*Journal of the Korean Data & Information Science Society*
*2014,* **25**(5), 1151–1160

http://dx.doi.org/10.7465/jkdi.2014.25.5.1151
한국데이터정보과학회지

# A modified partial least squares regression for the analysis of gene expression data with survival information[†]

So-Yoon Lee[1] · Myung-Hoe Huh[2] · Mira Park[3]

[1]Credit Bureau Business Department, NICE Information Service
[2]Department of Statistics, Korea University
[3]Department of Preventive Medicine, Eulji University

Received 30 June 2014, revised 11 August 2014, accepted 22 August 2014

**Abstract**

In DNA microarray studies, the number of genes far exceeds the number of samples and the gene expression measures are highly correlated. Partial least squares regression (PLSR) is one of the popular methods for dimensional reduction and known to be useful for the classifications of microarray data by several studies. In this study, we suggest a modified version of the partial least squares regression to analyze gene expression data with survival information. The method is designed as a new gene selection method using PLSR with an iterative procedure of imputing censored survival time. Mean square error of prediction criterion is used to determine the dimension of the model. To visualize the data, plot for variables superimposed with samples are used. The method is applied to two microarray data sets, both containing survival time. The results show that the proposed method works well for interpreting gene expression microarray data.

*Keywords*: Gene expression, mean square error of prediction, partial least squares regression, survival time.

## 1. Introduction

In gene expression microarray studies, the number of genes far exceeds the number of samples, which is known to be "large p, small n" problem. Moreover, gene expression measures are highly correlated. For the gene expression data related to the response variable, several dimensional reduction methods such as principal components regression (PCR), partial least squares regression (PLSR) and sliced inverse regression (SIR) could be applied.

We focus on PLSR. There have been several studies for analyzing gene expression data using PLSR. Nguyen and Rocke (2002a) proposed using PLSR for the dimensional reduction as a preliminary step to classification. Kim (2003) proposed the unified procedure which

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2062848).
[1] Associate manager, NICE Information Service Co., Ltd., Seoul 150-973, Korea.
[2] Professor, Department of Statistics, Korea University, Seoul 136-701, Korea.
[3] Corresponding author: Professor, Department of Preventive Medicine, Eulji University, Daejeon 301-832, Korea. E-mail: mira@eulji.ac.kr

combine the PCR, PLSR and OLSR procedure. Fort and Lambert-Lacroix (2005) proposed the method combining PLSR and ridge penalized logistic regression for data-mining of microarray data. Dai *et al*. (2006) compared PLSR, SIR, and PCR as tools for the classification with gene expression data. Although PLSR can handle a large number of genes, they utilize linear combination of all genes. Therefore selecting a subset of the genes is necessary not to be contaminated by irrelevant genes.

For the case where the response variable is survival time, most of the proposed methods handle the survival time using Cox's proportional hazards model. Park *et al*. (2002) proposed the approach in which the full likelihood for proportional hazard (PH) regression model is reformulated as the likelihood of a Poisson model. The method using extracted PLSR gene components as a covariate in PH regression model is proposed by Nguyen and Rocke (2002b). After dimensional reduction by PLSR, they use the extracted PLSR components as covariates in a proportional hazard regression to predict the survival probabilities. During PLSR procedure, they treat censoring time as if it is a survival time, and thus it makes some bias in predicting the survival probability. Bøvelstad *et al*. (2007) compared the performances of seven methods for survival prediction. More recently, Nguyen and Rojo (2009) proposed rank-based PLSR.

In this study, we consider PLSR approach to predict survival time based on gene expression data. We develop a new gene selection method based on PLSR. Mean square error of prediction (MSEP) criteria is used to select the final model. To predict the survival probability, we follow the similar way to Nguyen and Rocke (2002b). However, to avoid the bias as in PLS-PH model, we construct the iteration procedure for the imputing censoring time. We also suggest plots visualizing both genes and response variable. We call these procedures as a modified PLS-PH method. It could be considered a variant of PLS-PH method of Nguyen and Rocke (2002b) with the imputation and variable selection procedure added. These methods are applied to two well-known microarray data sets; lymphoma and breast carcinoma data.

## 2. Methods

We introduce the traditional PLS algorithms first, and then describe the modified PLS-PH method. The modification accomplished by two ways: gene selection and censoring time imputation. We also suggest how to visualize the results.

### 2.1. Basic algorithms of PLSR

Let $X$ be the $n \times p$ data matrix with $n$ samples and $p$ predictor variables. Let $y$ be the $n \times 1$ response vector. Objectives of PLSR is to maximize the covariance between $y$ and a linear combination of columns in $X$, and thus to find the optimal weight vector $w_1$ satisfying

$$w_1 = \text{argmax } Cov(Xw_1, y)$$

subject to the constraint $w_1'w_1 = 1$ (Helland, 1988). The solution is given by

$$w_1 = X'y/\left|X'y\right|.$$

The first PLSR component can be constructed as $t_1 = Xw_1$. It is also a linear combination of the original variables and is called a score vector. The $X$ and $y$ using first PLSR component

can be predicted by

$$\widehat{X}^{(1)} = t_1 b_1'$$

and

$$\widehat{y}^{(1)} = t_1 c_1',$$

respectively, where $b_1^{'} = (t_1' t_1)^{-1} t_1' X$ and $c_1' = (t_1' t_1)^{-1} t_1' y$. Here $b_1'$ and $c_1'$ are called x-loading and y-loading vector, respectively.

The subsequent components can be obtained based on the residuals of $X$ and $y$ from the previous stage instead of the original data. Let $X^{(1)} = X$ and $y^{(1)} = y$. For $k^{th}$ statge, $X^{(k)}$ and $y^{(k)}$ are updated by

$$X^{(k)} = X^{(k-1)} - \widehat{X}^{(k-1)}$$

and

$$y^{(k)} = y^{(k-1)} - \widehat{y}^{(k-1)}$$

respectively, for $k = 2, \cdots, K$. The maximum number of components, $K$, is determined by the number of non-zero eigenvalues ($K \leq \min(n, p)$). Subsequent weight vector $w_k$'s are found as

$$w_k = \text{argmax } Cov(X^{(k)} w_k, y^{(k)})$$

subject to $w_k' w_k = 1, w_k' X^{(k)'} X^{(j)} w_j = 0$, for all $1 \leq j < k \leq K$. The score vector $t_k = X^{(k)} w_k$ is calculated. All the components are uncorrelated and ordered. Predicted values of $X$ and $y$ using $K$ components are given by

$$\widehat{X} = t_1 b_1' + t_2 b_2' + \ldots + t_k b_K'$$

and

$$y = t_1 c_1' + t_2 c_2' + \ldots + t_k c_K'.$$

First few components are usually enough to explain the most of the covariance. Thus they are retained as the new predictors.

### 2.2. Modified PLS-PH method

#### 2.2.1. Gene selection procedure

To select a subset of genes, we consider the backward elimination procedure and MSEP criteria for gene selection. The gene selection procedure we considered is as follows:

1) Apply classical PLSR procedure for the dataset containing all genes.

2) Determine optimal number of components $k$ from the cross-validation technique.
For $k^{th}$ component, the mean square error of prediction is calculated

$$MSEP(k) = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2 / n$$

where $y_i$ is the observed response value of the $i^{th}$ sample and $\widehat{y}_{(i)}$ is predicted response value fitted with all but the $i^{th}$ sample ($i = 1, \ldots, n$). Then $k$ is chosen as the optimal number of components at which MSEP achieves the minimum value.

3) Squared covariance is calculated:

$$C^2 = \sum_{j=1}^{k} (Cov(y_j, X_{j(l)} w_{j(l)})^2,$$

where $X_{j(l)}$ and $w_{j(l)}$ are gene expressions and weight vectors, respectively, where $l^{th}$ gene is deleted.

4) Remove the gene associated to maximum $C^2$ and compute new weight vector

$$w_{j(l)} = X_j(l)' y_j / \left| X_j(l)' y_j \right| .$$

5) Iterate steps 3) and 4) until $k$ variables are left.

6) Apply PLSR procedure to all gene subsets with $p^*$ genes for $k \leq p^* \leq p$.

7) Choose $p^*$ which minimizes MSEP and select corresponding subset of genes.

At step 1), to select optimal number of PLSR components $k$, we use a leave-one-out cross validation method with MSEP criteria. The MSEP criteria used again at step 7) to choose optimal gene subset among candidates. Proportion of the response variation explained by the reduced model is also considered as a secondary criterion. We may use forward selection procedure instead of backward elimination.

### 2.2.2. Censoring data estimation

During the PLSR procedure, the PLS-PH method was derived by Nguyen and Rocke (2002b) to treat censoring time as if it were a survival time. However, this results in underestimates of the survival probability. To avoid such a bias, we consider the iteration procedure for the imputing censoring time. Once PLSR is applied to all genes without considering censoring information, the censoring time $y_i$ is changed to predicted value $\widehat{y}_i$ from previous PLSR procedure if $y_i \leq \widehat{y}_i$. Then PLSR procedure is performed again to updated data with all genes. Now the optimal number of components $k$ is determined.

Gene selection procedure explained in 2.2.1 is then applied while keeping the minimum MSEP value and the corresponding gene subset. After updating the censoring time by predicted value from the previous PLSR results, these procedures are repeated until there is no significant change in minimum MSEP value. Finally, by applying proportional hazard

regression using the final PLSR components as covariates, we estimate the patient survival probability. See Figure 2.1. Note that we use the original survival and censoring time in this PH regression procedure. Note also that the updated censoring time is used only to construct PLSR components.

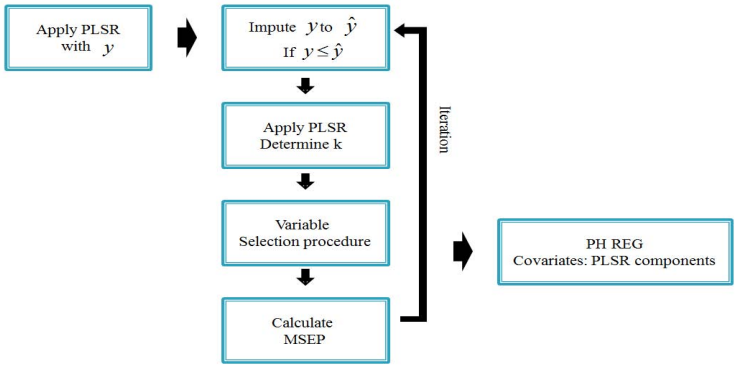



**Figure 2.1** Map of modified PLS-PH procedure

### 2.2.3. Data visualization

Instead of plotting x-loadings according to variable index as in classical PLS regression, we propose to plot the first two (or three) columns of

$$P_j = (x_j' t_1^*, x_j' t_2^*, \ldots, x_j' t_K^*)$$

to represent the gene $x_j$ where $t_j^* = t_j / |t_j|$. Similarly, to represent response variable $y$, we can plot the first two columns of

$$Q = (y' t_1*, y' t_2*, \ldots, y' t_K*).$$

Although each point of $P_j$ is the same as traditional x-loadings, this type of plot can provide a new interpretation by superimposing two plots.

We can interpret two genes that are located closely to have similar expression patterns while we can interet two genes located at a distance to have distinct expression patterns. Moreover, the genes directed toward the response variable are positively related to the survival time whereas the genes directed away from the response variable are negatively related to the survival time. For the plot of samples, using the same number of PLSR components, we interpret that the samples located near similar position tend to have a similar patterns. Samples pointing to the response variable should be positively related to the survival time.

## 3. Data analysis

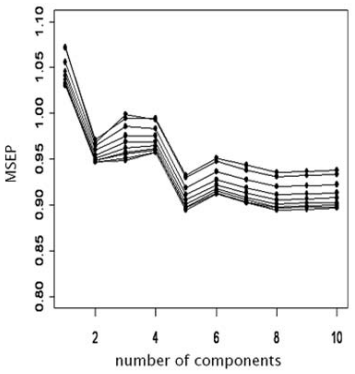
### 3.1. Lymphoma data

This data came from a cDNA experiment and has 40 diffuse large B-cell lymphoma (DLBCL) samples of 4026 genes (Alizadeh *et al.*, 2000). There are two subtypes of DLBCL samples; germinal centre B-like cell and activated B-like cell. Of the 40 samples, 18 samples

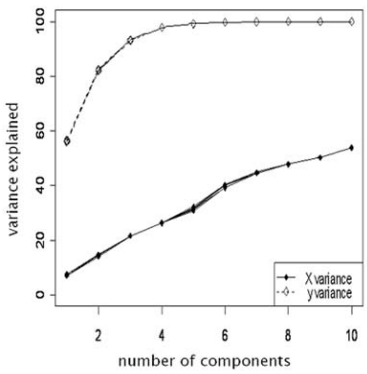
have censored survival time. The data were centered and standardized for each gene across the array.

Figure 3.1(a) is the graph of MSEP versus the number of components for the first 10 iteration stages, which represented by the lines from the top. For all stages, the MSEP is lowest when the number of components is 5. The corresponding cumulative response variation explained for 5 components is over 95% (Figure 3.1(b)). On the other hand, Figure 3.1(c) represents MSEP at each stage. Since resulting MSEP is lowest at the $9^{th}$ iteration stage, we take the stage 9 and corresponding subset of 180 genes (Figure 3.1(d)). We have 18 censored samples, among them 6 samples are imputed. The other 12 censored samples are not modified.

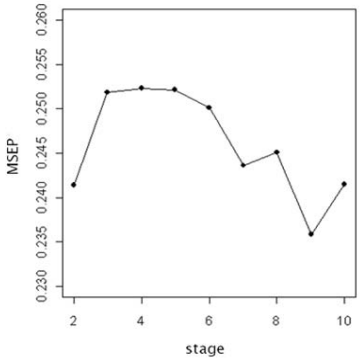
Figure 3.2 shows the loadings of genes and the locations of samples. Figure 3.2(a) displays 180 genes and the survival response. Because the survival time $y$ is located at the right upper part of the graph, genes in the first quadrant are positively related to the survival time, whereas the genes in left lower part are negatively related to survival time. Genes near the origin, being unrelated to the survival time, have been eliminated. For observation plot, we can interpret that samples lying on similar positions have similar expression patterns and the samples with positive projections on the line of 45 degree slope tend to have relatively long survival time (Figure 3.2(b)). Thus we can interpret that the germinal center B-cells tend to have long survival times than the activated B-like cells.



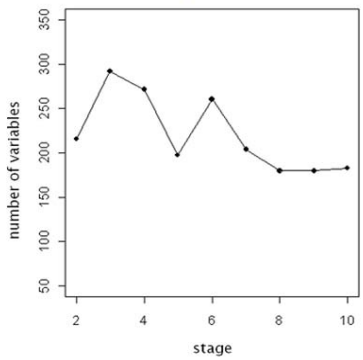

(a) MSEP vs. number of components




(b) variance explained vs. number of components




(c) MSEP vs. iteration stages

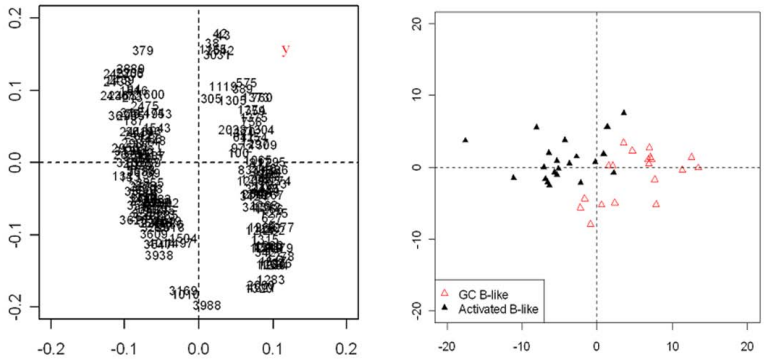



(d) number of variables vs. iteration stages

**Figure 3.1** Summary graphs for the lymphoma data

We fit the proportional hazard regression model using the final 5 PLSR components as the covariates. Patient survival probabilities are predicted by PH regression. Figure 3.3 shows the estimated survival curves obtained for the group-averaged component profiles. One can find that the predicted survival probability for activated B-like samples is distinctly lower than that for the GC B-like samples.

Note that Alizadeh *et al.* (2000) identified these two groups using hierarchical clustering. They selected 50 genes by applying two sample t-tests and provided survival curves of the two group from unadjusted survival times. Nguyen and Rocke (2002) estimated survival probabilities using the first two PLS components as covariates. They obtained PLS components based on 2000 genes. Since we made the survival curve from the PLSR including variable selection and imputing procedures, the estimated survival curves are different from two previous methods though the general patterns are similar.




(a) genes and survival response plot (b) sample plot

**Figure 3.2** PLSR loading plots for the lymphoma data

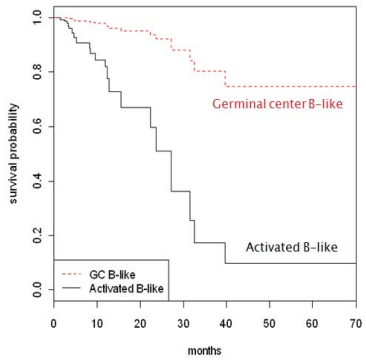
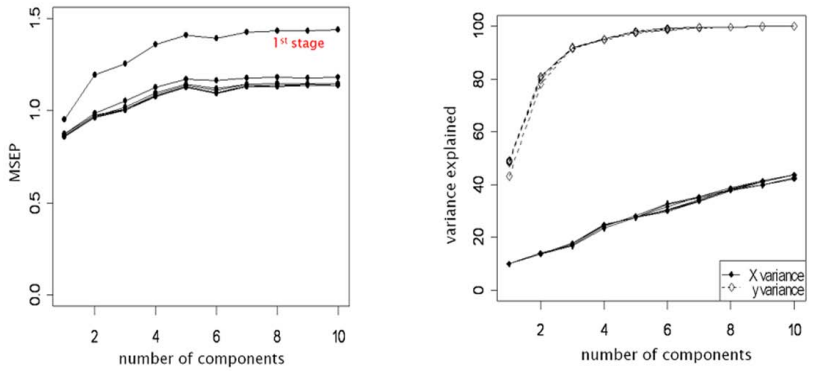



**Figure 3.3** Estimated survival curves for the lymphoma data

### 3.2. Breast carcinoma data

Data from the gene expression measurements of breast carcinoma are obtained from 49 samples with 456 genes. Sorlie *et al.* (2001) originally identified 6 clusters corresponding

to basal-like, ERBB2+, normal breast-like, luminal subtype A, B and C. But because they combined luminal subtypes B and C as one group (luminal B+C), we now have five classes for analysis. For survival time, 25 samples are censored. The data were centered and standardized for each gene across the array. As in the lymphoma data, we fit modified PLS-PH model and estimate the survival probability for each subgroup.




(a) MSEP vs. number of components (b) explained variance vs. number of components

**Figure 3.4** Summary graph for the breast carcinoma data

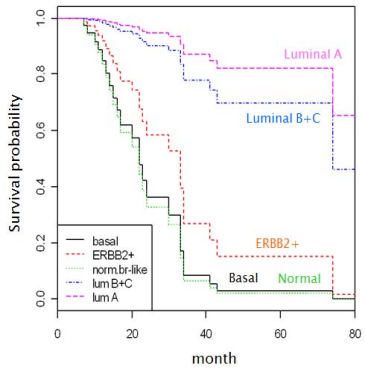



**Figure 3.5** Estimated survival curves for breast carcinoma data

The MSEP of the first stage is distinctly higher than those of the other stages (Figure 3.4(a)). It means that the iteration procedure for imputation and gene selection should be effective. The MSEP is lowest at the number of component of one, although the explained response variance is not enough. Thus we decided to use two components (Figure 3.4(b)). In this case, the $8^{th}$ stage has the lowest MSEP, enabling us to select 43 genes, conclusively. Figure 3.5 shows the estimated survival curve from the mean profiles of the five groups. We can see that the predicted survival probabilities for luminal A and luminal B+C groups are distinctively higher than those for other groups. The basal-like and normal breast-like cells have similar survival probabilities. Sorlie *et al.* (2001) provided a dendrogram of the hierarchical clustering for gene expression without the information of survival time. In their paper,

all the luminal cells (A, B, C) are clustered in one group whereas basal-like cells, normal breast-like cells and ERBB2+ cells are clustered as another group, which are consistent to our result.

## 4. Discussion

In this study, we proposed the modified version of the partial least squares regression to achieve the dimensional reduction for highly correlated gene expression data with a survival time.

There are numerous suggested methods for variable selection. The methods can be categorized into three main categories: filter methods, wrapper methods, and embedded methods (Saeys *et al*., 2007). The filter methods initially fit the PLSR; then they identify a subset of important variables using some measure of relevancy obtained from the output. Loading weights or variable importance in projection (VIP) can be considered as the filter measure. As wrapper methods, many methods including genetic algorithm, uninformative variable elimination (UVE) and backward variable elimination are developed. Wrapper methods mostly use the filter methods in an iterative way and are based on some supervised learning approach. In this case, the model refitting is wrapped within the variable selection algorithm (Mehmood *et al.*, 2012). On the other hand, embedded methods combine variable selection and modeling. In this study, we choose the backward elimination procedure and MSEP criteria for gene selection. One could choose another criterion instead. For more details about the the merits and demerits of the various methods, see Mehmood *et al.* (2012).

In traditional PLSR, a response variable is complete and the variable selection procedure is not considered within the process. PLS-PH method suggested by Nguyen and Rocke (2002b) did not distinguish the censoring time and the survival time. With the proposed modified PLS-PH method, we impute the censored observations and add variable selection procedure inside of the PLSR process. For visualization, y-loading and observation plot as well as x-loading plot are explored, which reveals the relationship between the gene expressions and the survival time. The methods work well and seem to have better properties than other methods making it suitable for the exploratory data analysis.

## References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Broldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Bøvelstad, H. M.,Nygård, S., Størvold, H. L., Aldrin, M., Borgan, O, Frigessi, A. and Lingjærde, O. C. (2007). Predicting survival from microarray data - A comparative study. *Bioinformatics*, **23**, 2080-2087.

Dai, J. J., Lieu, L. and Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, **5**, article 6.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21**, 1104-1111.

Helland, I. (1988). On the structure of partial least squares regression. *Communications in Statistics-Simulation and Computation*, **17**, 581-607.

Kim, J. D. (2003). Unified non-iterative algorithm for principal component regression, partial least squares and ordinary least squares. *Journal of the Korean Data & Information Science Society*, **14**, 355-366.

Mehmood, T., Liland, K., Snipen, L. and SæbøA, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **118**, 62-69.

Nguyen, D. V. and Rocke, D. M. (2002a). Tumor classification by partial least squares using gene expression data. *Bioinformatics*, **18**, 39-50.

Nguyen, D. V. and Rocke, D. M. (2002b). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625-1632.

Nguyen, T. S. and Rojo, J. (2009). Dimension reduction of microarray gene expression data: The accelerated failure time model. *Journal of Bioinformatics and Computational Biology*, **7**, 939-954.

Park, P. J., Tian, L. and Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120-127.

Saeys, Y., Inza, I. and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., *et al*. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**, 10869-10874.