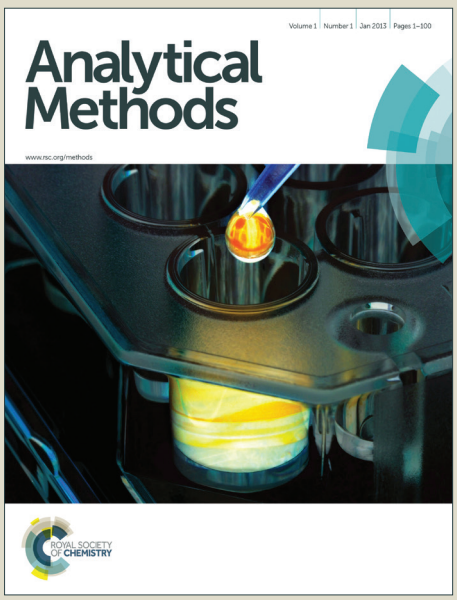
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Application of sparse linear discriminant analysis for metabolomics data

Meilan Ouyang [a], Zhimin Zhang [a], Chen Chen[a], Xinbo Liu[a], Yizeng Liang[a,*]

[a]Research Center of Modernization of Traditional Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China

Abstract

To discover the potential biomarkers that may be closely related to diseases is a major purpose of metabolomics data analysis. Hence, it is expected to explore some effective methods for screening these informative metabolites from large amounts of dataset. In this paper, we propose an effective strategy named sparse linear discriminant analysis (SLDA) which can perform classification and variable selection simultaneously to analyze complicated metabolomics datasets. Compared with two other approaches partial least squares discriminant analysis (PLS-DA) and competitive adaptive reweighted sampling (CARS), SLDA relatively obtains better results and can select some informative metabolites which are proved to be in consistent with the biochemical study. Furthermore, by building a model based on the selected features SLDA could be applied to the high dimensional small sample cases where linear discriminant analysis (LDA) fails to work. To sum up, SLDA is a very useful method to explore and process metabolomics data.

Keywords: sparse linear discriminant analysis; variable selection; classification; metabolomics; potential biomarkers

1. Introduction

Metabolomics is a systemic approach for analyzing metabolites released by living organisms

Abbreviations: SLDA, sparse linear discriminant analysis; PLS-DA, partial least squares discriminant analysis; CARS, competitive adaptive reweighted sampling;

*Correspondence to: Department of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China. Tel.:86-731-88830831. E-mail address: yizeng_liang@263.net.

during the metabolic process [1]. All the low molecular weight metabolites in special physiological period produced by an organism or a cell can be studied qualitatively and quantitatively simultaneously[2]. So the main purpose of metabolomics data analysis is to discover the potential biomarkers that provide information about disease diagnosis and treatment, drug toxicity and new drug development and many other fields [3]. While the information contained in metabolomics datasets becomes more and more complicated with the widely use of advanced analysis instruments such as GC-MS [4, 5], LC-MS [6, 7] and NMR [8] and so on. Hence, it is in sore need of a great variety of analytical tools to screen valuable metabolites from the complex information.

So far, numerous statistical methods and chemometrics approaches like principal component analysis (PCA) [9, 10], partial least squares discriminant analysis (PLS-DA) [11], competitive adaptive reweighted sampling (CARS) [12], subwindow permutation analysis (SPA) [13] and model population analysis random forests (MPA-RF) [14] have been applied in metabolomics work. Whereas, it still exists many problems. For example, supervised approach PLS-DA may perform poorly when the variance or covariance of some features is large in spite of their little contribution to the classification. These uninformative and noise features perhaps lead to the absence of optimal variables in complex cases [15]. For CARS, the weight coefficient of a feature will change when we repeat the operation process, so the variable importance will be unstable. And this may cause the consequence that irrelevant features are chosen while informative features are lost [13].

In order to deal with these issues, we propose an effective method called sparse linear discriminant analysis (SLDA) by performing discriminant analysis with penalty coefficients to get sparsity [16, 17]. The main idea of SLDA is to perform the classification by the selected variables which are obtained via the sparse discriminant vectors. When it is applied to three metabolomics datasets, the results show that SLDA is an effective classification approach and can be also suitable for the situation where the features is more than the samples. Meanwhile, it can screen some potential biomarkers which may be related with the disease.

2. Materials and methods

2.1 Description of metabolomics datasets

In order to compare the performance of three methods PLS-DA, CARS and SLDA, three metabolomics datasets are used. In these models all variables are normalized to have mean zero and standard deviation one. All these three methods are implemented by MATLAB.

Dataset 1: "ESTE" dataset, which is made up of endogenous substrates to enzymes from 6 wild-type mice and 6 mice lacking the enzyme fatty acid amide hydrolase (FAAH), each with 409 variables. The ESTE samples are profiled using a liquid chromatography-mass spectrometry (LC-MS). The dataset is taken from the paper [18].

Dataset 2: "TLS" dataset, which is made up of the urinary metabolite profiles of 25 patients with mild tubulointerstitial lesions and 25 patients with severe tubulointerstitial lesions, each with 200 variables. The urinary samples are obtained from the Department of Nephrology of the University Hospital of Ioannina, and profiled using a ^{1}H NMR. The dataset is taken from the paper [19]. In this article, the author mentioned that "All study participants gave informed consent for the investigation, which was approved by the Ethical Committee of the University Hospital of Ioannina."

Dataset 3: "CHOB" dataset, which is made up of the metabolic profiles of 16 healthy children and 13 overweight children, each with 30 variables. The plasma samples are got from the Xiangya Hospital in Changsha City, Hunan Province, China, and profiled using a GC–MS. The dataset is taken from the paper[20]. In this article, the author made a statement that" All clinical experiments were approved by Xiangya Institutional Human Subjects Committee."

2.2 Linear discriminant analysis

In this study, LDA we mentioned is seen as typical Fisher's discriminant analysis [21]. We define a data matrix $\mathbf{X}_{n\times p}$ which has n observations belonging to one of k classes and p features. And let $\mathbf{x}_{ij}$, j=1,2,3,..., n_i, denote the vector falls into the ith class, set $\overline{\mathbf{x}_i}$ be the mean of the ith

class and $\bar{\mathbf{x}}$ be the mean of the whole data. Then the within-class covariance matrix is

$$\mathbf{\Sigma}_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \quad (1)$$

And the between-class covariance matrix is

$$\mathbf{\Sigma}_b = \sum_{i=1}^{k} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \quad (2)$$

Fisher's discriminant problem is to find appropriate discriminant vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_{k-1}$ which are able to make the between-class covariance matrix is maximal relative to the within-class covariance matrix:

$$\max_{\boldsymbol{\beta}_i}\left\{\boldsymbol{\beta}_i^T \mathbf{\Sigma}_b \boldsymbol{\beta}_i\right\}, \text{ subject to } \boldsymbol{\beta}_i^T \mathbf{\Sigma}_w \boldsymbol{\beta}_i = 1 \quad (3)$$

The problem (3) can be settled by considering it as the eigenvalue problem. Since the upper bound for the rank of the matrix $\mathbf{\Sigma}_b$ is min (p, k-1), so there are no more than q ($q \leqslant$ min(p, k-1)) discriminant vectors. Fisher's discriminant analysis can reduce the dimension of $\mathbf{X}_{n\times p}$ by projecting it onto the q-dimensional space. Hence it is very conducive to the classification and the visualization of the original data matrix.

In this study, we use a series of scoring [22] to transform classification analysis into regression analysis. We define a data matrix $\mathbf{Y}_{n\times k}$ consists of dummy variables, $\boldsymbol{\theta}_i$ is a k-vector of scores for the k classes, $\boldsymbol{\beta}_i$ is a p-vector representing variable coefficients for the p features. So the criterion of the optimal scoring problem is defined as follows:

$$\min_{\boldsymbol{\beta}_i, \boldsymbol{\theta}_i}\left\{\| \mathbf{Y}\boldsymbol{\theta}_i - \mathbf{X}\boldsymbol{\beta}_i \|^2\right\} \quad (4)$$

Since $\boldsymbol{\beta}_i$ in formula (4) is proportional to that in formula (3) [23], so we can see $\boldsymbol{\beta}_i$ as the discriminant vector.

2.3 Sparse linear discriminant analysis

Although LDA is widely used because of its simplicity and predictive ability [24], it fails to work when the features is more than the observations [25]. To cope with this problem, we can

impose a penalty coefficient on the L_1 norm of discriminant vectors [16, 26]. So the lasso is defined as follows:

$$\min_{\boldsymbol{\beta}} \left\{ \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|_1 \right\} \quad (5)$$

When λ is large sufficiently, some values in $\boldsymbol{\beta}$ will be shrunk to zero and the sparse discriminant vector will be obtained. So it can realize variable selection through this way. To better reveal the grouping information of correlated features which lasso couldn't do, we also impose a penalty γ on the L_2 norm of $\boldsymbol{\beta}$ which is named the elastic net [17].

$$\min_{\boldsymbol{\beta}} \left\{ \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \| \boldsymbol{\beta} \|_1 + \gamma \| \boldsymbol{\beta} \|^2 \right\} \quad (6)$$

It is proved to have great ability to select variable and high prediction accuracy [27]. Combining the optimal scoring with the elastic net, the sparse linear discriminant analysis (SLDA) is defined as:

$$\min_{\boldsymbol{\beta}_i, \boldsymbol{\theta}_i} \left\{ \| \mathbf{Y}\boldsymbol{\theta}_i - \mathbf{X}\boldsymbol{\beta}_i \|^2 + \lambda \| \boldsymbol{\beta}_i \|_1 + \gamma \| \boldsymbol{\beta}_i \|^2 \right\} \quad (7)$$

The problem (7) is solved by using an iterative algorithm [23]. Figure 1 shows how exactly the SLDA works.

(Insert Figure 1)

2.4 Partial least squares discriminant analysis

Partial least squares discriminant analysis(PLS-DA) is a discriminant analysis that **y** consists of dummy variables corresponding to the category of the observations in **X**[28]. The original goal of PLS is not for discriminant analysis, but there are connections between PLS and DA which have been discussed in the literature[29]. PLS-DA is widely used for classification and the stability of classification results is also assessed in studies [30]. The PLS-DA is implemented in the following steps. First, we recode the vector **y** with dummy variables which are consistent with the response categories. Then, the PLS model is built with the training set obtained from cross validation. After that, we compute the predicted category variables of the test set through this model. By comparing the predicted category variable and the actual variable we get the error rates of different numbers of components. So, according to the minimum error we can choose the optimal set of components. And the approach realizes dimensional reduction through these components. PLS can be applied to classification in spite of the situations where there are more features than observations.

2.5 Competitive adaptive reweighted sampling

Competitive adaptive reweighted sampling (CARS) is a promising approach for building a predictive calibration model and it can be applied for variable selection of different datasets like genomics data, metabolomics data and so on [12]. The steps of CARS are briefly introduced as below. To begin with, we build a PLS model using the samples chosen by Monte Carlo strategy in every sampling run. This strategy is proposed by Stanislaw Ulam in the late 1940s and it is demonstrated to be a successful method for selecting the suitable model [31]. Then we use the exponentially decreasing function (EDF) to perform the variable selection. The process encompasses the following two steps. Firstly, EDF deletes the features of no or little information which have relatively small weights obtained from the PLS model. At the second step, the ratio of the remaining features can be computed by the formula:

$$r_i = ae^{-ki} \tag{8}$$

Where *a* and *k* are two constants calculated as below:

$$a = (p/2)^{1/(N-1)} \tag{9}$$

$$k = \ln(p/2)/(N-1) \tag{10}$$

Where *p* and *N* are the number of features and the sampling runs. After that, we use the adaptive reweighted sampling (ARS) to select features further. In this procedure, the larger the weight of the feature, the greater probability it will be selected. At last, we compute the root mean squares error of cross validation (RMSECV) of the *N* subsets of features and choose the optimal one which has the lowest RMSECV.

3. Results and discussion

3.1 Classification results calculated by different methods

This section illustrates classification results on three different metabolomics datasets. Because the sizes of these samples are small, so we directly use 10-fold cross-validation to determine the training data and the test data and select the best parameters for the three methods

PLS-DA, CARS and SLDA. Then we build these classification models with the best parameters to evaluate the predictive ability. The selected variables and the results namely accuracy, sensitivity, specificity and AUC values for each method are listed in Table 1. And the receiver operating characteristic (ROC) curves are shown in Figure 2.

(Insert Table 1)

(Insert Figure 2)

As it is described in Table 1, the SLDA method gets the best results as a whole. Then, also from the Table 1 one can firstly see that the models with feature selection can get better results than those which are established without choosing variables. Because the prediction assessment parameters including the accuracy, sensitivity, specificity got by PLS-DA is not as good as CARS and SLDA. Furthermore, from the ROC curves shown in Figure 2, one can see that the area under the curve (AUC) obtained by PLS-DA is smaller than the other two methods. As for ROC curves, a model whose AUC values is 0.5 has no predictive ability just like random guess. The closer AUC value is to 1, the better is the predictive ability. So it can be seen from the above description, CARS and SLDA can obtain better results than PLS-DA. It is worth noting that SLDA seems to be more effective comparatively compared with CARS. In Table 1, even though the accuracy, sensitivity, specificity and AUC values computed by these two methods for the ESTE and TLS datasets are the same, SLDA is more stable and can select more appropriate variables than the CARS (see the next section for more detail). For the CHOB dataset, SLDA achieves a higher predictive accuracy, specificity and AUC values than CARS. Consequently, the SLDA method has better discriminant ability relatively.

What's more, SLDA can obtain satisfied classification results for the datasets of the features being much more than the observations[25]. As it is known to all, some methods such as LDA can lead to satisfied results in low dimensional case but it fails to have good classification performance when the number of samples is less than the number of features relatively. However, the SLDA method performs well in the high dimensional small sample case. Since the SLDA model can get a low dimensional representation of the original dataset without missing much useful information by means of electing some important variables for classification and exclude many unrelated features. As is shown in Table 1, the three metabolomics datasets which have more features than samples especially in the ESTE and TLS dataset the number of variables is very

large relative to samples obtain good classification results. In conclusion, SLDA is a powerful feature selection and classification method.

3.2Comparison of the models' stability

As is mentioned above, the CARS method sometimes can achieve satisfied results but it is unstable. The reason is that the model selects samples randomly, so it obtains different variables in different runs. The significance of each variable will also change. When CARS runs a few times, the results are different and we may not get the best result through the running process. Table 2 shows three results for CHOB dataset by running the CARS program three times with the same parameters. And the corresponding feature importance is also shown in Figure 3. From both Table 2 and Figure 3, one can see that the variables chosen by CARS are different at different times. And the importance of some variables will increase and others will decrease. Because of the instability of variable selection caused by the CARS method, the results in Table 1 and Table 2 are the optimum values by running the CARS program with the number of Monte Carlo Sampling is 1000.

(Insert Table 2)

(Insert Figure 3)

When the CARS program is implemented by running much more times, we can get satisfied results. But the procedure is time-consuming. For example, running the CARS program 1000 repeats and 200 runs in a repeat for CHOB dataset which consists of a matrix X of size 29×30 will take 3378.936s. As the dataset grows larger, the procedure becomes increasingly complex and it also takes more and more time. The metabolomics datasets become more and more complicated and high-throughputs with the widely use of advanced analysis instruments. So the CARS method is time consuming and leads to low efficiency when we run it thousands repeats to obtain distribution of each selected variables.

Compared with the CARS method, the SLDA method is more stable in the whole process. The reason is that when the data matrix of the variables is singular, the beta value won't be an exact number by means of the straight inversion. However, the beta value will be stable with the use of the penalty coefficients to get sparsity. Therefore the SLDA model can obtain a stable result

and take less time, which is conducive to the analysis of the metabolomics datasets.

3.3 Biomarkers of the metabolomics datasets

In this part, we use the variable importance of each variable which is calculated as the absolute value of the sparse coefficient divide by the sum of all variables to analyze the variable selection of classification results and the biomarkers of the metabolomics datasets. The informative variables are selected as biomarkers and other variables are seen as disturbing variables and uninformative variables. The results obtained by CARS and SLDA are shown in Figure 4 and 5 respectively.

(Insert Figure 4 and 5)

As is shown in figures 4 and 5, for the ESTE dataset, the potential biomarker selected by both the CARS and SLDA method is Anandamide. In the TLS dataset, the informative metabolites discovered by CARS are Citrate and variable 32 which is unidentified in the biochemical work. While we perform SLDA to obtain the possible biomarkers: 1-Methylhistidine, Citrate and Proteins. For the CHOB dataset, three potential biomarkers Glyceric acid, Serine and Tyrosine are selected by CARS. But SLDA shows different metabolites which are Glyceric acid and Palmitic acid. Since the SLDA method can get better results including accuracy, sensitivity, specificity and AUC values relatively, it is reasonable to deduce that the potential biomarkers selected by SLDA will be more precise. As it turns out, the potential biomarkers identified by SLDA are in accordance with the conclusion got by biochemical study.

For the ESTE dataset, SLDA provides the potential biomarker which is Anandamide. The study [18], reported that Anandamide is the endogenous substrate to fatty acid amide hydrolase (FAAH). Both targeted method using selected ion monitoring (SIM) and untargeted method termed discovery metabolite profiling (DMP) can demonstrate that Anandamide is the endogenous substrate. In the TLS dataset, we by means of SLDA to obtain the potential biomarkers: 1-Methylhistidine, Citrate and Proteins. The research [19], reported that the start of the tubulointerstitial lesions is characterized by decreased excretion of Citrate, and Proteinuria often is a feature of renal patients. So according to these metabolites it is able to distinguish the patients with mild tubulointerstitial lesions and the patients with severe tubulointerstitial lesions. For the

CHOB dataset, Glyceric acid and Palmitic acid are selected as possible biomarkers by SLDA. The study [20], reported that several phosphate derivatives of Glyceric acid are significant biochemical intermediates of lipid metabolism. These derivatives such as 2-phosphoglyceric acid and 3-phosphoglyceric acid are relative to overweight have been demonstrated [32, 33]. Palmitic acid is discovered by Edmond Fremy in palm oil [34], and it is demonstrated to cause the brain insulin resistance then lead to obesity [35].

Hence, these three results further imply that SLDA can screen biologically meaningful biomarker and build reasonable and predictive models with better performance.

4. Conclusion

In this article, we introduce an effective method called sparse linear discriminant analysis (SLDA) to discover informative metabolites in complicated metabolomics datasets by performing variable selection and classification simultaneously. From the above description, one can see that SLDA have obtained satisfied classification results comparatively. Introducing sparsity by means of the penalty coefficients in the discriminant vectors is very useful to select some biologically meaningful features. Hence the SLDA algorithm can build stable and predictive classification model, which can avoid the overfitting problem. What's more, in this way SLDA can get a very good performance in the high dimensional small sample case where the number of variables is very large relative to the samples. In addition, the potential biomarkers identified by SLDA during the variable selection procedure are of good correlation with the biochemical study and also in accordance with the conclusion reported in the literatures. Therefore, SLDA can screen biologically meaningful biomarker and build reasonable and predictive models with better performance, which have broad applications in metabolomics.

Acknowledgment

This work is financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 21075138, Grants No. 21105129, Grants No. 21175157, Grants No. 21275164

and Grants No. 21305163), China Hunan Provincial science and technology department (Grants No. 2012FJ4139), Hunan Provincial Natural Science Foundation of China (Grants No. 14JJ3031). The studies meet with the approval of the university's review board. We are grateful to all employees of this institute for their encouragement and support of this research.

References

1 J. K. Nicholson, J. C. Lindon and E. Holmes, *Xenobiotica*, 1999, **29**, 1181.

2 R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell, *Trends Biotechnol*, 2004, **22**, 245.

3 J. K. Nicholson, J. Connelly, J. C. Lindon and E. Holmes, *Nat Rev Drug Discov*, 2002, **1**, 153.

4 H. Kanani, P. K. Chrysanthopoulos and M. I. Klapa, *J Chromatogr B Analyt Technol Biomed Life Sci*, 2008, **871**, 191.

5 A. Garcia and C. Barbas, *Methods Mol Biol*, 2011, **708**, 191.

6 B. Zhou, J. F. Xiao, L. Tuli and H. W. Ressom, *Mol Biosyst*, 2012, **8**, 470.

7 S. Becker, L. Kortz, C. Helmschrodt, J. Thiery and U. Ceglarek, *J Chromatogr B Analyt Technol Biomed Life Sci*, 2012, **883-884**, 68.

8 D. S. Wishart, *TrAC Trends in Analytical Chemistry*, 2008, **27**, 228.

9 Z. Ramadan, D. Jacobs, M. Grigorov and S. Kochhar, *Talanta*, 2006, **68**, 1683.

10 G. Nyamundanda, L. Brennan and I. C. Gormley, *BMC Bioinformatics*, 2010, **11**, 571.

11 E. Szymanska, E. Saccenti, A. K. Smilde and J. A. Westerhuis, *Metabolomics*, 2012, **8**, 3.

12 H. Li, Y. Liang, Q. Xu and D. Cao, *Anal Chim Acta*, 2009, **648**, 77.

13 H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu and D.-S. Cao, *Metabolomics*, 2010, **6**, 353.

14 J. H. Huang, J. Yan, Q. H. Wu, M. Duarte Ferro, L. Z. Yi, H. M. Lu, Q. S. Xu and Y. Z. Liang, *Talanta*, 2013, **117**, 549.

15 D. Yuan, Y. Liang, L. Yi, Q. Xu and O. M. Kvalheim, *Chemometrics and Intelligent Laboratory Systems*, 2008, **93**, 70.

16 R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, **58**, 267.

17 H. Zou and T. Hastie, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, **67**, 301.

18 A. Saghatelian, S. A. Trauger, E. J. Want, E. G. Hawkins, G. Siuzdak and B. F. Cravatt, *Biochemistry*, 2004, **43**, 14332.

19 N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos and E. T. Bairaktari, *Journal of proteome research*, 2007, **6**, 3760.

20 M. Zeng, Y. Liang, H. Li, M. Wang, B. Wang, X. Chen, N. Zhou, D. Cao and J. Wu, *J Pharm Biomed Anal*, 2010, **52**, 265.

21 R. A. FISHER, *Annals of eugenics*, 1936, **7**, 179.

22 T. Hastie, A. Buja and R. Tibshirani, *The Annals of Statistics*, 1995, **23**, 73.

23 L. Clemmensen, T. Hastie, D. Witten and B. Ersbøll, *Technometrics*, 2011, **53**, 406

24 J. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, *Pattern Recognition Letters*, 2005, **26**, 181.

25 J. Shao, Y. Wang, X. Deng and S. Wang, *The Annals of Statistics*, 2011, **39**, 1241.

26 H. Zou, *Journal of the American Statistical Association*, 2006, **101**, 1418.

27 Q. Li and N. Lin, *Bayesian Analysis*, 2010, **5**, 151.

28 M. Perez-Enciso and M. Tenenhaus, *Hum Genet*, 2003, **112**, 581.

29 M. Barker and W. Rayens, *Journal of chemometrics*, 2003, **17**, 166.

30 D. V. Nguyen and D. M. Rocke, *Bioinformatics*, 2002, **18**, 39.

31 Q. S. Xu and Y. Z. Liang, *Chemometrics and Intelligent Laboratory Systems*, 2001, **56**, 1.

32 M. W. Hulver, J. R. Berggren, R. N. Cortright, R. W. Dudek, R. P. Thompson, W. J. Pories, ... and J. A. Houmard, *American Journal of Physiology-Endocrinology and Metabolism*, 2003, **284**, 741.

33 J. He, S. Watkins and D. E. Kelley, *Diabetes*, 2001, **50**, 817.

34 P. K. Stumpf, *Annual review of biochemistry*, 1969, **38**, 159.

35 S. C. Benoit, C. J. Kemp, C. F. Elias, W. Abplanalp, J. P. Herman, S. Migrenne, A. L. Lefevre, C. Cruciani-Guglielmacci, C. Magnan, F. Yu, K. Niswender, B. G. Irani, W. L. Holland and D. J. Clegg, *J Clin Invest*, 2009, **119**, 2577.

Table 1.classification results from three datasets using three different methods

Dataset	Variable	Prediction assessment parameters (%)				Variable choose
		Accuracy	sensitivity	specificity	AUC	
ESTE	PLSDA	91.67	83.33	100	83.33	All(409)
	CARS	100	100	100	83.33	52 115
	SLDA	100	100	100	83.33	115
TLS	PLSDA	96	92	100	92	All(200)
	CARS	100	100	100	96	32 112 141 142 149 182
	SLDA	100	100	100	96	32 33 105 112 126 138 141 142 149 182
CHOB	PLSDA	82.76	87.5	76.92	75.96	All(30)
	CARS	86.21	87.5	84.62	78.85	12 13 19 22 23
	SLDA	89.66	87.5	92.31	79.81	2 3 4 5 12 13 14 15 16 17 19 20 21 22 23 27 29

*Note: For these three datasets, in PLSDA the maximal numbers of components for cross validation are all 10; in CARS the numbers of Monte Carlo Sampling are all 1000; in SLDA the desired numbers of variables are 409, 200 and 30, respectively. And the γ and the maximal number of iteration are all 1e-6 and 50.

Table 2.the classification results of CHOB dataset using the CARS program three times with the same parameters

Dataset	Method	Prediction assessment parameters (%)				Variable choose
		Accuracy	sensitivity	specificity	AUC	
CHOB	CARS	68.97	68.75	69.23	62.98	5 12 19 29
	CARS	75.86	75.00	76.92	77.40	5 10 17 22 23 27
	CARS	86.21	87.5	84.62	78.85	12 13 19 22 23

*Note: In CARS the numbers of Monte Carlo Sampling are all 1000.

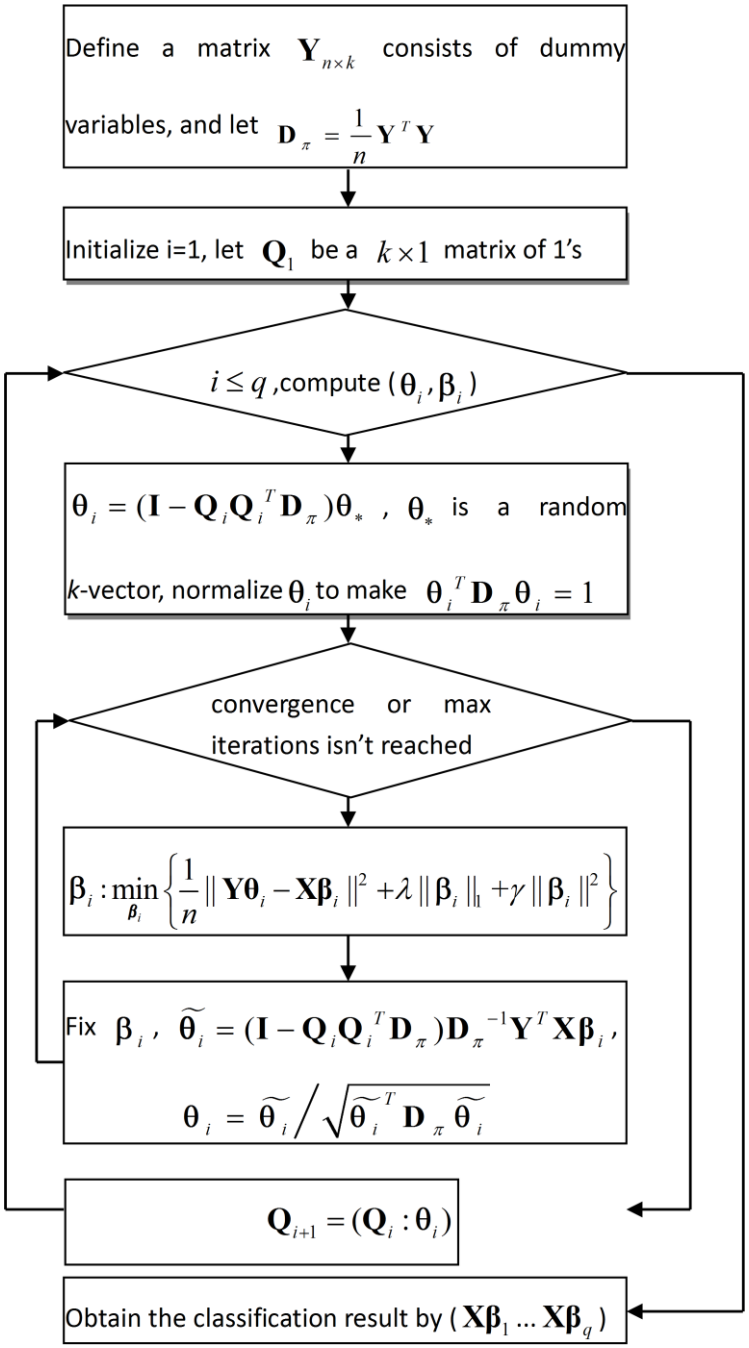


Figure 1 The process of SLDA algorithm is described in detail.

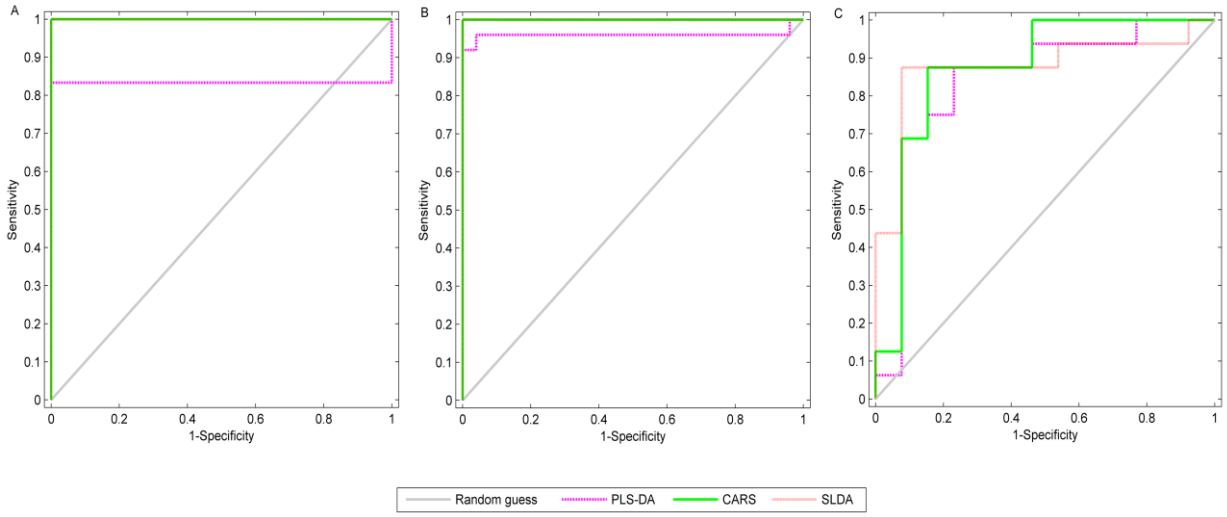


Figure 2 The ROC curves for three metabolomics datasets by using different discriminant methods PLS-DA, CARS and SLDA. (A)ESTE dataset;(B)TLS dataset;(C)CHOB dataset.

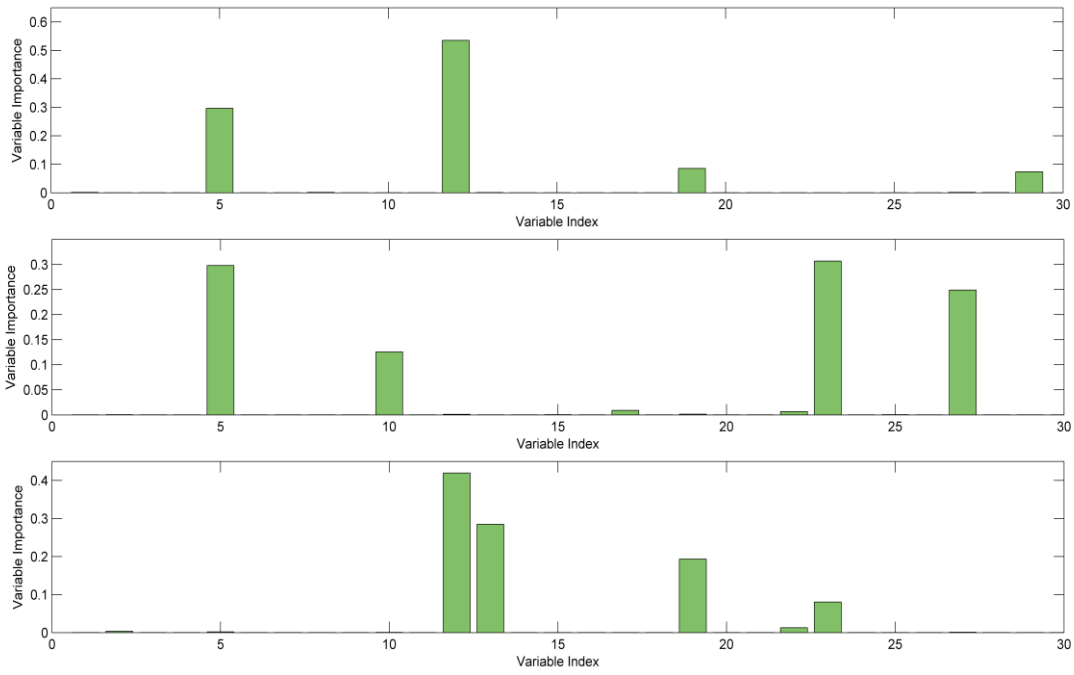


Figure 3 Three results for CHOB dataset' feature importance by running the CARS program three times with the same parameters.

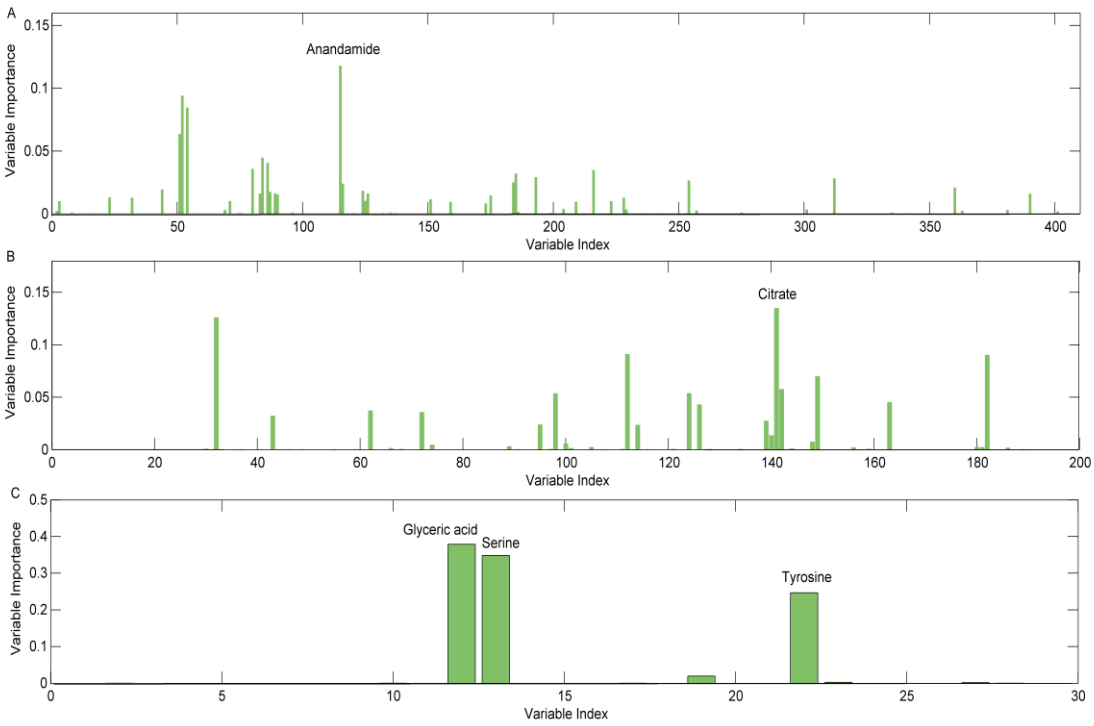


Figure 4 The variable importance and the potential biomarkers obtained by CARS for three metabolomics dataset. (A)ESTE dataset;(B)TLS dataset;(C)CHOB dataset.

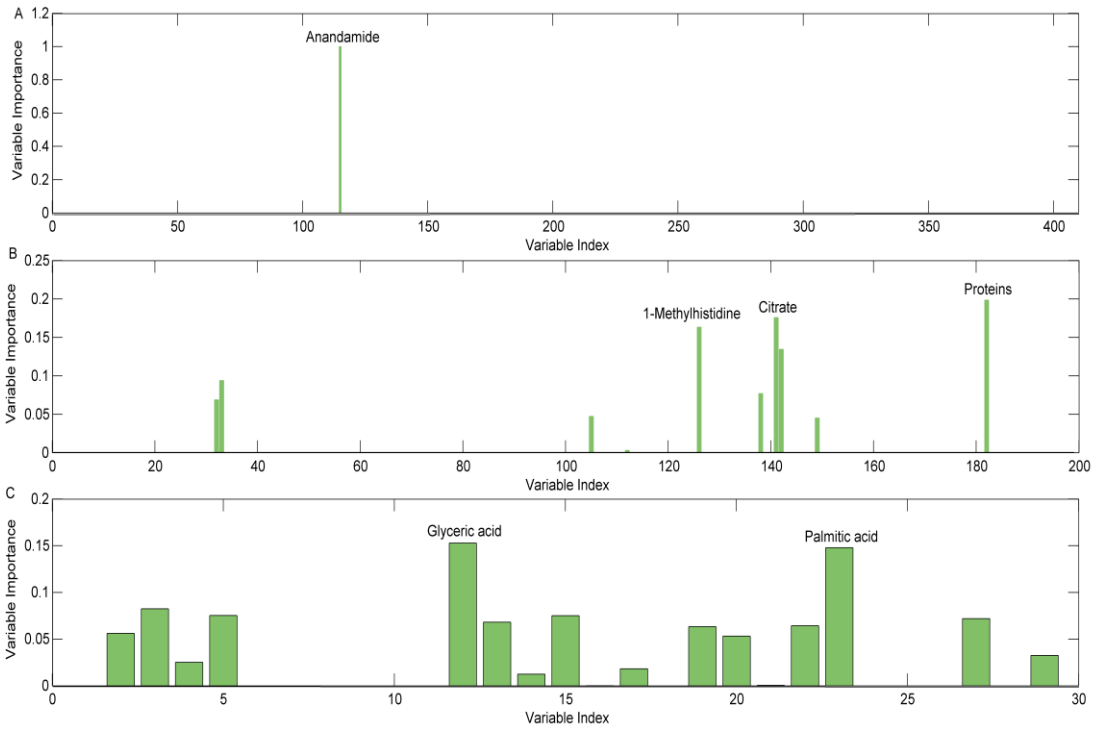


Figure 5 The variable importance and the potential biomarkers obtained by SLDA for three metabolomics dataset. (A)ESTE dataset;(B)TLS dataset;(C)CHOB dataset.