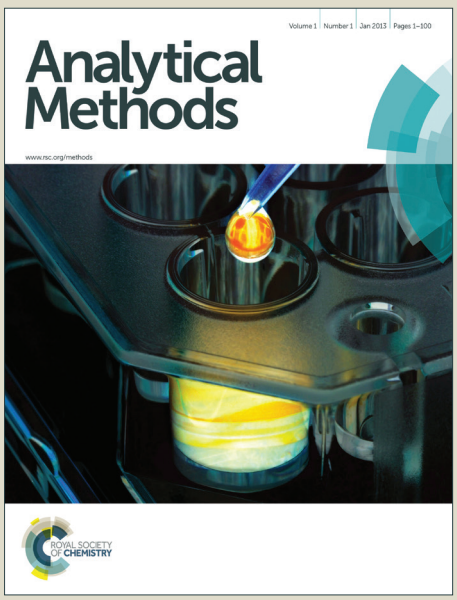
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Analytical Methods

RSCPublishing

ARTICLE

Variable Selection Based on Information Tree for Spectroscopy Quantitative Analysis

Cite this: DOI: 10.1039/x0xx00000x

Hui Cao[a,b] , Xingyu Yan[a,b], Shuzhi Sam Ge[b], Hongliang Ren[b]

Received 00th January 2015,
Accepted 00th January 2015

DOI: 10.1039/x0xx00000x

www.rsc.org/

Spectroscopy is a fast and efficient component analysis method, and full spectrum prediction model may be redundant and inaccurate. This paper proposes a variable selection method based on information tree for spectroscopy quantitative analysis. Firstly, a feature training set that indicates the information of the selected variables is generated. Then, the partial least squares (PLS) is performed on the spectral calibration set, and root-mean-square error of cross-validation is used to evaluate the feature training set. According to the corresponding evaluation results, the information gain of each wavelength is calculated. The wavelength with maximum information gain is defined as the root node, and an information tree is built based on the information gain where each leaf node represents a wavelength. The final selection result is a conjunction path of the leaf nodes that has bigger information gain. The full spectrum PLS, the uninformative variable elimination with PLS method, the genetic algorithm with PLS method and the proposed method are conducted on the real spectral dataset of flue gas, and the effectiveness of the methods are compared and discussed. The experimental results verify that the prediction precision and the compression ability of the proposed method is higher.

1 Introduction

Spectroscopy studies the spectrum data to tell the related information precisely and swiftly and is used in a wide range of applications, such as colorimetric thermometer[1] and quantitative analysis, et al.[2-6] One pivotal task of spectral quantitative analysis is to build up a model that takes the spectral data on different wavelengths as inputs and consequently predict the amount of chemical species according to regression results.[7,8] The most commonly used regression modelling method for spectroscopy is partial least squares (PLS) as it can solve the multicollinearity problem between the variables to a certain extent.[9] Nevertheless, the PLS model with all wavelength data included is not able to filter the useless information that complicates the program and lowers the precision.[10]

In order to deal with the useless information and meanwhile simplify the quantitative calibration model, several methods have been proposed for the purpose to study the characteristic wavelengths instead of using the full spectrum.[11-13] The Correlation coefficient method was used on spectral data to search for the relevant similarity from data with noise[14] and the sulphur emissions with spectral data was detected by combining correlation coefficients.[15] Competitive adaptive reweighted sampling method for the key wavelength selection was proposed for wavelength selection.[16] A randomization test method for wavelength selection was proposed.[17] An method which can select the variable by an index of stability that is defined as the absolute value of regression coefficient divided by its standard deviation was used for variable selection.[18] Latent projective graph method was proposed to find the informative variables.[19] An influential variables method was proposed for multivariate calibration.[20] Uninformative variable elimination with PLS (UVE-PLS) method[21] has long been used and studied by researchers. UVE-PLS was adopted and estimated on near-infrared spectral quantitative analysis, and a successive projection algorithm of UVE-PLS was presented to find the effective variables in the pesticide spectral data.[22] Moreover, as wavelength selection could be considered as a combinatorial issue, genetic algorithm, which mimics the natural selection and genetic mechanism, together with PLS (GA-PLS), is widely used for spectral quantitative analysis.[23] the method was used for wavelength selection of visible and near-infrared spectral calibration to construct robust and predictive regression models[24] and a more advanced parallel GA-PLS was presented for wavelength selection.[25] All these methods show their ability of searching and approaching the optimal solution in a certain period of time (or a certain number of iterations). However, these algorithms may easily lead the solutions to a local optimum and a big number of iterations are time consuming. Meanwhile the convergence rate and optimal solutions will vary from time to time according to the randomness of these algorithms. Decision tree method is a kind of inductive learning algorithm that is able to reason out a tree-like classification rule from unordered data-sets without randomness. Tree uses the descending speed of information entropy as an index to select the attributes as former nodes.[26] However, the decision tree method cannot be directly applied to

[a] *State Key Laboratory of Electrical Insulation and Power Equipment, School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China*

[b] *Faculty of Engineering & Advanced Robotics Centre, National University of Singapore, 117580, Singapore.*

wavelength selection because too many attributes will bring immense calculation amounts.

In this paper, a variable selection method based on information tree (IT) for the spectral quantitative analysis is proposed. This method firstly generates a feature training set that contains the information of the filtered variables. PLS calculation is applied on the original data set with respect to the feature training set to evaluate the performance of the feature training set. Then the entropy index of tree is calculated and the information tree and the wavelengths are selected. The experiments on real flue gas was conducted and the results will be compared in terms of the effectiveness and precision of the other three methods: PLS, UVE-PLS method and GA-PLS method.

The structure of the rest of this paper is as follows. In section 2, the relevant methods are illustrated in detail. Section 3 discusses the experimental results of a real spectral data of flue gas. Section 4 concludes the paper.

2 Relevant Methods

2.1 Partial Least Squares (PLS)

In spectroscopy analysis, PLS decomposes the principal factors of the independent variable matrix X and the dependent variable matrix Y, respectively. Then r principal factors are obtained, dividing the contributing components and errors.[27]

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{E}_Y = \sum_{i=1}^{r} \boldsymbol{u}_i \boldsymbol{q}_i^T + \boldsymbol{E}_Y \quad (1)$$

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}_X = \sum_{i=1}^{r} \boldsymbol{t}_i \boldsymbol{p}_i^T + \boldsymbol{E}_X \quad (2)$$

where U and Q are the score matrix and loading matrix of the density matrix, respectively. T and P are the score matrix and loading matrix of the independent variable matrix, respectively. E_X and E_Y are the error matrices of the independent variable matrix and the dependent variable matrix, respectively.

Then build the linear relation between T and U:

$$\boldsymbol{U} = \boldsymbol{T}\boldsymbol{B} \quad (3)$$

where B is the regression coefficient of the latent model and the solution of B can be obtained by least squares:

$$\boldsymbol{B} = (\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{U} \quad (4)$$

While using PLS for prediction, we firstly calculate the score matrix of the new independent matrix T_{new}. Then with the model below, predictions are realized.

$$\boldsymbol{Y}_{\mathrm{new}} = \boldsymbol{T}_{\mathrm{new}}\boldsymbol{B}\boldsymbol{Q} \quad (5)$$

2.2 Uninformative Variable Elimination method (UVE)

UVE method is based on the theory of regarding the regression coefficient as the index to weigh the importance of the variable. The specific steps are as follows. [28]

Step 1: Make a noise matrix R (n × p) and combine the spectral input matrix of the calibration set X (n × p) and R to get a matrix X':

$$X'_{(\mathrm{n}\times 2\mathrm{p})} = [X\ R]$$

Step 2: For X and Y, conduct interaction verification each time PLS eliminates a sample. The n PLS regression coefficients form the matrix B (n ×2p);

Step 3: Calculate the standard deviation s (1 × 2p) and the mean vector Mean (1 × 2p) out of the column of B (n × 2p). Then find: $h_i = \mathrm{Mean}_i / s_i$, i=1,2, …, 2p;

Step 4: In the interval [m + 1, 2m], find the Maximal absolute value of h: h_{max} = max [abs (h)];

Step 5: In the interval [1, m], eliminate the variables that corresponding $h < h_{max}$ in the matrix X. Get the new matrix X_{UVE}.

Swiftly and practically, UVE method combines the noise and density together while selecting the wavelengths. So UVE is very commonly used for wavelength selection.

2.3 Genetic Algorithm-PLS (GA-PLS)

GA method is a process of imitating the natural selection and genetic mechanism in biosphere. As variable selection is a combinatorial issue, GAPLS is the very frequently used for spectral data set. [29] In GA method, a population of chromosomes are randomly generated. Each chromosome is coded by a binary string, which indicates that the wavelength is selected or dropped, respectively. The length of a chromosome equals the number of the all wavelengths. Then PLS, as an evaluation method, will calculate the fitness of the initial chromosome. Then GA-PLS creates a new population by selecting, protecting, exchanging and making variations or mutations on the different fragments of the chromosome. Thus an evolving mechanism is established and is repeated until a termination condition which is the number of evolution cycles or a pre-defined fitness value. Finally, the chromosome with the lowest fitness is the variable selection result.

2.4 Information Tree-PLS (IT-PLS)

IT-PLS uses a random feature training matrix to evaluate the importance of each wavelength. In this method, we firstly generate a random feature training matrix M(c × p), where c is the size of M and p is the number of wavelengths. The number of c will be discussed in the practical experiment. The value of each element of M is "0" or "1", which represents the wavelength been discarded or selected, respectively. The ratio of 0 and 1 is random. Each row of M is a plan of wavelength selection, and the number of the wavelength selection plans is c. Then we use the variables selected according to each row of M to calculate the prediction values by PLS and find the root-mean-square of cross-validation (RMSECV) value of each row. Find the median value of the RMSECV values to determine if a row of M can meet the requirement. For each row of M, if RMSECV (i) < med, mark this row with "good". If else, mark this row with "bad". Add this column to the training matrix and get M'(c × p + 1). This generating process can be described more precisely as follows:

Step 1: Generate a random feature training matrix M;

Step 2: Use the variables according to the input M to calculate the prediction values by PLS;

Step 3: Calculate the RMSECV of the each row of M and find the median of the RMSECV values;

Step 4: Obtain a new matrix M'(c×p+1) according to the contrast with the median values in Step 3;

As M' is built up, the information gain of the each variable can be calculated with the set:

$$I(j,k) = -\frac{j}{j+k}\log_2\frac{j}{j+k} - \frac{k}{j+k}\log_2\frac{k}{j+k}$$

where j is the number of "good" s in the last column of M' and k is the number of "bad" s in the last column of M'. I (j, k) is the expected information needed to generate the message.

Then calculate the information entropy of the variable a, E(a), which is:

$$E(a)=\frac{j_1+k_1}{j+k}I\left(j_1,k_1\right)+\frac{j_0+k_0}{j+k}I\left(j_0,k_0\right)$$

where j_1 and k_1 are the number of the "good" s and number of "bad" s, respectively in a subset of F_a1 which assembles all the rows whose number in the ath column is 1. Similarly, j_0 and k_0 are the number of 1s and the number of 0s, respectively in a subset of F_a0, which assembles all the rows whose number in the ath column is 0.

Then, find the information gain of the variable a, gain (a), which is:

$$gain(a) = I(j,k) - E(a)$$

In the same way, the information gain of each variable could be evaluated. Then build the information tree by using the variable that has a biggest information gain as the root node, and the rest variables are represented by the leaf nodes.[30] After successively connect the root node and the leaf nodes based on the wavelength selection plans. Finally, find the path with the biggest information gain in this information tree that can lead to best results, namely, the variable selection is completed. The flowchart of IT-PLS can be described in Fig. 1, in which the importance of each wavelength is evaluated in the process.

[Fig. 1 is about here]

Fig. 1 Flow chat of IT-PLS

3 Experimental results

Real flue gas samples were used in the experiments. These samples consist of 98 mixtures of different densities of sulphur dioxide (SO_2), nitrogen monoxide (NO) and nitrogen dioxide (NO_2). The density ranges of SO_2, NO and NO_2 in the gas mixtures were 0-1500ppm, 0-3000ppm and 0-500ppm, respectively. A spectrometer (USB2000t fibre optic spectrometer, Ocean Optics) was used to measure the absorbance value of the gas mixtures on each wavelength. The wavelength range was from 187.87nm to 1026.97nm with an interval of 0.35 nm and it includes 2048 wavelengths. The size of spectral matrix was 98 × 2048. A spectrum of the gas mixture is shown in Fig. 2. The first several wavelengths before 200nm consist of some noise information. In order to verify the robustness, they are not deleted and used in the experiments.

[Fig. 2 is about here]

Fig. 2: An initial spectrum of the flue gas

With respect to the shutter grouping strategy, the data set was divided into calibration set and validation set.[23] One sample in five was placed into validation set and the rest into calibration set. So there are 80 samples in the calibration set for training and 18 samples in the validation set.[31] The calibration set and the validation set were for building the prediction models and estimating the effectiveness, respectively. PLS, UVE-PLS, GA-PLS and IT-PLS method were applied on the calibration set and the performances were tested on the validation set. Since these methods are based on PLS, the number of latent variables is determined by the minimum of the root mean-squared error of leave-one-out cross validation (RMSECV). For GA-PLS, the crossover rate and the mutation rate were set to be 60% and 10%, respectively. [32] Moreover, as indicators of analytical methods, the root-mean-square error of prediction (RMSEP), squared correlation coefficient of calibration ($\boldsymbol{R_c^2}$), squared correlation coefficient of prediction ($\boldsymbol{R_p^2}$), and squared correlation coefficient of cross validation ($\boldsymbol{R_{cv}^2}$) were used to compare the predictive ability of the four methods.

The parameter c, which is the number of rows of *M*, will affect the results to a certain extent. In the experiments, a range of c from 10 to 34 is tested. Fig. 3 shows how RMSECV values vary with different c values. RMSECV of the SO_2 with a c number of 29 is the best. Similarly, we chose 24 as the value of c for NO_2 and NO. The results the experiments are recorded in Table1, Table2 and Table3. The indicators include RMSEP, $\boldsymbol{R_p^2}$, RMSECV, $\boldsymbol{R_{cv}^2}$ and compression ratio (CR).

[Fig. 3 is about here]

Fig. 3: Relations of c and RMSECV: (a) SO_2, (b) NO_2 and (c) NO.

Table 1 shows a comparison of predictive ability of the four methods for SO2. The most accurate prediction was realized via IT-PLS method with a RMSEP value of 53.1888 and the most inaccurate prediction was produced by UVE-PLS. Furthermore, a considerable CR was also obtained with IT-PLS method. Therefore, the effectiveness of the IT-PLS method was the highest. The prediction value and measured value scatters of the four methods are shown in Fig. 4. For the values below 750, the scatters in PLS and UVE-PLS are distributed not as close to the diagonal line as GA-PLS and IT-PLS. For the values above 750, IT-PLS has the best performance as the scatters are closest to the line on both sides. PLS is relatively better than UVE-PLS and GA-PLS but not as good as IT-PLS. The prediction ability of IT-PLS is the best for SO_2.

Table 1: Experiment results for SO_2

Methods	RMSEP	R_p^2	RMSECV	R_{cv}^2	CR
Partial Least Squares	65.6796	0.9791	105.1932	0.9771	0
Uninformative Variable Elimination with Partial Least Squares	73.8438	0.9745	83.8151	0.9728	0.9302
Genetic Algorithm with Partial Least Squares	61.9160	0.9876	54.1680	0.9889	0.5000
Information Tree with Partial Least Squares	**53.1888**	0.9835	54.0141	0.9602	**0.8779**

[Fig. 4 is about here]

Fig. 4: Predicted value vs. measured value scatter diagram of different methods for SO2. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

Table 2 shows a comparison of predictive ability of the four methods for NO_2. The most accurate prediction was done via IT-PLS method with a relatively lower RMSEP value of 151.0093 and the most inaccurate prediction was produced by PLS, which is 400.5299. Furthermore, $\boldsymbol{R_p^2}$ of IT-PLS was the highest and the CR is 0.8779. Therefore, the effectiveness of the IT-PLS method was also the highest in the NO_2 case. The prediction value and measured value scatters of the four methods are shown in Fig. 5. The PLS can not hold the points close to the diagonal line as the points are distributed on a large area on the diagram. UVE-PLS and GA-PLS have better performances but some of the points break away from the diagonal line. The prediction result of IT-PLS is the best as no point is far from the diagonal line.

Table 2: Experiment results for NO_2

Methods	RMSEP	R_p^2	RMSECV	R_{cv}^2	CR
Partial Least Squares	400.5299	0.1985	454.2047	0.4865	0
Uninformative Variable Elimination with Partial Least Squares	263.0552	0.6872	233.1586	0.8481	0.9902
Genetic Algorithm with Partial Least Squares	259.7527	0.6350	247.7038	0.8294	0.4902
Information Tree with Partial Least Squares	**141.1331**	0.8933	183.6444	0.8513	**0.8779**

[Fig. 5 about here]

Fig. 5: Predicted value vs. measured value scatter diagram of different methods for NO_2. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

Table 3 presents a comparison of predictive ability of the methods for NO. The most accurate prediction was also achieve via IT-PLS method with a lowest RMSEP value of 19.3806 and the most inaccurate prediction was produced by GA-PLS, which is 34.8568. $\boldsymbol{R_p^2}$ of IT-PLS was the highest and a considerable CR of 0.9268 is obtained. Therefore, the effectiveness of the IT-PLS method was also the highest in the NO case. The prediction value and measured value scatters of the four methods are shown in Fig. 6. The points of PLS and GA-PLS break away from the diagonal line. While the points of UVE-PLS and IT-PLS stay close with the line. For values below 150, UVE-PLS has a good performance as some points are directly on the line, and IT-PLS has a more robust prediction as the points are distributed equally on both sides of the diagonal line. For values above 150, IT-PLS has the best prediction ability as the distance from the point to the line won't fluctuate as much as the other methods.

Table 3: Experiment results for NO

Methods	RMSEP	R_p^2	RMSECV	R_{cv}^2	CR
Partial Least Squares	31.9052	0.7096	44.4247	0.5239	0
Uninformative Variable Elimination with Partial Least Squares	21.3421	0.8784	23.4832	0.8637	0.8638
Genetic Algorithm with Partial Least Squares	34.8568	0.6829	22.7365	0.8630	0.4927
Information Tree with Partial Least Squares	**19.3806**	0.9055	21.4831	0.8299	**0.9268**

[Fig. 6 is about here]

Fig. 6: Predicted value vs. measured value scatter diagram of different methods for NO. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

The variables selected by UVE-PLS, GA-PLS and IT-PLS for SO_2, NO_2 and NO are shown in Fig. 7. The number of variables selected by GA-PLS is so large that the prediction model is more complex and the effectiveness is affected. Although the number of variables selected by UVE-PLS is the smallest, some informative variables may be eliminated and the predication capability could be limited. The variable selection result of IT-PLS is reasonable and IT-PLS has a more accurate prediction.

[Fig. 7 is about here]

Fig. 6: Variable selection results for three components of the flue gas. (a) SO_2. (b) NO_2. (c) NO.

4 Conclusions

A wavelength selection method based on information tree is proposed and is combined with PLS for predicting the various components of real spectral datasets. The proposed method was has some of advantages over the conventional variable selection methods as follows. First, the IT-PLS method can be more advanced as it utilized the information entropy with the tree concept to select the wavelengths. Second, for the various components of flue gas dataset, the prediction model based on the wavelengths selected by IT-PLS has a higher prediction precision. Third, the robustness of the ITPLS method is better as the prediction results of IT-PLS has a steady result and compress the variable number greatly. The experiments results verify that the proposed method has higher predicative ability. The RMSEP of the IT-PLS method for sulphur dioxide was 19.02%, 27.97% and 14.10% lower than that of PLS, UVE-PLS and GA-PLS, respectively. The RMSEP of the IT-PLS method for nitric dioxide was 64.76%, 46.35% and 45.67% lower than that of PLS, UVE-PLS and GA-PLS, respectively. And the RMSEP of the IT-PLS method for nitric monoxide was 39.26%, 9.19% and 44.4% lower than the PLS, UVE-PLS and GA-PLS results, respectively. Therefore, IT-PLS is an efficient variable selection method and can help to achieve more accurate predictions and can be implemented to different types of spectra.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 61375055, the Program for New Century Excellent Talents in University under Grant NCET-12-0447, the Natural Science Foundation of Shaanxi Province of China under Grant 2014JQ8365, the Fundamental Research Funds for the Central University and the A*STAR Industrial Robotics Program of Singapore under grant R-261-506-007-305 and R-261-506-008-305.

References

[1] Yuanjing Cui, Wenfeng Zou, Ruijing Song, Jiancan Yu, Wenqian Zhang, Yu Yang and Guodong Qian. *Chemical Communications*, 2014, **50**, 719–721.

[2] Jiajia Shan, Tetsuhito Suzuki, Diding Suhandy, Yuichi Ogawa and Naoshi Kondo. *Engineering in Agriculture, Environment and Food*, 2014, **7**, 139-142.

[3] Hongguang Zhang, Qinmin Yang and Jiangang Lu. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2013, **120**, 625-629.

[4] Nanyoung Kim, Miao Yu, Dong Young Lee, Young Hee Hahn, Young Choong Kim, Sang Hyun Sung and Seung Hyun Kim. *Analytical Letters*, 2013, **46,** 1289–1298.

[5] M Lindkvist and C Grönlund. *Spectroscopy Letters*, 2015, **48**, 170-172.
[6] Philippa Alice Hayes, Signe Vahur and Ivo Leito. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2014.
[7] Jerome J Workman Jr. *Applied Spectroscopy Reviews*, 1996, **31**, 251–320.
[8] Benoit Igne, James B Reeves III, Gregory McCarty, W Dean Hively, Eric Lund and Charles R Hurburgh Jr. *Journal of Near Infrared Spectroscopy*, 2010, **18**, 167–176.
[9] Yong Hu, Silong Peng, Jiangtao Peng and Jiping Wei. *Spectra. Talanta*, 2012, **94**, 301–307.
[10] Brenchley J. M., Horchner U. and Kalivas J. H. *Applied Spectroscopy*, 1997, **51**, 689-699.
[11] Mulang Chen, Swanand Khare, Biao Huang, Haitao Zhang, Eric Lau and Enbo Feng. *Industrial & Engineering Chemistry Research*, 2013, **52**, 7886–7895.
[12] AX Zhao, XJ Tang, ZH Zhang and JH Liu. *Spectroscopy and Spectral Analysis*, 2014, **34**, 1836–1839.
[13] Wei Li, Ling Lin, and Gang Li. *Analytical Methods*, 2014, **6**, 1082–1089.
[14] Peter R Griffiths and Limin Shao. *Applied spectroscopy*, 2009, **63**, 916–919.
[15] P Siozos, A Philippidis, M Hadjistefanou, C Gounarakis and D Anglos. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 2013, **87**, 86–91.
[16] Li Hongdong. *Analytica Chimica Acta*, 2009, **648**, 77-84.
[17] Heng Xu, Zhichao Liu, Wensheng Cai and Xueguang Shao. *Chemometrics & Intelligent Laboratory* Systems, 2009, **97**, 189-193.
[18] Kaiyi Zheng, Qingqing Li, Jiajun Wang, Jinpei Geng, Peng Cao, Tao Sui, Xuan Wang and Yiping Du. *Chemometrics & Intelligent Laboratory Systems,* 2012, **112**, 48-54.
[19] Xueguang Shao, Guorong Du, Ming Jing and Weisheng Cai. *Chemometrics & Intelligent Laboratory Systems*, 2012, **114**, 44-49.
[20] Xueguang Shao, Mu Zhang and Wensheng Cai. *Anal Methods*, 2012, **4**, 467-473.
[21] Vítězslav Centner, Désiré-Luc Massart, Onno E de Noord, Sijmen de Jong, Bernard M Vandeginste and Cécile Sterna. *Analytical chemistry*, 1996, **68**, 3851–3858.
[22] Guo Tang, Xiangzhong Song, Jing Hu, Hong Yan, Kaixian Qiu, Kuangda Tian, Yanmei Xiong and Shungeng Min. *Analytical Letters,* 2014, **47**, 2570–2579.
[23] Delphine Jouan-Rimbaud, Desire-Luc Massart, Riccardo Leardi and Onno E De Noord. *Analytical Chemistry*, 1995, **67**, 4295–4301.
[24] Riccardo Leardi and Amparo Lupiáñez González. *Chemometrics and Intelligent Laboratory Systems*, 1998, **41**, 195–207.
[25] Olivier Devos and Ludovic Duponchel. *Chemometrics and Intelligent Laboratory Systems*, 2011, **107**, 50–58.
[26] J. Ross Quinlan. *Machine learning*, 1986, **1**, 81– 106.
[27] Beatriz Álvarez-Sánchez, Feliciano Priego-Capote, Juan García-Olmo, María C. Ortiz-Fernández, Luis A. Sarabia-Peinador and María D. Luque de Castro. *Journal of Chemometrics*, 2013, **27**, 221–232.
[28] Wensheng Cai, Yankun Li and Xueguang Shao. *Chemometrics and intelligent laboratory systems*, 2008, **90**, 188–194.
[29] Masamoto Arakawa, Yosuke Yamashita and Kimito Funatsu. Genetic. *Journal of Chemometrics*, 2011, **25**, 10–19.
[30] P. N. Tan, M. Steinbach and V. Kumar. *Introduction to data mining*. Addison Wesley Higher Education, USA, 2006.
[31] Tormod Naes, Tomas Isaksson and Bruce Kowalski. *Analytical Chemistry*, 1990, **62**, 664–673.
[32] Zou Xiaobo, Zhao Jiewen, Mao Hanpin, Shi Jiyong, Yin Xiaopin and Li Yanxiao. *Applied Spectroscopy*. 2010, **64**, 786-794.

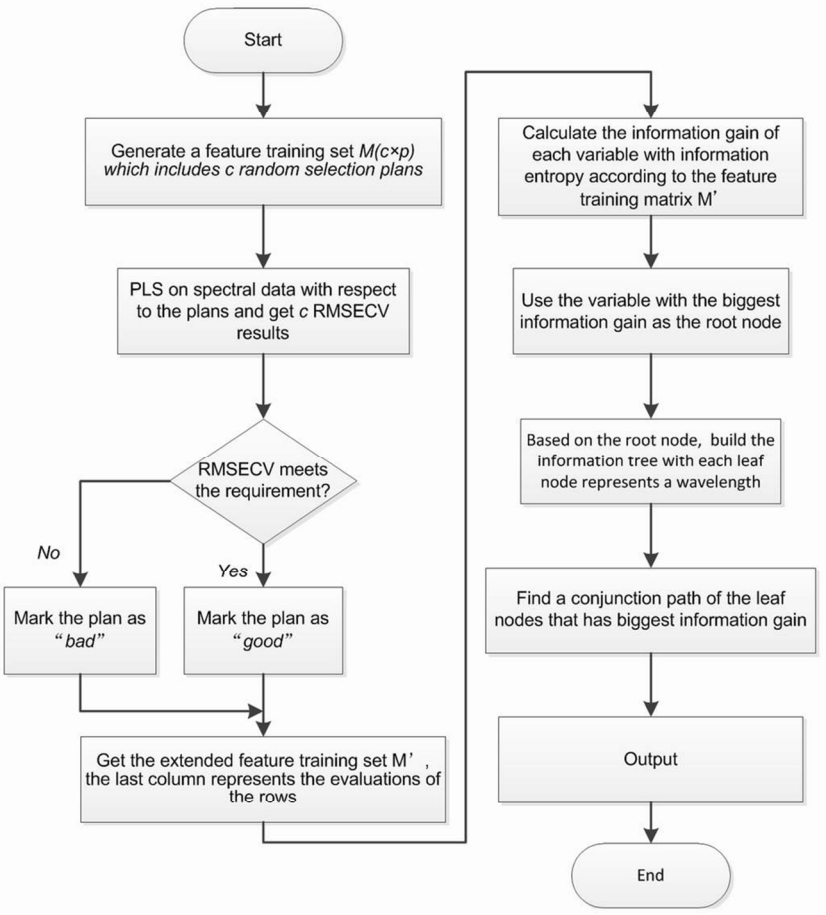


Fig. 1 Flow chat of IT-PLS

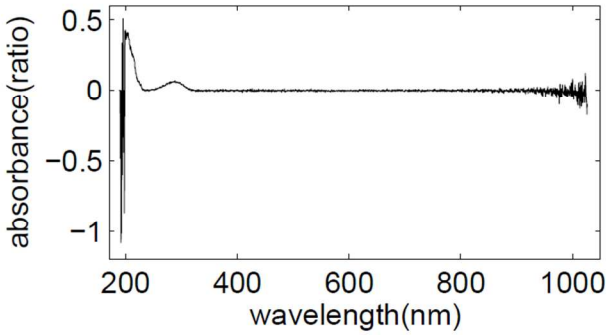


Fig. 2: An initial spectrum of the flue gas

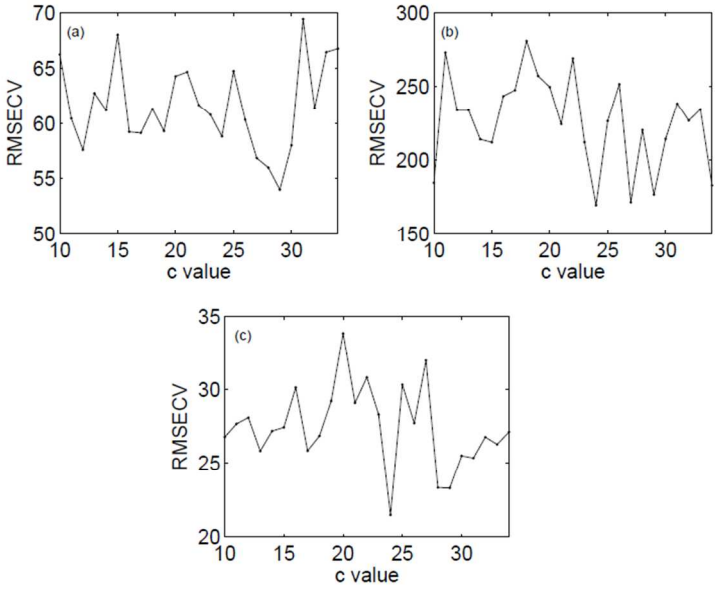


Fig. 3: Relations of c and RMSECV. (a) SO_2. (b) NO_2. (c) NO.

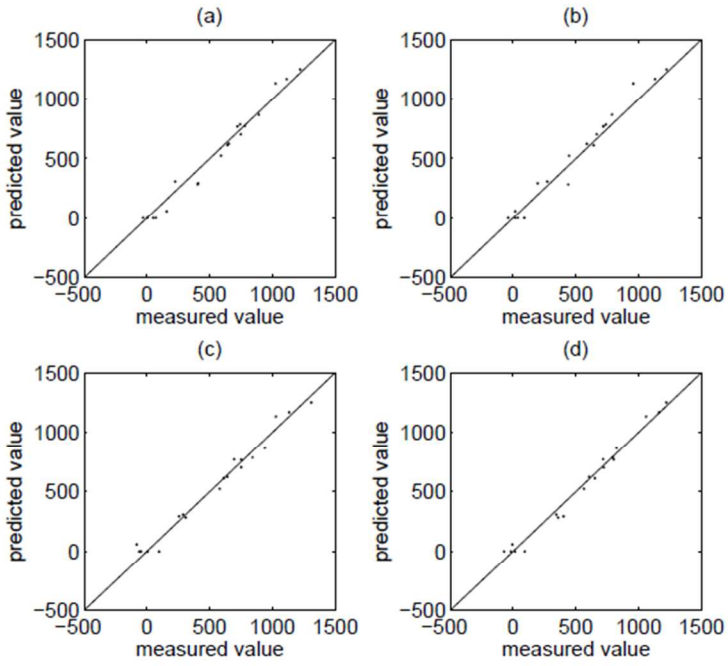


Fig. 4: Predicted value vs. measured value scatter diagram of different methods for SO2. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

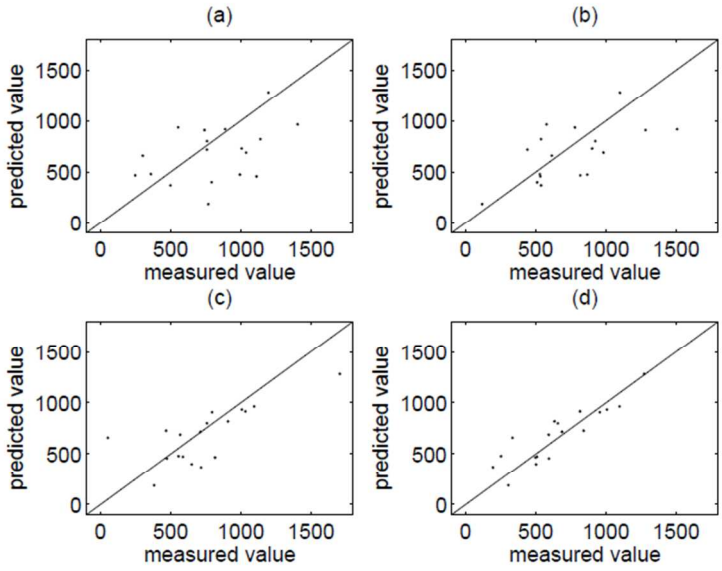


Fig. 5: Predicted value vs. measured value scatter diagram of different methods for NO2. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

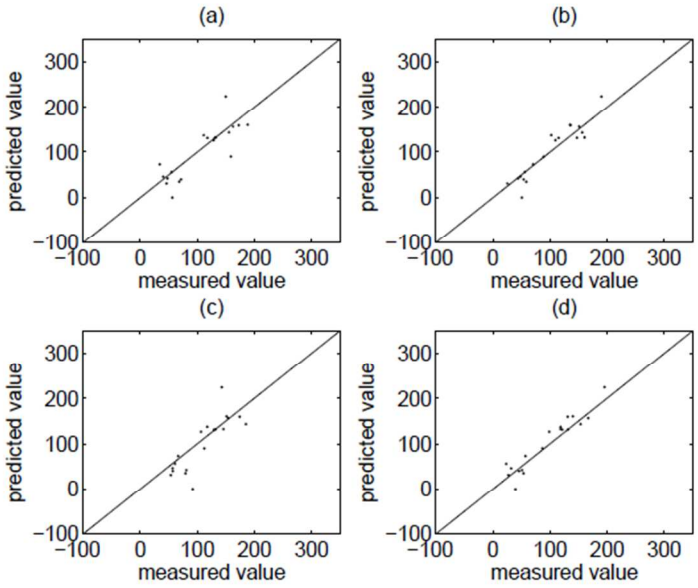


Fig. 6: Predicted value vs. measured value scatter diagram of different methods for NO. (a) PLS. (b) UVE-PLS. (c) GA-PLS. (d) IT-PLS.

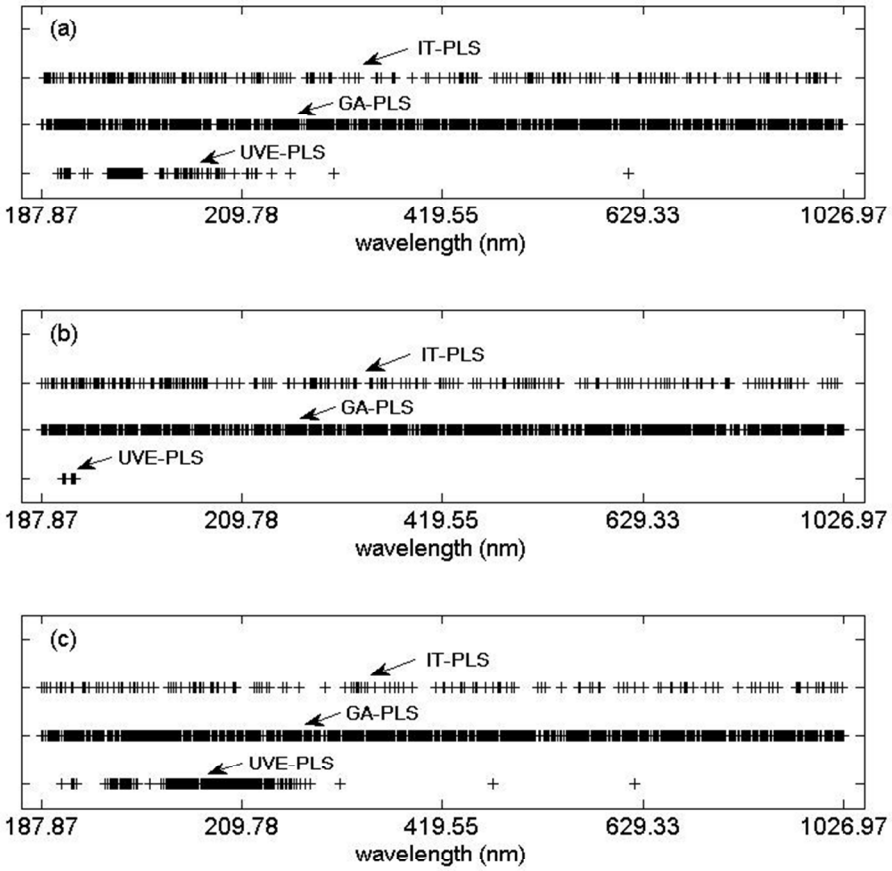


Fig. 7: Variable selection results for three components of the flue gas. (a) SO2. (b) NO2. (c) NO.