The Model Adaptive Space Shrinkage (MASS) Approach: A New Method for Simultaneous Variable Selection and Outlier Detection Based on Model Population Analysis

Ming Wen[a,b], Bai-Chuan Deng[c], Dong-Sheng Cao[a*], Yong-Huan Yun[b], Rui-Han Yang[b], Hong-Mei Lu[b†], Yi-Zeng Liang [b]

[a] *School of Pharmaceutical Sciences, Central South University, Changsha 410013, PR China*

[b] *College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China*

[c] *College of Animal Science, South China Agricultural University, Guangzhou 510642, P.R. China*

Abstract

Variable selection and outlier detection are important processes in chemical modeling. Usually, they affect each other. Their performing orders also strongly affect the modeling result. Currently, many studies perform them separately and in different orders. In this study, we discussed the interaction between outliers and variables, and compared the modeling procedures performed in different variable selection and outlier detection orders. Because the order of outlier detection and variable selection can affect the interpretation of the model, it is hard to decide which order is better when the predictability (prediction error) of different orders is relatively close. To handle this problem, a simultaneous variable selection and outlier detection approach

[*] Corresponding author. *E-mail address:* oriental-cds@163.com (Dong-Sheng Cao)

[†] Corresponding author. *E-mail address:* hongmeilu@csu.edu.cn (Hong-Mei Lu)

called Model Adaptive Space Shrinkage (MASS) was developed. This proposed approach is based on model population analysis (MPA). Through weighted binary matrix sampling (WBMS) from model space, a large number of partial least square (PLS) regression models were built, and the elite part of models were selected for statistically reassigning the weight of each variable and sample. Then, the whole process repeated until the weights of variables and samples were converged. Finally, MASS adaptively found a high performance model which consisted of the optimized variable subset and sample subset. The combination of these two subsets could be considered as the cleaned dataset used for the chemical modeling. In the proposed approach, the problem of the order of variable selection and outlier detection is avoided. One near infrared spectroscopy (NIR) dataset and one quantitative structure-activity relationship (QSAR) dataset were used to test this approach. The result demonstrated that MASS is a useful method in data cleaning before building a predictive model.

Key words: outlier detection, variable selection, model population analysis, shrinkage, model space

1. Introduction

With the development of modern analytical instruments, numerous data which contain a large number of variables and samples can be obtained through high-throughput experimental method. Multivariate regression techniques such as multivariate linear regression (MLR) [1], partial least square regression (PLS) [2], support vector regression (SVR) [3] and random forest (RF) [4] are useful tools to analyze those data and have been applied in different fields. However, the applications of a built model are seriously affected by the quality of the model. To build a robust and reliable model, variable selection and outlier detection method have been wildly used to improve the performance of regression models.

In general, variable selection methods can be divided into three categories. One is classical methods such as forward selection method [5] and backward elimination method [6], without considering the combination effect of variables [7]. One is artificial intelligence-based method like genetic algorithm (GA) method [8], artificial neural network (ANN) method [9] and particle swarm optimization (PSO) method [10] which have been applied to search the optimal subset of variables. One is statistical method such as uninformative variable elimination (UVE) [11], variable iterative space shrinkage approach (VISSA) [8] and iteratively retaining informative variables (IRIV) [12]. They select variables by statistically evaluating some values of a model.

Detecting outlier is troublesome especially when several outliers coexist. Diagnostics and robust regression are two methods to deal with outliers [13]. In the diagnostic method, outliers are identified first, and the rest of samples are used to build model. Monte-Carlo (MC) method is a typical diagnostic method. It uses Monte-Carlo sampling method to build a large number of models. Each sample is predicted by all models. The standard deviation and mean value of predictive error are calculated. The sample with large standard deviation or large mean value could be considered as outliers. In the robust regression method, a regression model is constructed to fit the majority of the data. Outliers are detected by examining the residuals which are predicted by the built model. Some representative methods include least median of squares (LMS) [14], robust principal component regression (RPCR) [15] and robust partial least squares (RPLS) [16] and so on.

Before building a model, variable selection and outlier detection must be carefully considered, especially their interactions (i.e., their performing orders). It is worth to note that outlier detection and variable selection can influence each other [17]. Different results may be obtained by performing these two tasks in the opposite order. Thus, the order of variable selection and outlier detection will intensively influence the application of a model. It is therefore necessary to consider variable selection and outlier detection simultaneously. Jennifer Hoeting proposes a method for simultaneous variable selection and outlier identification in linear regression, which is an early research on this aspect. The approach is based on posterior model probabilities. A Markov chain Monte Carlo approach is used to approximate the Bayesian model average over the space of all possible variables and outliers under consideration. For more detail information see reference [18]. Later some GA-based methods [17, 19, 20] are proposed for this task and have been applied in different fields. J. Tolvi et al. uses an ordinary genetic algorithm for outlier detection and variable selection in linear regression [17]. Patrick Wiegand combines a robust outlier determination method with a genetic algorithm for variable selection [19]. Rachel Cavill et al. develops a genetic algorithm approach which simultaneously selects sub-sets of samples and spectral regions (variables) in metabonomics data. Their results indicate that simultaneous sample and variable selection method improved model performance by over 9% compared with those separated method [20]. Rajiv S. Menjoge gives a diagnostic method for simultaneous feature selection and outlier identification in linear regression [21]. The method performs by adding a dummy variable set to the data matrix and running backward selection on the augmented matrix. The sequences of feature-outlier combinations are identified. Another method proposed by Sung-Soo Kim et al [22] consists of two procedures, first identifying the potential outliers (mean-shift outlier model), then exhaustively searching the possible subset regressions for the mean-shift outlier model. A recent method is Monte-Carlo Outlier and Variable Screening approach (MCOVS) [23]. MCOVS builds a series of sub-regression models and simultaneously evaluates the importance of variables and location of outliers statistically.

Model Population Analysis (MPA) [24] , proposed by Li et al., is a general framework for designing new types of chemometrics and bioinformatics algorithms[24]. In MPA, firstly, randomly produce N sub-training datasets using sampling methods from the original dataset. Secondly, establish a sub-regression model on each sub-training dataset. Finally, statistically analyze interesting outputs of all established N sub-regression models. Many methods such as MCUVE, VISSA, and IRVR are developed based on MPA.

Here, we proposed a strategy based on MPA called Model Adaptive Space Shrinkage (MASS). It was applied to select variables and remove outliers simultaneously. MASS aims to find a high performance model based on a clean dataset in the model space through a weighted iteration strategy. The variable and sample subsets are simultaneously obtained. In addition, MASS considers the outlier masking effect and variable combination effect through its random sample procedures. In this study, MASS coupled with PLS was tested on different data. Comparison with other existing popular methods or method combination showed that MASS is a useful method to select variables and outliers simultaneously. It should be noted that MASS can also be coupled to other modeling methods such as artificial neural network (ANN), support vector regression (SVR).

2. Theory and method

2.1 Definition of Model Space

After obtaining a data with *N* samples and *P* variables, a model space is defined as a set of models which are constructed by all possible combinations of samples and variables. Fig. 1 is the sketch of model space. The combination of a variable subset and a sample subset forms a sub-training dataset, and the sub-training dataset is used to building a regression model. The built model is a member of model space. In Fig. 1, #1 and #2 are two models (members) in the model space. The number of all the possible combinations for variables is 2^P-1 (variable space) and for samples is 2^N-1(sample space). The model space is the combination of variable space and sample space. It has $(2^P-1) \times (2^N-1)$ models (combinations).

(Insert Figure 1)

2.2 The interaction between variables and outliers

Outliers depend on the variables used for characterization [23]. A sample can be seen as an outlier when its location represented by variables is far away from the bulk of samples. As is shown in Fig. 2a, all samples can be well fitted only using one variable x1. But in Fig. 2b, when added a variable x2, sample 1 turns into an outlier since its location is far away from other samples. In addition, in the dataset with outliers, more variables are needed to reduce the influence of outliers. As is shown in Fig. 2a, with the outlier (sample 1) in dataset, variable x2 is needed to build a model (the red dotted line) to reduce the impact of this outlier. This explicitly indicates that, on the one hand, different variables can lead to different outliers in the sample set; on the other hand, different samples need different variables to build the best model. Building a high performance model not only needs to consider the effects from variable selection and outlier detection separately but also needs to consider their interactions.

(Insert Figure 2)

2.3 BMS and WBMS

Binary matrix sampling (BMS) is a new strategy for random sampling which is proposed by Yun and Deng et.al [8, 12, 25, 26]. It can ensure that all the variables have the same overall frequency of sampling in the sub-regression models. A final sampling matrix with a special variable frequency is consisted of a number of sub-binary matrices which have different variable frequency. Weighted binary matrix sampling method (WBMS) [8, 25] is a modified BMS which ensure important variables and samples to have high selected probability in each iteration. In the BMS strategy sample sampling ratio and variable sampling ratio should be manually set according to the real problem.

2.4 Model Adaptive Space Shrinkage (MASS)

By combining MPA and WBMS, a novel method called Model Adaptive Space Shrinkage (MASS) was proposed to select variables and detect outliers simultaneously. The flowchart of MASS is depicted in Fig. 3.

(Insert Figure 3)

Firstly, through BMS, a number of sub-training datasets was sampled from the original dataset. That is to say, the samples with specific sampling ratio (e.g., 0.95) and the variables with specific sampling ratio (e.g., 0.5) were randomly selected to construct one sub-training dataset from the original dataset. Initially (i.e., in the first iteration), the frequency of each variable or each sample appearing in these models is somewhat equal according to their sampling ratio. For example, for each sub-regression model, the sub-training dataset is consisted of 95% samples and 50% variables. Thus, these sub-training datasets were used to build sub-regression models which are evenly distributed in model space. Then, these models were sorted by the coefficient of determination of cross-validation Q_{CV}^2 (eq. 1) [27, 28].

$$Q^2_{CV} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

where n is the number of samples in the model, $\hat{y}_i$ is the prediction property value of the ith sample, and $\bar{y}_i$ is the mean property value of sub-training dataset. The models with large Q^2_{CV} were extracted. Then, the frequency of each variable and each sample in the selected models were counted. The weight (ω) of variable i and sample j were obtained by eq.2 and eq.3, respectively.

$$\omega_i = \frac{\rho_i}{\mathrm{K_{best}}} \quad (2)$$

$$\omega_j = \frac{\rho_j}{\mathrm{K}_{best}} \quad (3)$$

Where ρ_i and ρ_j are the frequency of variable i and sample j in the selected sub-regression models respectively, K_{best} is the number of extracted models, and ω is a number between 0 and 1, which represents the ratio of a sub-regression model that contain variable i or sample j in the next iteration. In other words, large ω_i and ω_j indicate that the variable i and sample j are more important and have more chance to appear in sub-regression models. So far, the first iteration finished.

In the next step, WBMS was used to build a number of new sub-regression models by using the weight of variables (ω_i) and sample (ω_j) obtained from last iteration. Unlike the models evenly distributed in model space in the first iteration, the models gradually focus on the high performance model in the next iteration, and the model space is also gradually shrinkage. The procedure for obtaining new weights of variables and samples was repeated until the weights of all variables and samples were constant (either 1 or 0). Thus the best model was obtained; the variables and samples that constructed the best model are simultaneously selected.

MASS aims to find a high performance model in the model space through a continuous model space shrink procedure. In the beginning, all variables and samples have the same weight. In each WBMS step, the sampling method focuses on the variables and samples with larger weight until the weight is up to 1. Thus, the extent of the best model space shrinks continuously until we find the best model. The MATLAB codes for implementing MASS are freely available at the Supporting Information.

3. Datasets

To illustrate the performance of our proposed method, two online available datasets were used to evaluate the MASS approach.

3.1 Wheat kernel dataset

This dataset represents 43 different varieties or variety mixtures from two different locations, and consists of 415 samples and 100 variables. Each sample was analyzed at the range of 850-1050 nm, and 100 wavelengths were recorded as variables. This data is freely available at http://www.models.life.ku.dk/wheat_kernels.

3.2 ACE dataset

This is a commonly used real QSAR dataset for testing the proposed approach. This dataset consists of 114 angiotensin converting enzyme (ACE) inhibitors originally taken from the work of Depriest et al and 56 descriptors. Activities are spread over a wide range, with each inhibitor pIC50 values ranging from 2.1 to 9.9 [29].

4. Results and discussion

4.1 The comparison of Wheat kernel and ACE dataset on different methods.

In this study, for comparison of different approaches, Monte-Carlo sampling (MCS) method and variable iterative space shrinkage approach (VISSA) were used to detect the outliers and select compact subset of variables, respectively. MCS method[13] is an outlier detection method based on MPA. It inherently provides a feasible way to detect different kinds of outliers by establishment of many cross-predictive models. MCS has been demonstrated as a practical outlier detection method by a series of works [30-32]. VISSA, proposed by our group, is a new variable selection method based on MPA. Unlike most of the existing optimization approaches for variable selection, VISSA statistically evaluates the performance of each model and makes full use of the information obtained in each model to iteratively find the best subset of variable. Its acceptability has been proved by comparing with other popular methods [8]. Furthermore, the combination of MCS and VISSA were employed to improve the prediction power of the model. Two strategies were considered: removing outliers with MCS followed by variable selection with VISSA (MCS + VISSA) and selecting variables with VISSA followed by outlier detection with MCS (VISSA + MCS). Finally, MASS was compared with PLS, VISSA, MCS, VISSA+MCS, MCS+VISSA.

All these methods were executed 20 times and used the same parameters to build models: the optimal number of PLS component was obtained by five-fold cross validation and was used for building models. The sampling number used in VISSA and MASS was 2000, and the ratio of selected best sub-regression models was 0.05 (that is 100 models). The initial weight of variables in VISSA is 0.5. In MASS, the initial weight of variables is 0.5 and the initial weight of samples is 0.95. In addition, all data were pretreated by mean-center method before modeling. The coefficient of determination of calibration set (R^2) and coefficient of determination of cross validation (Q_{CV}^2) were used to assess model performance. The number of selected samples and variables was recorded as Sam and Var. The number of optimal latent variables (optPC) was also recorded.

The results of wheat kernel dataset and ACE dataset performed by PLS, VISSA, MCS, VISSA +MCS, MCS+VISSA and MASS were listed in Table 1 and Table 2, respectively. As shown in Table 1 and 2, PLS has the worst prediction performance among all these approaches. It gives R^2 value of 0.880 and Q_{CV}^2 value of 0.869 for wheat kernel dataset and gives R^2 value of 0.745 and Q_{CV}^2 value of 0.623 for ACE dataset. MCS (with R^2 value of 0.899 and Q_{CV}^2 value of 0.889 for wheat kernel dataset and R^2 value of 0.819 and Q_{CV}^2 value of 0.729 for ACE dataset) and VISSA (with R^2 value of 0.894 and Q_{CV}^2 value of 0.886 and R^2 value of 0.775 and Q_{CV}^2 value of 0.694 for ACE dataset) yield better prediction accuracy than original PLS model, which indicates that PLS is strongly sensitive to outliers and uninformative variables. Furthermore, the two combination approaches, VISSA+MCS and MCS+VISSA, obtained similar prediction accuracy. The results are better than those obtained from single MCS and single VISSA approach. This indicates that variable selection and outlier detection method are two interactively promoted methods and are indispensable in data modeling process. As seen in Table 1 and 2, MASS achieves the best prediction accuracy. It gives R^2 value of 0.921 and Q_{CV}^2 value of 0.913 for wheat kernel dataset and gives R^2 value of 0.865 and Q_{CV}^2 value of 0.823 for ACE dataset. Compared with PLS which building the model with all the samples and variables, the R^2 and Q_{CV}^2 of MASS increased 4.51% and 5.18% for wheat kernel dataset and 16.1% and 32.1% for ACE dataset (P value < 0.05 MASS versus PLS), respectively. Compared with other methods, the R^2 and Q_{CV}^2 of MASS for both datasets are also increased considerably (P value < 1×0.05 MASS versus MCS, P value < 1×10e-3 MASS versus MCS+VISSA, P value < 1×10e-3 MASS versus MCS+VISSA, P value < 1×10e-5 MASS versus VISSA+MCS).

(Insert Table 1)

(Insert Table 2)

The accuracies of different orders of variable selection and outlier detection were similar, but the outliers detected and variables selected by different orders varied dramatically. Fig. 4 is the outlier detection plot of wheat kernel dataset. It was detected by MCS method (MCS+VISSA), two blue dash lines separate the picture into 4 areas, the samples located in the lower left are normal samples, the samples located in other areas are outliers [13]. The locations of the dash line are determined by

3 times of the standard deviation of the mean error Mean and mean error STD [33]. In addition, MCS is a very robust outlier detection method and the outlier detection plots are near the same in 20 times execution. Fig. 5 is the frequency of a wheat kernel sample located in the outlier area in the VISSA+MCS order in 20 times. As is shown in Fig. 4, the samples enclosed by red ellipse (sample number 38, 58, 157, 158 and 404) are located in the lower left area. These samples are normal samples. However, as is shown in Fig. 5, they turned into outliers after variables selection. As is shown in Fig. 4, the samples enclosed by green ellipses (sample number 18, 25, 104 and 408) are located in the lower right. These samples are outliers. However, as is shown in Fig. 5, they become normal samples after variables selection. Similarly, for ACE dataset, Fig. 6 is the outlier detection plot detected by MCS method (MCS+VISSA). Fig. 7 is the frequency of outlier detected in the VISSA+MCS order in 20 times. Sample number 18, 48, 63, 64 73 and 81 (enclosed by red ellipse) are normal samples in MCS+VISSA order whereas they became outliers in VISSA+MCS order. Sample number 12, 13, 15, 22, 26 and 52 are outliers in MCS+VISSA order whereas they turned into normal samples in VISSA+MCS order. This indicates that different variables can lead to different outliers in sample set. These two different process orders are acceptable if just considering the results of built models. If considering the interpretation of built model, these is a puzzle to decide the final variable selection and outlier detection order. Thus, when dealing with datasets with redundant variables and outliers, it is important to select variables and detect outliers simultaneously.

(Insert Figure 4)

(Insert Figure 5)

(Insert Figure 6)

(Insert Figure 7)

(Insert Figure 8)

(Insert Figure 9)

4.2 The visualization of the interaction between variables and outliers

Fig. 8 and Fig. 9 can fully explain the interaction between variables and samples in MASS iteration process. Fig. 8 and Fig. 9 are the plots of sample and variable weight against MASS iteration. The weight reveals the trend of sample and variable in the iteration. The weight is the probability of a sample or a variable to be selected to build a model. In other words, large weight indicates that the variable and sample are more important and have more chance to appear in sub-regression models. As shown in Fig. 8 and Fig. 9, each line represents the weight variation of a sample or a variable. There are three different weight variation types: 1), the lines which go down all the time till the weights reach to 0. This kind of variation indicates that these variables (or samples) are uninformative variables (or outliers) and there is no strong interaction between these variables and outliers, these outliers and variables can be easily detected and removed; 2), the lines which go up all the time till the weights reach to 1. These variables (or samples) are informative variable (or normal samples) and should be selected to build model; 3), the lines which go up at first, then go down; or go down at first, then go up till the weights reach to 1 or 0. This kind of variation indicates there is strong interaction between these variables and outliers. When the line goes down, it means that the variable (or sample) may be an uninformative variable (or outliers) with current samples (or variables). After several iterations, some outliers (variables) are removed, then the line goes up and the variable (or sample) became important with current samples (or variables).

With regard to wheat kernel dataset and the ACE dataset, MASS was converged after 31 and 30 iterations, respectively. The whole MASS iteration process can be separate into 3 parts: 1), the early iteration period (1-17 iteration for wheat kernel and 1-13 iterations for ACE dataset). In this period, the weights of most samples gradually reached to 1 except some weights of samples decreased step by step. At the same time, except the weights of some variables rose up to 1 and the weights of few variables went down to 0, the weights of most variables fluctuated dramatically up and down. However, the weights of samples and variables which went down to 0 at this period are without fluctuation or with small fluctuation. As is shown in Fig. 8(a), for wheat kernel dataset, sample number 199, 3, 25, 363, 158 and 71 were detected as outliers in this period. As is shown in Fig. 4 and Fig. 5, sample 199, 3, 363, 158 and 71 were also detected as outliers in both MCS+VISSA order and MCS+VISSA order. In Fig. 9(a), for ACE dataset, sample number 19, 53, 48, 91, 34 and 108 were detected as outliers in this period. As is shown in Fig. 6 and Figure 7, sample 19, 53, 91, 34 and 108 were also detected as outliers in both MCS+VISSA order and MCS+VISSA order. These outliers can be detected by both separate and simultaneous methods. This means that these samples are essentially far away from the main part of sample. These outliers are not affected by variables. One can easy detect these outliers without considering the impact of variables. 2), the middle iteration period (18-25 iteration for wheat kernel dataset and 14-25 iteration for ACE dataset). In this period, the variables and samples went ahead along with the tendency in the early period and most variables and samples arrived 0 or 1. The weights of samples and variables which went down to 0 at this period varied dramatically. As is shown in Fig. 9(a), 9 samples were detected as outlier and located in this period. Among them, none of them were detected in MCS (Fig. 6), while 4 of them detected in VISSA+MCS (Fig. 7). This means that these samples and outliers strongly affect each other in this period. When more variables are selected, some samples with high weights may not proper for current variables and result in weights decreasing, and vice versa; 3), in the ending iteration period (26-31 iteration for wheat kernel dataset and 26-30 for ACE dataset), the weights of the entire sample kept constant and all the outliers were detected and removed. The rest is to optimize the variable subset which could wonderfully support current selected samples. Finally, MASS converged to and found out the best model in the model space through this model space shrink iteration procedure.

4.3 Comparison with other methods

The comparison with the standard procedures for outlier detection and variable selection such as Williams plot of leverage values versus abnormal residuals (WP)[34], Variable importance in projection (VIP)[35] and genetic algorithm (GA)[36] were also performed, which were listed in Table 3. One can clearly see that the variable selection method based on VIP on these two dataset obtained the poor prediction statistics, and its results were similar [35]to those from original PLS models. VIP could effectively eliminate some uninformative or noisy variables and therefore obtained a relatively easy-to-interpret model. Compared to original PLS method without any variable selection, VIP only selected 59 variables for Wheat kernel dataset and 31 variables for ACE dataset, respectively. Compared to original PLS and VIP method, GA yields the better prediction results, and obtains the similar prediction performance to VISSA and our proposed MASS. The final variable number used in the regression model is also sharply reduced for these three methods. Compared to the commonly used outlier detection method WP, from the Williams plot of leverage versus abnormal residuals, there are 22 outliers (3, 25, 52, 71, 83, 104, 114, 199, 18, 33, 158, 208, 221, 231, 341, 363, 371, 397, 406, 408, 409, 411) in wheat kernel dataset and 4 outliers (8, 19, 53, 91) in ACE dataset. All the outliers detected in the first period of MASS (Fig. 8 and Fig. 9) in wheat kernel dataset and ACE dataset were also detected by WP, whereas the remaining outliers detected by these methods vary dramatically. In ACE dataset, besides these four outliers no more outliers where detected by WP but another 11 outliers were detected by MASS. In wheat kernel dataset, the outliers detected by WP embodied most outliers in MASS (12 out of 16) and other 10 outliers. After removing outliers, the R^2 and Q_{CV}^2 values WP were listed in Table 5. From Table 5, compared with WP, MASS is also increased considerably for both datasets.

4.4 The effect of MASS parameters

In our proposed MASS method, three important parameters related to Monte-Carlo sampling need to be set. The number of Monte-Carlo experiments seems an important parameter which affects the quality of the distribution. Theoretically, the fewer samples are selected randomly from the calibration samples, the more repeats are needed. Whereas, it has been proven that the number of number of Monte-Carlo experiments equal to n^2 (n is the number of the total samples) is generally enough to make Monte-Carlo strategy better performance. Larger sampling number tends to generate more accurate and stable results. However, the accuracy improvement is very small, indicating that MASS is insensitive to this parameter. To save the computing source, in practice, the number of Monte-Carlo sampling is manually set to 2000. By means of Monte-Carlo method, the computational complexity could be reduced substantially. Similar to VISSA, the initial weights of variables (i.e., the variable sampling ratio) were set to 0.5 in MASS, that is, each variable has 50% probability to be selected in one sub-model in the beginning. Under the circumstance without any prior information, it is a relatively natural choice to set the initial weights of variables to 0.5. Another important parameter is the initial weight of samples (i.e., the sample sampling ratio). To evaluate the influence of the initial weight of samples on MASS, different experiments were carried out on the wheat kernels dataset. These results are shown in Fig.10 and listed in Table 5. From Table 5, one can see that with the decreasing of the initial samples weights, the number of iteration tends to slightly increase and the number of outliers and variables changed a little (when the initial weight is very small such as 0.5, the number of iterations, outliers or variables may vary relatively large). The accuracy of MASS decreased a little when the initial weights of samples decreased. Moreover, from Fig. 10, with different initial weight, the variations of sample weight are similar. This indicates that MASS is insensitive to the initial weights of samples to some extent. Given a dataset without any prior information, we could assume that the main parts of samples are normal and only small parts (e.g., 5%) of samples are outliers. Considering that, in my opinion, it is a relatively suitable choice to set the sample initial weights to 0.95.

4.5 The effect of variable and sample combinations

Given a dataset, the number of combinations of variables and samples is extremely high. Assume that we have a data set with n = 100 and p = 60, all combinations from model space will be (2100-1)×(260-1), and this will be an extremely high number for computer simulation. Therefore, it is almost impossible for current computer simulation to enumerate all model combinations. Alternatively, were randomly chosen some combinations from all possible combinations by Monte-Carlo sampling strategy and then use the best part of generated models to represent the distribution of important variables and samples. In general, Monte-Carlo approach can be used to generate such a distribution of some statistic of interest by repeatedly calculating that statistic randomly selected portions of the data because of its good asymptotic properties. Through this sampling procedure, though the model combination is usually high, only part of combinations was used which can dramatically reduce the modeling time. Take a wheat kernel dataset for an example, we only used about sixty thousands (2000×31 (31 iterations)) combinations to shrink to a relatively good solution. We calculated the elapsed time of MASS on this dataset, which is listed in Table 4. Although computation time of MASS is slightly higher than those from other variable selection or outlier detection programs including MCS and VISSA, it is worthy to waste somewhat more time to obtain a clean dataset and a higher performance model. Simultaneous variables selection and outlier detection is usually a hard task. We applied a computing-intensive method and therefor a little more time was required. After MASS were performed, no additional codes between variables selection and outlier detection was needed. From an overall perspective, MASS takes less time in the model building process.

Conclusion

In this study, we proposed MASS to simultaneously detect outliers and select variables before building a final prediction model. The proposed method is based on MPA which iteratively and smoothly shrinks the model space to obtain the best model. MASS is a mild stepwise optimization method. The model space shrinks smoothly which reduce the risk of eliminating informative variables and normal samples. The weights variation of variables and outliers illustrate the cross interaction between variables and outliers: if the weights of variables and samples go down to 0 in the first period, these variables and outliers do not interact with each other and they can be easily identified. If the weights of variable and samples go down to 0 in the middle period, these samples and outliers strongly affect each other. In the last period, the weights of samples were constants, and the rest is to optimize the variable subset which can wonderfully support current selected samples. The performance of the new algorithm was compared with several other outlier detection and variable selection methods and methods combination. The results clearly indicate that: when outlier detection and variable selection performed separately, there is a great opportunity to obtain a wrong model that fails to reflect the true relationship between variables and outliers. To avoid this failure, it is recommended to do these tasks simultaneously. The results demonstrated that MASS is a useful method in data cleaning before building a predictive model.

Acknowledgement

The authors gratefully thank the National Natural Science Foundation of China for support of the projects (Grant No. 81402853, 21175157 and 21375151) and also supported by the Fundamental Research Funds for the Central University of Central South University (Grants No. 2015zzts163). The studies meet with the approval of the university's review board.

References

1. G. C. Reinsel and R. P. Velu, in *Multivariate Reduced-Rank Regression*, Springer, 1998, pp. 1-14.
2. W. W. Chin, *Modern methods for business research*, 1998, **295**, 295-336.
3. A. J. Smola and B. Schölkopf, *Statistics and computing*, 2004, **14**, 199-222.
4. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *Journal of chemical information and computer sciences*, 2003, **43**, 1947-1958.
5. F. G. Blanchet, P. Legendre and D. Borcard, *Ecology*, 2008, **89**, 2623-2632.
6. J. M. Sutter and J. H. Kalivas, *Microchemical journal*, 1993, **47**, 60-66.
7. M. Shahlaei, *Chemical reviews*, 2013, **113**, 8093-8103.
8. B.-c. Deng, Y.-h. Yun, Y.-z. Liang and L.-z. Yi, *Analyst*, 2014, **139**, 4836-4845.
9. M. Gevrey, I. Dimopoulos and S. Lek, *Ecological Modelling*, 2003, **160**, 249-264.
10. R. C. Eberhart and Y. Shi, 2001.
11. W. Cai, Y. Li and X. Shao, *Chemometrics and intelligent laboratory systems*, 2008, **90**, 188-194.
12. Y.-H. Yun, W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu and Q.-S. Xu, *Analytica chimica acta*, 2014, **807**, 36-43.
13. D. S. Cao, Y. Z. Liang, Q. S. Xu, H. D. Li and X. Chen, *Journal of computational chemistry*, 2010, **31**, 592-602.
14. P. J. Rousseeuw, *Journal of the American statistical association*, 1984, **79**, 871-880.
15. P. Filzmoser, *Computer data analysis and modeling. Robust and computer intensive methods. Belarusian State University, Minsk*, 2001, 132-137.
16. J. A. Gil and R. Romera, *Journal of chemometrics*, 1998, **12**, 365-378.
17. J. Tolvi, *Soft Computing*, 2004, **8**, 527-533.
18. J. Hoeting, A. E. Raftery and D. Madigan, *Computational Statistics & Data Analysis*, 1996, **22**, 251-270.
19. P. Wiegand, R. Pell and E. Comas, *Chemometrics and intelligent laboratory systems*, 2009, **98**, 108-114.
20. R. Cavill, H. C. Keun, E. Holmes, J. C. Lindon, J. K. Nicholson and T. M. Ebbels, *Bioinformatics*, 2009, **25**, 112-118.
21. R. S. Menjoge and R. E. Welsch, *Computational Statistics & Data Analysis*, 2010, **54**, 3181-3193.
22. S.-S. Kim, S. H. Park and W. Krzanowski, *Journal of Applied Statistics*, 2008, **35**, 283-291.
23. D. Cao, Y. Liang, Q. Xu, Y. Yun and H. Li, *Journal of computer-aided molecular design*, 2011, **25**, 67-80.
24. H.-D. Li, Y.-Z. Liang, D.-S. Cao and Q.-S. Xu, *TrAC Trends in Analytical Chemistry*, 2012, **38**,

154-162.

25. B.-C. Deng, Y.-H. Yun, P. Ma, C.-C. Lin, D.-B. Ren and Y.-Z. Liang, *Analyst*, 2015, **140**, 1876-1885.
26. Y.-H. Yun, W.-T. Wang, B.-C. Deng, G.-B. Lai, X.-b. Liu, D.-B. Ren, Y.-Z. Liang, W. Fan and Q.-S. Xu, *Analytica chimica acta*, 2015, **862**, 14-23.
27. R. Kohavi, 1995.
28. N. J. Nagelkerke, *Biometrika*, 1991, **78**, 691-692.
29. J. J. Sutherland, L. A. O'Brien and D. F. Weaver, *Journal of Medicinal Chemistry*, 2004, **47**, 5541-5554.
30. J. B. Wang, D. S. Cao, M. F. Zhu, Y. H. Yun, N. Xiao and Y. Z. Liang, *Journal of Chemometrics*, 2015.
31. D. S. Cao, Q. S. Xu, Y. Z. Liang, X. Chen and H. D. Li, *Journal of Chemometrics*, 2010, **24**, 584-595.
32. E. Pourbasheer, S. Shokouhi Tabar, V. Masand, R. Aalizadeh and M. Ganjali, *SAR and QSAR in Environmental Research*, 2015, **26**, 461-477.
33. N. Xiao, D.-S. Cao and Q.-S. Xu, 2015.
34. B. S. Everitt, *American Mathematical Monthly*, 1998, 387-388.
35. L. Eriksson, E. Johansson, H. Antti and E. Holmes, *Multi- and Megavariate Data Analysis*, 2005.
36. R. Leardi and A. L. González, *Chemometrics & Intelligent Laboratory Systems*, 1998, **41**, 195-207.

Tables:

Table 1 The results of wheat kernel dataset performed on different methods.

Method	Sam	Var	R^2	Q^2_{CV}	optPC	Iteration
PLS	415	100	0.880	0.868±0.005	10	
VISSA	415	32±2	0.894±0.001	0.886±0.003	9	13±2
MCS	402±0	100	0.899±0	0.889±0.003	10	
VISSA + MCS	404±2	32±2	0.909±0.002	0.902±0.002	9	13±2
MCS+ VISSA	402±0	31±3	0.911±0.001	0.904±0.003	9	13±2
MASS	398±2	31±5	0.921±0.003	0.913±0.005	10	31±4

Table 2 The results of ACE dataset performed on different methods.

Method	Sam	Var	R^2	Q^2_{CV}	optPC	Iteration
PLS	114	56	0.745±0	0.623±0.038	10	-
VISSA	114	23±10	0.775±0.023	0694±0.036	10	12±3
MCS	102	56	0.819	0.729±0.033	10	-
VISSA + MCS	106±3	23±10	0.837±0.017	0.772±0.044	10	12±3
MCS+ VISSA	102	30±13	0.841±0.017	0.775±0.031	10	12±3
MASS	102±3	26±10	0.865±0.021	0.823±0.027	10	24±5

Table 3 The results of Wheat kernel and ACE dataset performed on VIP and GA.

Dataset	Methods	Sam	*Var*	R^2	Q^2_{CV}
	WP	393	100	0.913	0.904
Wheat kernel dataset	**VIP**	415	59	0.877	0.861
	GA	415	34	0.891	0.881
	WP	110	56	0.774	0.691
ACE dataset	**VIP**	114	31	0.728	0.624
	GA	114	13	0.772	0.703

Table 4 The elapsed time of MCS, VISSA and MASS.

	Wheat kernel		ACE	
Methods	**Time (second)**	Q^2_{CV}	**Time (second)**	Q^2_{CV}
MCS	30	0.889	12	0.729
VISSA	810	0.889	688	0694
MASS	1260	0.917	981	0.823

Table 5 The performance of MASS with different sample initial weight.

Initial weight	0.95	0.9	0.8	0.7	0.6	0.5
Number of iterations	34	40	37	49	39	54
Number of outliers	17	13	14	18	13	29
Number of variables	34	28	36	36	13	17
$\mathbf{Q^2_{CV}}$	0.9173	0.9109	0.9111	0.9155	0.8783	0.9168

Figure captions:

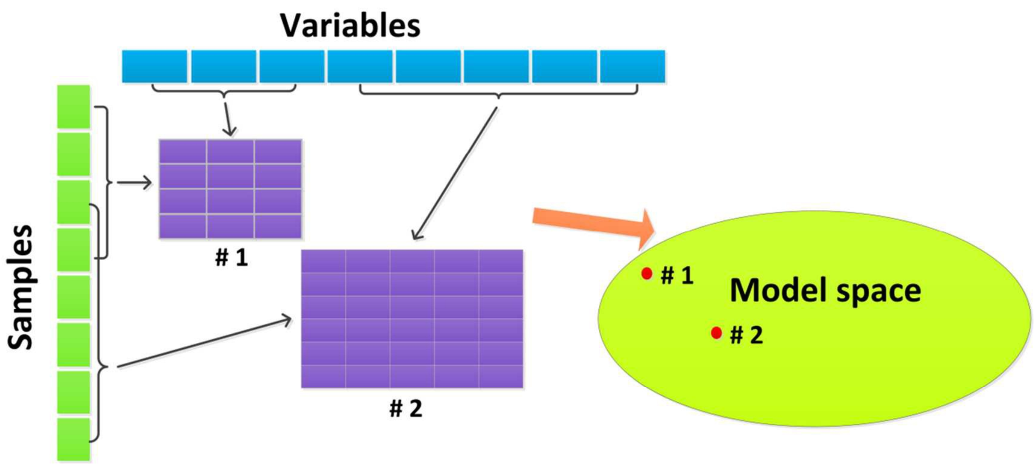

Fig. 1 The sketch of model space. A model is constructed by the combination of some variables and samples like #1 and #2, all the combinations make up the model space.

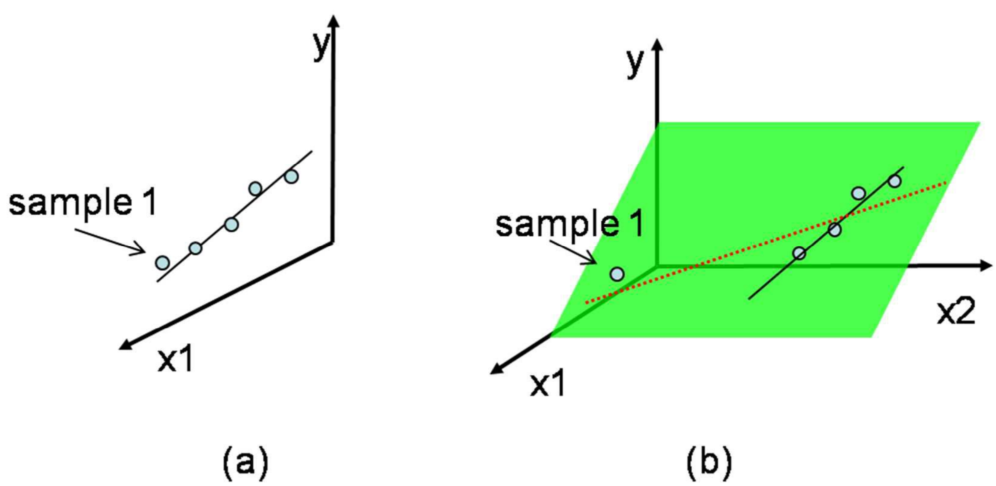

Fig. 2 The interactions between variables and outliers. (a) With only one variable $x1$, all samples (including sample 1) can be well fitted. (b) When variable $x2$ was added, sample 1 turns into an outlier.

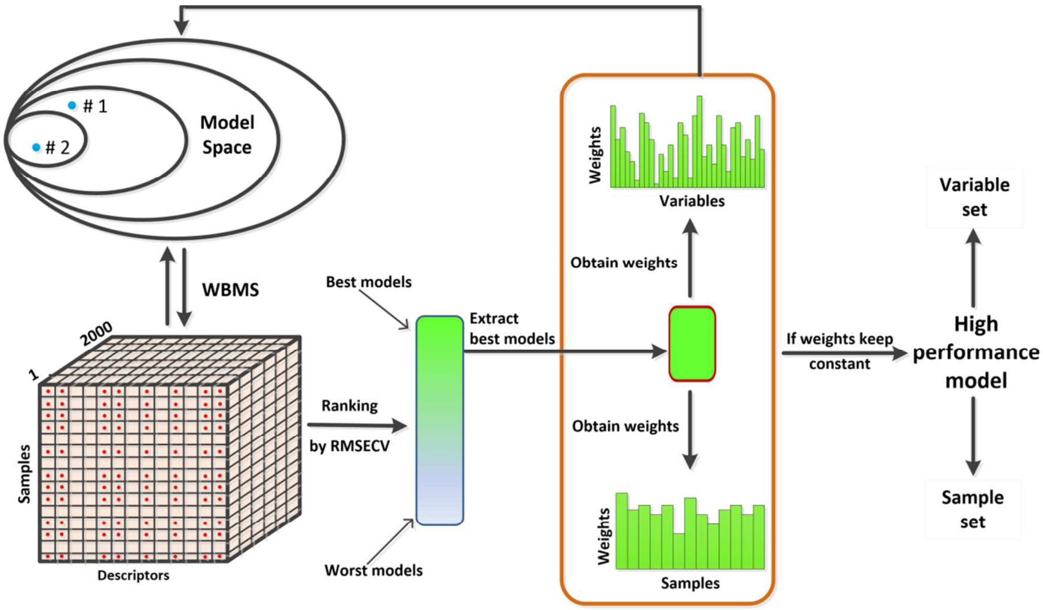

Fig. 3 The framework of MASS. Firstly, a number of sub-training datasets were sampled from the original dataset and build sub-regression models. The frequency of each variable and each sample in the best part models were counted. Then, a number of new sub-training datasets were sampled using the weight of variables and sample obtained from last iteration. The procedure for obtaining new weights of variable and sample was repeated until the weights of all variables and samples were constant (either 1 or 0). Thus the best model was obtained; the variables and samples that constructed the best model are simultaneously selected.

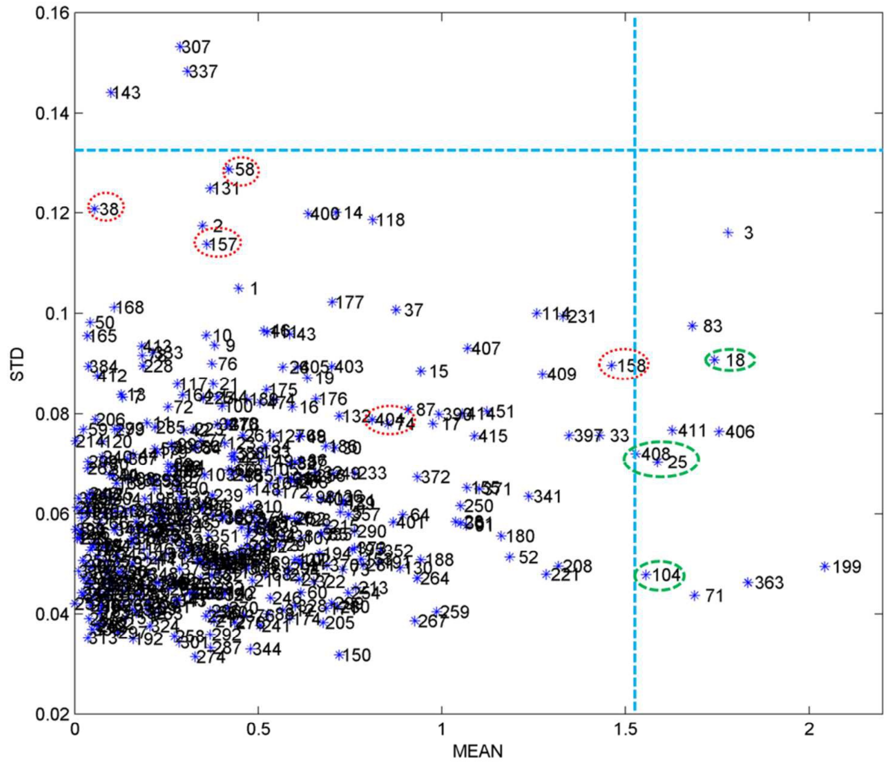

Fig. 4 Outlier detection plot of wheat kernel dataset detected by MCS+VISSA order. The samples enclosed by red ellipse located in lower left area and they are normal samples. Whereas they turned into outliers after variables selection (see Fig. 5). The samples enclosed by green ellipses are outliers but they become normal samples after variables selection (see Fig. 5)

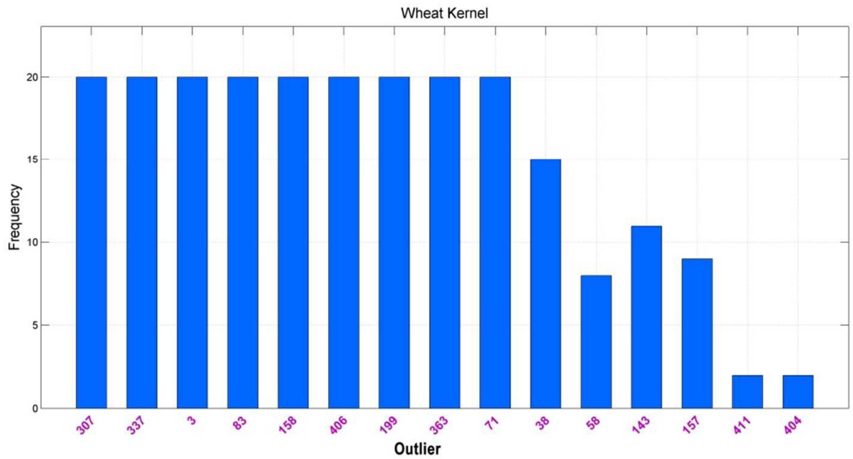

Fig. 5 Frequencies of outliers detected in wheat kernel dataset in 20 times in VISSA+MCS order.

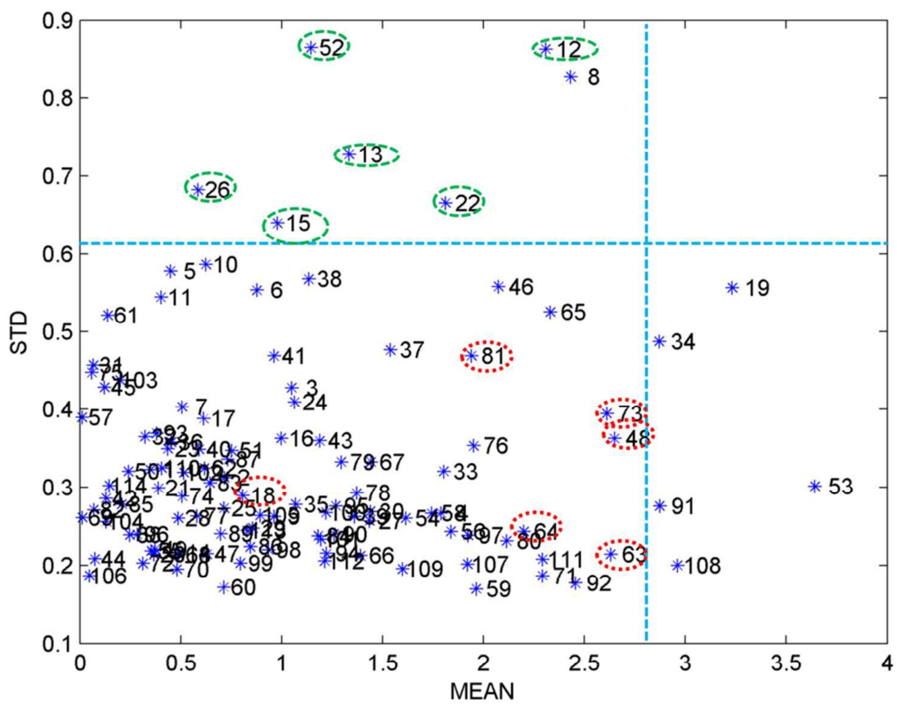

Fig. 6 Outlier detection plot of ACE dataset detected by MCS+VISSA order. The samples enclosed by red ellipse located in lower left area and they are normal samples. Whereas they turned into outliers after variables selection (see Fig. 7). The samples enclosed by green ellipses are outliers but they become normal samples after variables selection (see Fig. 7)

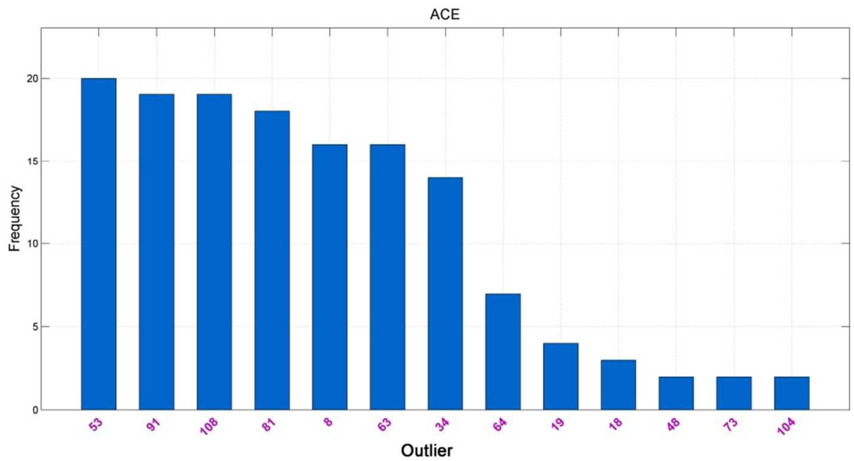

Fig. 7 Frequencies of outliers detected in ACE dataset in 20 times in VISSA+MCS order.

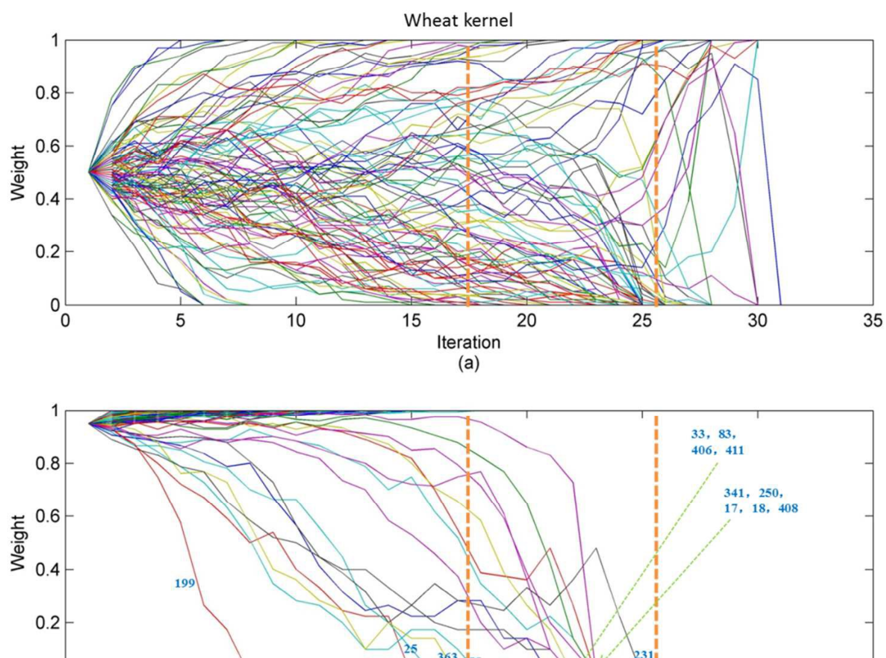

Fig. 8 The weight variation of (a) variables and (b) samples of wheat kernel dataset. Each line represents the weight variation of a sample or a variable.

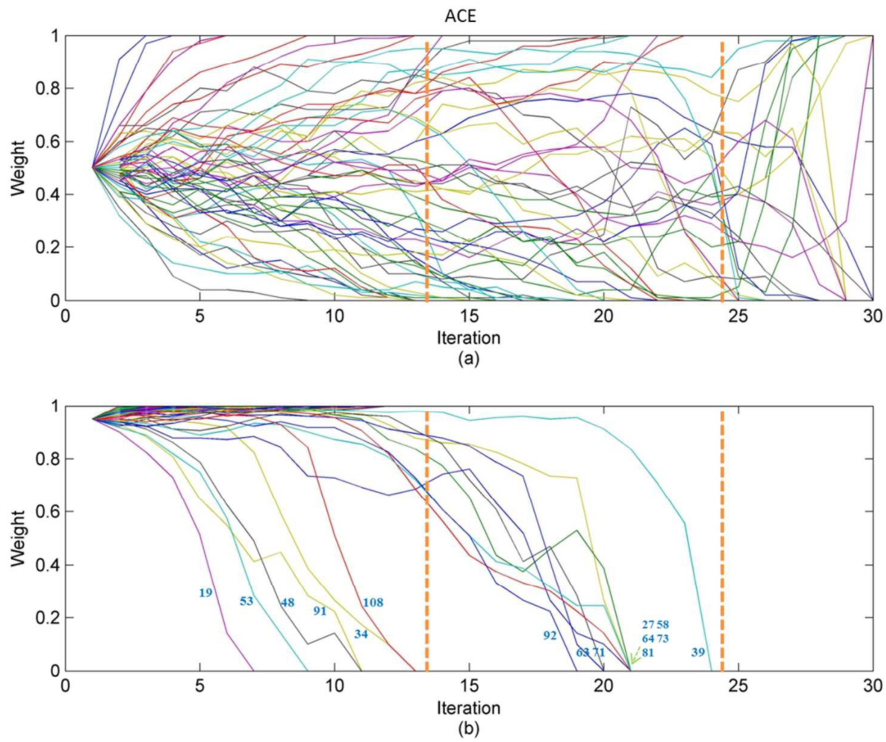

Fig. 9 The weight variation of (a) variables and (b) samples of ACE data set. Each line represents the weight variation of a sample or a variable.

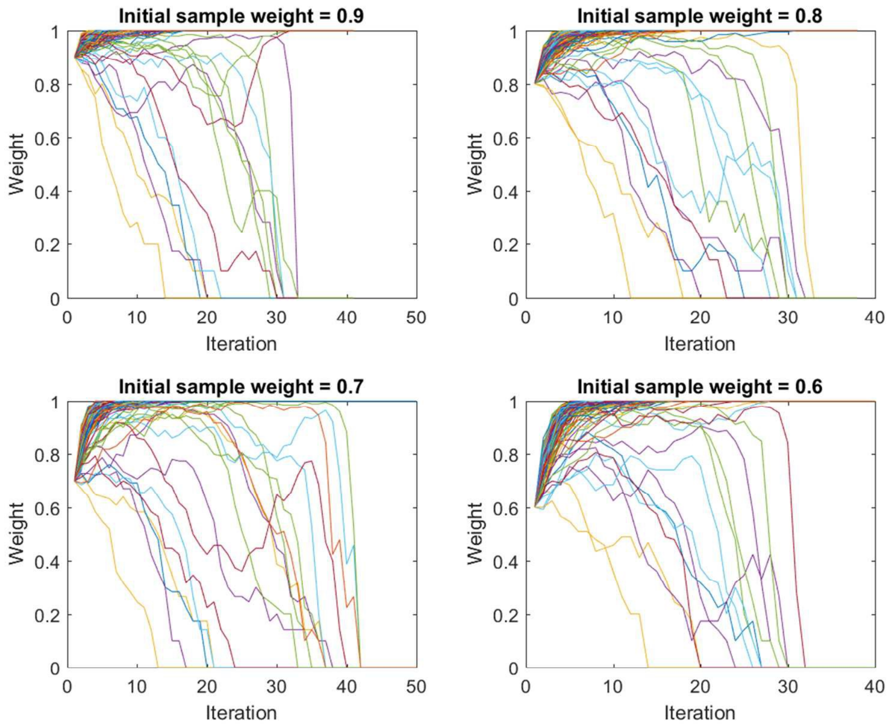

Fig. 10 The weight variation of samples of wheat kernel data set with different sample initial weight. Each line represents the weight variation of a sample.