Special Issue Article

Journal of CHEMOMETRICS

Received: 26 July 2009, Revised: 9 January 2010, Accepted: 22 January 2010, Published online in Wiley InterScience: 27 May 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1300

Model population analysis for variable selection

Hong-Dong Li[a], Yi-Zeng Liang[a]*, Qing-Song Xu[b] and Dong-Sheng Cao[a]

To build a credible model for given chemical or biological or clinical data, it may be helpful to first get somewhat better insight into the data itself before modeling and then to present the statistically stable results derived from a large number of sub-models established only on one dataset with the aid of Monte Carlo Sampling (MCS). In the present work, a concept model population analysis (MPA) is developed. Briefly, MPA could be considered as a general framework for developing new methods by statistically analyzing some interesting parameters (regression coefficients, prediction errors, etc.) of a number of sub-models. New methods are expected to be developed by making full use of the interesting parameter in a novel manner. In this work, the elements of MPA are first considered and described. Then, the applications for variable selection and model assessment are emphasized with the help of MPA.

Supporting information may be found in the online version of this article

Keywords: model population analysis; variable selection; Monte Carlo sampling; biomarker discovery

Live with the data before you plunge into modeling.

Leo Breiman, Statistical Science, 2001(16), Page 201

1. INTRODUCTION

Statistical learning and modeling plays a central role in the fields of data mining and artificial intelligence, intersecting with many disciplines. By learning from data in a supervised or unsupervised way, one can derive a mathematical model which in itself possesses some useful information on the patterns or trends of the data. For an established model, what concerns us more is, in most of the cases, it is not the fitting ability [1] but the prediction ability which indicates to what extent the model can generalize well when fed with new samples that have never been seen by the model. To obtain a well-generalized model, the number of variables included in the model is, from the parsimonious perspective, required to be as small as possible because overfitting may be brought about if a model contains too many redundant and/or uninformative variables.

Recently in the 11th Scandinavian Symposium on Chemometrics (SSC11), Professor Bro performed a live simulation experiment and got very counter-intuitive result. He first randomly produced a data matrix of size 100×200. Then 50 samples are also randomly chosen and assigned the class label '0', while the remaining 50 samples are assigned the class label '1'. Finally, partial least squares linear discriminant analysis (PLSLDA) [2–5] was employed to classify the two classes of samples. Surprisingly, for each replicate simulation, the two classes of samples are classified very well. The fact that randomly produced binary classification data with no between-class difference can be classified very well aroused the interests of the SSC11 participants significantly. What is the philosophy of the classification? Is it a pitfall or is there anything we did not pay much attention to [6,7]? Is there anything connecting with the overfitting, since, in his example, the number of variables (200) is much larger than that of samples (100).

However, such kind of high-dimensional data are always encountered by us in practice. In the field of OMICS study, such as genomics [8–10], proteomics [11,12] and metabolomics [13,14], the routinely produced analytical data are usually of very high dimension. Let us see some examples first: (1) a microarray experiment on only one gene chip can produce the expression profile of over 1 000 000 genes, (2) the data of a protein mixture sample subjected to MALDI-TOF experiment can be of over 10 000 dimensions and (3) in metabolomics study, the NMR [13] data or liquid chromatography/mass spectroscopy (LC/MS) or Gas chromatography/mass spectroscopy (GC/MS) are also high dimensional. As known, one heated topic in OMICS is to identify the potential biomarker pattern [15–18], which can be used to discriminate the patients from the controls or to qualitatively characterize the physiological state of a patient [19,20]. However, it is usually very difficult to screen out the informative biomarkers from such a large pool of candidates due to the fact that there exists too much redundant and/or interfering information in such kind of data. Therefore, seeking an efficient variable selection method to reduce the redundancy of the data is an immediate need for establishing a model with high performance because

* *Correspondence to: Y.-Z. Liang, Research Center of Modernization of Traditional Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China.*
E-mail: yizeng_liang@263.net

a *H.-D. Li, Y.-Z. Liang, D.-S. Cao*
Research Center of Modernization of Traditional Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China

b *Q.-S. Xu*
School of Mathematic Sciences, Central South University, Changsha 410083, P. R. China

fewer variables could greatly reduce the risk of overfitting. Moreover, how to comprehensively assess the performance of the selected variables is also of great interest [21].

The classical method for variable selection consists of, but is not limited to, forward step-wise, backward elimination and stage-wise methods. Some recently developed methods, such as Lasso [22], least angle regression (LAR) [23] and elastic net (EN) [24], also prove to be effective for variable selection for high-dimensional data. One common feature of these methods is that they try to select a fixed subset of variables for a given dataset, without considering the influence of variation in samples for the selection of variables. Thus, it is possible that there could be a problem due to a random selection of variables. One feasible remedy to this problem is to establish variable selection methods by incorporating the Monte Carlo Sampling (MCS) technique. Based on MCS, some useful methods, e.g. Monte Carlo uninformative variable selection (MCUVE) [25,26] and randomization test (RT) [27] method, are developed and got successful applications. These MCS-incorporated methods mainly work in three steps: (1) randomly drawing a great number of subsets of samples, e.g. 1000 subsets, by sampling with or without replacement from the original training set; (2) building a calibration model for each sub-dataset and (3) collecting the interesting output of all the models for further analysis. Clearly, such variable selection methods share one common feature that they are rooted in the analysis of a 'population' of MCS-derived models, which may be a key factor that accounts for their outstanding performance. Therefore, one may infer that a 'population' of models may contain some comprehensive information on the data. So, it could be expected that more comprehensive results for variable selection or for model assessment could be achieved if one examines the interesting outputs of the 'population' of models with a statistical eye.

Motivated by the idea of statistically analyzing the outputs of MCS-derived 'population' of models, a new concept called model population analysis (MPA) is developed in the present work. It is expected that the MPA-based method could provide some comprehensive insights into the data because it allows for analyzing some interesting outputs of a large number of models in a statistical way. One typical MPA-based method developed for outlier detection could be seen in Reference [28] where outliers could be identified by examining the distribution of prediction errors of each sample. The current work is focused on MPA for variable selection and assessment. The elements of MPA are first generalized and described followed by two applications of MPA to illustrate (1) the necessity of variable selection and (2) the necessity of model comparison by examining prediction errors' distribution.

2. THE ELEMENTS OF MPA

To give an overview of MPA, the outline of MPA is first introduced which is shown in Figure 1. It works mainly in three successive steps: (1) obtain a sub-dataset by MCS; (2) establish a sub-model for each sub-dataset; (3) statistically analyze some interesting outputs of all the sub-models.

2.1. Monte Carlo sampling for a sub-dataset

Sampling is a key tool in statistics which allows data analysts to repeatedly sample, with or without replacement, from the original dataset to create replicate datasets from which the interesting unknown parameters could be estimated. Generally, sampling can be divided into two types. One is sampling without replacement. The other is sampling with replacement, which has another more popular name 'bootstrap'.

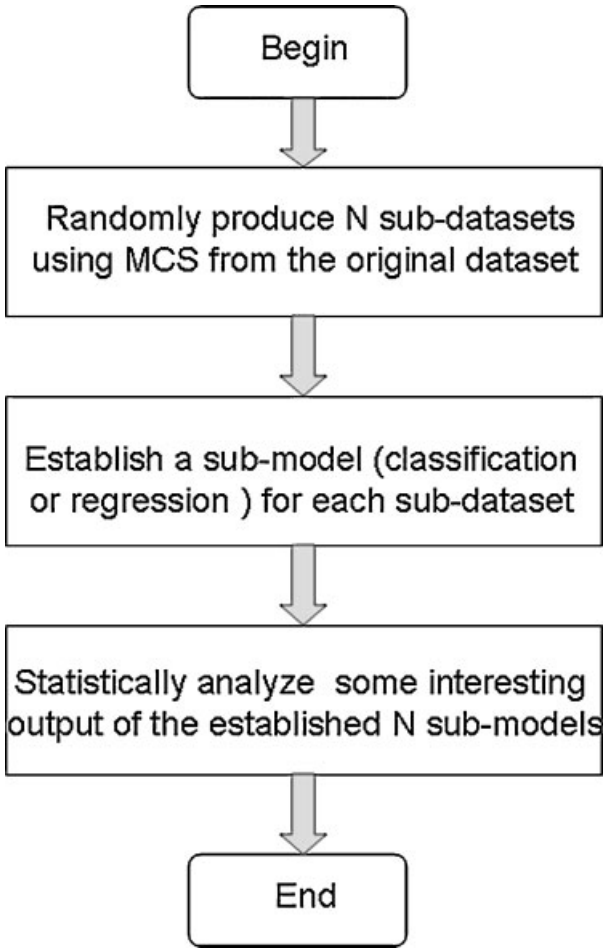

Figure 1. The outline of MPA. Briefly, MPA consists of three steps: (1) obtain a sub-dataset by MCS; (2) establish a sub-model for each sub-dataset; (3) statistically analyze some interesting output of all the sub-models. *N* is the predefined integer, e.g. 1000.

For a given dataset (**X**, **y**), suppose the design matrix **X** contains *m* samples in the rows and *p* variables in the columns and the response vector denoted by **y** is of order $m \times 1$. The number of MCS for an algorithm is set to *N*. With such a setting, one can draw *N* sub-datasets from *N* MCS using either sampling without replacement or bootstrap strategy. The sampled *N* sub-datasets are denoted by $(\mathbf{X}_{sub}, \mathbf{y}_{sub})_i$, $i = 1, 2, 3, \ldots, N$. Note that sampling only serves as a tool for drawing sub-datasets in the context of MPA. The key point of MPA lies in the third element which will be discussed below.

2.2. Establishing a sub-model for each sub-dataset

For each sub-dataset $(\mathbf{X}_{sub}, \mathbf{y}_{sub})_i$, one can establish a sub-model, say $f_i(\mathbf{X})$. Then, the collection of all the sub-models can be denoted as

$$\mathbf{C} = (f_1(\mathrm{X}), f_2(\mathrm{X}), f_3(\mathrm{X}), \ldots, f_N(\mathrm{X})) \quad (1)$$

It can be inferred that each sub-model may provide some local information on the data because it is built only on a part of samples (maybe we can call it a data window) drawn from the

whole dataset. Therefore, all the sub-models may jointly provide some comprehensive information on the whole data.

2.3. Statistically analyzing some interesting outputs of all the sub-models

After training a 'population' of sub-models, we come to the most important point of MPA that how can one statistically analyze the outputs of all the sub-models to extract some information for achieving some special goal, e.g. outlier detection or variable selection. It should be admitted first that it is really difficult for us to give a clear outline on how to aggregate and analyze the interesting outputs of each sub-model because it is deemed to be a problem-dependent procedure. That is to say, one should make different strategies to analyze the outputs of all the sub-models when faced with different problems. The readers are referred to our previous work [28] where the proposed strategy for outlier detection, based on statistically analyzing the distribution of the prediction errors of each sample, could be well reformulated into the framework of MPA.

3. DATASETS

3.1. Simulated data

In order to investigate the rationale behind the unbelievable classification result obtained by PLSLDA from the design of Prof. Bro, 10 groups of datasets of the same number of samples but of 10 different dimensions are randomly produced as Prof. Bro did. Each dataset contains 100 samples. The number of variables of each dataset is set to 5, 10, 20, 50, 100, 200, 500, 1000, 2000 and 5000. The class label for each sample is randomly assigned. For each dimension, 100 replicates are simulated to avoid coincidence and obtain the statistically stable results based on MPA. This study is aimed at investigating how the classification performance of PLSLDA is influenced by the dimension by systematically analyzing the 'population' of models established on the datasets of different sizes.

3.2. Colorectal cancer data

The colorectal cancer data include 64 cancer and 48 control samples. The measured MALDI-TOF serum protein profiles contain 16331 *m/z* values covering the domain of 960–11163 Dalton. The mass spectra for all the 112 samples are shown in Figure 2. See Reference [29] for detailed information on these data.

4. RESULTS AND DISCUSSION

4.1. Simulated data

Theoretically speaking, the misclassification error of any classifier on all the simulated datasets should be about 0.50 due to the fact that there is no between-class difference for each variable. For the balanced datasets simulated in the present work, any classifier which achieves a fitted error significantly lower than 0.5 is overfitted. But how can one detect whether the classifier is overfitted? In the present study, Monte Carlo Cross Validation (MCCV) [30] is taken to detect whether the built PLSLDA classifier is overfitted or not because MCCV possesses some advantages, e.g. asymptotical consistence, compared to ordinary cross validation.

For each dimension, the MCS technique is used to first randomly obtain a 'population' of datasets (100 replicate datasets here). Then for each replicate dataset, MCCV is taken to determine the optimal number of latent variables (nLV). The nLV, MCCV error and Q^2 are recorded. Then PLSLDA is utilized to classify the data. The fitting errors, defined as the ratio of the incorrectly classified samples to the total samples, and the corresponding R^2 are also recorded. Finally, the mean value and standard deviation for each performance parameter of the 100 replicate data of the same dimension are calculated.

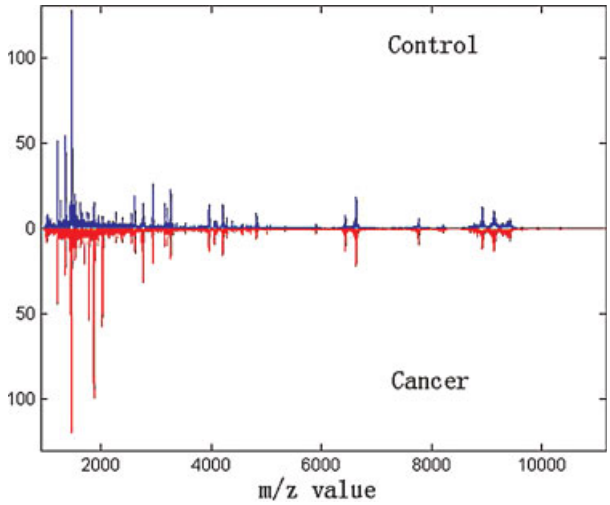

Figure 2. The MALDI-TOF serum profiles of control and cancer samples.

Table I presents all the results of the 10 different groups of datasets. From Table I, it can be found that the fitting errors/R^2 increase monotonously when adding the dimension. This is an indication that the data are becoming more and more seriously overfitted as the number of variables increases. A more important finding from this study is that overfitting will obviously decrease if the variable to sample ratio is smaller than 1:3, which is in agreement with the empirical rule in variable selection that 'the number of samples should be at least three times larger than that of variables'. Therefore, the results may serve as numerical evidence of the rule. Meanwhile, attention should be paid to the nLV used in the PLSLDA model. The nLV for all the data is very small, about 2 to 3. This fact may reflect that PLSLDA is so powerful for extracting the **y**-correlated latent variable that overfitting can easily occur.

To intuitively view how overfitting occurs when the dimension increases, the 2D score plots of six different dimensional data are shown in plots A, B, C, D, E and F of Figure 3, respectively. Apparently, the intrinsically inseparable data can be classified very well as long as they contain sufficient variables. Specifically in the case of dimension equal to 2500, all the samples can be classified into only one latent variable direction. It may be inferred from our results that including too many variables into a model is dangerous in that overfitting may be caused. Therefore, it is recommended that variable selection should be done before plunging into modeling your data.

Besides, one can see from Table I that the MCCV error for all the dataset is equal to or slightly lower than 0.50. This result demonstrates that MCCV is an effective method for estimating the prediction error and further for model assessment although the estimated error is somewhat optimistic. It should be noted that probably ordinary cross validation, e.g. leave-one-out CV, would work as well for the simulated datasets. However, MCCV possesses some advantages, such asymptotical consistency, over the ordinary cross validation [30].

Table I. The results on the simulated datasets using PLS-LDA with the number of MCS set to 100

Dimension	Fitting errors	R^2	MCCV errors	Q^2	nLV
5	0.41 ± 0.04	0.04 ± 0.03	0.49 ± 0.06	0.01 ± 0.02	2.01 ± 1.18
10	0.38 ± 0.04	0.07 ± 0.04	0.51 ± 0.05	0.01 ± 0.02	1.97 ± 1.16
20	0.31 ± 0.04	0.15 ± 0.06	0.49 ± 0.05	0.01 ± 0.01	2.22 ± 1.22
50	0.20 ± 0.06	0.37 ± 0.14	0.49 ± 0.05	0.01 ± 0.02	2.34 ± 1.33
100	0.08 ± 0.07	0.71 ± 0.23	0.48 ± 0.05	0.01 ± 0.01	2.40 ± 1.29
200	0.04 ± 0.04	0.87 ± 0.16	0.49 ± 0.05	0.01 ± 0.01	2.32 ± 1.30
500	0.01 ± 0.01	0.98 ± 0.04	0.49 ± 0.05	0.01 ± 0.02	2.32 ± 1.28
1000	0.00 ± 0.00	1.00 ± 0.01	0.49 ± 0.06	0.01 ± 0.02	2.11 ± 1.21
2000	0.00 ± 0.00	1.00 ± 0.00	0.49 ± 0.05	0.01 ± 0.01	2.16 ± 1.18
5000	0.00 ± 0.00	1.00 ± 0.00	0.49 ± 0.05	0.01 ± 0.02	1.91 ± 1.09

Summing up, by systematically analyzing the 'population' of models established on the datasets of different sizes, it is illustrated that variable selection is necessary for avoiding overfitting, especially when dealing with the high-dimensional data which are usually faced by the OMICS practitioners. Also, the appropriate variable to sample ratio is indicated by this study.

4.2. Colorectal cancer data

These proteomic data are utilized to demonstrate that better performance could be achieved by only including a subset of variables when modeling. In the present work, two methods are employed to perform variable selection coupled with PLSLDA. One is MCUVE, the other is the recently proposed variable

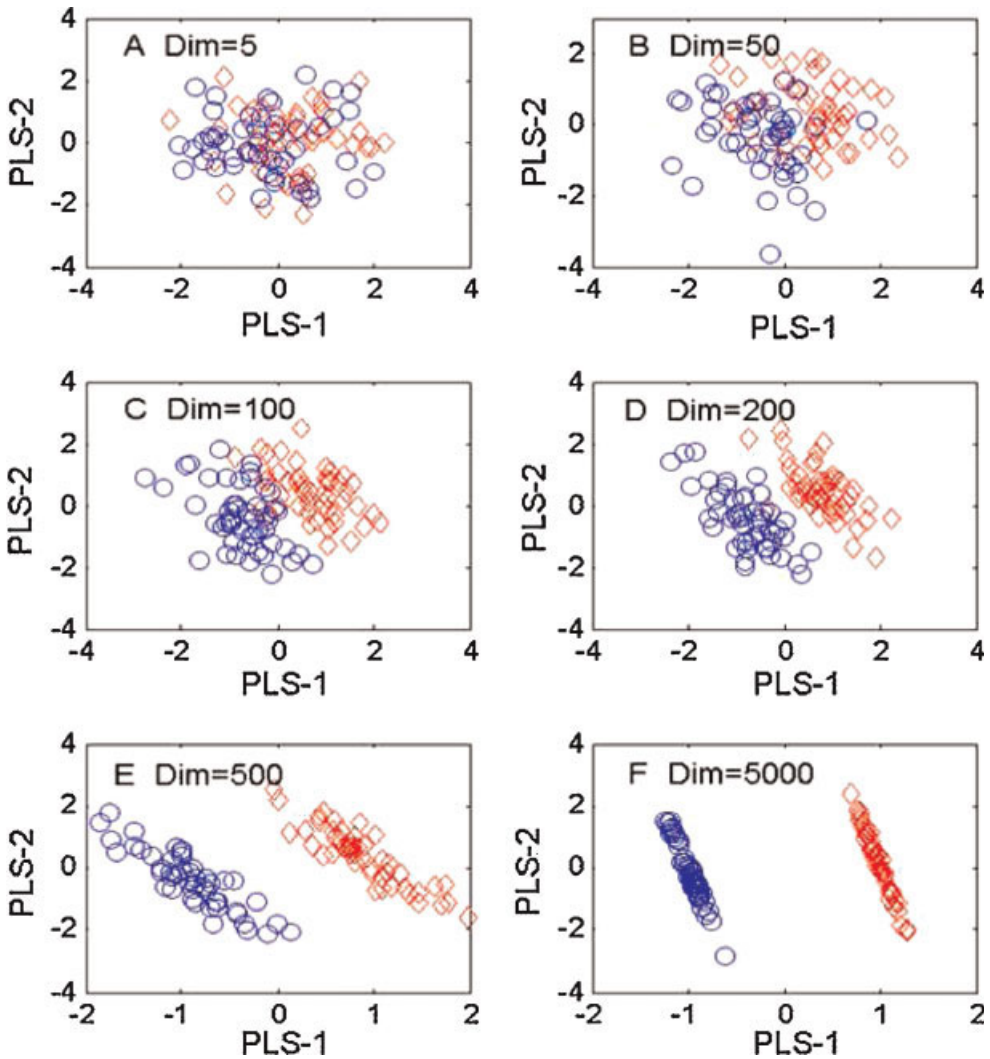

Figure 3. The score plots for the simulated data of six-different dimensions using PLSLDA.

selection procedure, called competitive adaptive reweighted sampling (CARS) [31]. To begin with, the algorithms for both MCUVE-PLSLDA and CARS-PLSLDA are briefly introduced first.

MCUVE-PLSLDA works mainly in three steps: (1) randomly draw N sub-datasets from the original data; (2) build a PLSLDA model for each sub-dataset and (3) collect the variable coefficients (interesting parameters) of all the N models and calculate a Reliability Criterion (RC) value for each variable. Finally the RC value, serving as a variable importance index, is taken to rank the variables.

Assuming that the number of sampling runs is set to N, then in the ith sampling run, CARS-PLSLDA works in the following three sequential steps:

(1) A predefined ratio (e.g. 80%) of the samples is randomly chosen to build a PLSLDA model using the retained variables in the $(i-1)$th sampling run.
(2) Based on the regression coefficients of the constructed model, only a ratio, denoted as r_i, of variables with large absolute coefficients is first retained. Here, r_i is computed using the exponentially decreasing function (EDF) $r_i = ae^{-ki}$, where a and k are two constants, which can be automatically determined. Then, adaptive reweighted sampling (ARS) technique is employed to further remove some uncompetitive variables from the EDF-based retained subsets. Finally, a subset of variables, denoted as V_i, is selected.
(3) Compute the misclassification error of V_i using K-fold cross validation.

Repeat the above procedure for N times, CARS-PLSLDA can sequentially select N subsets of variables, i.e. $(V_1, V_2, \ldots, V_n)$, from N MCS runs in an iterative and competitive manner. However, it should be mentioned that CARS-PLSLDA, like MCUVE-PLSLDA, also suffers from the disadvantage that it cannot guarantee to reproduce the result with the same tuning parameters. The source codes for implementing CARS-PLSLDA in both MATLAB and R (for Linux and Windows) can be freely available at: http://code.google.com/p/cars2009/.

First, MCCV, taking all the 16331 m/z values as input, is employed to evaluate the prediction ability of the PLSLDA classifier and simultaneously determine the optimal number of latent variables (nLV). Then, the PLSLDA classifier is established to distinguish the cancer samples from the controls. The sensitivity (Se), specificity (Sp) and overall accuracy (Acc) of both fitting and MCCV are presented in Table II. It could be seen that all these performance parameters, including Se, Sp and Acc, are relatively high, which might be an indication that the data are not overfitted or are only overfitted to a slight extent. The reason might be that (1) the full MALDI-TOF spectrum includes highly discriminating variables and (2) PLS has the intrinsic capability to extract the **y**-relevant latent variable while simultaneously suppressing the interference brought about by the uninformative or noisy variables. But whether a more parsimonious also with better performance model can be found by variable selection?

Here, MCUVE-PLSLDA and CARS-PLSLDA are applied to select the potential discriminating variables, respectively. All the performance parameters are presented in Table II. By MCUVE-PLSLDA, 17 variables are selected. Although the fitting results based on the 17 variables are equal to those of the full spectrum model, the predictions are improved. Using CARS-PLSLDA, only five variables are selected from such a large pool of 16 331 candidates. Clearly, both MCUVE-PLSLDA and CARS-PLSLDA outperform the full spectrum model in terms of Se, Sp and Acc. The variable subset selected by CARS is more discriminating, which may imply that the full spectrum contains many uninformative variables and these variables have negative effect on the prediction ability of the model. Moreover, compared to the results reported by Alexandrov *et al.* [29] (Acc: 0.973, Se: 0.984 and Sp: 0.958), our results are also better. Considering the high discriminating ability, the five variables may jointly serve as biomarker pattern to facilitate the diagnosing of the disease.

Further, in order to comprehensively compare the performance of the full spectrum model, MCUVE-PLSLDA and CARS-PLSLDA, 500 sub-datasets (95% of all the 112 samples) are first randomly produced. For each sub-dataset, the prediction error based on MCCV is computed. Then, the histograms of MCCV-based prediction ability in terms of accuracy, sensitivity and specificity are calculated and shown in Figure 4. For the sake of robustness, the median and the difference between 0.25-quantile and 0.75-quantile of MCCV-based prediction accuracy, sensitivity and specificity for the three cases are given in pair: 0.943(0.006), 0.943(0.008), 0.944(0.007) for full spectrum classifier, 0.959(0.005), 0.970(0.007), 0.946(0.008) for the MCUVE-PLSLDA and 0.987(0.002), 0.997(0.002), 0.975(0.004) for CARS-PLSLDA, respectively. Obviously, the performance of the PLSLDA classifier using the variables selected by MCUVE and CARS is improved. The result further implies that conducting variable selection before modeling is very necessary for building a model with better performance.

Note that the distributions of Acc, Se and Sp for different methods overlap to some extent (especially the distribution of Sp between all variable models and MCUVE), which suggests that wrong conclusions may be obtained by chance if one compares the performance of different methods based on only one sub-dataset or a single splitting of the data. However, this problem could be overcome by examining the distribution of prediction errors resulted from a 'population' of sub-models because the distribution is statistically stable. Therefore, it might be more appropriate to perform model assessment or comparison by examining the distribution of prediction errors.

Table II. The results on the MALDI-TOF serum protein profile data using MCUVE-PLSLDA and CARS-PLSLDA

methods	nVAR	Fitting			MCCV			nLV
		Acc	Se	Sp	Acc	Se	Sp	
PLSLDA	16331	0.973	0.984	0.958	0.946	0.945	0.947	3
MCUVE-PLSLDA	17	0.973	0.984	0.958	0.962	0.970	0.951	3
CARS-PLSLDA	5	0.991	1.000	0.979	0.988	0.998	0.976	2

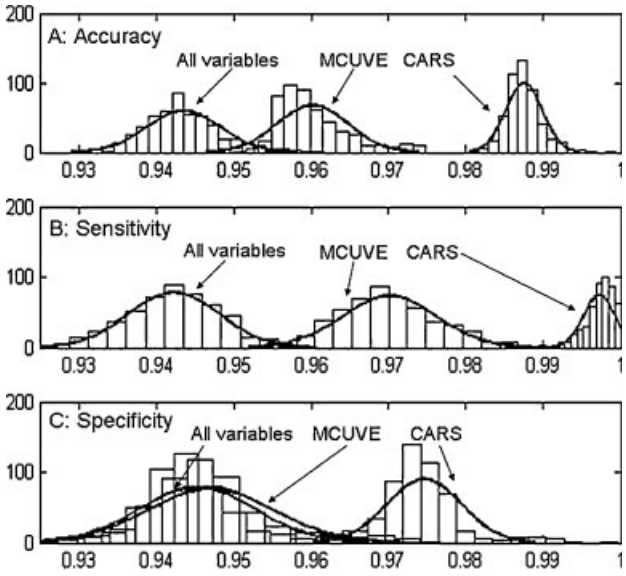

Figure 4. The distribution of prediction errors resulted from MCCV on 500 sub-datasets in terms of accuracy, sensitivity and specificity for full spectrum, MCUVE-PLSLDA and CARS-PLSLDA, respectively.

5. CONCLUSIONS

In the present work, a new concept MPA is developed and the elements of MPA are introduced first followed by two case studies. The necessity of variable selection is demonstrated in the simulation study. More importantly, the simulation study suggests that overfitting will obviously decrease if the variable to sample ratio is smaller than 1:3, which is in agreement with the empirical rule in variable selection that 'the number of samples should be at least three times larger than that of variables'. The proteomic study indicates that it may be more appropriate to perform model comparison by examining the distribution of predictive errors resulted from a large number of randomly produced sub-dataset. Our later work is focused on establishing new variable selection methods by strictly implementing the idea of MPA. It is expected that MPA will find its applications in the fields of variable selection and model assessment.

Acknowledgements

This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 20875104 and No. 10771217), the International Cooperation Project on Traditional Chinese Medicines of Ministry of Science and Technology of China (Grant No. 2007DFA40680). The studies were performed with the approval of the university's review board.

REFERENCES

1. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* 2003; **43**: 579–586.
2. Barker M, Rayens W. Partial least squares for discrimination. *J. Chemom.* 2003; **17**: 166–173.
3. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab.* 2001; **58**: 109–130.
4. De Jong S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab.* 1993; **18**: 251–263.
5. Yi L-Z, He J, Liang Y-Z, Yuan D-L, Chau F-T. Plasma fatty acid metabolic profiling and biomarkers of type 2 diabetes mellitus based on GC/MS and PLS-LDA. *FEBS Lett.* 2006; **580**: 6837–6845.
6. Cronin MTD, Schultz TW. Pitfalls in QSAR. *J. Mol. Struct: Theochem* 2003; **622**: 39–51.
7. Golbraikh A, Tropsha A. Beware of q2! *J. Mol. Graph. Model.* 2002; **20**: 269–276.
8. Tyers M, Mann M. From genomics to proteomics. *Nature* 2003; **422**: 193–197.
9. Teramoto R, Aoki M, Kimura T, Kanaoka M. Prediction of siRNA functionality using generalized string kernel and support vector machine. *FEBS Lett.* 2005; **579**: 2878–2882.
10. Lavine BK, Davidson CE, Rayens WS. Machine learning based pattern recognition applied to microarray data. *Comb. Chem. High Scr.* 2004; **7**: 115–131.
11. Domon B, Aebersold R. Review-Mass spectrometry and protein analysis. *Science* 2006; **312**: 212–217.
12. Nesvizhskii AI, Vitek O, Aebersold R. Aebersold, Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* 2007; **4**: 787–797.
13. Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.* 2000; **484**: 169–174.
14. Kamleh MA, Hobani Y, Dow JAT, Watson DG. Metabolomic profiling of Drosophila using liquid chromatography Fourier transform mass spectrometry. *FEBS Lett.* 2008; **582**: 2916–2922.
15. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom. Intell. Lab.* 2009; **95**: 35–48.
16. Liu QZ, Sung AH, Qiao MY, Chen ZX, Yang JY, Yang MQ, Huang XD, Deng YP. Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics* 2009; **10**.
17. Niwa T. Biomarker discovery for kidney diseases by mass spectrometry. *J. Chromatogr. B* 2008; **870**: 148–153.
18. Oh JH, Lotan Y, Gurnani P, Rosenblatt KP, Gao J. Prostate cancer biomarker discovery using high performance mass spectral serum profiling. *Comput. Meth. Prog. Bio.* 2009; **96**: 33–41.
19. Abdel-Aal RE. GMDH-based feature ranking and selection for improved classification of medical data. *J. Biomed. Inform.* 2005; **38**: 456–468.
20. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. *Artif. Intell. Med.* 2007; **41**: 251–262.
21. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* 2005; **33**: 5914–5923.
22. Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 1996; **58**: 267–288.
23. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann. Stat.* 2004; **32**: 407–499.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 2005; **67**: 301–320.
25. Centner V, Massart D-L, de Noord OE, de Jong S, Vandeginste BM, Sterna C. Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* 1996; **68**: 3851–3858.
26. Cai W, Li Y, Shao X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intell. Lab.* 2008; **90**: 188–194.
27. Xu H, Liu Z, Cai W, Shao X. A wavelength selection method based on randomization test for near-infrared spectral analysis. *Chemom. Intell. Lab.* 2009; **97**: 189–193.
28. Cao D-S, Liang Y-Z, Xu Q-S, Li H-D, Chen X. A new strategy of outlier detection for QSAR/QSPR. *J. Comput. Chem.* 2010; **31**: 592–602.
29. Alexandrov T, Decker J, Mertens B, Deelder AM, Tollenaar RAEM, Maass P, Thiele H. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics* 2009; **25**: 643–649.
30. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemometr. *Chemom. Intell. Lab.* 2001; **56**: 1–11.
31. Li H-D, Liang Y-Z, Xu Q-S, Cao D-S. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 2009; **648**: 77–84.

