Multitask Learning Using Partial Least Squares Method

Wen-Cong Lu, Nian-Yi Chen
Department of Chemistry
Shanghai University
200436, Shanghai
China
wclu,nychen@mail.shu.edu.cn

Guo-Zheng Li, Jie Yang
Inst. of Image Processing & Pattern Recognition
Shanghai Jiaotong University
200030, Shanghai
China
lgz,jieyang@sjtu.edu.cn

Abstract – *In the machine learning field, feature selection is used to discard the redundant information and improve the learning accuracy. In this paper, the redundant information is reused in the learning of partial least squares method within the frame of multitask learning. This newly proposed method is used to solve the multivariate calibration problem, a classic problem in the analytical chemistry field. Results on three data sets collected using fluorescence spectroscopy show that multitask learning can help to improve the prediction accuracy of partial least squares method greatly.*

Keywords: Partial Least Squares, Multitask Learning, Feature Selection, Multivariate Calibration

1 Introduction

Feature selection is important during the machine learning process, because it can help to select the relevant features and improve the accuracy of learning machines [1, 2, 3]. After feature selection, how to treat the features not selected is still a problem.

In the past few years, a concept named multitask learning(MTL) [4, 5] was proposed to use the redundancy information. This MTL method uses some features not selected as extra output targets and obtains better results than other methods does which either uses all the features as input or does not use the features not selected any more. Yet, in the previous work, the base learning method used are mainly k-nearest neighborhood, or artificial neural networks, accuracy improved by MTL is so slight that the researchers claimed MTL was only proper to be used in the cases that even slight improvements were needed[5].

Partial least squares(PLS) method [6, 7] which can build linear regression model has proved to be useful in situations when the number of observed features is significantly greater than the number of observations and high multicollinearity among the features exists. This method is especially suit for the chemical data and is a frequently used method in the chemometrics field. But PLS are also prone to overfitting especially when the number of features are greatly more than the number of examples, thus, feature selection methods also play an important role in the learning process of PLS. However, in the previous work, features not selected are always discarded and not input into the later PLS model any longer. Motivated by this, we try use the MTL concept to improve the accuracy of PLS method.

Multivariate calibration is a classic problem in the analytical chemistry field[8, 9], which provide a convenient way to determine several components in a mixture within only one experimental step, without the tedious operation of pre-separation of these components[10]. In these multivariate calibration problems, there are many redundant features collected in the experiments, they will speed up the overfitting phenomena of the learning machine, so feature selection is used. The usually used methods is genetic algorithm, but it computes heavily, then clustering methods as Kohonen neural networks are used, these methods can effectively eliminate the redundant features [11, 12]. How about MTL combined with PLS to treat the multivariate calibration problems?

In this paper, we are going to combine MTL with the PLS regression method to address the multivariate calibration problems in the analytical chemistry. The rest of this paper is arranged as follows. Section 2 briefly describes the learning methods and the feature selection method used in this paper; Section 3 introduces the experimental data sets; Section 4 gives the computation results of the comparative learning methods using the leave-one-out cross validation method; this paper is ended up with discussions in Section 5.

2 Learning methods

2.1 Partial Least Squares Method

Partial least squares(PLS) regression algorithm can build linear regression models, and it has proved to be useful in situations when the number of observed features is significantly greater than the number of observations and high multicollinearity among the features exists. Since PLS regression method is not widely known in the fusion field, we will describe it in brief first.

Suppose the input features $X \subset \mathbb{R}^n$ and output targets $Y \subset \mathbb{R}^m$, PLS proposed by Wold[6, 13], uses a robust procedure, a nonlinear iterative partial least squares(NIPALS) algorithm [14], to solve a singular value decomposition problem. A modification [15, 16] of the classic PLS method can be described as in Fig. 1.

In Fig. 1, there are two loops. The inner loop is used to extract the score vector $\mathbf{t}$ and its corresponding latent vector $\mathbf{u}$. The outer loop is used to sequentially extract the latent

S1	Randomly initialize $\mathbf{u}$
S2	$\mathbf{w} = \mathbf{X^T u}$
S3	$\mathbf{t} = \mathbf{Xw}, \mathbf{t} \leftarrow \mathbf{t}/\|\mathbf{t}\|$
S4	$\mathbf{c} = \mathbf{Y}^T\mathbf{t}$
S5	$\mathbf{u} = \mathbf{Yc}, \mathbf{u} \leftarrow \mathbf{u}/\|\mathbf{u}\|$
S6	Repeat steps S2.–S5. until convergence
S7	Subtract $\mathbf{X}, \mathbf{Y}$ matrices: $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{tt^T X}, \mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{tt^T Y}$
S8	Repeat S1.–S7. until the rank of $\mathbf{X}$ is reached

Fig. 1: A nonlinear iteration partial least squares algorithm

vectors $\mathbf{t}, \mathbf{u}$ and the weight vectors $\mathbf{w}, \mathbf{c}$ from $\mathbf{X}$ and $\mathbf{Y}$ matrices in decreasing order of their corresponding singular values.

The PLS regression model can be written in matrix form as [16]

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}$$

where $\mathbf{B}$ is an $(n \times m)$ matrix of the regression coefficients and $\mathbf{F}$ is an $(N \times m)$ matrix of residuals. The matrix $\mathbf{B}$ has the form [16]

$$\mathbf{B} = \mathbf{X^T U}(\mathbf{T^T X X^T U})^{-1}\mathbf{T^T Y}$$

where the $\mathbf{T}$ and $\mathbf{U}$ are $(N \times p)$ matrices of the extracted p latent vectors, N is the number of cases.

As kernel methods is becoming an ad hoc topic, researchers have kernelized the PLS method to make it treat nonlinear data. Compared with other kernel methods as ν-support vector machines, kernel principal component regression and kernel ridge regression methods, experiments showed kernel-PLS performed better[16]. More and more people besides the chemometrics researchers are interested in the study of PLS method[17].

2.2 Multitask Learning

Multitask learning(MTL) [4, 5] is a form of inductive transfer that is applicable to any learning method that can share part of what is learned between multiple tasks. In this paper, we demonstrate will PLS method to address the multivariate calibration problem within the frame of MTL.

The basic idea of MTL is to use the selected features as the input feature set, and combine the target values with some of the discarded features as the target output. In the previous study, it has proved to help improve the accuracy of learning process. The terms used here is according to the previous work [5], they are arranged as in Table 1.

MTL has used many popular learning algorithms as the base learning machine such as k-nearest neighborhood, artificial neural networks, even support vector machines, etc., but obtain slight improvements [5]. Here we will apply it on the PLS method, to see if it can help to obtain better performance.

In this work we will show three real world problems that have benefited from using some of the features that feature selection have discarded as Extra Inputs or Extra Outputs instead. The procedure we will use is shown in Fig. 2.

Table 1: The terms used in the multitask learning procedure

Term	Explanation
Main Task	The output target values to be learned
Selected Inputs	The features selected as inputs in all experiments
Extra Features	The features selected from the discarded features
Extra Inputs	The extra features selected from the discarded features when used as inputs
Extra Outputs	The same extra features selected from the discarded features when used as outputs
STD	Standard PLS using the Selected Inputs as inputs and only the Main Task as outputs
STD+IN	Uses the Extra Features as Extra Inputs to learn the Main Task
STD+OUT	Uses the Extra Features as Extra Outputs in paralleled with the Main Task using the Selected Inputs as inputs

S1	Select the Selected Inputs and Extra Features using the Kohonen feature selection method in subsection 2.3
S2	Train an STD model using Selected Inputs as inputs and only Main Task as outputs
S3	Train an STD+IN model using Selected Inputs plus Extra Features as inputs and only Main Task as outputs
S4	Train an STD+OUT model using Selected Inputs as inputs and Main Task plus Extra Features as outputs

Fig. 2: The multitask leaning procedure

2.3 Feature Selection Method

Feature selection is an important issue in the machine learning field, it can remove the irrelevant features and make the learning more efficiently and accurately. Many feature selection methods have been proposed [1, 2, 3] in which unsupervised methods using clustering methods are popular methods, they can obtain the representative features as the selected subset [12, 18] and compute efficiently.

Kohonen neural network [19] has been used as the feature selection method for multivariate calibration problem [12] to improve the prediction accuracy. In this method, data is clustered according to the Euclidean distance of each feature vectors using the Kohonen neural network, then the features near the center of clusters are selected as Selected Inputs, and the other not selected features are ranked according to the closeness with the Selected Inputs. The algorithm is shown in Fig. 3, in which p denotes the number of Selected Inputs, and e, the number of Extra Features, is defined as

$$\max(\min(p - m, n - p - 20, N - m), 0),$$

S1	Clustering using Kohonen neural network
S2	Select the features near the center of the clusters as Selected Inputs, usually the number of Selected Inputs is less than the number of the clusters
S3	Rank the features not selected according to closeness with the selected p features, and select e Extra Features with the highest closeness

Fig. 3: The feature selection algorithm for multitask learning using Kohonen neural networks

This expression is considered of many factors such as the PLS model and the noise features, in which $p - m$ and $N - m$ are used to limit the total number of outputs because the number of outputs can not exceed either the number of inputs or the number of cases, $n - p - 20$ is used to discard some noise features, we choose the minimal of three numbers, and must keep it nonnegative.

This method is implemented using the neural network toolbox of MATLAB [20], Euclidean distance is used in the Kohonen neural networks.

3 Experimental data sets

The data sets used in this work consists of three different multivariate calibration data sets, all three are collected by fluorescence spectrometry[21], the number of features, cases and target values are listed in Table 2. It is interesting that there are two characteristics on all the data sets.

- There are many redundant features, and the features of data set I are shown in Fig. 4;
- Because of the cost on the experiments, the number of cases are always rather less than the number of features;

Table 2: The property of the multivariate calibration data sets

data set	num of feat.	num of cases	num of targets
I	211	23	3
II	141	17	2
III	116	17	2

3.1 Assessment of Regression Quality

Since the size of each data set is small, we use the leave-one-out cross validation(LOOCV) technique to evaluate the above learning methods with the common used measures *root mean square error* (RMSE), for the jth component, it is defined as

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{ij}^e - y_{ij})^2},$$

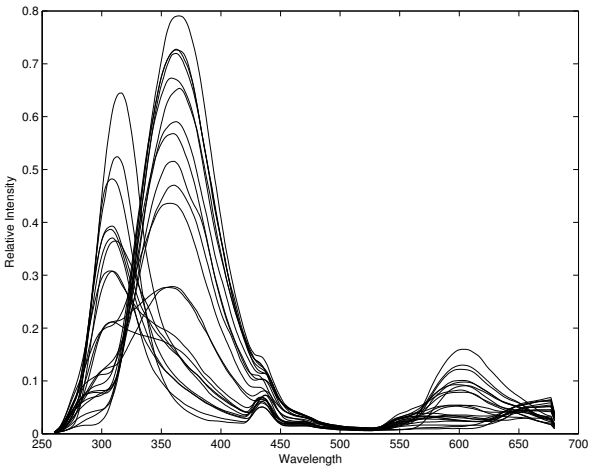

Fig. 4: Fluorescence Spectra of the first Data set. (Unit of the horizontal axis is nm)

and for the whole, it is

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\mathrm{RMSE}_j^2},$$

where y_{ij}^emeans the jth predicted target value of ith example, y_{ij}means the jth real target value of ith example, N denotes the number of examples and m denotes the number of target values in the Main Task of each example, which is 2 or 3 as in Table 2.

4 Results of computation

4.1 MTL on different number of Selected Inputs

Data sets in section 3 have been processed. Firstly, feature selection using Kohonen neural networks have been performed, then, PLS is used to perform a LOOCV computation under three cases as the procedure in Fig. 2.

As the number of the Selected Inputs increases, results of the RMSE on all three data sets in different cases are plotted on Fig. 5-7. Some statistical results are listed in Table 3, in which the smallest error in three cases is defined as RMSE_{ms}, RMSE_{mi} and RMSE_{mo} respectively, the average error is defined as RMSE_{as}, RMSE_{ai} and RMSE_{ao} respectively, the maximal distance of RMSE between the cases of STD and STD+IN, STD and STD+OUT, STD+IN and STD+OUT is defined as RMSE_{mdsi}, RMSE_{mdso} and RMSE_{mdio} respectively, the average distance of RMSE between the cases of STD and STD+IN, STD and STD+OUT, STD+IN and STD+OUT is defined as RMSE_{adsi}, RMSE_{adso} and RMSE_{adio} respectively.

From Fig. 5-7 and Table 3, we can see that on the multivariate calibration problems,

- Most features, about nine out of ten, are redundant, PLS with about one tenth of features can obtain the highest accuracy;
- Feature selection can improve the accuracy to some degree compared with the total feature subset, but the improvements are slight;

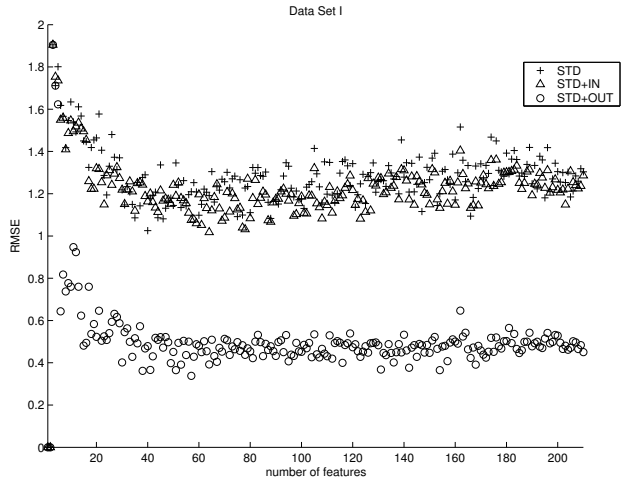

Fig. 5: Results of RMSE as the number of Selected Inputs increases on Data Set I

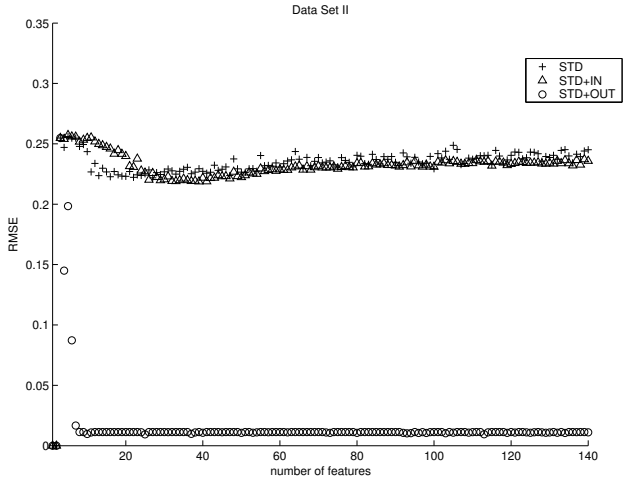

Fig. 6: Results of RMSE as the number of Selected Inputs increases on Data Set II

- Multitask learning can help to improve the prediction accuracy greatly, the error of STD+OUT is about one third of that of STD or STD+IN cases on Data Set I, even one twentieth on Data Set II, and one tenth on Data Set III.

4.2 MTL on different number of Extra Outputs

We select the feature subset in which the number of Selected Inputs is about one third of N, the number of cases, using Kohonen feature selection method, then we use the Selected Inputs as input and use the Extra Features plus the Main Task as outputs to build the PLS model. As the number of Extra Ouputs increases in STD+OUT case results of RMSE are plotted on Fig. 8-Fig. 10, Results of RMSE are also computed in the cases of STD and STD+IN and plotted on Fig. 8-Fig. 10.

From Fig. 8-Fig. 10, we can see that when the Extra Features are used as Extra Outputs in STD+OUT case, RMSE is rapidly decreasing as the number increases and then kept stable until 10 plus features are used. However, when the

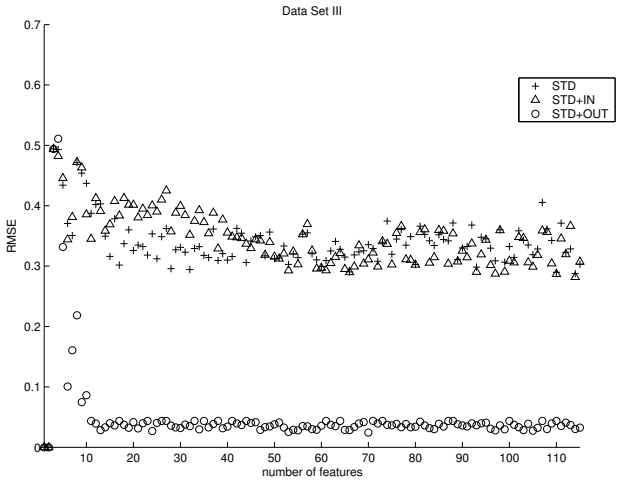

Fig. 7: Results of RMSE as the number of Selected Inputs increases on Data Set III

Table 3: Some statistical results for different learning methods

Metric	DataSet I	DataSet II	DataSet III
RMSE_{ms}	1.0253	0.2219	0.2877
RMSE_{mi}	1.0175	0.2188	0.2819
RMSE_{mo}	0.3379	0.0094	0.0244
RMSE_{as}	1.2861	0.2350	0.3411
RMSE_{ai}	1.2311	0.2326	0.3468
RMSE_{ao}	0.5107	0.0158	0.0513
RMSE_{mdsi}	0.2607	0.0131	0.0514
RMSE_{mdso}	1.0057	0.2381	0.3792
RMSE_{mdio}	1.0145	0.2451	0.3885
RMSE_{adsi}	0.0550	0.0024	-0.0057
RMSE_{adso}	0.7754	0.2192	0.2898
RMSE_{adio}	0.7203	0.2168	0.2955

Extra Features are used as Extra Inputs in STD+IN case, slight improvements and no explicit rules on the improvements can be obtained.

5 Discussions

Beyond our imagination, on multivariate calibration problems multitask learning using partial least squares(PLS) method can obtain so inspiring results. This owns to the special algorithm of PLS, which uses a robust NIPALS [14] to solve the singular value decomposition of the product X^TY of the input matrix and output matrix. So when PLS method uses the Extra Features as the output target values in the multivariate calibration problems, the Extra Outputs exert constraints on the PLS regression model and depress the overfitting, then PLS can obtain more precision models and give less error prediction values on the test examples.

The same procedure has been performed on the same data sets using other learning methods like artificial neural networks, which does not show much improvement on the prediction.

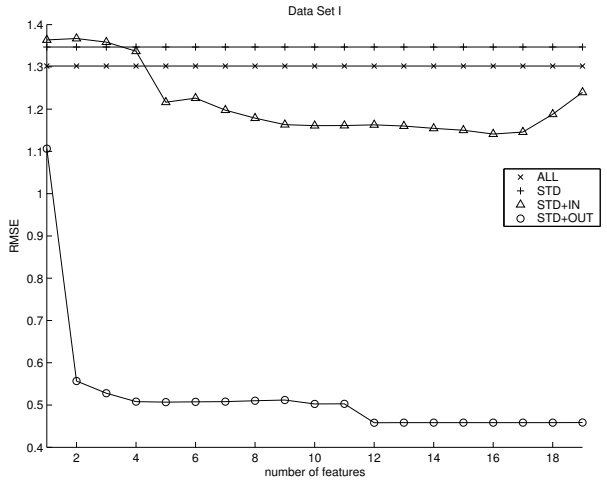

Fig. 8: Results of RMSE as the number of Extra Ouputs increases on Data Set I

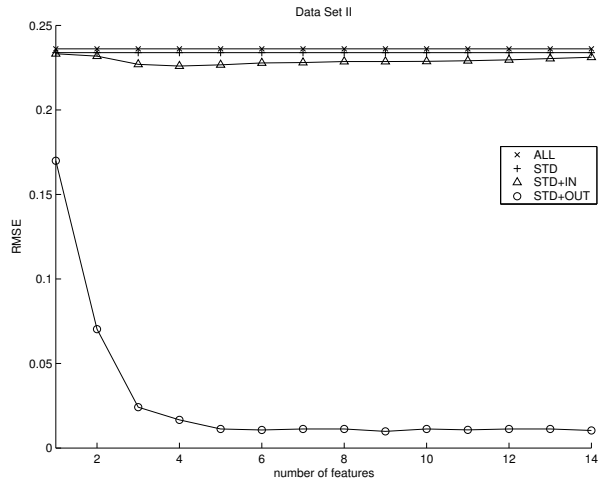

Fig. 9: Results of RMSE as the number of Extra Ouputs increases on Data Set II

In fact we have performed feature selection by other methods like clustering method [18], and embedded algorithm using multiple ridge regression [3]. Though the improvements is not as obvious as it shows in this paper, we find that MTL can greatly improve the prediction accuracy of PLS method, while MTL can only slightly improve the accuracy of artificial neural networks.

Although we obtain some exciting results in this work, there are still much work to do, i.e. which number is best for the Selected Inputs and the Extra Outputs and how about (kernel) PLS on other problems?

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China under grant number 50174038. Dr. Yong-Gang Wang, Jun Lu and Yu-Hai Liu have given valuable advices to this work.Thanks also go to the anonymous reviewers for their valuable advices.

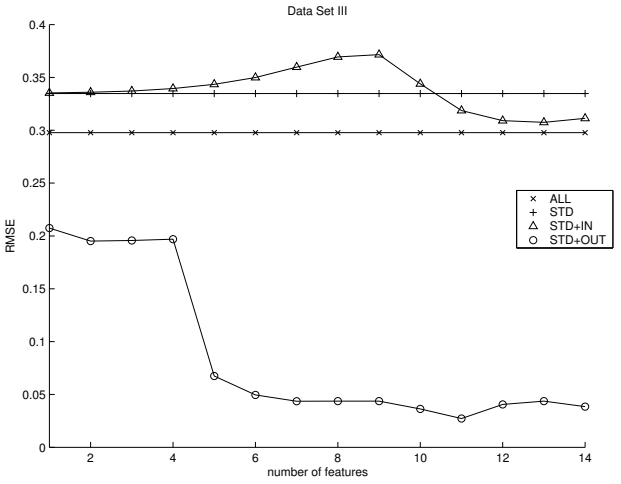

Fig. 10: Results of RMSE as the number of Extra Ouputs increases on Data Set III

References

[1] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[2] Ron Kohavi and John H George. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[3] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182, 2003.

[4] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[5] Rich Caruana and Virginia R. de Sa. Benefiting from the variables that variable selection discards. *Journal of machine learning research*, 3:1245–1264, 2003.

[6] Herman Wold. *Perspectives in probability and statistics, papers in honor of M.S. Bartlett*, chapter Soft modeling by latent variables; the nonlinear iterative partial least squares approach, pages 520–540. Academic Press, London, 1975.

[7] A Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1988.

[8] Marko Peussa, Satu Härkönen, Janne Puputti, and Lauri Niinistö. Application of PLS multivariate calibration for the determination of the hydroxyl group content in calcined silica by DRIFTS. *Journal of Chemometrics*, 14:501–512, 2000.

[9] Brian D. Marx and Paul H. C. Eilers. Multivariate calibration stability: A comparison of methods. *Journal of Chemometrics*, 16:129–140, 2002.

[10] H Martens and T Ns. *Multivariate Calibration*. John Wiley and Sons, Chichester, 1989.

[11] Jahanbakhsh Ghasemi, Ali Niazi, and Riccardo Leardi. Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: Application on copper and zinc mixture. *Talanta*, 59:311–317, 2003.

[12] L F Capitán-Vallvey, N Navas, M Del Olmo, V Consonni, and R Todeschini. Resolution of mixtures of three non-steroidal anti-inflammantory drugs by fluorescence using partial least squares multivariate calibration with previous wavelength selection by kohonen artificial neural networks. *Talanta*, 52:1069–1079, 2000.

[13] Svante Wold, H Ruhe, Herman Wold, and W J Dunn III. The collinearity problem in linear regression. the partial least squares(pls) approach to generalized inverse. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.

[14] Herman Wold. *Multivariate Analysis*, chapter Estimation of principal components and related models by iterative least squares, pages 391–420. Academic Press, New York, 1966.

[15] P J Lewi. Pattern recognition, reflection from a chemometric point of view. *Chemometrics and Intelligent Laboratory Systems*, 28:23–33, 1995.

[16] Roman Rosipal and Leonard J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 2:97–123, 2001.

[17] Roman Rosipal, Leonard J. Trejo, and Bryan Matthews. Kernel pls-svc for linear and nonlinear classification. In *Proceedings of the Twentieth International Conference on Machine Learning(ICML-2003)*, Washington DC, 2003.

[18] Pabitra Mitra, C A Murthy, and Sankar K Pal. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

[19] Jure Zupan, Marjana Novič, and Itziar Ruisánchez. Kohonen and counter-propagation artificial neural networks in analytical chemistry. *Chemometrics and Intelligent Laboratory System*, 38:1–23, 1997.

[20] H Demuth and M Beale. *Neural Network Toolbox User's Guide for Use with MATLAB, (4th Ed.)*. the Mathworks Inc., 2001.

[21] Guo-Zheng Li, Jie Yang, Yaping Ding, Qingsheng Wu, and Nianyi Chen. Comparative study of support vector machines and artificial neural networks for the multivariate calibration of spectrofluorimetric simultaneous determination of aromatic amino acids. *Journal of Chemometrics*, submitted.