Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 6 (2011) 273–278

Procedia Computer Science

Complex Adaptive Systems, Volume 1
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2011- Chicago, IL

Partial Least Squares (PLS) Applied to Medical Bioinformatics

Walker H. Land, Jr. [a*], William Ford [a*], Jin-Woo Park [a*], Ravi Mathur [a*], Nathan Hotchkiss [a*], John Heine [b*], Steven Eschrich [b*], Xingye Qiao [c*] and Timothy Yeatman [b*]

a Dept. of Bioengineering, Binghamton Univ., Binghamton NY, 13902-6000, USA.*
b Moffitt Cancer Center and Univ. of South Florida, Tampa FL, 33612, USA.*
c Dept. of Mathematical Sciences, Binghamton Univ., Binghamton NY, 13902-6000, USA.*

Abstract

PLS initially creates uncorrelated latent variables which are linear combinations of the original input vectors X_i, where weights are used to determine linear combinations, which are proportional to the covariance. Secondly, a least squares regression is then performed on the subset of extracted latent variables that lead to a lower and biased variance on transformed data. This process, leads to a lower variance estimate of the regression coefficients when compared to the Ordinary Least Squares regression approach. Classical Principal Component Analysis (PCA), linear PLS and kernel ridge regression (KRR) techniques are well known shrinkage estimators designed to deal with multi-collinearity, which can be a serious problem. That is, multi-collinearity can dramatically influence the effectiveness of a regression model by changing the values and signs of estimated regression coefficients given different but similar data samples, thereby leading to a regression model which represents training data reasonably well, but generalizes poorly to validation and test data. We explain how to address these problems, which is followed by performing a PLS hypotheses driven preliminary research study and sensitivities analysis by not doing a combinatorial analysis as PLS will eliminate the unnecessary variables using a microarray colon cancer data set. Research studies as well as preliminary results are described in the results section.

Keywords: Partial Least Squares; biomarker research; statistical learning theory;complex adaptive systems;colon cancer;microarrays

1. INTRODUCTION

One of the most popular current topics in bioinformatics is gene selection from microarray data because this problem involves both statistical processing as well as biological concepts. The statistical problems are daunting because of the large number of represented genes relative to the number of samples; this provides a prime opportunity to significantly overfit the data when model building. Biology is a significant component because identifying significant genes representative of a given clinical endpoint is a critical step toward understanding the

1877–0509 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.procs.2011.08.051

biological process. Several consequences result because of the statistical overfitting problem, some of which are: (1.)Very large area under the Receiver Operating Characteristic (ROC) curve, A_z (or AUC) values, can be achieved on training (and validation) data, but the results provided by these trained complex adaptive systems (CAS) frequently fail to generalize to other data sets nor do these CAS system designs necessarily operate on similar data sets with larger representative samples. Also, different CAS solutions may produce different gene sets from the same set microarray data set. Consequently, any CAS should first attempt to achieve some sort of repeatability, and (2.) Secondly, because of the overfitting problem described above, each proposed feature (or gene) reduction CAS generally is based on a unique theoretical analysis, which means that how these separate CAS are connected is not understood. Consequently, this difficulty results in the same problem stated above: different algorithms will generate different prognostic gene sets using the **same** microarray data. This means that developing an underlying theory for feature selection (to reduce gene dimensionality) would help to understand these algorithms as well as classify which of these are the "most" valid for gene selection**.** Song [1] presents a BAHSIC algorithm which claims to address this unifying algorithm principle proposal. BAHSIC defines a class of backward (BA) elimination feature selection algorithms that use kernels and the Hilbert-**S**chmidt Independence **C**riterion (HSIC) [2]. Song [1] demonstrates that the BAHSIC algorithm encompasses the following well-known feature selection algorithms: (1) Pearson's correlation coefficient [van't Veer, *et al.*; 3,4], (2) Ein-Dor, *et al.*; 5], (3) t-test [Tusher, *et al.*, ; 6], (4) signal-to-noise ratio [Golub, *et al.*, 7], (5) Centroid [Bedo, *et al.*, 8 and Hastie, *et al.*, ;9], (6) Shrunken Certroid [Tibshirani, *et al.*, 10,11], and finally, (7) ridge regression [Li and Yang, 12]. These collective results suggest the Evolutionary Programming derived Support Vector Machine (EP-SVM) [Land, W, *et al.* 13, 14] with a choice of similarity, sum and product kernels might be a good wrapper / classification candidate for gene selection. This paper adapts a method, summarized in the methods section, originally developed for the social sciences and subsequently adapted to chemometrics, called Partial Least Squares (PLS) to this "feature-rich /case-poor" environment, as subsequently described, by theoretically attempting to eliminate those features which do not contribute to the "best" chromosome marker for colon cancer.

2. DATA BASE DESCRIPTION AND COURSE FEATURE REDUCTION DESCRIPTION

For this study, DNA microarray data was collected from 104 Colon cancer patients who were treated at Moffitt Cancer Center & Research center. All tissue samples were collected during surgical resection (tumor is completely removed) by an IRB approved protocol and each sample was arrayed on the Affymetrix HG-U133+ GeneChip, that included 54,675 probes. Furthermore, each sample was classified as good or poor prognosis based on the time the patient was monitored and their status (dead or alive) at last contact. If the patient was monitored for less than 36 months and their status at last contact was dead, then they were classified as poor prognosis. All other scenarios resulted in good prognosis. The 36-month cut-off is a standard clinical measurement for prognosis. All Duke Stages (A through D) are represented in this cohort.

A major difficulty in analyzing microarray data is that they have many more features (genes) than cases (patients), which leads to the problem of overfitting (i.e."feature-rich / case- poor"). Many patterns can be found that appear to accurately distinguish the clinical classes, but these patterns often fail on new cases. By choosing to focus on gene sets we increase the risk of overfitting by opening the door to exploiting chance variances in the data.

The set of 54,675 probes (or features) were reduced to 271, using a t-test, variance pruning and adding biological relevant genes. The t-test selected features that accurately distinguished between good and poor prognosis by comparison of the distribution means, whereby the features were reduced to 5,157 using a p-value cut-off of 0.05. (Note: A high variance of the expression for the probe must exist between the two cases). Then variance pruning reduced the feature set further to 203, using a variance cut-off of 0.45. Finally, 68 known biologically relevant genes were added to the final feature set based on their importance in the prognosis of colorectal cancer. However, with this reduced feature set, over-fitting is still a major issue because the ratio of features to cases is close to 3 to 1 and the number of possible biomarker sets is 10^{81}. PLS is an excellent solution candidate to appliy at this point because of its ability to address overfitting problems of this type.

3. METHODS

This section contains a brief, heuristic overview of Partial Least Squares (PLS). PLS is an extension of least squares regression (LR). In LR, the response variable y is predicted from *p* coordinates and *n* observations, denoted by $X = \{x_1, x_2, \ldots x_n\}^{\mathrm{T}}$, where each $x_i \in R^p$. PLS finds "new variables" through the construction of specific

combinations of the original coordinates. These "latent variables" explain both the y response as well as the covariate space and are denoted by the following expressions:

$$\boldsymbol{X} = t_1 p_1 + t_2 p_2 + \ldots + t_s p_s + \varepsilon$$
$$\boldsymbol{y} = t_1 q_1 + t_2 q_2 + \ldots + t_s q_s + \zeta$$

where:

- t_S = latent variables (or conjugate vector directions; *n* by 1 column vectors). Generally most of the variability is characterized by *M* latent variables with a maximum of *M*=5 required for most problems.
- p_s and q_s=weights (p_s are 1 by *p* row vectors, q_s are scalar).
- ε, ζ=small errors in the remaining parts not explained by the latent variables

For this microarray data set, we began with 271 features and reduced this set to a minimum of 1 latent variable and a maximum of 5 latent variables (see *Results* section). Therefore, the principle advantage of PLS for a problem of this type is its ability to handle a very large number of features: ***a fundamental problem of a feature-rich/case-poor data set***. PLS then performs a least-squares fit (LSF) onto these latent variables, where this LSF is a linear combination that is highly correlated with the desired y response while, at the same time, accounting for the feature space variability.

Secondly, what are some of the advantages of PLS?

- PLS algorithms are very resistant to over-fitting, they are fast and reasonably easy to implement when compared to least squares regression.
- For most problems with few data points and high dimensionality, least squares fails, where PLS excels.
- PLS may be considered a better principal component analysis (PCA).
- PLS regression maps the original data into a lower-dimensional space using a W projection matrix and computes a least squares solution in this space.
- The first key difference from PCA is that PLS computes an orthogonal factorization of the $\vec{X}$ input vector and response $\vec{y}$ (note: y can also be a vector) response in the process of computing the projection matrix W
- The second key difference from PCA is that the least squares model for kernalized PLS (K-PLS) is based on approximation of the input and response data, not the original data (Note: k-PLS is simply the PLS process kernalized).
- PLS and PCA use different math models to compute the final regression coefficients. Specifically, the difference between PCA and PLS is that a new set of basis vectors (similar to the eigenvectors of $\mathbf{X^T X}$ in PCA) is NOT a set of succession of orthogonal directions that explain the largest variance in data, but rather are a set of CONJUGATE GRADIENT VECTORS in the correlation matrices which span a Krylov space.
- What makes PLS especially interesting for biomedical applications and data mining applications is its extension using kernels, similar to support vector machines.

Finally, a summary to the PLS paradigm as implemented in the paper follows:

Algorithm 1: $X_0 = X, y_0 = y$

For *m=1 to M*, where M = number of latent variables, *do*:

- Compute direction of maximum variance $w_m = (X_m)^T y_m$
- Project *X* onto $\vec{w}$ $t_m = X_m w_m$
- Normalize t $t_m = t_m / |t_m|$
- Deflate *X* $X_{m+1} = X_m - t_m (t_m)^T X_m$
- Deflate *Y* $y_{m+1} = y_m - t_m (t_m)^T y_m$
- Normalize *Y* after deflation $y_{m+1} = y_{m+1} / |y_{m+1}|$

Finally, compute the regression coefficients using latent variables: $\beta = W(T^T X W)^{-1} T^T y$

where: w_m is the mth column vector of w, t_m is the mth column vector of T, X_m *and* y_m are the input matrix and response vector that are being deflated, and β are the linear regression coefficients. A geometric representation of part of the algorithm (deflation) is depicted in Figure 2.

4. Results

To demonstrate the efficacy of using PLS to overcome the problem of over-fitting, and determine if PLS can be used to determine which features in the data set are important to predicting the patient prognosis, the goal of our experiments were twofold: 1) Use PLS to find genes indicative of a good/poor prognosis, and 2) Use PLS to build a classifier that will overcome the over-fitting problem and consistently predict good/poor prognosis from the data. ***By hypothesis****, and to address the first goal, we say the norm of the weight vector is a measure of 'importance' in that direction, and the magnitude of the vector component a measure of importance of the corresponding feature (gene).* Since PLS is a constructive process that finds the direction of maximum variance between X and y (before the variability explained by that latent variable 'deflation' occurs), the *first five or so* (at a maximum) weight vectors should explain most of the variability, and hence, have the largest norm. In fact, we found that *only* the first weight vector to be most important, as expected. Specifically, Figure 1 depicts the norm of the first 10 weight vectors, averaged over 500 trials.

Furthermore, the more important a gene is for making a correct classification, the larger the contribution of that feature will be to the first weight vector $\mathbf{w_1}$. To determine if PLS could be used to rank the significance of genes in the original data set, we conducted numerous independent trials, splitting the original dataset randomly in half, using half to develop the PLS model (see **Algorithm 1**), and half as a validation set. After each PLS model was constructed, each feature was given a score, with the largest vector component receiving a score of 30, the next largest a score of 29, and so on, with all but the largest 30 receiving a score of zero. After all trials were complete, the scores for each feature were summed, and normalized (Figure 4). Table 1shows the index and name of the 30 most important genes based on this method.

In determining the probes that are more influential than the others, a cut-off of thirty probes was chosen based on our previous work on *this* dataset (Mathur *et al.*, 2010 and 2011). For this research, a genetic algorithm-support vector machine (GA-SVM) hybrid was applied as a *wrapper feature selection algorithm*. The total length of probe subsets considered by Mathur *et al.* was shown to be thirty, which provided the GA with sufficient information to adequately adapt the probe subset length by the subset size operator (Schaffer et al, 2005) in the CHC GA (Eschelman, 1991). From the finding in this work we have adapted the same maximum length of the probes that we will consider more influential compared to the others.

To support the ***hypothesis that PLS can be used to find stable biomarkers***, we compared the 'important' genes found using PLS to those found using the GA-SVM method. Several genes were common. For example, genes MMP12, IGH, PRKAA1, GDAP1, ROBO2, LOC38983, ZNF207, UGGT2, and DLEU2 were also identified by Mathur *et al.* (2011) as being important indicators of prognosis. The biological significance of these genes is currently being investigated further.

To address the second goal, we again ran numerous trials, evenly dividing the data into training and validation sets, using the training data to construct the PLS classifier, and the validation set to measure the generalization capabilities of the classifier. The well-known metric used to measure the fitness of the PLS classifier was the A_z value. We found that using 3 latent variables could achieve an A_z value of around 0.71 for the validation set, indicating fairly good generalization (see Figure 3).

Finally, we removed all but the 30 best genes from the data set, and repeated the experiment. We found that the validation A_z peaked at one latent variable, and was about 10% higher (around 0.77) than the models using all 271 features (see Figure 4). *This is further evidence that PLS can be used to rank the importance of features.*

The data presented herein support the goals which we stated at the beginning of this section. We have found certain genes that explain much of the variability in the data, and are therefore important predictors of prognosis. This is further supported by the fact that many of these genes were also found by a completely different technique (GA-SVM). We achieved the second goal of overcoming the over-fitting problem by fairly high validation A_z values.

Furthermore, since the average validation A_z values increased when only the 30 most important genes were used as predictors further supports the first goal.

Future work will concentrate on using the non-linear 'kernalized' version of PLS to determine if better classifiers can be constructed from non-linear PLS regression models.

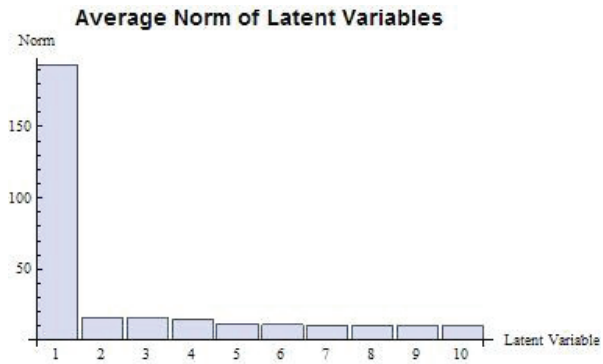

Figure 1 – Average Norm of Weight Vectors from 500 Independent Trials.

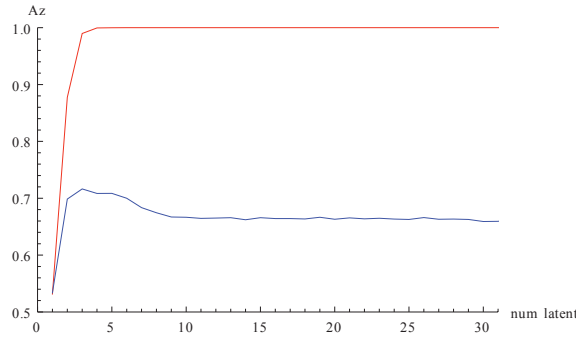

Figure 3 – ROC A_z Average as a Function of the Number of Latent Variables Used to Construct the Linear Regression Model (Training: Red Circle; Validation: Blue, Square).

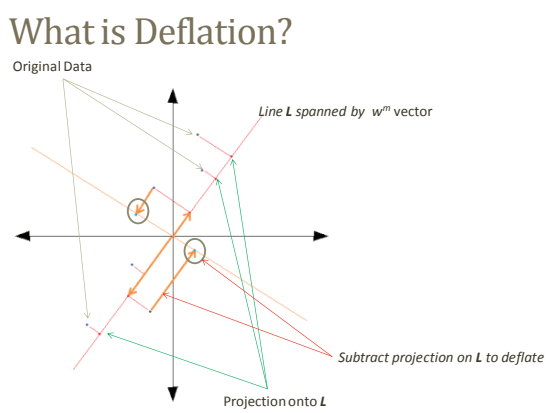

Figure 2 – Geometric Explanation of 'Deflation'

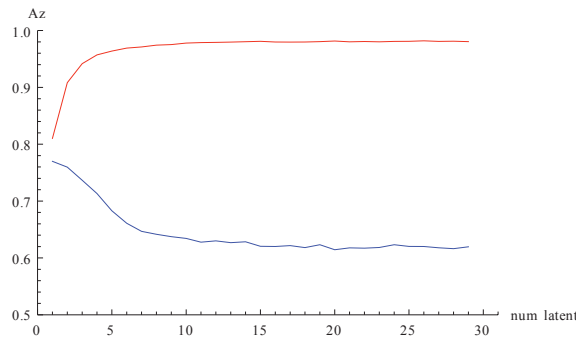

Figure 4 – ROC A_z Average as a Function of the Number of Latent Variables Used to Construct the Linear Regression Model Using Only the Top 30 Genes (Training: Red Circle; Validation: Blue, Square).

Probe Index	Gene Name
4	MMP12
26	IGH
27	IGH
44	PRKAA1
57	ADAMTS5
61	N4BP2L2
62	HNRNPA1
69	PCDHB10

Probe Index	Gene Name
87	GDAP1
88	N/K
91	ROBO2
93	DOCK11
96	LOC727820
100	N/K
102	LOC389831
104	N/K

Probe Index	Gene Name
107	ZNF207
108	RBM15
110	JMJD1C
122	ZNF207
127	ZCCHC7
153	UGGT2
159	N/K
163	N/K

Probe Index	Gene Name
166	CBFA2T
167	C11orf80
197	N/K
198	DLEU2
232	CXCR6
252	FLT1

Table 1 – Thirty Most Influential Genes Based on the Norm of the First Weight Vector

5. Conclusions

This research effort provided the following preliminary conclusions:

- To support the ***hypothesis that PLS can be used to find stable biomarkers***, we compared the 'important' genes found using PLS to those found using the GA-SVM method. Several genes were common. For example, genes MMP12, IGH, PRKAA1, GDAP1, ROBO2, LOC38983, ZNF207, UGGT2, and DLEU2 were also identified by Mathur *et al.* (2011) as being important indicators of prognosis. The biological significance of these genes is currently being investigated further.
- We found that the validation A_z peaked at one latent variable, and was about 10% higher (around 0.77) than the models using all 271 features (see Figure 4). *This is further evidence that PLS can be used to rank the importance of features.*
- *The data presented herein support the goals which we stated at the beginning of this section. We have found certain genes that explain much of the variability in the data, and are therefore important predictors of prognosis. This is further supported by the fact that many of these genes were also found by a completely different technique (GA-SVM). We achieved the second goal of overcoming the over-fitting problem by fairly high validation A_z values.* Furthermore, since the average validation A_z values increased when only the 30 most important genes were used as predictors further supports the first goal.

References

[1] Song, L., Bedo, J., Borgwardt, K.M., Gretton, A., Smola, A., "Gene Selection via the BAHSIC family of algorithms", Bioinformatics, vol. 00, no. 00, 2007, pp. 1-9

[2.] Gretton, A., Bousquet, O., Smola,A., Scholkopf,B., "Measuring Statistical dependence with Hilbert-Schmidt norms", In Proc. Intl. Conf. on Algorithmic learning theory, pp.63-78, 2005

[3.] van de Vijver, M.J., He, Y.D., van't Veer L.J., *et al.*, "A gene-expression signature as a predictor of survival of breast cancer", N. Engl. J. Med. 2471999-2009 (2002)

[4.] van't Veer L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., et. Al, "Gene expression profiling predicts outcome of breast cancer", Nature, 415, pp. 530-536, 2002

[5.] Ein-Dor, L., Zuk, O.,and Domany, E., "Thousands of samples are needed to generate robust gene list for predicting outcome of cancer", Proc. Natl. Acad. Sci.., USA, 103(15), pp 5923-5928, 2006

[6.] Tusher, V.G., Tibshirani, R., and Chu, G., " Significance analysis of microarrays applied to ionizing radiation response", Proc. Natl. Acad. Sci., USA, 98(9) pp 5516-5121, 2001

[7.] Golub, T.R., Solnim, D.K., Tamayo, P., *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", Science, 286(5439) pp. 531-537, 1999

[8.] Bedo, J., Sanderson, C., and Kowalczyk, A., "An efficient alternative to SVM based recursive feature selection with application in natural language processing and bioinformatics", In artificial intelligence (to appear)

[9.] Hastie, T., Tibshirani, R., and friedman, J., "The elements of statistical Learning", Springer, New York

[10.] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., " Diagnosis of multiple cancer types by shrunken centroids of gene expression", In National Academy of sciences, volume 99, pp. 6567-6572, 2002

[11.] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G., "Class prediction by nearest shrunken centroids, with application to dns microarrays", Stat. Sci. 18, pp.104-117, 2003

[12.] Li, F., and Yang, Y., "Analysis of recursive gene selection approaches from microarray data", Bioinformatics, 21(19), pp. 3741-3747, 2005

[13.] Land, Walker H. Jr., Heine, J., Tomko, G., Mizay, A., Gupts, S., and Thomas, R., " Performance Evaluation of Evolutionary computational and conventionally trained support vector machines" [6560-30], *Intelligent Computing: Theory and Applications* V, Edited by Kevin L. Priddy and Emre Ertin, Proc. SPIE Vol. 6560,65600W, (2007) 0277-786X/07/ $18 doi: 10.1117/12.716543

[14.] Land, Walker H., Jr., Heine, J., Tomko, G., and Thomas, R., " Evaluation of Two Key Machine Technologies"[6560-28], *Intelligent Computing: Theory and Applications V*, Edited by Kevin L. Priddy and Emre Ertin, Proc. SPIE Vol. 6560,65600W, (2007) 0277-786X/07/ $18 doi: 10.1117/12.716543

[15.] Mathur,R.,*et al.*, "Evolutionary computation with noise perturbation and cluster analysis to discover biomarker sets*", CAS,2011*