Orthogonal projection weights in dimension reduction based on Partial Least Squares

Xue-Qiang Zeng

School of Computer Science and Engineering,
Shanghai University, Shanghai 200072, PR China

Computer Center,
Nanchang University, Nanchang 330006, China
E-mail: stamina_zeng@shu.edu.cn

Guo-Zheng Li*

Department of Control Science and Engineering,
Tongji University, Shanghai 201804, PR China
E-mail: drgzli@gmail.com
*Corresponding author

Mary Qu Yang

National Human Genome Research Institute,
National Institutes of Health,
US Department of Health and Human Services,
Bethesda, MD 20852, USA and also Oak Ridge, D.O.E.
E-mail: yangma@mail.NIH.GOV

Geng-Feng Wu

School of Computer Science and Engineering,
Shanghai University, Shanghai 200072, PR China
E-mail: gfwu@shu.edu.cn

Jack Y. Yang

Department of Radiation Oncology,
Massachusetts General Hospital and Harvard Medical School,
Harvard University,
Boston, Massachusetts 02114, USA
E-mail: jyang@bwh.Harvard.edu

Abstract: Dimension reduction is important during the analysis of gene expression microarray data, because the high dimensionality in the data set hurts the generalisation performance of classifiers. Partial Least Squares Based Dimension Reduction (PLSDR) is a frequently used method, since it is specialised in handling high dimensional data set and leads to satisfying classification performance. However, the previous works exist an ambiguous usage of projection weights in PLSDR. To assure the orthogonality of projected components, the usually used project weights

are nonorthogonal. Here, we propose to use orthogonal project weights for PLSDR. Experimental results on four microarray data sets show our proposed orthogonal project weights are better than the previous used to help improve the generalisation performance of classifiers.

Keywords: dimension reduction; partial least squares.

Reference to this paper should be made as follows: Zeng, X-Q., Li, G-Z., Yang, M.Q., Wu, G-F. and Yang, J.Y. (2009) 'Orthogonal projection weights in dimension reduction based on partial least squares', *Int. J. Computational Intelligence in Bioinformatics and Systems Biology*, Vol. 1, No. 1, pp.100–115.

Biographical notes: Xue-Qiang Zeng received his MS Degree from Jiangxi Normal University, China. He is currently a PhD candidate in computer science and engineering at Shanghai University, China. He has published several papers and his research interest is bioinformatics and machine learning.

Guo-Zheng Li received his PhD Degree from Shanghai JiaoTong University. He is currently an Associate Professor in the School of Computer Science and Engineering at Shanghai University, China. He is also currently serving on the Committees at CAAI Machine Learning Society, CAS Pervasive Computing Society, International Society of Intelligent Biological Medicine and IEEE Computer Society. He is a Principle Investigator of machine learning and bioinformatics research projects under the grants of Nature Science Foundation of China. He has published more than 30 refereed papers in professional journals and conferences. He has written three book chapters and translated one professional book from English into Chinese. He is an Associate Editor of IJCIBSB and JCIB, Vice Chair of CSC07, ICAI07, and Program Committee Member of ICMLC07, IEEE 7th BIBE, IEEE BIBM07 and OSB07. He was a recipient of the Best Paper Awards at PRICAI 2006 and ICAI 2007.

Mary Qu Yang is an American Computer Scientist and Biologist, she is Editor-in-Chief of *International Journal of Computational Biology and Drug Design*. She received her PhD, MSECE and MS Degrees, all from Purdue University main campus in West Lafayette and her post-doctoral training from NIH main campus in Maryland. She also completed the research specialist training from NIH, US Department of Health and Human Services and Oak Ridge, DOE and received training in biostatistics and bioinformatics from Johns Hopkins University. She was a visiting scholar of Dr. Jun S. Liu's statistical and computational genomics laboratory of Harvard University in Cambridge. She was a recipient of the Outstanding Interdisciplinary Bilsland Dissertation Fellow for Computer Engineering (Advisor: Dr. Okan K. Ersoy) and Biophysics (Advisor: Dr. Albert W. Overhauser) Dual Degrees and NIH Fellow for the National Human Genome Research. She works in both engineering practice and translational medicine and was trained as a combined experimental and computer scientist with more than 15 years of teaching, research and engineering practice experience. She was the scientific review, advisory and steering committee chair of IEEE Bioinformatics and Bioengineering at Harvard Medical School in 2007 and advisory committee chair of Biocomp and IJCBS, as well as an honorary consulting editor of *International Journal*

of Functional Informatics and Personalised Medicine, an official journal of *International Society of Intelligent Biological Medicine*. She has been an editor of a number of journals and proceedings books including *Journal of Supercomputing (Springer Science), International Journal of Pattern Recognition and Artificial Intelligence (World Scientific), IEEE Bioinformatics and Bioengineering (IEEE), International Journal of Bioinformatics Research and Applications (InderScience) and International Conference on Bioinformatics and Computational Biology*. She developed hybrid intelligent systems and contributed to the research in bidirectional promoters, cancer genomics, transmembrane proteins and their disease relevancies. She is known for Intelligent Computing Research and Education and was a recipient of a number of outstanding achievement and best paper awards including IEEE Computer Society Bioinformatics and Biomedicine Distinguished Workshop Keynote Lecturer, IEEE Bioinformatics and Bioengineering Outstanding Achievement Award, Artificial Neural Networks in Engineering Best Paper Award, World Congress on Computer Science, Computer Engineering and Applied Computing Outstanding Achievement Award. She has published more than 100 peer-reviewed scientific papers and book chapters and has delivered many invited talks including a number of keynote lectures to promote the emerging field of translational bioinformatics and personalised medicine. She specialises in genomics and machine learning.

Geng-Feng Wu is currently Professor and Deputy Dean in the School of Computer Engineering and Science at Shanghai University, China. He is also currently serving on the Executive Committee at Shanghai Computer Society. He is a Principle Investigator of artificial intelligence research project under the grants of Shanghai Technique Committee and National Institute of Earthquake. He has published more than 100 refereed papers in professional journals and conferences. He was Programme Chair of CIT01 and committee member of numerous conference.

Jack Y. Yang is a Harvard scientist, and chair of board of directors of *International Society of Intelligent Biological Medicine* (ISIBM). Dr. Yang is the Editor-in-Chief of *International Journal of Functional Informatics and Personalised Medicine*. He received his PhD and MS Degrees both from Purdue University, West Lafayette main campus and his post doctoral training was from Harvard Medical School and Indiana University School of Medicine. He also received training in biostatistics and bioinformatics from Johns Hopkins University, and in computer science from University of Illinois at Urbana-Champaign. He was trained as a combined experimental and computer scientist with more than 15 years of teaching, research and engineering practice experience in computer science and biomedical engineering. He has been an editor of more than a dozen journals and proceedings books and was the General Chair of the IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School and Co-PI of US National Science Foundation grant. He is also a consultant to IJCBS. He has published more than 100 papers. He specialises in cancer biology and artificial intelligence. http://en.wikipedia.org/wiki/Jack_Yang

1 Introduction

DNA microarray experiments are used to collect information from tissue and cell samples regarding gene expression differences for tumor diagnosis (Golub et al., 1999; Alon et al., 1999; Dudoit et al., 2002). The output of microarray experiment is summarised as an $n \times p$ data matrix, where n is the number of tissue or cell samples, p is the number of genes. Here, p is always much larger than n, which hurts the generalisation performance of most classification methods. To overcome this problem, we can either select a small subset of interesting genes (gene selection) or construct K new components summarising the original data as well as possible, with $K < p$ (dimension reduction, feature extraction).

Gene selection has been studied extensively in the last few years. The most commonly used procedures of gene selection are based on a score which is calculated for all genes individually and genes with the best scores are selected. Gene selection procedures output a list of relevant genes which can be experimentally analysed by biologists. These methods are often denoted as univariate gene selection, whose advantages are its simplicity and interpretability. However, much information contained in the data set is lost when genes are selected solely according to their individual capacity to separate the samples, since interactions and correlations between genes are omitted, as are of great interest in system biology.

Dimension reduction is an alternative to gene selection to overcome the problem of curse of dimensionality. Unlike gene selection, dimension reduction projects the whole data into a low dimensional space and constructs the new dimensions (components) by analysing the statistical relationship hidden in the data set. Researchers have developed different dimension reduction methods in applications of bioinformatics and computational biology (Antoniadis et al., 2003; Nguyen et al., 2004; Dai et al., 2006), among which Partial Least Squares Based Dimension Reduction (PLSDR) is one of the most effective methods (Dai et al., 2006).

PLS was firstly developed as an algorithm performing matrix decompositions by Wold (1975), and then was introduced as a multivariate regression tool in the context of chemometrics (Wold et al., 1984). A detailed chronological introduction of PLS was given in Martens (2001), some comprehensive overviews of PLS were given in Helland (1988), Wold et al. (2001), Helland (2001) and Boulesteix and Strimmer (2006). Only in recent years, PLS has been found to be an effective dimension reduction technique (Nguyen and Rocke, 2002a, 2002b).

Nguyen and Rocke (2002a, 2002b) proposed to use PLS for dimension reduction as a preliminary step for binary and multi-class classification. A numerical simulated study on total predictor variance explained by PLS was also carried out by Nguyen et al. (2004). Experiments on microarray data proved that PLSDR is better than Principle Component Analysis (PCA) based dimension reduction. Barker and Rayens (2003) explained the relationship between PLS and Canonical Correlation Analysis (CCA) in a formal statistical manner. They clarified that PLS is superior to PCA when dimension reduction is needed. Boulesteix (2004) compared PLS with some of state-of-the-art classification methods and investigated some interesting properties of PLSDR. Dai et al. (2006) provided a comparative study of three dimension reduction techniques: PLSDR, sliced inverse regression (SIR) and PCA, which evaluated the predictive accuracy and computational efficiency of classification procedures incorporating those methods. Zeng et al. (2007) introduced PLS into the field of text classification as

a text representation method. All these works have demonstrated the outstanding performance of PLSDR.

There are two series of projection weights in PLS method denoted as W and V respectively. The difference between W and V is slight but significant: W are project weights which related to residual matrix E_k and V are the modified versions of W which linked with original matrix X. By the notion that using latent components constructed by PLS as new predictors, V is the natural choose projection weights. Though the previous works show that PLSDR is much faster than PCA and leads to accurate classification (Dai et al., 2006; Barker and Rayens, 2003; Boulesteix, 2004), the ambiguous usage of projection weights W and V in PLSDR has not been clarified yet.

In this paper, we concentrated on the orthogonality of the projection weights for PLSDR. Since dimension reduction is a certain kind of coordinates transformation, it is important to consider the orthogonality among the projection weights, furthermore an orthogonal space is much popular than an nonorthogonal one. The difference of these two series of weights has not been mentioned before and the choice of W and V has not been clarified. It is difficult to make sure which projection weights were used in previous works, while the investigation of this problem has great sense for the standard usage of PLSDR. Therefore, we propose to investigate the classification performance affected by dimension reduction with W and V.

This paper is organised as follows. Some essential notions are given in Section 2. In Section 3, PLS is shortly introduced and then PLSDR is presented in detail. The difference between W and V is also discussed. Experiments and discussions on four biological data sets are described in Section 4. Finally, conclusions are given in Section 5.

2 Notions

Expression levels of p genes in n microarray samples are collected in an $n \times p$ data matrix $X = (\mathbf{x}_{ij}), 1 \leq i \leq n, 1 \leq j \leq p$; of which an entry $\mathbf{x}_{ij}$ is the expression level of the jth variable gene in the ith microarray sample.

Here we consider binary classification problem, the labels of the n microarray samples are collected in vector $\mathbf{y}$. When the ith sample belongs to class one, the element $\mathbf{y}_i$ is 1; otherwise it is -1.

Besides, $\| \bullet \|$ denotes the length of a vector. X^T represents the transpose of X, X^{-1} represents the inverse matrix of X.

Note that X and $\mathbf{y}$ used in Section 3 are assumed to be centred to zero mean by each column.

3 Partial Least Squares Based Dimension Reduction

3.1 Principle

Partial Least Squares (PLS) is a class of techniques for modeling relations between blocks of observed variables by means of latent variables. The underlying assumption of PLS is that the observed data is generated by a system or process which is driven

by a small number of latent (not directly observed or measured) variables. Therefore, PLS aims at finding uncorrelated linear transformations (latent components) of the original predictor variables which have high covariance with the response variables. Based on these latent components, PLS predicts response variables $\mathbf{y}$ and reconstruct original matrix X at the same time.

Let matrix $T = [\mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathbb{R}^{n \times K}$ represents the n observations of the K components which are usually denoted as Latent Variables (LV) or scores. The relationship between T and X is defined as:

$$T = XV \tag{1}$$

where $V = [\mathbf{v}_1, \ldots, \mathbf{v}_K] \in \mathbb{R}^{p \times K}$ is the matrix of projection weights. PLS determines the projection weights V by maximising the covariance between the response and latent components.

Based on these latent components, X and $\mathbf{y}$ are decomposed as:

$$\begin{aligned} X &= TP^T + E \\ \mathbf{y} &= TQ^T + \mathbf{f} \end{aligned} \tag{2}$$

where $P = [\mathbf{p}_1, \ldots, \mathbf{p}_K] \in \mathbb{R}^{p \times K}$ and $Q = [\mathbf{q}_1, \ldots, \mathbf{q}_K] \in \mathbb{R}^{1 \times K}$ are denoted as loadings of X and $\mathbf{y}$ respectively. Generally, P and Q are computed by Ordinary Least Squares (OLS). E and $\mathbf{f}$ are residuals of X and $\mathbf{y}$ respectively.

By the decomposition of X and $\mathbf{y}$, response values are decided by the latent variables not by X (at least not directly). It is believed that this model would be more reliable than OLS because the latent variables are coincided with the true underlying structure of original data.

The major point of PLS is the construction of components by projecting X on the weights V. The classical criterion of PLS is to sequentially maximising the covariance between response $\mathbf{y}$ and latent components. There are some variants of PLS approaches to solve this problem (Wold et al., 2001). Ignoring The miner differences among these algorithms, we demonstrate the most frequently used PLS approach: PLS1 (Helland, 1988; Wold et al., 2001).

PLS1 determines the first latent component $\mathbf{t}_1 = X\mathbf{w}_1$ by maximising the covariance between $\mathbf{y}$ and $\mathbf{t}_1$ under the constraint of $\|\mathbf{w}_1\| = 1$. The corresponding objective function is:

$$\mathbf{w}_1 = \arg \max_{\mathbf{w}^T\mathbf{w}=1} (Cov(X\mathbf{w}, \mathbf{y})). \tag{3}$$

The maximisation problem of equation (3) can be easily solved by the Lagrange multiplier method.

$$\mathbf{w}_1 = X^T\mathbf{y}/\|X^T\mathbf{y}\|). \tag{4}$$

To extract other latent components sequentially, we need to model the residual information of X and $\mathbf{y}$ which could not be explained by previous latent variables. So, after the extraction of the score vector $\mathbf{t}_1$, PLS1 deflate matrices X and $\mathbf{y}$ by subtracting their rank-one approximations based on $\mathbf{t}_1$. The X and $\mathbf{y}$ matrices are deflated as:

$$\begin{aligned} E_1 &= X - \mathbf{t}_1\mathbf{p}_1^T \\ \mathbf{f}_1 &= \mathbf{y} - \mathbf{t}_1\mathbf{q}_1^T \end{aligned} \tag{5}$$

where $\mathbf{p}_1$ and $\mathbf{q}_1$ are loadings determined by OLS fitting:

$$\begin{aligned} \mathbf{p}_1^T &= (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T X \\ \mathbf{q}_1^T &= (\mathbf{t}_1^T \mathbf{t}_1)^{-1} \mathbf{t}_1^T \mathbf{y} \end{aligned} . \tag{6}$$

As an iterative process, PLS1 constructs other latent components in turn by using the residuals E_1 and $\mathbf{f}_1$ as new X and $\mathbf{y}$.

$$\begin{aligned} \mathbf{w}_k &= E_{k-1}^T \mathbf{f}_{k-1} / \| E_{k-1}^T \mathbf{f}_{k-1} \| \\ \mathbf{t}_k &= E_{k-1} \mathbf{w}_k \\ \mathbf{p}_k^T &= (\mathbf{t}_k^T \mathbf{t}_k)^{-1} \mathbf{t}_k^T E_{k-1} \\ \mathbf{q}_k^T &= (\mathbf{t}_k^T \mathbf{t}_k)^{-1} \mathbf{t}_k^T f_{k-1} \\ E_k &= E_{k-1} - \mathbf{t}_k \mathbf{p}_k^T \\ \mathbf{f}_k &= \mathbf{f}_{k-1} - \mathbf{t}_k \mathbf{q}_k^T . \end{aligned} \tag{7}$$

For the convenient of expression, matrices X and $\mathbf{y}$ are often denoted as E_0 and $\mathbf{y}_0$ respectively. The number of components is a parameter of PLS which can be fixed by user or decided by a cross-validation scheme. In general, the maximal number of latent components is the rank of matrix X which have non-zero covariance with $\mathbf{y}$.

It is obvious that the deflation scheme guarantees mutual orthogonality of the extracted score vectors T, that is, $T^T T = I$. By the arguments of Hoskuldsson (1988), it can be seen that the weights $W = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}^{p \times K}$ are also orthogonal. Furthermore, the relation between V and W was demonstrated as Manne (1987):

$$V = W(P^T W)^{-1} \tag{8}$$

from which, we evade the iterative construction of latent components on residual matrix E_k, but relate T to X directly. In general, the loading vectors P and Q are not orthogonal (Some obscure variants of PLS provide the orthogonality of P or Q (Wold et al., 2001)). So deduced from equation (8), we can see that the projection weights V are not orthogonal.

A side result which comes up during the derivation of equation (8) is that the weights W span the same space as an orthogonal Krylov sequence which is defined as:

$$\kappa = (z, Az, \ldots, A^{p-1} z) \tag{9}$$

where κ is the p-dimensional Krylov space of A and z.

3.2 Dimension reduction

PLS reduces the complexity of microarray data analysis by constructing a small number of new predictors, T, which are used to replace the large number of original gene expression measures. Moreover, obtained by maximising the covariance between the components and the response variables, the PLS components are generally more predictive than the principal components extracted by other unsupervised methods like PCA (Barker and Rayens, 2003).

The projection matrix V are approximate coordinates of the original data space. After projecting and representing each sample in the new space, PLS build models on some latent variables by using the ordinary least squares algorithm. When the number of latent components is the same as the rank of matrix X, all the information of X is preserved and PLS exhibits the same as OLS does on X. But, we need not stick to using OLS in the transformed space. Other statistical learning may also be used in this space, such as Support Vector Machine (SVM), Logistic Discrimination (LD) etc. That is, we may just use PLS as a dimension reduction method instead of a classification/regression model (Nguyen and Rocke 2002a, 2002b).

After dimension reduction, many statistical methods may be used for classification based on these new predictors. But the new space has one problem that the projection weights V are nonorthogonal. As the independent assumption (orthogonality) of input variables (latent components projected by V) is important for OLS regression, PLS keep the orthogonality of components T by modifying projection weights from orthogonal (W) to nonorthogonal (V). When it came to the application of dimension reduction, the orthogonality of projection directions is more desired than the orthogonality of projected components.

Additionally, it needed be clarified that the lengths of columns of V are not unit. Due to the deflation scheme of PLS, the significance of components produced iteratively are in the descending order. That is, the tail components are less informative than the initial components. Reflected by V, the lengths of these projection weights are in the descending order too. V instinctively punish the uninformative projection weights by reducing the corresponding vector lengths.

Consequently, when casting classification on the dimensions created by V, the performance of classifiers is hardly influenced by adding tail components to gene expression. This would be a problem when we are interested in these 'important components', because in some situations, similar cancers can only be distinguished by certain miner genes. It is hard to say that weighted projection weights are better than united ones to help improve the generalisation performance.

Though V is a natural choice of projection weights, we advocate using W to replace V. As for the vector length of W, the length of each projection weight is unit which is guaranteed by the PLS algorithm.

It is noted that the latent components projected by W is not the same as original PLS latent components T. The orthogonality of latent components is not preserved as well, while we consider the modification of T is trivial, since we just use PLS as a dimension reduction tool with W and V.

4 Experiments

4.1 Date sets

Four real microarray data sets are used in our study which are briefly described as below.

Leukemia

The acute leukemia data set was published by Golub et al. (1999). The original training data set consists of 38 bone marrow samples with 27 ALL and 11 AML (from adult

patients). The independent (test) data set consists of 24 bone marrow samples as well as 10 peripheral blood specimens from adults and children (20 ALL and 14 AML). Four AML samples in the independent data set are from adult patients. The gene expression intensities are obtained from Affymetrix high-density oligonucleotide microarrays containing probes for 7129 genes.

Colon

Alon et al. (1999) used Affymetrix oligonucleotide arrays to monitor expressions of over 6,500 human genes with samples of 40 tumour and 22 normal colon tissues. Using two-way clustering, Alon et al. were able to cluster 19 normal and five tumour samples into one group and 35 tumour and three normal tissues into the other. Expression of the 2000 genes with highest minimal intensity across the 62 tissues were used in the analysis.

Prostate

Singh et al. (2002) used microarray expression analysis to determine whether global biological differences underlie common pathological features of prostate cancer and to identify genes that might anticipate the clinical behaviour of Prostate tumours. In Singh's experiments, the training set contains 52 prostate tumour samples and 50 non-tumour (labeled as 'Normal') prostate samples with around 12600 genes. An independent set of test samples is also prepared, which is from a different experiment and has a nearly ten-fold difference in overall microarray intensity from the training data. After removing extra genes, 25 tumour and 9 normal samples were left in the test samples.

Central Nervous System

Pomeroy et al. (2002) developed a classification system based on DNA microarray gene expression data derived from 99 patient samples of Embryonal tumours of the Central Nervous System (CNS). Only data set C is used in our study. The data set contains 60 patient samples, 21 are survivors and 39 are failures. Survivors are patients who are alive after treatment whiles the failures are those who succumbed to their disease. There are 7129 genes in the data set.

4.2 Experimental settings

For each data set, 100 random partitions into a training data set $\mathcal{L}$ containing $n_{\mathcal{L}}$ observations and a test data set $\mathcal{T}$ containing the $n - n_{\mathcal{L}}$ remaining observations are generated. The class distribution of the training and test data set is the same as the original data set. If the data set was split already, we construct a whole observation collection by pooling them together. This scheme is widely used in the comparative studies of classification methods for microarray data (Dudoit et al., 2002). It is more reliable than leave-one-out cross-validation (Ambroise and McLachlan, 2002). We fix the partition ratio $n_{\mathcal{L}}/n$ at 0.5.

For each partition $\{\mathcal{L}, \mathcal{T}\}$ the gene expressions are transformed to have zero mean and standard deviation one across samples on $\mathcal{L}$. In the test set $\mathcal{T}$, data expressions are transformed according to the means and standard deviations of the corresponding training set $\mathcal{L}$. As no gene selection is performed, all genes of the original data set are used in our study.

We compare the classification performance with features extracted by W, V and V^*, where V^* is a normalised form of V. In order to avoid bias we predict the observations in $\mathcal{T}$ using three classical classifiers: SVM, LD and Ridge Regression (RR). These models have been widely used for binary classification problems (Dudoit et al., 2002).

We compare dimension reduction with the following cases in our experiments:

- W: Original projection weights W produced by PLS iteratively.
- V: The modified projection weights V which related to X directly. In order to preserve the orthogonality of T, the orthogonality of these projection weights is lost during the optimisation of PLS.
- V^*: We eliminate the punishing effect on vector length of V by normalising the projection weights to unit. The emended projection weights are denoted as V^*.

In our experiments, the number of project weights is retained as a meta-parameter. In order to examine how the classification performance varies with the dimension of latent components increasing, we vary the dimension of reduced space from 1 to 30 for all the three series of projection weights.

The mean classification success (accuracy) rate (SUC) is used to evaluate the different performance between W, V and V^*. The definition of SUC is given by

$$\text{SUC} = \frac{1}{100}\sum_{j=1}^{100}\frac{1}{n_{\mathcal{T}j}}\sum_{i=1}^{n_{\mathcal{T}j}} I(\widehat{Y}_i = Y_i) \tag{10}$$

where I is the standard indicator function ($I(A) = 1$ if A is true, $I(A) = 0$ otherwise). Note that the linear version of SVM is used and the parameter C of SVM is set to 100 in our study. All the classification models have been applied with the same partitions and data preprocessing.

4.3 Results and discussions

The SUC results on four different data sets are shown in Figures 1–4 respectively, the SUC results are averaged on 100 random partitions, from which we can see that the classification performance with W is much better than those with the other two weights V and V^*. Several further observations are made as follows.

- The corresponding SUC scores with W are higher than those with V on all data sets with respect to different classifiers. The top value with W is 1.6%, 0.8%, 1.9% and 0.7% better than those with V by three classifiers on the data sets of Leukemia, Colon, Prostate and CNS respectively, which show W is better than V for the dimension reduction to improve the generalisation performance of classifiers.
- The average value with V^* on each dimension is 0.1% 0.2% and 0.1% better than V on the data sets of Leukemia, Colon and Prostate respectively and no improvement is found on the data set of CNS, which show that V^* exhibit the same performance as V.

Figure 1 Statistical results by using three classifiers on the Leukemia data set (training set with 36 samples and 7129 genes, test set with the same size) (see online version for colours)

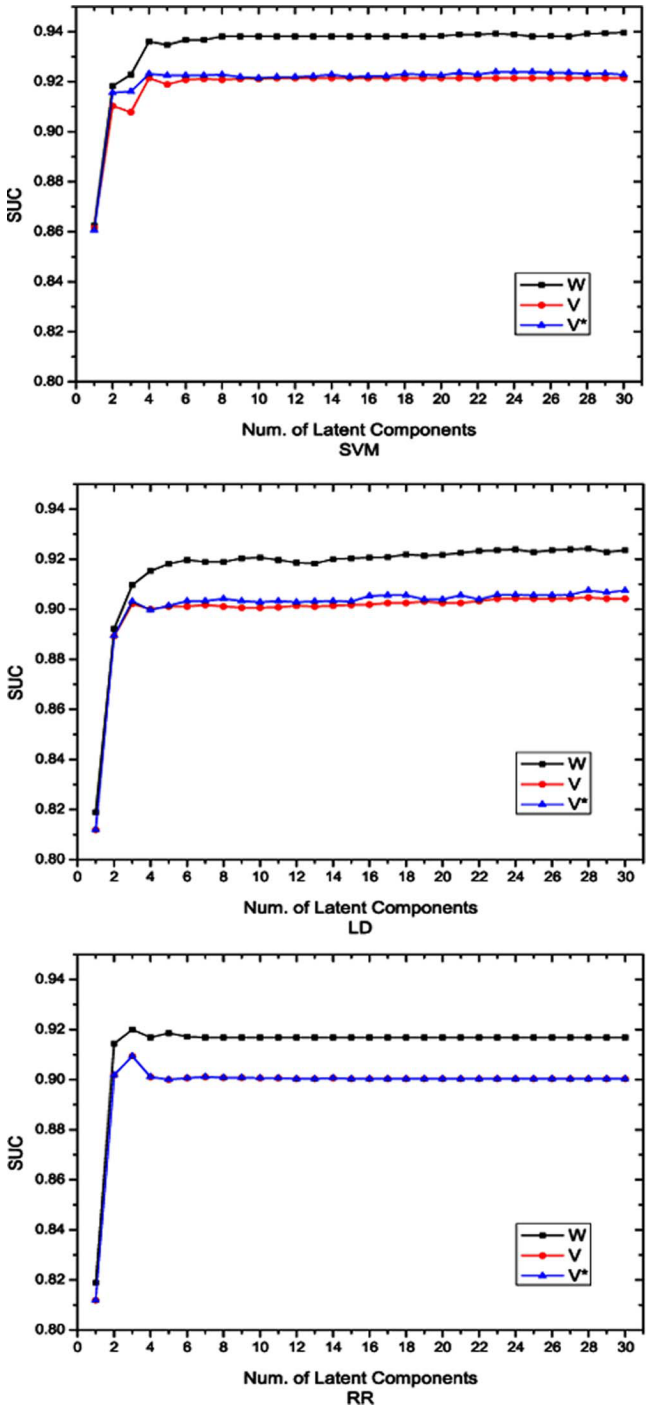


Figure 2 Statistical results by using three classifiers on the Colon data set (training set with 31 samples and 2000 genes, test set with the same size) (see online version for colours)

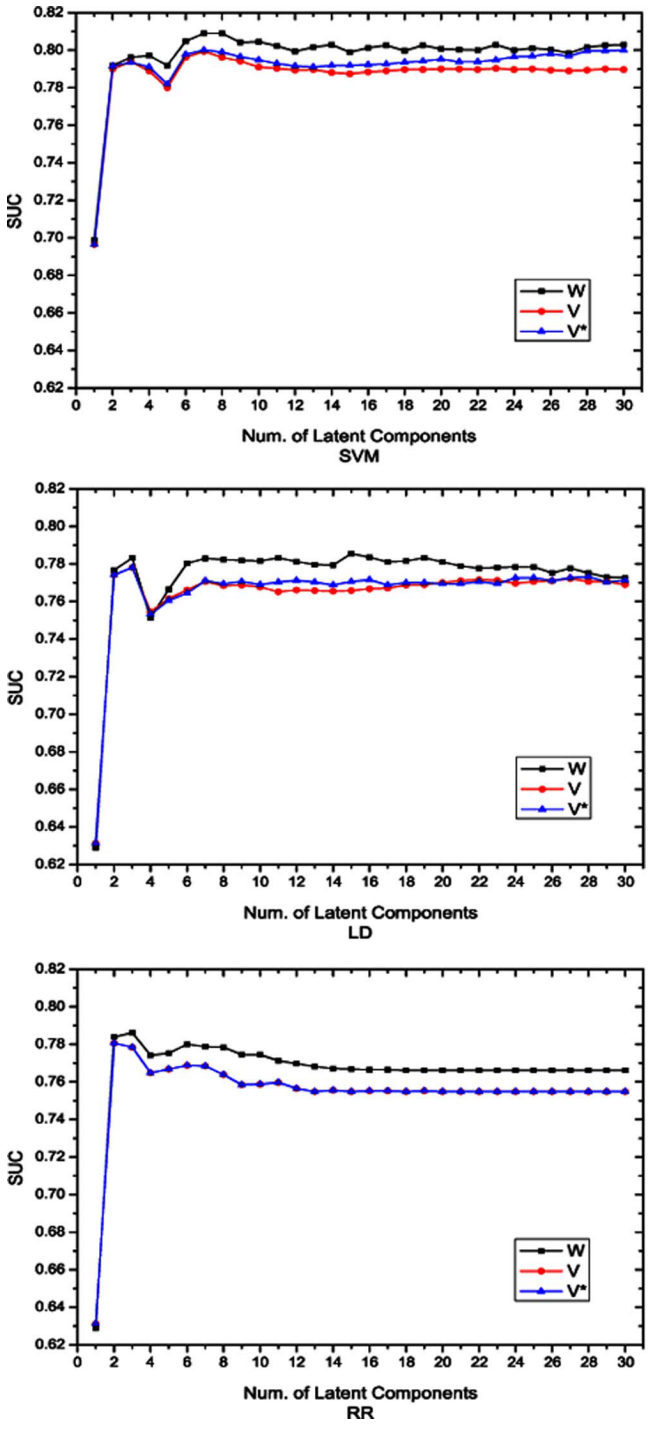


Figure 3 Statistical results by using three classifiers on the Prostate data set (training set with 68 samples and 12600 genes, test set with the same size) (see online version for colours)

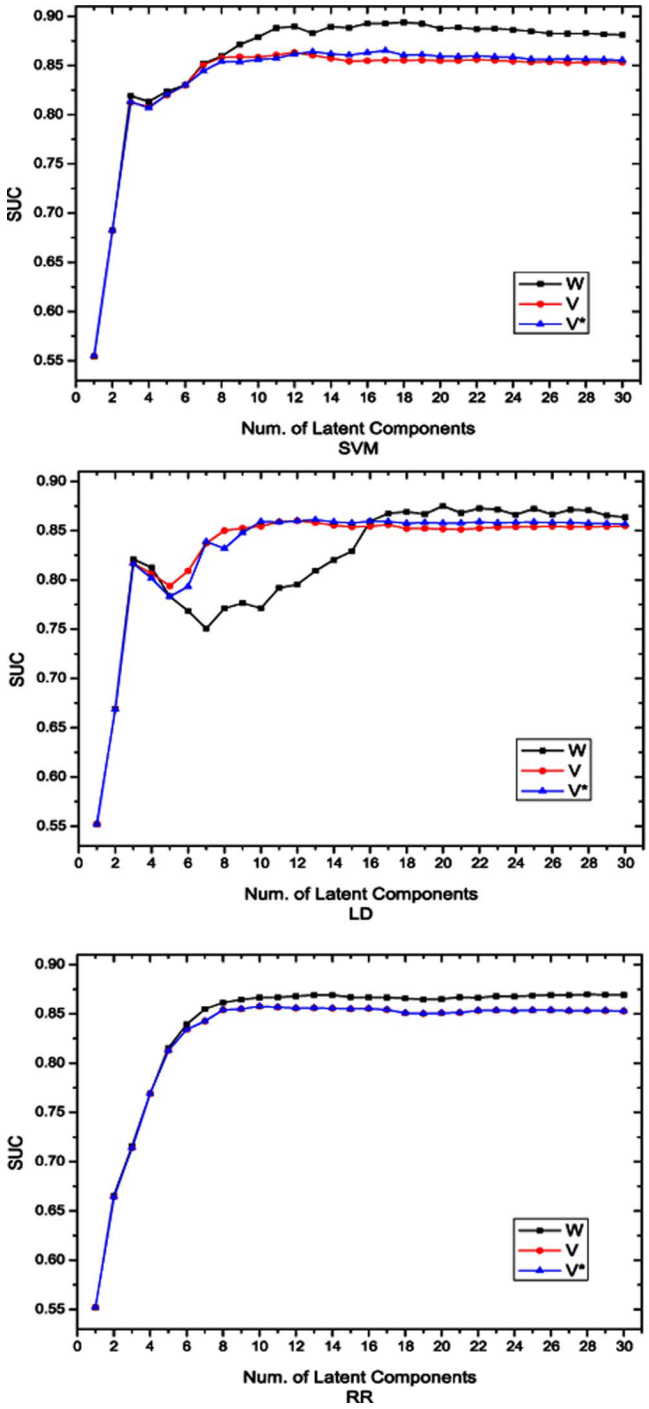
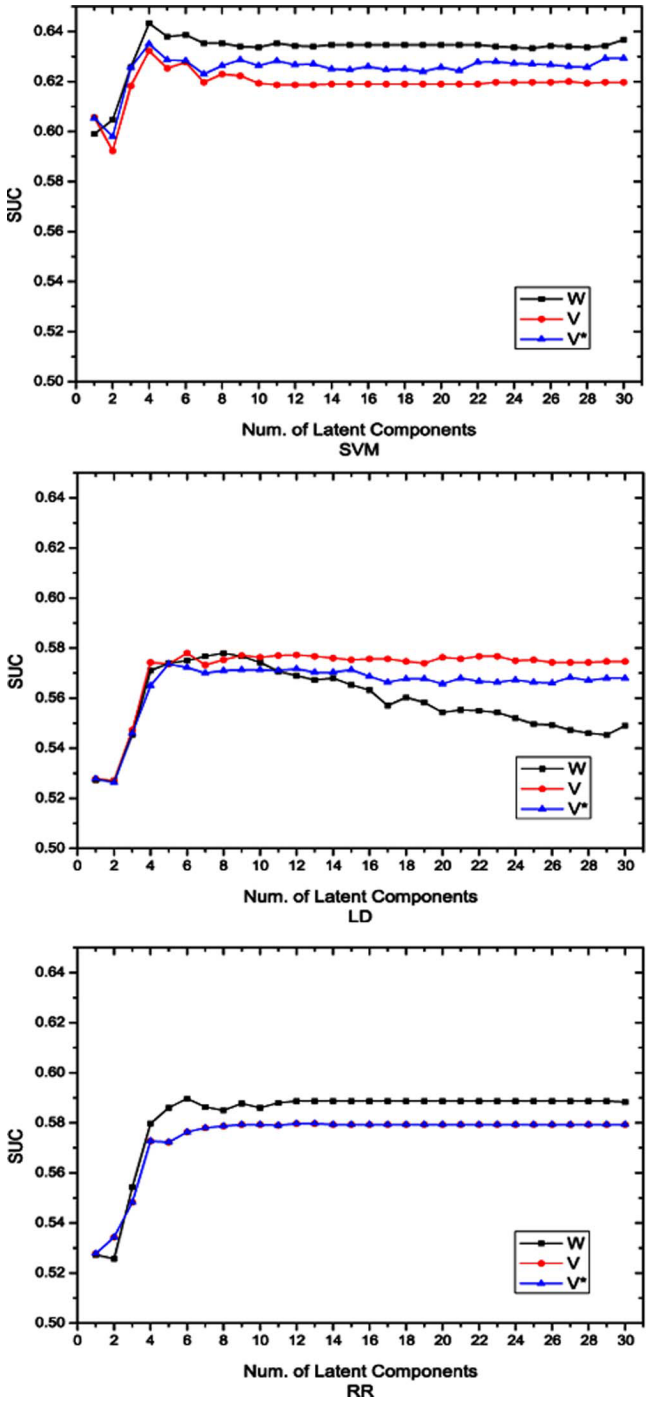


Figure 4 Statistical results by using three classifiers on the CNS data set (training set with 30 samples and 7129 genes, test set with the same size) (see online version for colours)



On the other hand, this also indicate that the classifiers used here can tolerate the imbalance of the scale of dimensions in some degree. In particularly, scores got by RR make no sense for the comparison between V and V^*. Due to the discriminative function, RR is insensitive to the weight of input variables.

5 Conclusions

This work investigates the difference of two series of projection weights in dimension reduction based on PLS from the view of orthogonality. W are the orthogonal projection weights related to residual matrix E_k, while V are the nonorthogonal weights linked with original matrix X directly.

We propose to use W instead of V as the projection weights in the dimension reduction for the orthogonality of W. Experimental results on four real microarray data sets proved our proposal that W is better than V to be used in dimension reduction for classification on high dimensional data set. We also examine the uniformity of vector length of V and find that the unit of direction length is not important for the classification of cancer.

Acknowledgement

Many thanks go to Prof. Yuehui Chen for his invitation. This work was supported by the Natural Science Foundation of China under Grant No. 20503015 and 60873129, the STCSM 'Innovation Action Plan' Project of China under Grant No. 07DZ19726, the Shanghai Rising-Star Program under Grant No. 08QA14032, Systems Biology Research Foundation of Shanghai University, and Scientific Research Fund of Jiangxi Provincial Education Departments under Grant No. 2007-57.

References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999) 'Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays', *Proceedings of the National Academy of Sciences of the United States of America*, pp.6745–6750.

Ambroise, C. and McLachlan, G. (2002) 'Selection bias in gene extraction on the basis of microarray gene-expression data', in Fienberg, S. (Ed.): *Proceedings of the National Academy of Sciences*, pp.6562–6566.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) 'Effective dimension reduction methods for tumour classification using gene expression data', *Bioinformatics*, Vol. 19, No. 5, pp.563–570.

Barker, M. and Rayens, W. (2003) 'Partial least squares for discrimination', *Journal of Chemometrics*, Vol. 17, No. 3, pp.166–173.

Boulesteix, A-L. (2004) 'Pls dimension reduction for classification of microarray data', *Statistical Applications in Genetics and Molecular Biology*, Vol. 3, No. 1.

Boulesteix, A-L. and Strimmer, K. (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.pdf.

Dai, J., Lieu, L. and Rocke, D. (2006) 'Dimension reduction for classification with gene expression data', *Statistical Applications in Genetics and Molecular Biology*, Vol. 5, No. 1, Article 6.

Dudoit, S., Fridlyand, J. and Speed, T. (2002) 'Comparison of discrimination methods for the classification of tumours using gene expression data', *Journal of the American Statistical Association*, Vol. 97, No. 457, pp.77–87.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression', *Bioinformatics and Computational Biology*, Vol. 286, No. 5439, pp.531–537.

Helland, I. (1988) 'On the structure of partial least squares regression', *Communications in Statistics. Simulation and Computation*, Vol. 17, No. 22, pp.581–607.

Helland, I. (2001) 'Some theoretical aspects of partial least squares regression', *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, pp.97–107.

Hoskuldsson, A. (1988) 'Pls regression methods', *Journal of Chemometrics*, Vol. 2, No. 3, pp.211–228.

Manne, R. (1987) 'Analysis of two partial-least-squares algorithms for multivariate calibration', *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, Nos. 1–3, pp.187–197.

Martens, H. (2001) 'Reliable and relevant modeling of real world data: a personal account of the development of pls regression', *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, pp.85–95.

Nguyen, D. and Rocke, D. (2002a) 'Multi-class cancer classification via partial least squares with gene expression profiles', *Bioinformatics*, Vol. 18, No. 9, pp.1216–1226.

Nguyen, D. and Rocke, D. (2002b) 'Tumor classification by partial least squares using microarray gene expression data', *Bioinformatics*, Vol. 18, No. 1, pp.39–50.

Nguyen, D., David, D.M. and Rocke, M. (2004) 'On partial least squares dimension reduction for microarray-based classification: a simulation study', *Computational Statistics and Data Analysis*, Vol. 46, No. 3, pp.407–425.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerovak, L., Blackk, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T. and Wetmo, C. (2002) 'Prediction ofcentral nervous system embryonaltumouroutcomebased on geneexpression', *Nature*, Vol. 415, No. 6870, pp.436–442.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. and Sellers, W. (2002) 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell*, Vol. 1, No. 2, pp.203–209.

Wold, H. (1975) *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, chapter Path models with latent variables: the NIPALS, approach, p.307.

Wold, S., Ruhe, A., Wold, H. and Dunn, W. (1984) 'Collinearity problem in linear regression. The partial least squares (pls) approach to generalized inverses', *SIAM Journal of Scientific and Statistical Computations*, Vol. 5, No. 3, pp.735–743.

Wold, S., Sjostrom, M. and Eriksson, L. (2001) 'Pls-regression: a basic tool of chemometrics', *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, No. 22, pp.109–130.

Zeng, X-Q., Wang, M-W. and Nie, J-Y. (2007) 'Text classification based on partial least square analysis', *The 22nd Annual ACM Symposium on Applied Computing, Special Track on Information Access and Retrieval*, pp.834–838.