{y}_{i(i)}$, the predictive residual reflects the difference between the actual observed value y_i and the estimated expected value $\hat{y}_{i(i)}$ and hence can help to reduce the risk caused by the masking effect without involvement of the fitted values $\hat{y}_i$.

Moreover, in the cross-validation cases, $n-1$ cases are generally used for predicting the "new" nth case, we can obtain the estimated variance of $e_{i(i)}$:

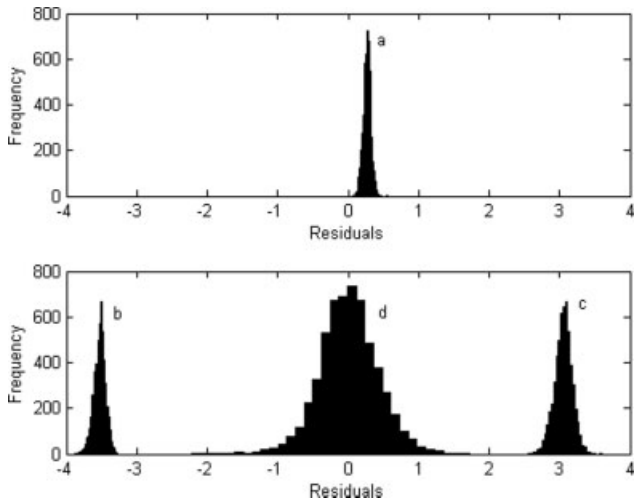

Figure 2. Histogram of sample residuals calculated by the MC method for a data set randomly generated using a multivariate normal distribution (a) normal sample; (b) *y* outlier; (c) model outlier; (d) ***X*** outlier.

$$s^2\{e_{i(i)}\} = \mathrm{MSE}_{(i)}(1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^t\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^t) \quad (4)$$

where $\boldsymbol{x}_i$ is the *ith* observation vector, $\mathrm{MSE}_{(i)}$ is the mean square error when the *i*th case is omitted in fitting the regression function, and $\boldsymbol{X}_{(i)}$ is the $\boldsymbol{X}$ matrix with the *i*th case deleted. An algebraically equivalent expression for $s^2\{e_{i(i)}\}$ is:

$$s^2\{e_{i(i)}\} = \mathrm{MSE}_{(i)}/(1 - h_{ii}) \quad (5)$$

From the eqs (4) and (5), we can see that the variability of the sampling distribution of $e_{i(i)}$ is effected by how far $\boldsymbol{x}_i$ is from the centroid $\overline{X_{(i)}}$, through the term $\boldsymbol{x}_i(\boldsymbol{X}_{(i)}^t\mathbf{X}_{(i)})^{-1}\boldsymbol{x}_i^t$ or h_{ii}. The further $\boldsymbol{x}_i$ from $\overline{X_{(i)}}$ is, the greater the quantity is and the larger the variance of $e_{i(i)}$ is. Hence, the variation of $e_{i(i)}$ obtained from different cases will be greater when $\boldsymbol{x}_i$ is far from the mean value than the ones near the mean value.

From discussion earlier, we can clearly see that predictive residuals reflect the difference between normal samples and outliers. They can effectively identify the single outlier existing in the data set. Nevertheless, when multiple outliers coexist in the model, the case will be not as the same as the one discussed earlier. To detect all the outliers and overcome the masking effect, the Monte-Carlo method as sampling without replacement is employed to extract the information and used as statistical inference.[51–53] In general, the Monte-Carlo method 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[54,55] (see ref. 56 for more detail). Suppose that one does Monte-Carlo cross-validation many times, one may obtain some kind of distribution of predictive errors for every sample. What kind of information one could extract from these distributions?

Figure 2 shows such a situation, in which the histograms of prediction errors from a 130-sample dataset, containing three kinds of outliers (30%), randomly generated using Monte-Carlo cross-validation for every sample. From Figure 2, one could get a very clear impression that there exist three kinds of samples in the dataset investigated. Figure 2A shows a distribution of predictive errors from a normal sample, looking like a normal distribution with a small mean value and standard deviation. Whereas, Figure 2D shows a distribution of prediction errors for an ***X*** outlier, which has a small mean value but a large standard deviation with a tailing toward large mean value direction. The distributions of the prediction errors of outliers in ***y*** direction are shown in Figures 2B and 2C. Their distributions of prediction errors are far away from the origin and have small standard deviations.

The MC Method

Inspired by the above discovery, we developed a Monte-Carlo cross-validation procedure for detecting outliers by studying the distribution of prediction errors of each sample obtained from original data set. We refer to this method as Monte-Carlo method (MC). Figure 3 shows a flow chart for the complete

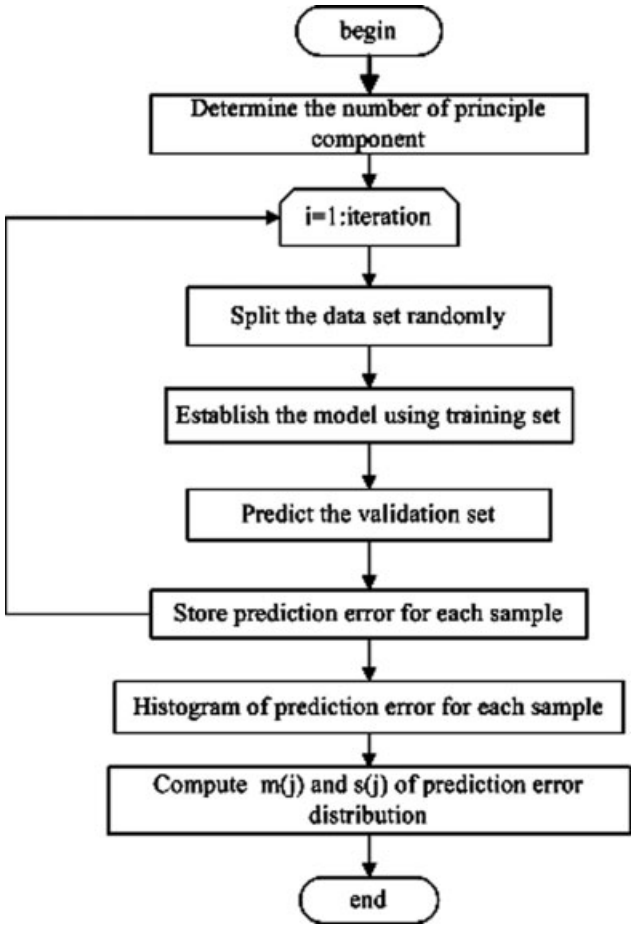

Figure 3. Flow chart showing the steps used in the MC method of outlier detection.

algorithm. The number of principle components was firstly determined using cross-validation in PLS and PCR methods (MLR method will be lack of this procedure). Then, with the help of the Monte-Carlo method, the whole data was randomly divided into two parts which are training set and validation set, respectively. Generally speaking, the size of the training set varies from 70 to 90% of all the data. Similar to other models, the training set was used to establish the model using the determined principle component. The validation set was used to predict and the prediction error would be obtained for each validation sample. This cycle was executed according to the predetermined number of times. Finally, the prediction error distribution for each sample was obtained. The histograms of these distributions were plotted and their statistic features were used to detect the outliers. Then, the mean value $m(j)$ and the standard deviation $s(j)$ of the error distribution for the jth sample were employed to describe this distribution.

$$m(j) = \frac{1}{k}\sum_{i=1}^{k} \text{error}(i) \tag{6}$$

$$s(j) = \left(\frac{1}{k-1}\sum_{i=1}^{k}(\text{error}(i) - m(j))^2\right)^{1/2} \tag{7}$$

Here k is the total times of which the jth sample was found in the validation set. The error (i) is the prediction error of the jth sample in the ith cycle.

The property of a sample in the model established can be reflected by its residual generated by the model.

$$\text{property}_i = f(e_i) \tag{8}$$

where property_i is the property of predictive sample i and e_i is the ith predictive sample residual generated by the model. (Here, different samples are selected to form the calibration set, thereby, they determine a model.) So, the distribution of the predictive residuals generated by many models can contain more sample information about whether this sample is an outlier or not. This is the important reason why we use Monte-Carlo to obtain multiple models. For a normal sample, its residual distribution will approach the origin and have a small uncertainty, that is to say, the predictive residual distribution will have a high and narrow peak close to the origin, because the normal samples occupy the bulk of all the data. Whereas, for a ***y*** or a model outlier, no matter how the models change, the predictive residual distribution of ***y*** outlier all have a large expectation value far away from the origin and moderate standard deviation. However, an ***X*** outlier is a case different from the ***y*** outlier. Because ***X*** outliers break away from the main body of all the samples, different models which are made up of different samples should predict a broad band of standard deviation for this ***X*** outlier. So, the distribution of its predictive residuals will have a wider peak around the origin. Here, we have used the empirically derived distribution, as is commonly done in bootstrapping.

In the MC method, the number (N) of Monte-Carlo experiments seems an important parameter which affects the quality of the predictive distributions. Given n (n is the number of the total samples) total observations contaminated by m outlier observations, more precisely, if k observations are set aside to generate the predictive residuals, the times N_i of selecting some observation i among N Monte-Carlo experiments is about given by:

$$N_i = \frac{C_{n-1}^{k-1}}{C_n^k} \times N = \frac{Nk}{n} \tag{9}$$

Notice that if N_i is big enough, the experiential distribution such obtained will be dominated by the inherent feature of the sample. On the other hand, the probability p of selecting at least one outlier observation for each Monte-Carlo experiment is given by:

$$p = \frac{C_m^1 C_{n-m}^{k-1} + C_m^2 C_{n-m}^{k-2} + \cdots + C_m^m C_{n-m}^{k-m}}{C_n^k} \tag{10}$$

Given $n = 100$, $m = 10$, $k = 20$, $N = 10{,}000$, we can get $N_i \approx 2000$, $p = 0.9049$. From these two numbers one could see that if we do 10,000 MC experiments, we could obtain around 2000 predictive residuals for each sample, which could be big enough to get some useful distribution parameters, say mean value and variance, to evaluate the features of the distribution. On the other hand, we could also obtain some information from outliers through each MC experiment, since we have the chance of 90 percent ($p = 0.9049$) to acquire at least one sample from the outliers for their predictive errors. Moreover, by means of the Monte-Carlo method, the computational complexity can be reduced substantially.

Theoretically, the fewer samples are selected randomly from the calibration samples, the more repeats are needed. Whereas, it has been proven that $N = n^2$ (n is the number of the total samples) is generally enough to make Monte-Carlo strategy better performance according to Zhang.[57] Besides, the size (n–k) of the calibration samples is also an important parameter according to our observation. If the calibration samples are too large, the p value may become somewhat small, that is to say, the chance of selecting at least one outlier observation is relatively small. So, the masking effect can be not disclosed well without the interaction of outlier observations, the division between ***X*** outliers and normal samples is not quite clear. So, to uncover the masking effect, it is very important to maintain a relative large p value. In practice, about 70–90% of the total samples are a relative suitable choice when used as calibration samples.

A visual diagnostic for the distribution of prediction errors is insufficient and complex. The MC method inherently provides a feasible way to detect all the outliers. To identify the outliers conveniently and directly, we offer a schema analogous to principle component plot. The statistic features of the distribution, say the absolute value of mean value and the standard deviation, were employed. Here, two histogram distributions were generated by the mean value and the standard deviation, respectively. Figure 4 shows the two distributions for the mean value and standard deviation of 1060 simulated samples. From Figure 4, it is shown that outliers and normal samples in two distributions have a significant difference. So, according to real circumstance, we can select a suitable tuning value such as 2–2.5 times of av-

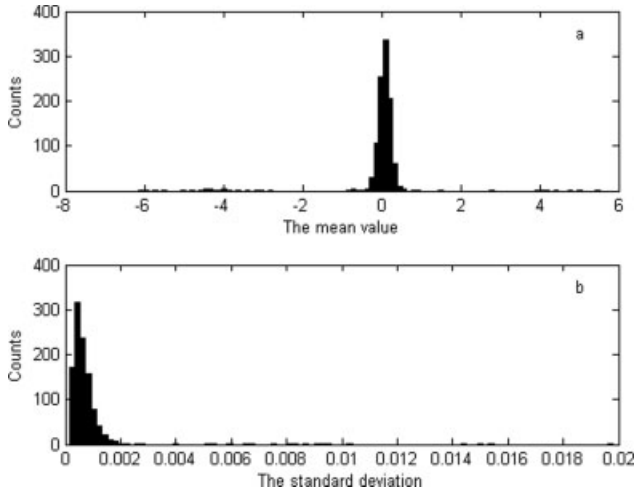

Figure 4. The distributions of the mean value and standard deviation for 1060 simulated samples. (a) mean value; (b) standard deviation.

erage value of the main body of the data set to differentiate outliers and normal samples. Thus, the schema was divided into four regions, each of which corresponds to one sort of outlier. Figure 5 shows the schematic diagram which differentiates outliers and normal samples. Among the four regions, the lower left region, which occupies most of all the data, is the normal samples which have small mean value and standard deviation. The upper left area is the sample outliers which have small mean value but large standard deviation. Conversely, the lower right region is the ***y*** outliers or model outliers which have large mean value but small standard deviation. Moreover, the upper right region is some extreme outliers in both sample and ***y*** direction, or the abnormal samples obtained due to recording errors, etc. This schema should provide an excellent and direct diagnostic to determine whether samples were checked to be the outliers.

To illustrate the performance of the proposed approach when dealing with different types of outliers, several examples often used for comparison in statistical and chemical literature together with an additional data sets from one of our research projects are used to test this MC method. Some of these data sets such as Stack Loss Plant Data, Hawkins-Bradu-Kass data not only have both ***X*** and ***y*** outliers but also exhibit extreme masking effects caused by multiple outliers.[58–61]

Simulated Data

For illustrative purposes, simulated data analogous to QSAR/QSPR model was designed to test this method (***X*** (100 × 10) and ***y*** (100 × 1) with normally distributed noise). In these examples discussed, matrix ***X*** contained independent columns (***X***), which represented some molecular descriptors, and a dependent column (***y***) being related to by $\boldsymbol{y} = f(\boldsymbol{X})$. 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 of chemical molecules. Different outliers are introduced into this data set.

1. ***y*** outliers: 10 additional ***y*** outliers with three-fold noise (about 20% variation of the normal y values) are added to this data set. In these ***y*** outliers, their corresponding independent variables derived from the main body of 100 normal samples.
2. ***X*** outliers: 10 additional ***X*** outliers with large Mahalanobis distance (two times larger than the average value of the leverage values from normal samples) are added to this data set. For these ***X*** outliers, they have the same functional relationship as the 100 normal samples.
3. Model outliers: 10 additional model outliers with different functional relationships are added to this data set. For these model outliers, their corresponding independent variables derived from the bulk of 100 normal samples. Moreover, one out of 10 model outliers, which also has a good functional relationship in two models due to the intersection of two models, should not be considered as model outliers.
4. Mixed outliers: the above 30 outliers mentioned are added to this data set.

Stack Loss Plant Data

Stack loss plant data set is an operational data of a plant for the oxidation of ammonia to nitric acid. This data includes 21 observations on three independent variables measuring flow of cooling air, cooling water inlet temperature, concentration of acid, and one dependent variable, stack loss.[58–60]

Hawkins-Bradu-Kass Data

Described and analyzed fully in the Supporting Information.[30,61]

QSAR/QSPR Data

Finally, two real QSAR/QSPR data sets were used to detect the method. The first data is boiling point data for the alkenes with

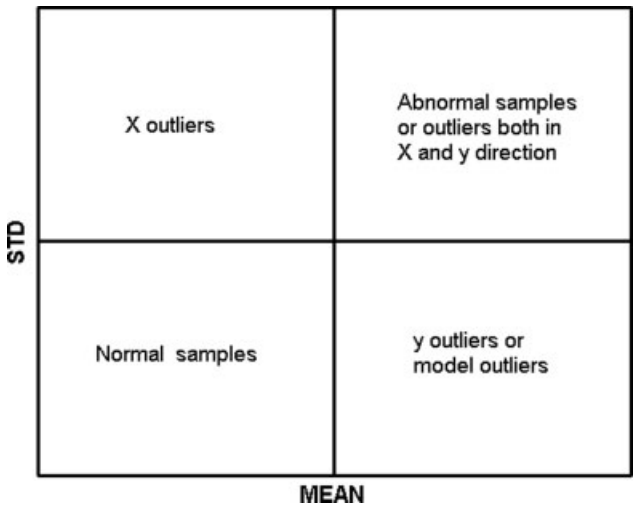

Figure 5. The schematic diagram which differentiates outliers and normal samples.

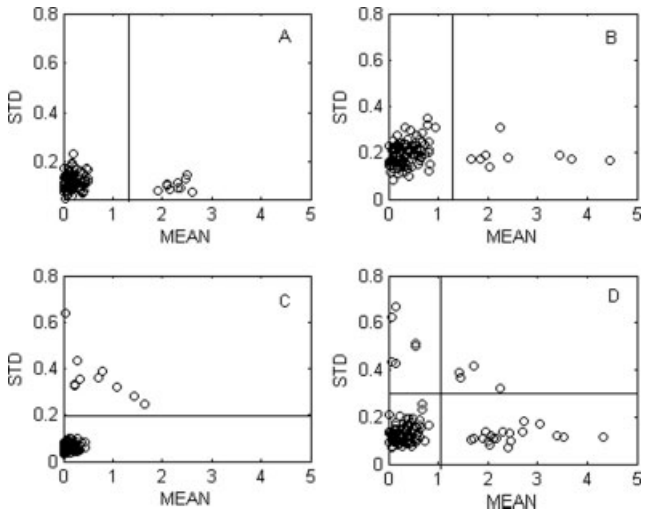

Figure 6. The result of variance of residuals versus mean of residuals on simulated data. (A) Data contaminated by *y* outliers; (B) Data contaminated by model outliers; (C) Data contaminated by ***X*** outliers; (D) Data contaminated by three type of outliers.

different branches. Sixty-two molecular descriptors which consist of connectivity indices and walk and path counts are used to characterize these alkenes. The range of experimental boiling points is between −42.10 and 525.04. The second data consists of 88 drug-like molecules on 33 molecular descriptors measuring different molecular information. Together, they form data set ***X***. The dependent variable ***y*** is the aqueous solubility of 88 drug-like molecules measured with accurate method as described in Ref. 62.

Results and Discussions

Simulated Data

Results for the MC method applied to simulated data are shown in Figure 6. In this simulated dataset, three types of outliers and mixed outliers are used to test this method. From Figure 6, we can see that the MC method can excellently identify all different types of outliers. Figure 6A shows the results for the data set including only outliers in ***y*** direction. These outliers have large mean value than normal samples. Similar to Figure 6A, Figure 6B shows the results for the data set including only model outliers. It should be mentioned here one sample (106) was not identified as model outliers since it has also a good functional relationship with the model investigated. Figure 6C shows the results for detecting ***X*** outliers. From this plot one could see that the entire datum is clearly divided into two parts. For three types of outliers above, the MC method all provides satisfactory results. More importantly, this method can also identify mixed outliers in which other diagnostic methods are not competent. From Figure 6D, The top left area is outliers in ***X*** direction which have a large standard deviation, and the lower right one gives outliers in ***y*** direction and model outliers, which have a large mean value. In addition, to demonstrate the influence that the outliers gave the model established, a data set without outliers was used to compare with the one with three types of outliers, which is shown in Figure 7A. From this plot, the results show that the samples are very close to each other with smaller mean values approaching zero and smaller deviations. However, when the data set contains mixed outliers the distribution even for normal samples (see lower left part of Fig. 7B) are much more diverse. That is to say, when the outliers contaminated the data set, they will have a large influence on the normal samples and hence lead to establish an inaccurate model.

For comparison of different methods, MCD, RHM, MVT, RPPSV, RPLS, and M-estimator were used to seek the outliers in the mixed data set. MVT used 23% trimming with iterations terminated after the covariance matrix stabilized. MCD sampled the data 1000 times and determined approximate squared Mahalanobis distances for all samples based on the subset of data having the smallest determinant. RHM used 650 samples and determined the relative frequencies of the samples having vector lengths in the upper 10% of vector lengths. RPPSV and M-estimator used the critical value equaled 2 and 2.24 to determine whether a sample is an outlier, respectively. The outlier detection ability of different methods is described using three indexes defined.[50]

$$\mathrm{MP} = \frac{\text{number of undetected true outliers}}{\text{number of true outliers (leverage points + high residual points)}}$$

$$\mathrm{LMP} = \frac{\text{number of undetected true leverage points}}{\text{number of true outliers}}$$

$$\mathrm{WP} = \frac{\text{number of normal points wrongly identified as outliers}}{\text{number of true outliers}}$$

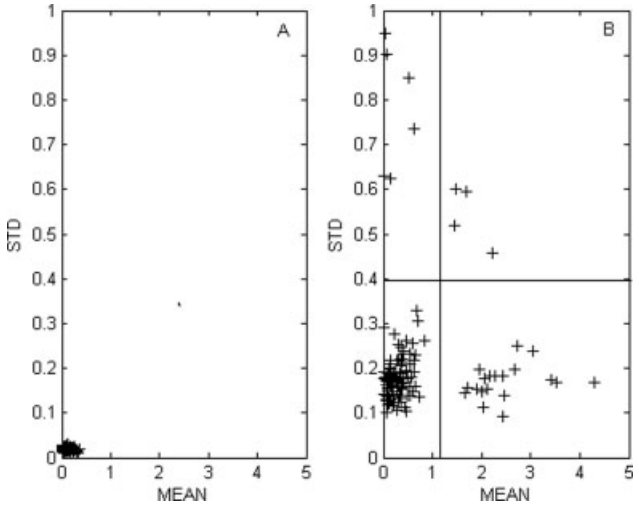

Figure 7. The result of variance of residuals versus mean of residuals on simulated data. A significant difference between A and B illustrates the influence of the outliers (A) data set without outliers; (B) data set with three types of outliers.

Table 1. Outlier's Detection Using Different Methods for Simulated Data.

Method	Outliers detected in regression	MP	LMP	WP
MC	101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 117 118 119 120 121 122 123 124 125 126127 128 129 130	0/29	0/29	0/29
MCD	103 105 106 121 122 123 124 125 126 127 128 129 130	16/29	0/29	16/29
MVT	105 106 121 122 123 124 125 126 127 128 129 130	17/29	0/29	17/29
RHM	110 106 105 121 122 123 124 125 126 127 128 129 130	16/29	0/29	16/29
M-estimator	101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 117 118 119 120 121 122 123 124 125 126	4/29	4/29	4/29
RPPSV	101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 117 118 119 120 121 122 123 124 126 129 130	3/29	3/29	3/29
RPLS	101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 117 118 119 120 121 122 123 124 125 126127 128 129 130	0/29	0/29	0/29

The true outliers in regression are from sample 101 to sample 130 except for 116.

Results for these methods applied to simulated data are shown in Table 1. The MC method identified all types of outliers exhaustively and obtained the best results among all methods. MCD, MVT, and RHM have similar ability to detect outliers, in addition, they have the ability to seek the ***X*** outliers but they are not competent for the ***y*** outliers. The reason why we obtained the results is that these three methods only made use of the information of samples but without considering the influence from ***y*** direction. M-estimator, which placed extra emphasis on finding the ***y*** outliers obtain the better detection results than the above three methods. It can identify all the ***y*** and model outliers and parts of ***X*** outliers. Moreover, similar to the MC method, sample 116 was not detected by the M-estimator method. RPPSV and RPLS can also diagnose the ***X*** outliers and ***y*** outliers and obtain a satisfactory result for the simulated data (except for sample 116). From this simulation study, it is concluded that the MC method is effective for different types of outliers and detects these outliers simultaneously.

Stack Loss Plant Data

Stack loss plant data set is an operational data of a plant which is used to illustrate robust regression techniques. It is known that this data contains several outliers which yield the masking effect. The classical Mahalanobis distance does not reveal them. Table 2 lists the detection results of different methods. From Table 2, we can see that the MC method and RPLS method obtained the best results among all the methods again and identified all the outliers simultaneously. MCD, MVT, and RHM diagnose all the ***X*** outliers such as samples 1, 2, 3. M-estimator finds all the ***y*** outliers such as samples 3, 4, 21. RPPSV obtains the poor result for the stackloss data and diagnoses samples 4 and 21, which are the ***y*** outliers. Figure 8 shows the detection result of MC method. It is shown that different type of outliers compactly clustered together, respectively. From Figure 8, Sample 3 is the ***y*** outlier who has a large mean value as well as the ***X*** outlier which has a large standard deviation. The MC method exhibited a good performance when multiple outliers coexisted. That is to say, the MC method can overcome the influence which the masking effect brings about.

Table 2. Outlier's Detection Using Different Methods for Stack Loss Data.

Method	Outliers detected in regression	MP	LMP	WP
MC	1 2 3 4 21	0/5	0/5	0/5
MCD	1 2 3 21	1/5	0/5	1/5
MVT	1 2 3 21	1/5	0/5	1/5
RHM	1 2 3	2/5	1/5	2/5
M-estimator	3 4 21	2/5	2/5	2/5
RPPSV	4 21	3/5	3/5	3/5
RPLS	1 2 3 4 21	0/5	0/5	0/5

The true outliers in regression are samples 1, 2, 3, 4, 21.

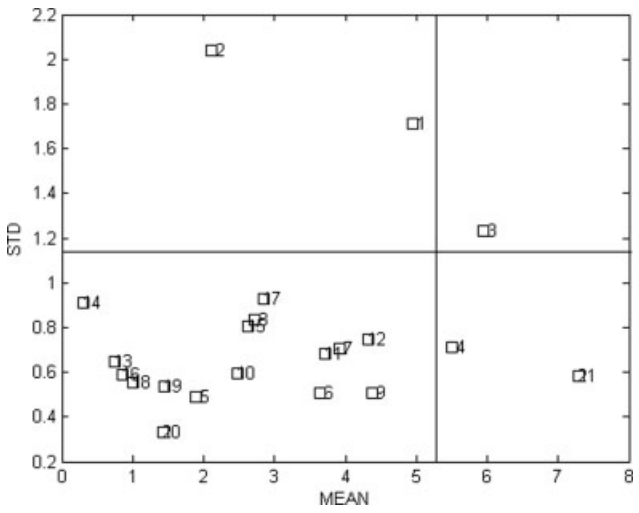

Figure 8. The result of variance of residuals versus mean of residuals on stack loss data.

Hawkins-Bradu-Kass Data

For Hawkins-Bradu-Kass data, the results similar to those of the Stack Loss Plant Data were obtained. MCD, MVT, RHM, RPLS, and M-estimator can obtain all the 14 outliers. Apart from this, RPPSV detected the first 10 outliers but the next four ones were nevertheless maintained in the model established. The MC method not only diagnosed all the outliers but also differentiated two types of outliers which have extreme masking effect.

QSAR/QSPR Data

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 (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, such as flash points,[63] critical Temperatures,[64] etc. 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. A large number of methods for estimating normal boiling points were previously reported in the literature. However, the boiling points of a few compounds in every model have been exceeded the chemical errors. That is to say, the prediction of boiling points for these compounds become inaccurate through the model established. The major reason may be the existence of the outliers caused by the diversity of the compounds.

Table 3 lists the predictive results of different methods for detecting the outliers. To compare all the methods impartially, the same number of outliers (52 molecules in our experiment) was deleted and the remaining ones were used to establish the predictive model. From Table 3, the MC method, RPLS, and RPPSV give quite similar results and the MC method somewhat outperform the two other methods. For M-estimator method, we can see that there is a large difference between RMSEF and RMSECV. A main reason is that latent outliers may coexist in the model established, because this method places extra emphasis only on the ***y*** outliers. To compare the performance of every method well, Figure 9 shows the RMSECV values of 500 Monte-Carlo cross-validations (20% samples were removed and used to predict) in PLS model. A good PLS model should have a small RMSECV value and small principal number. A big principal number may be caused due to the diversity of the molecules. That is to say, the training set was still contaminated by the outliers. From Figure 9, it is shown that MC and RPLS have similar principal numbers (about 11) and RMSECV values and the MC method gives the best results among all the methods. RHM, M-estimator and NONE (none of methods were used to detect the outliers) have relative small principal numbers but big RMSECV values. In addition, RPPSV have a very big principal number when it reaches a relative good RMSECV value. The possible reason may be that the training set was still contaminated by some latent outliers. According to Table 3 and Figure 9, we can clearly see that the MC method obtains the best performance among the above methods for the boiling point data.

Table 3. Outlier's Detection Using Different Methods for Alkanes.

Method	RMSEF	R^2	RMSECV	Q^2
MC	1.4237	0.9997	1.8974	0.9995
RHM	4.5758	0.9958	5.1677	0.9946
M-estimator	1.7298	0.9996	6.2533	0.9952
RPPSV	0.9739	0.9999	2.4280	0.9993
RPLS	1.5028	0.9997	1.9906	0.9994
NONE	5.6193	0.9967	8.4078	0.9926

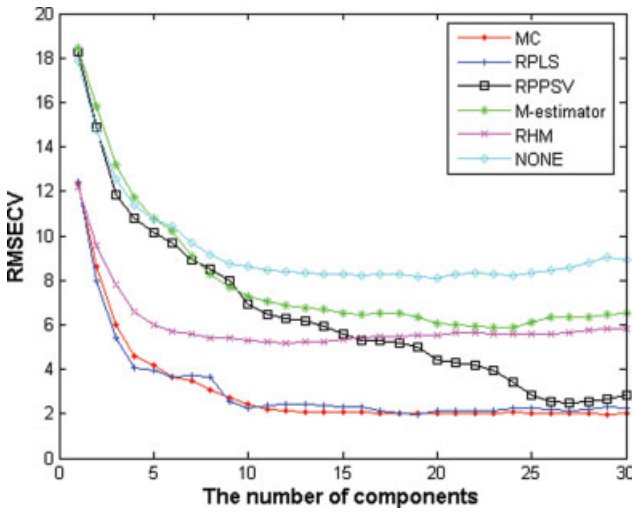

Figure 9. The RMSECV values of 500 Monte-Carlo cross-validations versus the number of principal components in PLS model.

Solubility Data

Solubility is a difficult property to predict, and one reason for this is the diversity of the drug-like molecules. That is to say, there are some special points which coexist in the data set and break away from the bulk of the data set, or single model is not adequate to relate the molecular descriptors to the aqueous solubility of drug-like molecules. These points will influence the quality of the model established. So, the discarding of these outliers will improve the predictive degree of accuracy for the model established. To obtain a high quality model, several outlier detection methods were employed to identify the outliers in the QSAR/QSPR model. Results for the MC method are shown in Figure 10. We can clearly see from Figure 10 that seven outliers were detected by the MC method. When these outliers were removed, the degree of accuracy for the model has a significant improvement (for the model including all samples, $R^2 = 0.9630$, $Q^2 = 0.8987$, RMSEF = 0.2222, RMSECV = 0.3680, while for the model after MC deleting outliers: $R^2 = 0.9745$, $Q^2 = 0.9302$, RMSEF = 0.1785, RMSECV = 0.2956). It is interesting to mention that sample 83 seems to be a very special point. From Figure 10, it is clearly an outlier in both ***X*** and ***y*** directions. However, the value of RMSECV decreased from 0.2956 to 0.2867 when it is included in the model. Moreover, MCD, MVT, and RHM also diagnosed the sample 83 as an ***X*** outlier. So, we should speculate

that sample 83 is possibly a good leverage point but far away from the main body of the data. The reason for this may be that a negligible variation in MC procedure, in which different samples are selected in the calibration set, may cause a large fluctuation for the predictive error of sample 83. The good leverage points and bad leverage points are somewhat inadequate to differentiate for the MC method. Similar to the boiling point data above, Figure 11 shows the RMSECV values of leave-one-out cross-validations in PLS model for seven detection methods. From Figure 11, we can see that big principal numbers are used to construct the PLS model due to the diversity of the drug-like molecules. For MCD, MVT, and RHM, the values of RMSEF and RMSECV are higher than the results for all samples (MCD: $R^2 = 0.9623$, $Q^2 = 0.8977$, RMSEF = 0.2199, RMSECV = 0.3623. MVT: $R^2 = 0.9616$, $Q^2 = 0.8737$, RMSEF = 0.2231, RMSECV = 0.4006. RHM: $R^2 = 0.9616$, $Q^2 = 0.8737$, RMSEF = 0.2231, RMSECV = 0.4006). This is because they only take account of the sample information from $\boldsymbol{X}$ direction and it is inadequate to generate a good model. M-estimator gives a similar result to the MC method and samples 29, 32, 47, 49 are diagnosed as $\boldsymbol{y}$ outliers again (M-estimator: $R^2 = 0.9745$, $Q^2 = 0.9313$, RMSEF = 0.1774, RMSECV = 0.2915). Different samples were diagnosed as outliers in RPPSV method and the results of the model give a little improvement (RPPSV: $R^2 = 0.9635$, $Q^2 = 0.9065$, RMSEF = 0.2171, RMSECV = 0.3475). Besides, RPLS method gives a similar result to the RPPSV method and six samples were detected as outliers (RPLS: $R^2 = 0.9631$, $Q^2 = 0.9012$, RMSEF = 0.2191, RMSECV = 0.3651). For this real QSAR/QSPR data set, the MC method obtains the best results among these methods mentioned.

It is worth noting that some robust methods, such as RPLS, M-estimator, etc., are data-structure dependent. From the results shown in Figures 9 and 11, one may see such a situation. For boiling point data, RPLS gives satisfactory results as MC method does (see Fig. 9), but this is not the case for solubility data of drug-like molecules (see Fig. 11). The same situation can be seen for M-estimator. For solubility data of drug-like molecules, M-estimator gives satisfactory results as MC method does (see Fig. 11), but not for boiling point data (see Fig. 9). However, according to our investigation in this study, the MC method can be considered as a robust and reliable detection method to a certain extent.

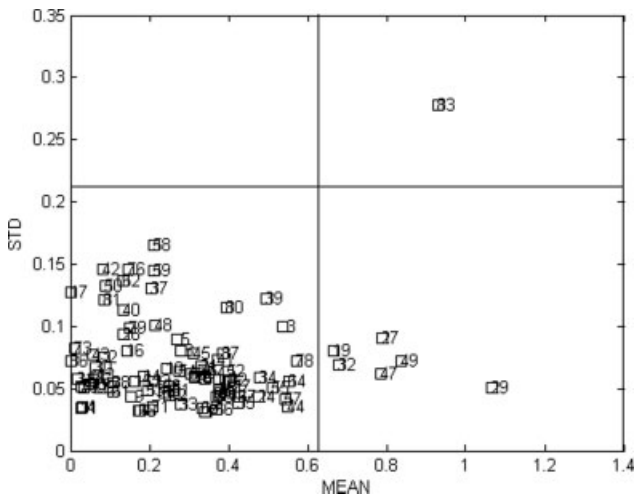

Figure 10. The result of variance of residuals versus mean of residuals on 88 drug compounds.

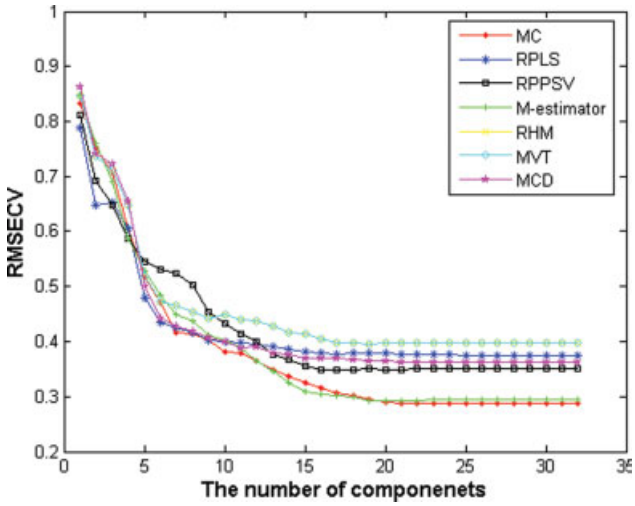

Figure 11. The RMSECV values of leave-one-out cross-validations versus the number of principal components in PLS model (MVT and MCD are overlapped).

Conclusions

Building a robust and reliable QSAR/QSPR model depends on a good outlier's detection method. In this article, we have put forward a new strategy of outlier detection for QSAR/QSPR. In a data set contaminated by outliers, the MC method can diagnose all types of outliers exhaustively and simultaneously. For a data set which exhibits extreme masking effects caused by multiple outliers, the MC method inherently provides a feasible way to detect different kinds of outliers by the use of establishment of many models. So, it should help to reduce the risk which the masking effect brings about and disclosing the reason why the masking effect yields to a certain extent. Moreover, it can cope with situations where there are more variables than objects with the help of PLS and PCR, etc.

Compared with the selected robust diagnostic methods, the MC method has high outlier detection ability, but it is somewhat inadequate to differentiate the good and bad leverage points. However, the deletion of those good leverage points does not influence the prediction ability of the calibration model. It can be concluded that the MC method performs well for outlier detection in regression models and that the calibration model after discarding of the outliers has a better prediction performance.

Acknowledgments

This work is financially supported by the National Nature Foundation Committee of People's Republic of China and the international cooperation project on traditional Chinese medicines of

ministry of science and technology of China. The studies meet with the approval of the university's review board.

References

1. Wessel, M. D.; Jurs, P. C. J Chem Inf Comput Sci 1995, 35, 841.
2. Carlton, T. S. J Chem Inf Comput Sci 1998, 38, 158.
3. Andrew, J. C.; Bernd, B.; Timothy, C. J Chem Inf Comput Sci 2001, 41, 457.
4. Karthikeyan, M. J Chem Inf Model 2005, 45, 581.
5. Christel, A. S. B.; Ulf, N.; Kristina, L.; Per, A. J Chem Inf Comput Sci 2003, 43, 1177.
6. Jorgensen, W. L.; Duffy, E. M. Bioorg Med Chem Lett 2000, 10, 1155.
7. Jarmo, H.; Jukka, R.; David, L. Eur J Med Chem 2000, 35, 1081.
8. Yan, A. X.; Gasteiger, J. J Chem Inf Comput Sci 2003, 43, 429.
9. Yan, A. X. J Comput Aided Mol Des 2004, 18, 75.
10. Vimon, T. J Pharm Biomed Anal 2005, 37, 411.
11. Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. J Chem Inf Model 2007, 47, 150.
12. Lei, D.-C.; Michael, W.; Manfred, K. Eur J Med Chem 2008, 43, 501.
13. Schultz, T. W.; Cronin, M. T. D. EnViron Toxicol Chem 2003, 22, 599.
14. Mazzatorta, P.; Vračko, M.; Jezierska, A.; Benfenati, E. J Chem Inf Comput Sci 2003, 43, 485.
15. Paolo, M.; Martin, S.; Elena, L. P.; Emilio, B. J Chem Inf Model 2005, 45, 1767.
16. Kirkpatrick, P.; Ellis, C. Nature 2004, 432, 823.
17. Kolossov, E.; Stanforth, R. SAR QSAR EnViron Res 2007, 18, 89.
18. Dmitry, A. K.; Lyndon, E. L.; Yvan, V. H.; Danny, C. J Chem Inf Model 2008, 48, 2081.
19. Dmitry, A. K.; Nigel, S.; Eric, D.; Yvan, V. H.; Danny, C. J Chem Inf Model 2008, 48, 370–383.
20. Rainer, G.; Torstem, S. J Comput Chem 2008, 29, 847.
21. Gerrit, S.; Ralf-Uwe, E.; Jingwen, C.; Bin, W.; Ralph, K. J Chem Inf Model 2008, 48, 2140.
22. Igor, V. T.; Iurii, S.; Anil, K. P.; Hao, Z.; Alexander, T.; Ester, P.; Tomas, O.; Roberto, T.; Denis, F.; Alexandre, V. J Chem Inf Model 2008, 48, 1733.
23. Romualdo, B.; Cecilia, B. J Chem Inf Model 2008, 48, 971.
24. Kutner, M. H.; Nachtsheim, C.; Neter, J. Applied Linear Regression Models; Academic Higher Education Press, 2005.
25. Liang, Y. Z.; Kvallheim, O. M. Chemom Intell Lab Syst 1996, 32, 1.
26. Mark, H.; Tunnell, D. Anal Chem 1985, 57, 1449.
27. Shak, N. K.; Gemperline, P. J. Anal Chem 1990, 62, 465.
28. Maesschalck, R. D.; Jouan-Rimbaud, D.; Massart, D. L. Chemom Intell Lab Syst 2000, 50, 1.
29. Gemperline, P. J.; Boyer, N. R. Anal Chem 1995, 67, 160.
30. Hoaglin, D. C.; Welsch, R. E. Am Stat 1978, 32, 17.
31. Rousseeuw, P. J.; Van Zomeran, B. C. J Am Stat Assoc 1990, 85, 633.
32. Rousseeuw, P. J.; Leroy, A. M. Robust Regression and Outlier Detection; Academic: New York, 1987.
33. Gnanadesikan, R.; Kettenring, J. R. Biometrics 1972, 28, 81.
34. Rousseeuw, P. J. In Mathematical Statistics and Applications; Grossmann, W.; Pflug, G.; Vincze, I., Eds.; Vol. B. Reidel: Dordrecht, 1985; pp. 283–297.
35. Rocke, D. M.; Woodruff, D. L. J Am Stat Assoc 1996, 91, 1047.
36. Rousseeuw, P. J. In Proceedings of the 4th Pannonian Symposium on Mathematics and Statistics; Grossman, W.; Pflug, G.; Vincze, I.; Wertz, W., Eds.; D. Reidel Publishing Co.: Dordrecht, the Netherlands, 1983; pp. 283–297.
37. Rousseeuw, P. J.; Katrien, V. D. Technometrics 1999, 41, 212.
38. Egan, W. J.; Morgan, S. L. Anal Chem 1998, 79, 2372.
39. Huber, P. J. Robust Statistics; Academic: Wiley: New York, 1981.
40. Rousseeuw, P. J. J Am Stat Assoc 1984, 79, 871.
41. Massart, D. L.; Kaufman, L.; Rousseeuw, P. J.; Leroy, A. Anal Chim Acta 1986, 187, 171.
42. Fung, W. J Am Stat Assoc 1993, 88, 515.
43. Yuzhu, H.; Smeyers-Berbeke, J.; Massart, D. L. Chemom Intell Lab Syst 1990, 9, 31.
44. Hettmansperger, T. P.; Sheather, S. J Am Stat 1992, 46, 79.
45. Walczak, B.; Massart, D. L. Chemom Intell LabSyst 1995, 27, 41.
46. Jdreskog, K.; Wold, H., Eds. Systems Under Indirect Observation: Causality, Structure, Prediction; North-Holland: Amsterdam, 1982.
47. Wakeling, I. N.; Macfie, H. J. H. J Chemom 1992, 6, 189.
48. Beaton, A. E.; Turkey, J. W. Technometrics 1974, 16, 147.
49. Randy, J. Pell Chemom Intell LabSyst 2000, 52, 87.
50. Zhang, M. H.; Xu, Q. S.; Massart, D. L. Chemom Intell Lab Syst 2003, 67, 175.
51. James, E. Gentle. Elements of Computational Statistics; Springer Science and Business Media, Inc: 2002.
52. Qing-Song, X.; Yi-Zeng, L. Chemom Intell Lab Syst 2001, 56, 1.
53. David, J. C. M. K. Information Theory, Inference, and Learning Algorithms; Cambridge university press: 2003.
54. Shao, J. J Am Stat Assoc 1993, 88, 486.
55. Xu, Q. S.; Liang, Y. Z.; Du, Y. P. J Chemom 2004, 18, 112.
56. Shao, J. J Am Stat Assoc 1996, 91, 655.
57. Zhang, P. Ann Statist 1993, 21, 299.
58. Brownlee, K. A. Statistical Theory and Methodology in Science and Engineering; Academic: New York, pp. 491–500.
59. Becker, R. A.; Chambers, J. M.; Wilks, A. R. The New S Language; Wadsworth & Brooks/Cole: 1988.
60. Dodge, Y. The Guinea Pig of Multiple Regressions. In: Robust Statistics, Data Analysis, and Computer Intensive Methods; In Honor of Peter Huber's 60th Birthday, 1996, Lecture Notes in Statistics 109, Springer-Verlag, New York.
61. Hawkins, D. M.; Bradu, D.; Kass, G. V. Technometrics 1984, 26, 197.
62. Antonio, L.; Robert, C. G.; Jonathan, M. G. J Chem Inf Comput Sci 2008, 48, 1289.
63. Satyanarayana, K.; Kakati, M. C. Fire Mater 1991, 15, 97.
64. Fisher, C. H. Chem Eng 1989, 96, 157.