RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Iteratively variable subset optimization for multivariate calibration

Weiting Wang,[a] Yonghuan Yun,[a,**] Baichuan Deng,[a,b] Wei Fan,[c] Yizeng Liang,[a,*]

[a]College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China

[b]Department of Chemistry, University of Bergen, Bergen N-5007, Norway

[c]Joint Lab for Biological quality and safety, College of Bioscience and Biotechnology, Hunan Agriculture University, Changsha 410128, P.R. China

Abstract: Based on the theory that a large partial least squares (PLS) regression coefficient on the autoscaled data indicates an important variable, a novel strategy for variable selection called iteratively variable subset optimization (IVSO), is proposed in this study. In addition, we take it into consideration that the optimal number of latent variables generated by cross-validation will make a great difference to the regression coefficients and sometimes the difference can even vary by several orders of magnitude. In this work, the regression coefficients generated in every sub-model are normalized to remove the influence. In each iterative round, the regression coefficients of each variable obtained from the sub-models are summed to evaluate its importance level. A two-step procedure including weighted binary matrix sampling (WBMS) and sequentially addition is employed to eliminate uninformative variables gradually and gently in a competitive way and reduce the risk of losing important variables. Thus,

∗Corresponding author. Tel.:86-0731-88830824; fax: +86-0731-88830831.

E-mail address: yizeng_liang@263.net (Yizeng Liang)

** Corresponding author. *E-mail address:* yunyonghuan@foxmail.com (Yonghuan Yun)

IVSO can achieve high stability. Investigated by one simulated dataset and two NIR datasets, IVSO shows much better prediction ability than another two outstanding and commonly used methods, Monte Carlo uninformative variable elimination (MC-UVE) and competitive adaptive reweighted sampling (CARS). The MATLAB code for implementing IVSO is available in the supplemental materials.

Keywords: Partial least squares, Regression coefficient, Weighted binary matrix sampling, Sequentially addition, Variable selection

1. Introduction

Nowadays, multivariate calibration models have been playing an essential role in multi-component spectral data, such as ultraviolet (UV), near infrared (NIR) and Raman spectroscopy. However, the spectral data obtained from these modern spectroscopic instruments usually contain hundreds or thousands of variables with high colinearity. Latent variable extraction techniques, such as principal component regression (PCR) and partial least squares (PLS) [1], provide a way to address the high colinearity problem. But the full spectrum used in these methods may bring negative influence on the performance of the calibration model due to the existing of uninformative variables. Many papers have demonstrated that it is critical to conduct variable selection in models instead of using full spectrum.[2-6] The advantages of variable selection have been concluded in the following three aspects: (1) improve the prediction accuracy of the model because of the elimination of uninformative variables that must lead to less precision as proved theoretically; (2) selecting wavelengths probably responsible for the property of interest makes the model more interpretative; (3) enhance the computational efficiency for modeling with a small amount of

variables.[7]

At present, many methods on variable selection have been employed in multi-component spectral data. In general, these methods can be classified into two categories as static and dynamic approach. The static approaches use one criterion for the whole data space, while the dynamic approaches take into account the result of each iteration. The static approaches includes t-statistics and Akaike information criteria (AIC), uninformative variable elimination (UVE),[8] Monte Carlo based uninformative variable elimination (MC-UVE),[9, 10] variable importance in projection (VIP),[11] selectivity ratio (SR),[12] and moving window partial least squares (MWPLS).[13] The dynamic approaches consists of optimized algorithm-based such as Genetic algorithm (GA),[14-16] particle swarm optimization (PSO),[17] firefly,[18] ant colony optimization (ACO),[19, 20] gravitational search algorithm (GSA),[21] and simulated annealing (SA).[22] The variable selection methods, such as Random forest,[23] successive projection algorithm (SPA),[24] iteratively retaining informative variables (IRIV),[25] variable combination population analysis (VCPA),[26] competitive adaptive reweighted sampling (CARS),[27] interval random frog (*i*RF),[28] are also the dynamic approaches. The theories of UVE, MC-UVE, CARS, and iRF comes from that the larger the absolute regression coefficient on the autoscaled data is, the more important the variable is.[8, 29] In addition to regression coefficient, Kvalheim et al. discussed the usage of SR that can assist in improved algorithm for variable selection in latent variable regression model.[30]

Among all the methods upon regression coefficient, MC-UVE and CARS are adopted extensively in multivariate calibration models for their better prediction. In both MC-UVE and CARS, Monte Carlo sampling technique is applied to the sample space to establish a large number of sub-models, which assures that the number of samples selected randomly for modeling is strictly the same, for example, 80% of all

samples is used to build the model. For MC-UVE, after *N* Monte Carlo sampling runs, one variable is evaluated according to a criterion which is equal to the ratio of the mean of the regression coefficients and its standard deviation. The variables with small criteria are eliminated. Unlike MC-UVE, in each iterative round, CARS removes the variables with small means of regression coefficients by the exponentially decreasing function (EDF) by force and adaptive reweighted sampling (ARS) competitively. However, it is the full spectrum in MC-UVE that is used to establish sub-models, which will lead to that the regression coefficients of the informative variables can be influenced by the uninformative variables. With regard to CARS, the enforced elimination of variables by EDF may lose important variables and further result in instability. Hence, in most cases the results achieved by MC-UVE and CARS are not satisfied enough.

In this study, a novel strategy for variable selection based on regression coefficient is proposed, called iteratively variable subset optimization (IVSO). At first, we introduce a new random sampling method, named weighted binary matrix sampling (WBMS),[31, 32] which is an improvement of the binary matrix sampling (BMS).[25, 33] Giving different weights to different variables, WBMS aims to make variables with larger weights more likely to be chosen. On the contrary, if the weight of one variable is small, it will be selected with little or even no possibility. Furthermore, combining WBMS with another strategy called sequentially addition, the variables with small criteria are deleted and a new variable subset is yielded. After *N* WBMS runs, *N* different variable subsets are obtained and the root mean squares error of

cross-validation (RMSECV) is used as the objective function to search for the best variable subset. In addition, the regression coefficients of one variable in all sub-models are summed to evaluate its importance level. This data fusion step is a good option, as the noise cancels out and the systematic information accumulates. However, we find that the optimal number of latent variables generated by cross-validation will make a great difference to the regression coefficients, which is consistent with the viewpoint in Reference.[34] Thus, the regression coefficients of the same variable in different sub-models cannot be calculated or compared directly due to the great difference. The strategy of normalization is applied to eliminate the influence. Tested on a simulated dataset and two NIR datasets, IVSO coupled with partial least squares (PLS) demonstrates better prediction ability and higher stability compared to the two outstanding methods above, namely MC-UVE and CARS. The result demonstrates that IVSO has the ability to eliminate uninformative variables gradually and gently in a competitive way, which can avoid those two problems of MC-UVE and CARS discussed above. It proves that IVSO is an efficient method for variable selection in multivariate calibration.

Additionally, it should be noted that IVSO is just evaluated by NIR datasets with PLS in this study, but it is a general strategy and can be combined with other regression and pattern recognition methods, and applied to other kinds of datasets, such as metabolomic and quantitative structure-activity relationship (QSAR).

2. Theory and algorithms

2.1. Notation

In this study, the matrix **X** with dimensionality $K \times P$ represents the observation matrix, in which K stands for the number of samples in rows and P is the number of variables in columns. Vector **y** with dimensionality $P \times 1$ denotes the measured property of interest, for example the concentration. In addition, the number of WBMS runs is set to N.

2.2. Weighted binary matrix sampling (WBMS)

In IVSO, a new method called WBMS is introduced for randomly sampling and further eliminating a part of uninformative variables, which is an improvement of binary matrix sampling (BMS). If the weight of one variable is small, the variable will be selected with little or even no possibility. Therefore, WBMS can eliminate variables competitively.

It works as follows: assume that the weight of the ith variable is w_i. At first, a binary matrix **M** with dimensionality $N \times P$ is generated, which contains either '1' or '0'. '1' represents that the responding variable is included for modeling, while '0' represents non-sampling for the variable. In each column, there are Nw_i '1' and the left ones are all '0'. The procedure is displayed in Fig. 1, where the row of **M** is set to 5 and the column is 7 for simplicity. When sampling, the weights of some variables are too small to be sampled in any column. The first and second columnns in Fig. 1 can represent this case. As the last column shows, if the weight of one variable is equal to 1, it will be sampled in each iterative round. Next, permuting each column in **M** generates a new binary matrix **NM**. Remarkably, after the permutation, the number of '1' or '0' in each column is kept unchanged.

In the matrix of **NM**, each row represents a sampling process for building a sub-model. It can be summarized that when Nw_i of one variable is less than 1, it will be

eliminated.

(Insert Figure 1)

2.3. Normalizing PLS regression coefficients

PLS is one of the most widely used methods for establishing the relationship between the observation matrix **X** and the properties of interest **y**. The scores matrix **T** is a linear combination of **X** with the combination coefficients **W**, and **c** is the regression coefficient vector of y against **T** by least squares.[1, 35] PLS can be expressed by by the following formulas:

$$\mathbf{T} = \mathbf{XW} \quad (1)$$

$$\mathbf{y} = \mathbf{Tc} + \mathbf{e} = \mathbf{XWc} + \mathbf{e} = \mathbf{Xb} + \mathbf{e} \quad (2)$$

where $\mathbf{b} = \mathbf{Wc}$ is the vector of PLS regression coefficients and **e** is the vector of residuals that cannot be explained by the model.

In addition, the matrix **X** needs to be autoscaled to guarantee that each variable has the same variance before modeling. It should be noted that the regression coefficients mentioned in this study have been changed into the absolute value before calculating. Afterwards, the larger the regression coefficient is, the more important the variable is.

Moreover, it is found that the optimal number of latent variables generated by cross-validation will make a great difference to the regression coefficients and sometimes the differences can even vary by several orders of magnitude.[34] Thus, the regression coefficients of the same variable in different sub-models may change a great deal and they cannot be calculated or compared directly. In this study, we employ the strategy of normalization to remove this influence. Assume that after building N sub-models, a regression coefficient matrix $\mathbf{B}$ ($\mathbf{B} = [\mathbf{b_1}, \mathbf{b_2}, \ldots \mathbf{b}_N]^{T)}$ is generated. The jth

row vector in **B**, denoted by $\mathbf{b}_j$ ($1 \leqslant j \leqslant N$) records the regression coefficients of the *j*th sub-model. The elements in the matrix **B** will be set to 0 if the responding variables are not included into the sub-models. The regression coefficient b_{ij} of the *i*th variable in the *j*th sub-model is normalized as follow:

$$c_{ij} = b_{ij} / \max(\mathbf{b}_j) \quad (3)$$

where $\max(\mathbf{b}_j)$ stands for the maximum of the row vector $\mathbf{b}_j$. The normalized regression coefficient matrix composed by all c_{ij} is denoted by **C**. The elements in **C** range from 0 to 1.

2.4. The criteria and weights of variables

For CARS, it is the mean of the regression coefficients of one variable that is considered as the criterion to determine its importance level. In IVSO, the normalized regression coefficient matrix **C** is summed in columns to generate a row vector **s** (**s** = [s_1, s_2,… s_P]), where s_i stands for the sum of the regression coefficients of the *i*th variable in all *N* sub-models. The sum s_i of the *i*th variable is regarded as its criterion. By this data fusion step, the noise can be cancelled out and the systematic information can be accumulated. In this way the difference between variables will become larger than that in CARS, which accelerating the iteration.

In each iterative round, the weight of the *i*th variable is defined as:

$$w_i = s_i / \max(\mathbf{s}),\ i = 1, 2, \ldots, P \quad (4)$$

where $\max(\mathbf{s})$ is the maximum of the vector **s**. The weights of the variables having been eliminated are set to zero automatically so that the weight vector **w** is always p-dimensional. Moreover, it should be mentioned that the weight vector only work for sampling by WBMS in the next iterative round.

2.5. Sequentially addition

In each iterative round, we combine WBMS with another strategy called sequentially addition to optimize the variable space. Firstly we use WBMS to eliminate variables in a competitive way. Denote L_1 as the number of the variables which can be sampled by WBMS. Then the L_1 variables are ranked based on their criteria. The variable space is further shrunk by sequentially addition. The L_1 variables are sequentially added step by step to establish L_1 PLS sub-models according to the rank and the performance of the sub-models is estimated by cross-validation. The first sub-model consists of only the first variable in the rank, and the second sub-model contains the first two variables, and the ith sub-model contains the first i variables. Repeat this process until the L_1 variables are all included into the last sub-model. When the RMSECV value of the sub-model reaches minimum with addition one by one, the corresponding variable subset in this best sub-model is chosen. The number of variables in this variable subset is denoted by L_2. The iterative process is continued with L_1 getting closer to L_2, until both L_1 and L_2 reach an equal value. One variable subset is yielded in one iterative round and finally many different variable subsets are generated. The RMSECV value is employed as the objective function to search for the best variable subset.

In each iterative round, sequentially addition can select a variable subset which contains informative variables. Thus, if some important variables are lost by WBMS, they still can be retained in the variable subset in the previous rounds by sequentially addition. When selecting the best variable subset among all iterative rounds, these lost

variables still have the opportunity to be included into the ultimate variable subset. In this way, no loss of important variables can be assured. For the same reason, IVSO possesses high stability. Overall, IVSO has the ability to eliminate variables gradually and gently in a competitive way and reduce the risk of losing important variables.

2.6. General description of IVSO

Fig. 2 shows the scheme of IVSO algorithm. The initial value of the weight of each variable is set to 1. It should be mentioned that the weights for sampling by WBMS are obtained from the previous iterative round. The detailed algorithm of IVSO is described as follows:

Step 1: Creating a binary matrix **NM** with dimensionality $N \times P$ for sampling by WBMS gives N sampling runs. In each column of **NM**, there are Nw_i '1' and the left ones are all '0'. If the Nw_i of one variable is less than 1, it will not be sampled in any row. Record the number of variables which can be sampled by WBMS, namely L_1.

Step 2: Build N PLS sub-models to calculate the regression coefficient matrix **B**. Each row of the matrix **B** is normalized to generate the matrix **C**, as Formula 2 did.

Step 3: Each column of the matrix **C** is summed as the criterion of the corresponding variable, denoted by the vector **s**. Rank the L_1 variables based on their criteria.

Step 4: Build L_1 sub-models through sequentially addition according to the rank. Take the variable subset in the sub-model with the lowest RMSECV value as the objective variable subset of this iterative round. Record this RMSECV value R and the length of this variable subset L_2.

Step 5: The vector **s** is normalized to calculate weights as Formula 3. The weights in this iterative round only work in the sampling of the next iterative round.

Step 6: Repeat the steps 1-5 until L_1 is equal to L_2, then many variable subsets are obtained. The variable subset with the lowest R value is chosen as the ultimate variable subset of the algorithm.

(Insert Figure 2)

3. Datasets and Software

3.1. Simulated dataset

This dataset, called SIMUIN, is simulated as described in Reference 18. SIMUIN contains 100 samples in rows and 200 variables in columns with exactly five latent variables. The relative eigenvalues by principal component analysis on the autoscaled data are 21.29%, 20.30%, 19.84%, 19.61%, 18.96%. The first 100 variables of SIMUIN are linearly relative with **y** but the last 100 ones are random numbers from 0 to 1, regarded as uninformative variables. The noises with normal distribution in the range from 0 to 0.005 are added to SIMUIN .

3.2. Corn moisture dataset

The corn dataset is available in the website: http://www.eigenvector. com/data/Corn/index.html. This dataset contains 80 samples of corn measured on three different NIR spectrometers. The spectrum has been recorded from 1100 - 2498 nm with 700 spectral points at intervals of 2 nm. In this study, the NIR spectrum of 80 corn samples measured on m5 instrument is used and the moisture value is considered as property of interest **y**.

3.3. Wheat dataset

This NIR dataset[4] consists of 100 wheat samples and the protein value is considered as property of interest **y**. The spectrum has been recorded from 1100 - 2500 nm with 701 spectral points at intervals of 2 nm. Due to the 'large p, small n' problem,[36,

[37] the original spectrum is compressed into a maximum of 200 points by an appropriate window size as did by Leardi.[14] Setting the window size to 4, this dataset is reduced to 175 variables with the average of every four original variables.

3.4. Software

All the computations are achieved in MATLAB on an ordinary computer configured to Intel Core i5 3.2 GHz CPU, 3G RAM, WIN7 Ultimate. The MATLAB code for implementing IVSO is available in the supplemental materials.

4. Results and Discussion

In this study, all the datasets are split into calibration set (80% of the dataset) and independent test set (20% of the dataset) based on Kennard - Stone (KS) method.[38] KS method aims at covering the multidimensional space by maximizing the Euclidean distances between each pair of the selected samples. The calibration set is used for variable selection and goodness of fit, and the independent test set is used for validation of the calibration model for prediction. When conducting variable selection on the calibration set, cross-validation is conducted. Furthermore, in order to evaluate the performance of IVSO, we compare it with another two outstanding methods based on the regression coefficient, namely MC-UVE and CARS. For MC-UVE, the number of Monte Carlo sampling runs is set to 1000, and 80% samples are randomly chosen for modeling in each sampling run. As to CARS, the number of Monte Carlo sampling runs is set to 100. For all methods, the maximum latent variable is limited to 10 and the number of latent variables is determined by 10-fold cross-validation. Each dataset is

277 autoscaled to have zero mean and unit variance before modeling. Besides, the root
278 mean square error of calibration set (RMSEC) and the root mean square error of
279 prediction of test set (RMSEP) are employed to assess the performance of the three
280 methods. In addition, because of the random sampling, these methods are conducted 50
281 times to obtain statistical results and compare the three methods fairly.

282 **4.1. The influence of the number of sampling by WBMS**

283 To investigate the influence of the number of sampling runs by WBMS, we
284 discuss four cases about the performance of IVSO, as shown in Fig. 3. The number of
285 sampling runs is set to 3000, 5000, 8000 and 10000, respectively, in the three datasets.
286 For these three datasets, their RMSEP values generated by full spectrum are 0.4043,
287 0.0237 and 0.2382, respectively. All of the results of the three datasets have good
288 stability. Overall, no significant influence on the results of IVSO has been found among
289 these four cases. For the dataset of wheat protein, the median values of the four RMSEP
290 values are the same, but the results with the parameter of 8000 shows the best stability.
291 Thus, the number of sampling by WBMS is set to 8000 in this study.

292 (Insert Figure 3)

293 **4.2. Simulated dataset**

294 This dataset is simulated to assess the ability of IVSO to select appropriate
295 variable subset. The first 100 variables are linearly relative with **y** and regarded as
296 informative ones, but the last 100 ones are noisy. The results obtained by conducting
297 different methods within 50 times are discussed in detail.

298 Table 1 includes the results of the three methods on the three datasets. The mean
299 and 95% confidence interval are given as well. The simulated dataset is investigated
300 by comparing with MC-UVE, CARS, the full spectrum and the first 100 informative

301 variables. Compared with the full spectrum, the RMSEC value of only the first 100
302 variables drops from 0.0644 to 0.0091 and the RMSEP value drops from 0.4043 to
303 0.0135, even using a smaller number of latent variable, 6. It demonstrates the
304 importance and necessity for variable selection in multivariate calibration.

305 The statistical features of the results by different methods can be observed visually
306 in the boxplot of Fig. 4, in which Fig. 4A displays the results of the simulated dataset.
307 As it can be seen from Table 1 and Fig. 4A, IVSO shows the best performance with
308 regard to the improving the prediction ability of the model and good stability. The 95 %
309 confidence interval of RMSEC and RMSEP results for each method shows that IVSO
310 has no overlap with other methods. In addition, the selected latent variable of IVSO is 6,
311 which is much smaller than that of the full spectrum.

312 The frequency distribution of selected variables within 50 times is displayed in Fig.
313 5. For different methods the selected variables all concentrate in the first 100 variables.
314 Both IVSO and MC-UVE can select variables with high frequencies. However, the
315 selected variables by CARS are of low frequencies and even no one variable can be
316 selected by all 50 times, which reveals its instability. The fact is just consistent with its
317 large confidence interval in Table 1 and standard deviation in Fig. 4A.

318 Fig. 6A and Fig. 6B show the changing trend of the number of variables sampled
319 by IVSO and CARS respectively. The arrow indicates the point reaching the optimal
320 variable subset. As to MC-UVE, it is the full spectrum that is used to establish the
321 sub-models, so no iterative round has ever occurred. For the simulated dataset in Fig.
322 6A, the number of sampled variables decreases to 100 in the 3th and 4th iterative
323 rounds, then the curve drops much more slowly. In stark contrast, the number of
324 sampled variables of the simulated dataset in Fig. 6B varies tremendously in the front
325 section of the curve. It is in the first iterative round that the number decreases to 97,

which means that just the first iterative round can removes not only uninformative but also informative variables. In the 11th iterative round CARS achieves its optimal variables subset containing only 26 variables. It can be concluded that CARS eliminates variables too quickly and thus lose informative variables, which may result from the enforced elimination of variables by EDF. On the contrary, IVSO has the ability to eliminate uninformative variables gradually and gently, and achieve much higher stability.

Fig. 7 shows the RMSECV value of the variable subset chosen by sequentially addition in each iterative round, which is corresponding to the sampling curve of the simulated dataset in Fig. 6A. The '0' iterative round stands for the process during which the weights of all variables are set to '1' as initial values but all these variables are conducted by sequentially addition. From Fig. 7A, the RMSECV value of the simulated dataset drops first because of the existing of some uninformative variables and begins to rise again due to the missing of some informative variables. It demonstrates the good ability of IVSO to eliminate uninformative variables and keep informative ones.

(Insert Table 1)

(Insert Figure 4)

(Insert Figure 5)

(Insert Figure 6)

(Insert Figure 7)

4.3. Corn moisture dataset

The results obtained by repeating the three different methods 50 times are reported in Table 1 and Fig. 4B. In Table 1, compared with the full spectrum, the RMSEC and RMSEP values of IVSO decrease by 98.4% and 98.6%, respectively. Clearly, IVSO

has highly improved the prediction performance. IVSO exhibits not only the best prediction ability in terms of the RMSEC and RMSEP values but also owns the best stability based on confidence interval. The number of latent variable is also the smallest, which means that it can generate more parsimony model.

The frequencies of variables selected by these methods are displayed in Fig. 8. Both CARS and IVSO mainly select variables of 1908nm and 2108nm, which have been discussed and proven as the key wavelengths by the literature of CARS experimentally and theoretically[27]. These two wavelengths are relative with the water absorption and the combination of O-H bond.[13] For CARS, it cannot select the key wavelength of 2108nm in every iterative round. However, except for these two key variables, MC-UVE selects too many other variables with high frequencies.

From the corn moisture dataset in Fig. 6A and Fig. 6B, we also can see that the number of variables sampled by IVSO in the previous rounds drops much more gradually and gently than that of CARS. In the latter rounds, though this number of IVSO changes more quickly, the key variables of 1908nm and 2108nm still can be retained in every iterative round due to sequentially addition. But CARS cannot do it.

In Fig. 7B, it reaches the optimal variable subset with the two key variables firstly in the 4th iterative round. Then the optimal variable subset keeps unchanged, so the RMSECV value is stable. From the RMSECV values, we can summary that the strategy of sequentially addition used in every iterative round makes the result stable.

(Insert Figure 8)

4.4. Wheat protein dataset

In Table 1 and Fig. 4C, IVSO can achieve better results with the smallest number of latent variable than the full spectrum. But after selecting variables, the RMSEP values of both MC-UVE and CARS get much worse. Fig. 9 displays the frequencies of

variables selected by these methods. The variables around 1144-1296nm can be selected by all methods, which is the same as the result of GA-PLS.[14] IVSO can select variables with quite high frequencies. As to MC-UVE and CARS, it selects many variables in other regions. Moreover, the frequencies of variables selected by CARS are not high and even quite a number of variables are selected by less than five times. From the wheat protein dataset in Fig. 6A and Fig. 6B, we also can see that the number of variables sampled by IVSO decreases much more slowly than that of CARS. The RMSECV value of the variable subset in Fig. 7C goes down at first with the decrease of the uninformative variables and then goes up because of increasingly deleting the informative variables.

(Insert Figure 9)

5. Conclusion

This paper presents a new method for variable selection based on the regression coefficient, called iteratively variable subset optimization (IVSO). Investigated by one simulated dataset and two NIR datasets, IVSO is proven to be a better variable selection method than another two methods, namely Monte Carlo uninformative variable elimination (MC-UVE) and competitive adaptive reweighted sampling (CARS). IVSO can eliminate uninformative variables gradually and gently, and achieve good prediction and stability. The outstanding performance of IVSO indicates that it is an efficient procedure and an alternative for variable selection.

Although IVSO is worked with partial least squares (PLS) to select variables in this study, it also can be coupled with other modeling methods on regression or

399 pattern recognition. Our future work will focus on investigating the application of
400 IVSO in other fields, such as metabolomic and quantitative structure-activity
401 relationship (QSAR).

402 **Acknowledgments**

403 This work is financially supported by the National Nature Foundation Committee
404 of P.R. China (grants no. 21275164 and 21465016) and also supported by the
405 Fundamental Research Funds for the Central Universities of Central South University
406 (grants no. 2014zzts014). The studies meet with the approval of the university's review
407 board.

408
409
410
411
412
413
414
415
416
417
418
419
420

References

1. S. Wold, M. Sjöström and L. Eriksson, *Chemometr. Intell. Lab.*, 2001, **58**, 109-130.
2. C. M. Andersen and R. Bro, *J. Chemometr.*, 2010, **24**, 728-737.
3. D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last and K. A. Prebble, *Anal. Chim. Acta.*, 1995, **304**, 285-295.
4. J. H. Kalivas, *Chemometr. Intell. Lab.*, 1997, **37**, 255-259.
5. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue and G. L. Coté, *Anal. Chem.*, 1998, **70**, 35-44.
6. Y. H. Yun, Y. Z. Liang, G. X. Xie, H. D. Li, D. S. Cao and Q. S. Xu, *Analyst*, 2013, **138**, 6412-6421.
7. I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157-1182.
8. V. Centner, D. L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste and C. Sterna, *Anal. Chem.*, 1996, **68**, 3851-3858.
9. W. Cai, Y. Li and X. Shao, *Chemometr. Intell. Lab.*, 2008, **90**, 188-194.
10. Q. J. Han, H. L. Wu, C. B. Cai, L. Xu and R. Q. Yu, *Anal. Chim. Acta.*, 2008, **612**, 121-125.
11. S. Favilla, C. Durante, M. L. Vigni and M. Cocchi, *Chemometr. Intell. Lab.*, 2013, **129**, 76-86.
12. O. M. Kvalheim, *J. Chemometr.*, 2010, **24**, 496-504.
13. J. H. Jiang, R. J. Berry, H. W. Siesler and Y. Ozaki, *Anal. Chem.*, 2002, **74**, 3555-3565.
14. R. Leardi, *J. Chemometr.*, 2000, **14**, 643-655.
15. R. Leardi and A. Lupiáñez González, *Chemometr. Intell. Lab.*, 1998, **41**, 195-207.
16. Y. H. Yun, D. S. Cao, M. L. Tan, J. Yan, D. B. Ren, Q. S. Xu, L. Yu and Y. Z. Liang, *Chemometr. Intell. Lab.*, 2014, **130**, 76-83.
17. J. Kennedy, in *Encyclopedia of Machine Learning*, eds. C. Sammut and G. Webb, Springer US, 2010, ch. 630, pp. 760-766.
18. M. Goodarzi and L. dos Santos Coelho, *Anal. Chim. Acta.*, 2014, **852**, 20-27.
19. F. Allegrini and A. C. Olivieri, *Anal. Chim. Acta.*, 2011, **699**, 18-25.
20. S. Tabakhi and P. Moradi, *Pattern. Recogn.*, 2015, **48**, 2798-2811.
21. J. Xiang, X. Han, F. Duan, Y. Qiang, X. Xiong, Y. Lan and H. Chai, *Appl. Soft. Comput.*, 2015, **31**, 293-307.
22. J. H. Kalivas, N. Roberts and J. M. Sutter, *Anal. Chem.*, 1989, **61**, 2024-2030.
23. L. Breiman, *Mach. Learn.*, 2001, **45**, 5-32.
24. M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame and V. Visani, *Chemometr. Intell. Lab.*, 2001, **57**, 65-73.
25. 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, *Anal. Chim. Acta.*, 2014, **807**, 36-43.
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, *Anal. Chim. Acta.*, 2015, **862**, 14-23.
27. H. Li, Y. Liang, Q. Xu and D. Cao, *Anal. Chim. Acta.*, 2009, **648**, 77-84.
28. Y. H. Yun, H. D. Li, L. R. E. Wood, W. Fan, J. J. Wang, D. S. Cao, Q. S. Xu and Y. Z. Liang, *Spectrochim.Acta.A.*, 2013, **111**, 31-36.
29. A. G. Frenich, D. Jouan-Rimbaud, D. L. Massart, S. Kuttatharmmakul, M. M. Galera and J. L. M. Vidal, *Analyst*, 1995, **120**, 2787-2792.

30. O. M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A. K. Smilde and J. A. Westerhuis, *J. Chemometr.*, 2014, **28**, 615-622.
31. B. C. Deng, Y. H. Yun, P. Ma, C. C. Lin, D. B. Ren and Y. Z. Liang, *Analyst*, 2015, **140**, 1876-1885.
32. B. C. Deng, Y. H. Yun, Y. Z. Liang and L. Z. Yi, *Analyst*, 2014, **139**, 4836-4845.
33. H. Zhang, H. Wang, Z. Dai, M.-s. Chen and Z. Yuan, *BMC Bioinformatics*, 2012, **13**, 298.
34. B. C. Deng, Y. H. Yun, Y. Z. Liang, D. S. Cao, Q. S. Xu, L. Z. Yi and X. Huang, *Anal. Chim. Acta.*, 2015, **880**, 32-44.
35. Q. S. Xu, Y. Z. Liang and H. L. Shen, *J. Chemometr.*, 2001, **15**, 135-148.
36. E. Candes and T. Tao, *Ann. Stat.*, 2007, 2313-2351.
37. H. Zou and T. Hastie, *J. Roy. Stat. Soc. B.*, 2005, **67**, 301-320.
38. R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137-148.

Table 1

Results of different methods on the three datasets.

Methods	nVar[a]	nLVs[b]	RMSEC		RMSEP	
			Min-Max	Average ± CI[c]	Min-Max	Average ± CI
Simulated						
PLS[d]	200	10	0.0644		0.4043	
PLS[e]	100	6	0.0091		0.0135	
MC-UVE	78.2 ± 6.3	6	0.0087-0.0093	$0.0089\pm8.3\times10^{-5}$	0.013-0.0135	$0.0132\pm5.8\times10^{-5}$
CARS	27.6 ± 4.8	6.3 ± 0.7	0.0080-0.0124	$0.0100\pm3.1\times10^{-4}$	0.0118-0.0197	$0.0155\pm5.3\times10^{-4}$
IVSO	68.1 ± 2.1	6	0.0090-0.0093	$0.0091\pm1.5\times10^{-5}$	0.012-0.0137	$0.0125\pm8.7\times10^{-5}$
Corn moisture						
PLS	701	10	0.017		0.0237	
MC-UVE	70.4 ± 2.6	10	2.4×10^{-3}-3.0×10^{-3}	$2.7\times10^{-3}\pm4.1\times10^{-5}$	2.8×10^{-3}-3.7×10^{-3}	$3.2\times10^{-3}\pm4.0\times10^{-5}$
CARS	3.4 ± 2.7	3.1 ± 1.9	2.4×10^{-4}-2.7×10^{-3}	$4.6\times10^{-4}\pm1.8\times10^{-4}$	3.4×10^{-4}-4.5×10^{-3}	$6.4\times10^{-4}\pm2.8\times10^{-4}$
IVSO	2.3 ± 0.8	2.3 ± 0.8	2.6×10^{-4}-2.8×10^{-4}	$2.8\times10^{-4}\pm1.1\times10^{-6}$	3.3×10^{-4}-3.6×10^{-4}	$3.4\times10^{-4}\pm1.4\times10^{-6}$
Wheat protein						
PLS	175	10	0.3923		0.2382	
MC-UVE	10.6 ± 1.3	9.9 ± 0.3	0.3370-0.3657	0.3475 ±0.0012	0.2466-0.2791	0.2532 ±0.0027
CARS	9.8 ± 2.8	8.2 ± 1.3	0.2501-0.3427	0.2969 ±0.0054	0.1818-0.3535	0.2432 ±0.0111
IVSO	14.8 ± 3.0	7.5 ± 1.0	0.2415-0.2695	0.2641 ±0.0030	0.2313-0.2363	0.2339 ±0.0009

[a] The number of selected variables

[b] The number of selected latent variables of PLS

[c] 95% confidence interval (CI)

[d] Results using the full spectrum with 200 variables by PLS

[e] Results using only the first 100 informative variables by PLS

Figure Captions

Fig.1. The process of generating binary matrix.

Fig.2. The scheme of iteratively variable subset optimization (IVSO) algorithm.

Fig.3. The boxplot for each dataset with the number of sampling runs by WBMS set to 3000, 5000, 8000 and 10000, respectively. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentile, the whiskers extend to the most extreme data points are the maximum and minimum, the "+" plotted individually represents outliers.

Fig.4. The boxplot of 50 RMSEP values for the three methods. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentile, the whiskers extend to the most extreme data points are the maximum and minimum, and the "+" plotted individually represents outliers.

Fig.5. The frequencies of variables selected by different methods within 50 times on the simulated dataset. (A) MC-UVE; (B) CARS; (C) IVSO.

Fig.6. The changing trend of the number of sampled variables by IVSO (A) and CARS (B).

Fig.7. The root mean squares error of cross-validation (RMSECV) of the variable subset chosen by sequentially addition in each iterative round. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset.

526

527 Fig.8. The frequencies of variables selected by different methods within 50 times on
528 the corn moisture dataset. (A) MC-UVE; (B) CARS; (C) IVSO.

529

530 Fig.9. The frequencies of variables selected by different methods within 50 times on
531 the wheat protein dataset. (A) MC-UVE; (B) CARS; (C) IVSO.

532

533

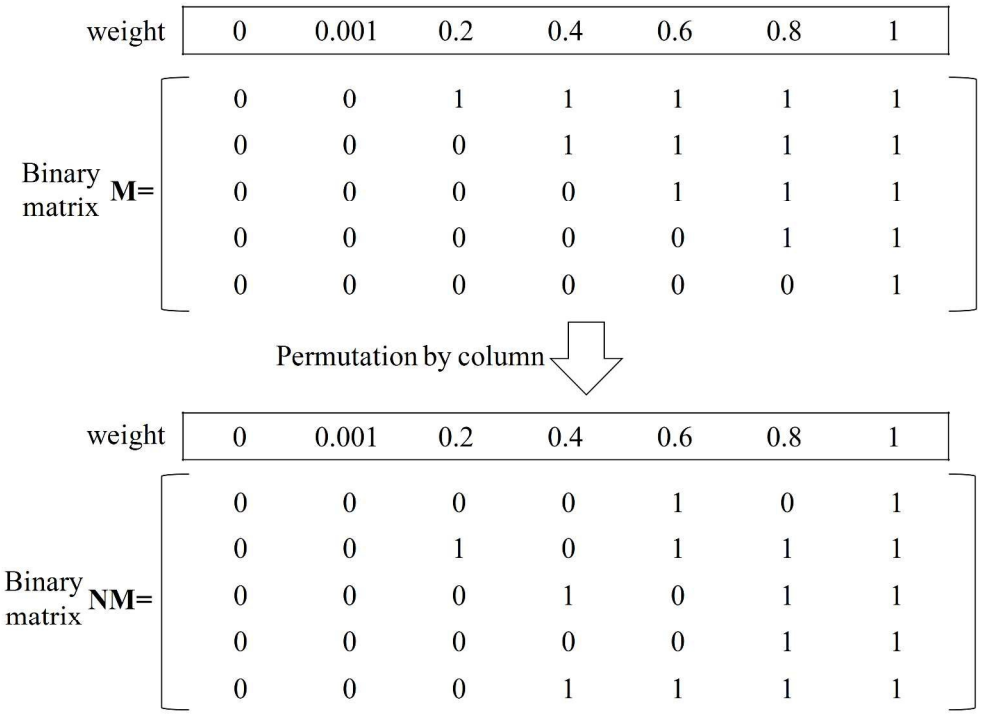

Fig.1. The process of generating binary matrix.
295x216mm (300 x 300 DPI)

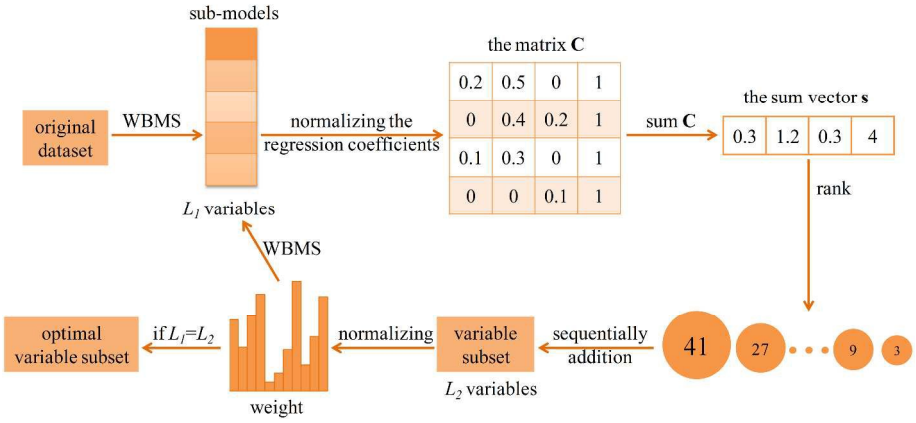

Fig.3. The boxplot for each dataset with the number of sampling runs by WBMS se
438x187mm (300 x 300 DPI)

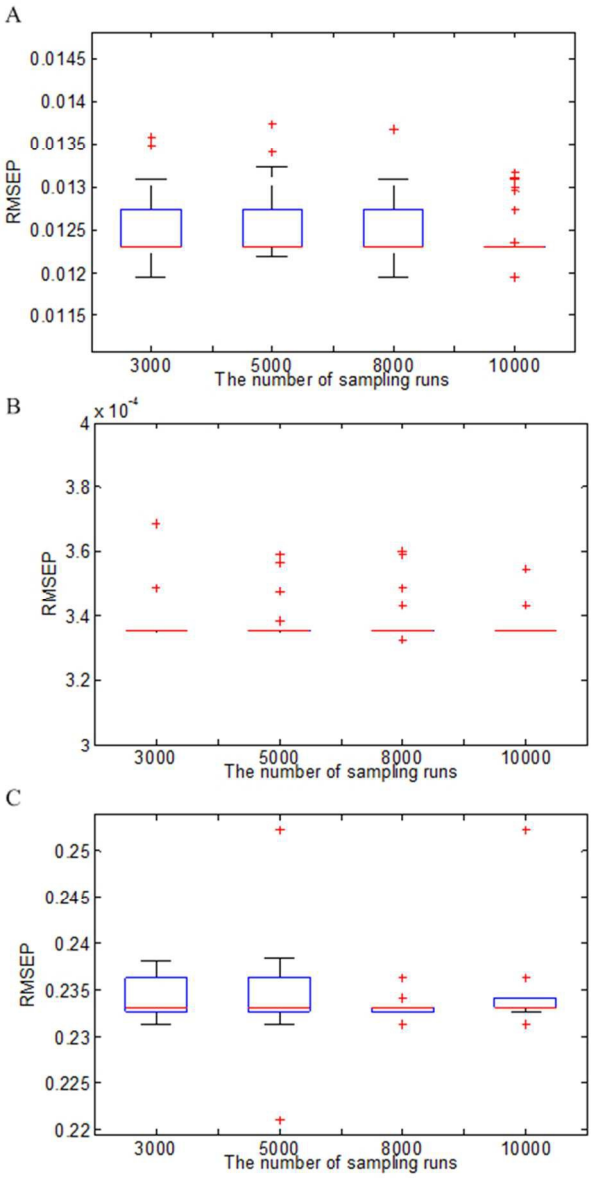

Fig.3. The boxplot for each dataset with the number of sampling runs by WBMS set to 3000, 5000, 8000 and 10000, respectively. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentile, the whiskers extend to the most extreme data points are the maximum and minimum, the "+" plotted individually represents outliers.
239x445mm (50 x 50 DPI)

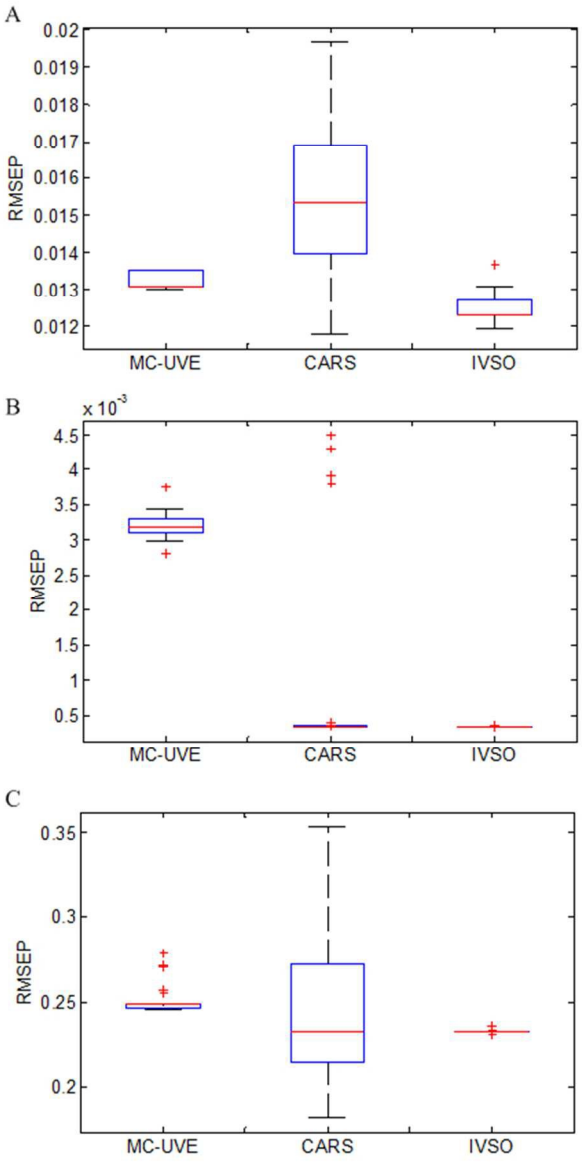

Fig.4. The boxplot of 50 RMSEP values for the three methods. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentile, the whiskers extend to the most extreme data points are the maximum and minimum, and the "+" plotted individually represents outliers.
236x443mm (50 x 50 DPI)

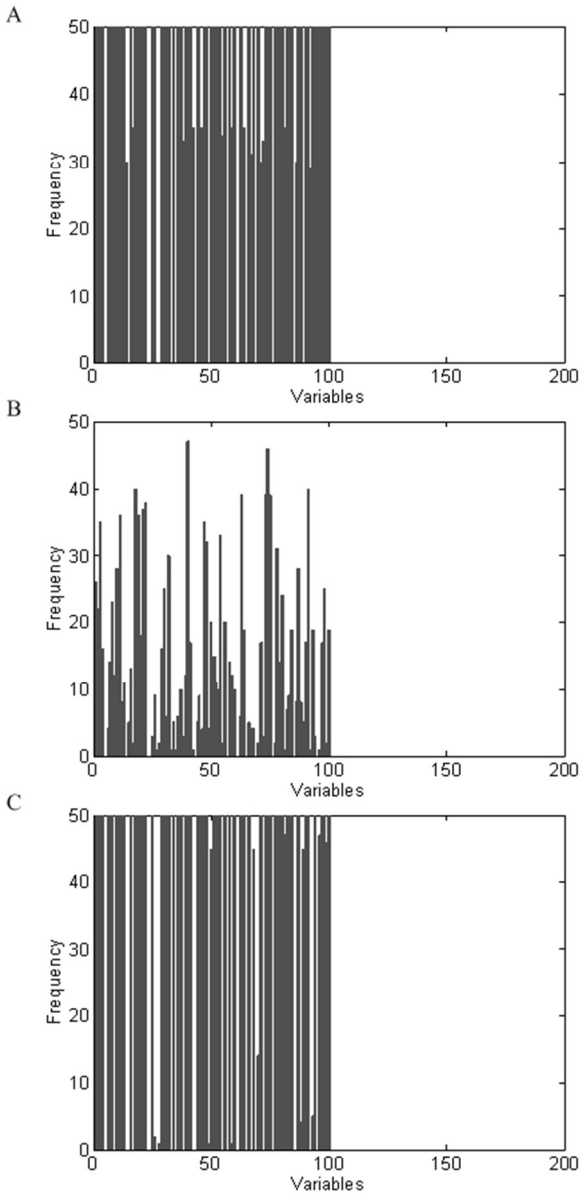

Fig.5. The frequencies of variables selected by different methods within 50 times on the simulated dataset. (A) MC-UVE; (B) CARS; (C) IVSO.
235x457mm (50 x 50 DPI)

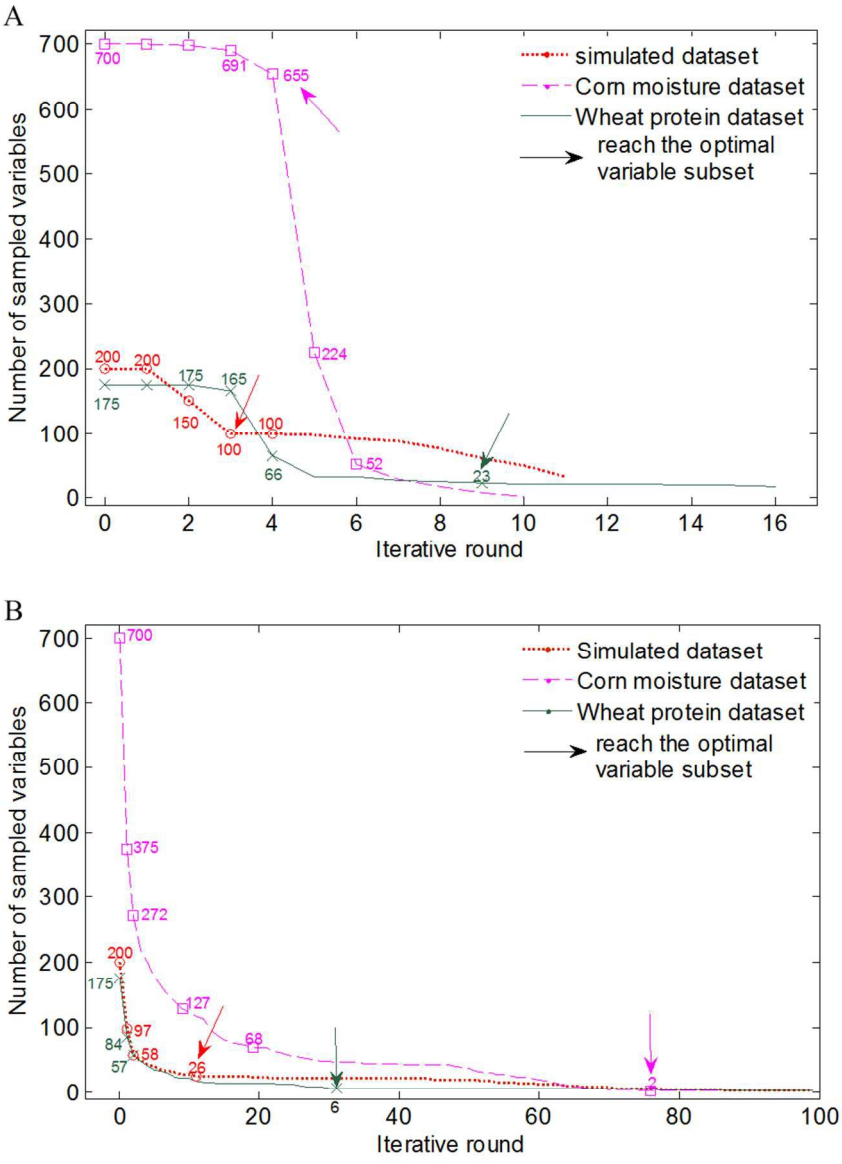

Fig.6. The changing trend of the number of sampled variables by IVSO (A) and CARS (B).
271x343mm (100 x 100 DPI)

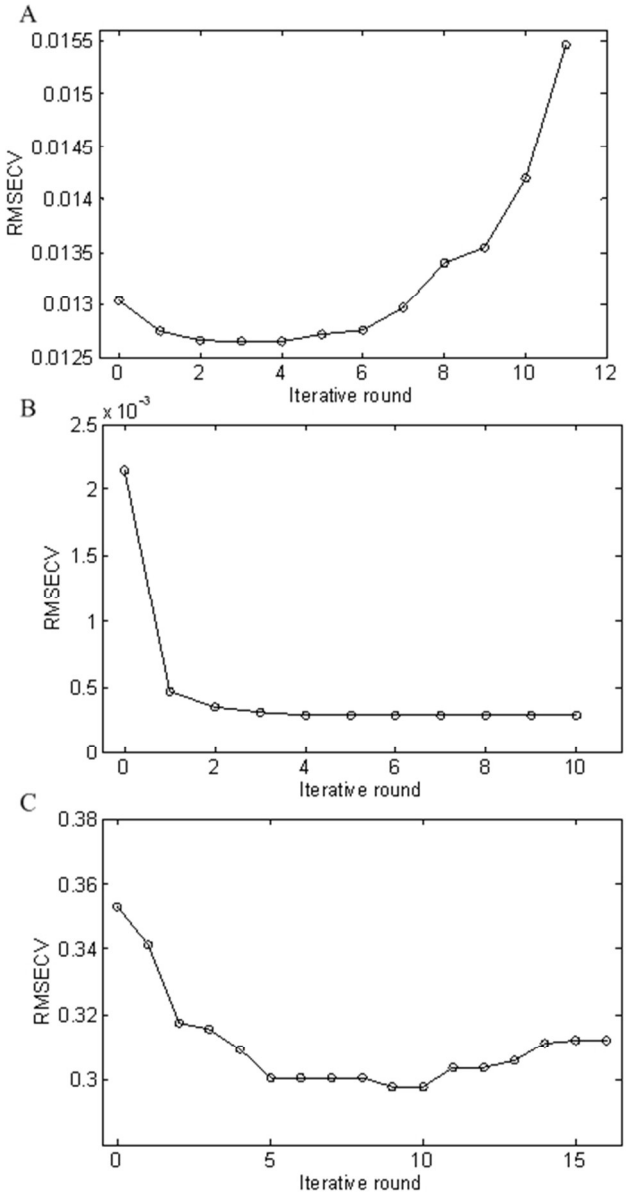

Fig.7. The root mean squares error of cross-validation (RMSECV) of the variable subset chosen by sequentially addition in each iterative round. (A) simulated dataset; (B) corn moisture dataset; (C) wheat protein dataset.
234x411mm (50 x 50 DPI)

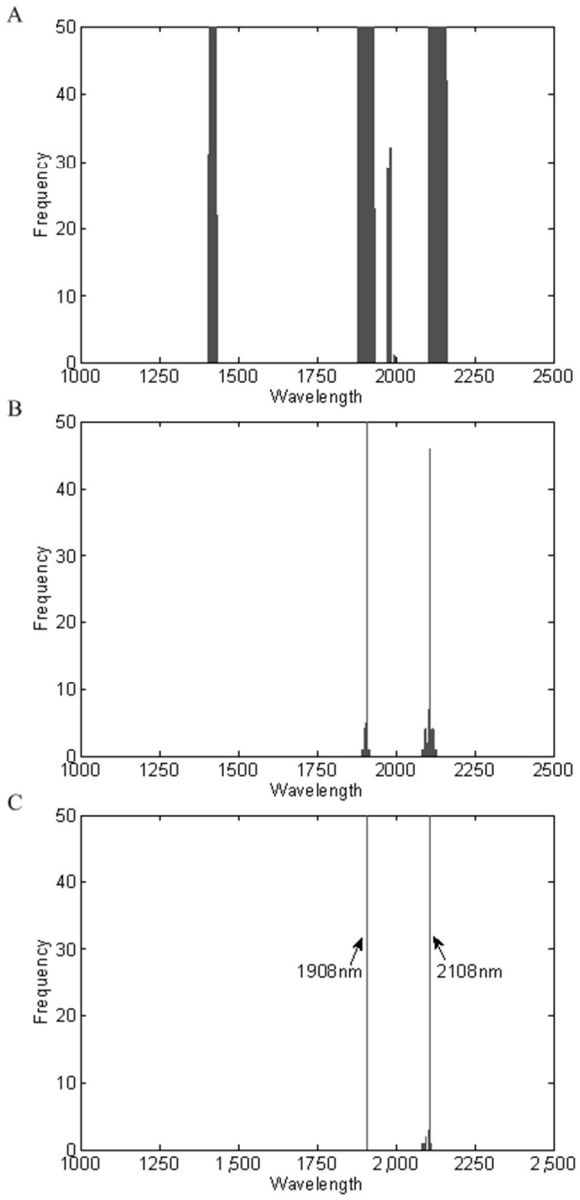

Fig.8. The frequencies of variables selected by different methods within 50 times on the corn moisture dataset. (A) MC-UVE; (B) CARS; (C) IVSO.
231x457mm (50 x 50 DPI)

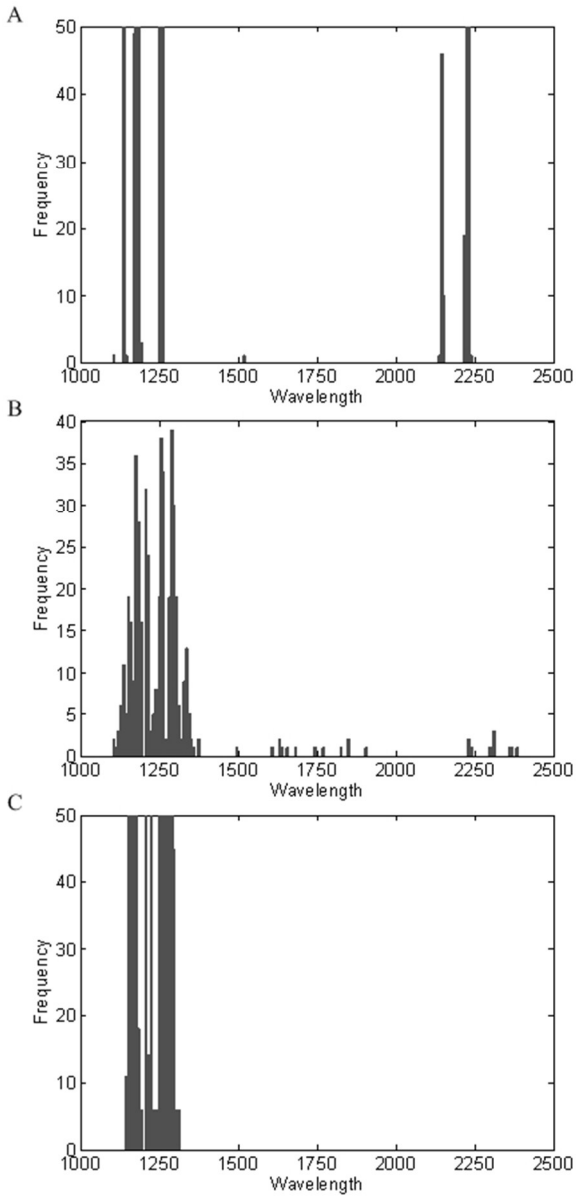

Fig.9. The frequencies of variables selected by different methods within 50 times on the wheat protein dataset. (A) MC-UVE; (B) CARS; (C) IVSO.

231x457mm (50 x 50 DPI)