Predictive model building for microarray data using generalized partial least squares model

Baolin Wu *
Division of Biostatistics
School of Public Health
University of Minnesota
A460 Mayo Building, MMC 303
Minneapolis, MN, 55455, USA

October, 2005; Revised August, 2006.

Abstract

Microarray technology enables simultaneously monitoring the expression of hundreds of thousands of genes in an entire genome. This results in the microarray data with the number of genes $p$ far exceeding the number of samples $n$. Traditional statistical methods do not work well when $n \ll p$. Dimension reduction methods are often required before applying standard statistical methods, popular among them are the principal component analysis (PCA) and partial least squares (PLS) etc. PCA is the de facto multivariate dimension reduction technique, and has been successfully used in analyzing the microarray data. PLS has been motivated and interpreted as a numerical algorithm for modeling continuous response variables like quantities of chemical products of reactions. For analyzing the microarray data, PLS has been commonly used to form linear predictor combinations, which are then used as inputs to traditional statistical methods for outcome prediction. In this paper we propose a novel maximum likelihood estimation based framework specially designed for data with $n \ll p$. Through the likelihood based approach, we propose the generalized PLS (gPLS) for modeling both continuous and discrete outcomes. For continuous outcome modeling, gPLS is shown closely related to the PLS. For discrete outcome, the proposed gPLS provides a more appropriate and natural modeling approach than the PLS. We further propose a penalized estimation procedure to achieve automatic variable selection, which is very useful for modeling high-dimensional data with $n \ll p$. Applications to public microarray data and simulation studies are conducted to illustrate the proposed methods.

*Email: baolin@biostat.umn.edu, Phone: (612)624-0647, Fax: (612)626-0660
1 Introduction

Microarray enables simultaneously monitoring the gene expression values in an entire genome. The resulting microarray data are typically characterized by many measured variables (genes) for very few samples (arrays). The number of genes $p$ on a single array is often in the magnitude of hundreds of thousands, while the number of samples $n$ is often around 10-100. This $n \ll p$ poses serious statistical challenges for microarray analysis. There are lots of statistical problems associated with microarray data analysis. The readers are referred to the books by Parmigiani et al. (2003) and Speed (2003) for comprehensive reviews. One of the central questions has been the molecular classification of samples using gene expression values (Golub et al., 1999; Alon et al., 1999; Ramaswamy et al., 2001; Singh et al., 2002, e.g.).

For microarray data, $n \ll p$ often calls for dimension reduction before traditional statistical methods are applied. The principal component analysis (PCA; Jolliffe, 2002) is the de facto multivariate dimension reduction technique, and has been successfully used to analyze the microarray data (Alter et al., 2000, 2003, e.g.). While PCA is an unsupervised dimension reduction approach, the partial least squares (PLS; Wold, 1976; Garthwaite, 1994, e.g.) regression can be seen as a supervised dimension reduction method. For analyzing the microarray data, PLS has been commonly used to form linear predictor combinations, which are then used as inputs to traditional statistical methods for outcome prediction. For example, Nguyen and Rocke (2002a,b) used PLS to create several predictors based on linear combination of genes, which are then applied to some discriminant methods, e.g., linear discriminant analysis, for cancer classification. And Nguyen and Rocke (2002c); Park et al. (2002) applied PLS to reduce dimension before linking the gene expression with patient survival time. In Huang et al. (2004) PLS is used to first obtain a linear regression model based on all the genes. Soft-thresholding is then applied to shrink the individual gene coefficients, which are then used to construct predictors to model the patient left ventricular assist device support time.
Another class of classification methods are derived from traditional discriminant methods with some simplicity assumptions. The diagonal linear/quadratic discriminant analysis (DLDA, DQDA; Dudoit et al., 2002), where correlations among genes are ignored, perform competitively compared to recently developed much more complicated machine learning methods, e.g., random forest (Breiman, 2001), bagging or boosting classification trees (see Friedman et al., 2000, e.g.) etc. The nearest shrunken centroid classifier (Tibshirani et al., 2002) incorporates shrinkage into the DLDA and proves to be very successful in empirical studies.

The PLS regression has been originally developed as a numerical algorithm for modeling continuous response variables like quantities of chemical products of reactions (Frank and Friedman, 1993). In this paper we study PLS from a statistical model perspective and develop the more appropriate generalized partial least squares (gPLS) procedure for modeling both discrete and continuous outcomes. We further propose a penalized estimation procedure for gPLS to achieve automatic gene selection for analyzing the high-dimensional microarray data.

The rest of the paper is organized as following: we first discuss the PCA and PLS dimension reduction methods from both the numerical algorithm and latent variable model perspectives. We will argue for the better performance of PLS over PCA heuristically. We then propose a maximum likelihood estimation based framework specially designed for data with $n \ll p$. Through the likelihood based approach, we propose the generalized PLS (gPLS) model for analyzing both continuous and discrete outcomes. For continuous outcome modeling, we show that gPLS is closely related to the PLS. For discrete outcome, the proposed gPLS provides a more appropriate and natural modeling approach than the PLS. A penalized estimation procedure incorporating automatic gene selection is then proposed specially for data with $n \ll p$. Application to public microarray data and simulation studies are conducted to compare several PCA and PLS based methods and illustrate the competitive performance.
of the proposed methods.

2 PCA and PLS

Throughout the discussion we use $Y$ to denote the random variable for the outcome of our interest, and $\{X_1, \cdots, X_p\}$ for the intensities of $p$ genes.

Let $x_{ij}$ denote the observed expression values for sample $i = 1, \cdots, n$ and gene $j = 1, \cdots, p$. Define the following observation matrix

$$X = (X_1, \cdots, X_p), \text{ where } X_j = (x_{1j}, \cdots, x_{nj})^T, \quad j = 1, \cdots, p,$$

where the subscript $T$ means the matrix transpose. Let $X_i^T = (x_{i1}, \cdots, x_{ip})$ denote the gene expression values for sample $i$, and denote the observed sample response vector as

$$Y = (y_1, \cdots, y_n)^T,$$

which could be categorical values, e.g., normal versus cancer tissues, or continuous measurements, e.g., blood pressure and age etc. Our goal is to model the response $Y$ as some function of the gene expression values $X$. We assume necessary pre-processing has been done (see Yang et al., 2002; Bolstad et al., 2003; Irizarry et al., 2003, e.g.).

Denote the column (variable) mean subtracted observation matrix as $\tilde{X}$. PCA seeks the linear combination of predictors to retain the maximum variation

$$\arg\max_{W_1} \| \tilde{X} W_1 \|^2, \text{ where } W_1 = (w_{11}, \cdots, w_{1p})^T, \quad \| W_1 \| = 1. \quad (1)$$

Here $\| \cdot \|$ is the $L_2$ norm, e.g., $\| W_1 \| = \sqrt{\sum_{i=1}^{n} w_{1i}^2}$. We can go on to find the subsequent orthogonal components. In summary the first $K$ components solve

$$\arg\max_W \text{tr} (W^T \tilde{X}^T \tilde{X} W), \quad \text{where } W^T W = I_K.$$

Here $I_K$ is an $K \times K$ identity matrix, and $\text{tr}(\cdot)$ calculates the trace of a matrix. $W$ can be derived from the matrix singular value decomposition (Healy, 2000): $\tilde{X} = UDV^T$, where
$U$ and $V$ are both orthogonal matrices, and absolute diagonal values of the diagonal matrix $D$ are ordered from large to small. We then have $W = V_K$, the first $K$ columns of $V$; and $\tilde{X}W = U_KD_K$, the scaled first $K$ columns of $U$, which are often called the principal components (PC), or meta/eigen genes in the microarray data analysis (Alter et al., 2000, 2003, e.g.).

Interestingly, the first PC can be equivalently derived from the following latent variable model

$$X_j = \alpha_j + \beta_jZ + \epsilon_j, \quad \text{where } Z \in \mathbb{R}^n, \quad \epsilon_j \in \mathbb{R}^n, \quad \mathbb{E}(\epsilon_j) = 0, \quad j = 1, \cdots, p. \quad (2)$$

For identifiability we require $\sum_{j=1}^p \beta_j^2 = 1$. When least squares is used for estimation:

$$\min_{\beta} \sum_{j=1}^p \|X_j - \alpha_j - \beta_jZ\|^2,$$

it is easily checked that $\beta_j = \pm w_{1j}$, and $Z = \sum_{j=1}^p \tilde{X}_j\beta_j$.

So the first PC can be interpreted as a latent variable. Using residuals from the regression model recursively as response, we can obtain subsequent PCs.

Notice that when deriving the latent variable $Z$, we did not use any information from $Y$. It is highly possible that $Z$ is not related to the outcome $Y$ of our interest. One remedy is to focus only on $Y$ related variables by pre-selecting those significantly associated with $Y$, which is the idea of the supervised PCA (Bair et al., 2006). Another more intuitive remedy is to directly replace $Z$ by $Y$ in model (2) to obtain the following inverse regression approach

$$X_j = \alpha_j + \beta_jY + \epsilon_j, \quad j = 1, \cdots, p, \quad (3)$$

which explicitly targets $Y$ as the latent variable, and may extract from $X$ more relevant information associated with $Y$. In the following we will formalize this inverse regression approach using a maximum likelihood estimation based framework to develop the generalized PLS (gPLS) model for analyzing both continuous and discrete outcomes. For continuous outcome, we will show that the proposed gPLS is closely related to the PLS. For discrete outcome, the proposed gPLS provides a more appropriate and natural modeling approach than the PLS. It is interesting to notice that the general idea of inverse regression also
underlies the slice inverse regression (SIR; see Li, 1991; Cook and Weisberg, 1991, e.g.) related approaches, which have been developed traditionally for data with \( n > p \) and often require \( n \gg p \) for asymptotic statistical inference.

The original PLS algorithm (see Wold et al., 1984; Frank and Friedman, 1993, e.g.) seeks linear predictor combinations as follows

(i) Standardize the response and predictors

\[
Y_0 = (Y - \bar{y})/s_y, \quad X_{j,0} = (X_j - \bar{x}_j)/s_j, \quad j = 1, \ldots, p,
\]

where

\[
\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad s_y = \frac{\|Y - \bar{y}\|}{\sqrt{n}}, \quad \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad s_j = \frac{\|X_j - \bar{x}_j\|}{\sqrt{n}}.
\]

(ii) For \( k = 1, \ldots, K \)

\[
Z_k = \sum_{j=1}^{p} \beta_{j,k-1} X_{j,k-1}, \quad \text{where} \quad \beta_{j,k-1} = \frac{X_{T,j,k-1} Y_{k-1}}{\|X_{j,k-1}\|^2},
\]

\[
X_{j,k} = X_{j,k-1} - \frac{X_{T,j,k-1} Z_k}{\|Z_k\|^2} Z_k, \quad Y_k = Y_{k-1} - \frac{Y_{T,k-1} Z_k}{\|Z_k\|^2} Z_k, \quad j = 1, \ldots, p.
\]

It is easily checked that \( Z_k \) are orthogonal to each other owing to the covariate adjustment in the second step. We define the constructed \( Z_k \) as the \textbf{PLS component (PLSC)} analogous to the PC. The final predictive model is derived by regressing \( Y \) on all the PLSCs. It is interesting to notice that the (normalized) linear coefficients of the first PLSC also solve (Frank and Friedman, 1993)

\[
\max_{\|W_1\| = 1} \text{corr}^2(Y, \tilde{X} W_1) \|\tilde{X} W_1\|^2,
\]

where \( \text{corr}(\cdot) \) calculates the sample correlation. It is a compromise between the PCA and regression.
From the PLS algorithm, it is easily checked that the first PLSC can be written as

$$Z_1 = \frac{1}{\sqrt{n}} \sum_{j=1}^{p} \frac{Y^T (X_j - \bar{x}_j)}{\|Y - \bar{y}\|} s_j^2 (X_j - \bar{x}_j).$$

(4)

In summary, PCA tries to recover the underlying latent variable (linear combination of the original variables) with minimum information loss, which may be unrelated to the outcome of our interest. In this sense PCA is an unsupervised dimension reduction method. Since our goal is to predict $Y$ from $X$, intuitively it is more appealing to construct the response-associated latent variable. PLS uses $Y$ information as a guidance to define the PLSC. It is a supervised dimension reduction method and might be more useful for outcome prediction.

Figure 1 compares the first two components for PCA, PLS and rgPLS (regularized generalized PLS, to be discussed later) for the prostate cancer data reported at Singh et al. (2002). We can see that the first two PCs, which preserve the maximum observation variations, do not necessarily have the best prediction power for the response (see Appendix for more detailed analysis results). The first two components for both the PLS and rgPLS have a good prediction power for the response (we are going to analyze this data in detail in the Application section).

Next we formalize the inverse regression approach (3) through an independence regression model and propose the generalized PLS approach for modeling both continuous and discrete outcomes.

3 Generalized partial least squares (gPLS) model

Consider the following independence regression model

$$\Pr(Y = y) = f(y, \theta), \quad \Pr(X_1, \cdots, X_p|Y = y) = \prod_{j=1}^{p} f_j\{X_j, \theta_j(y)\},$$

(5)
Figure 1: PCA, PLS and rgPLS comparison: the first two components are plotted. The data contains the gene expression values for 50 normal (labeled as class 1) and 52 prostate cancer samples (labeled as class 2) reported at Singh et.al (2002). The two classes are mixed together for PCA. They are reasonably separated for the PLS and rgPLS.

where parameter $\theta_j(y)$ is some function of the response. The conditional independence assumption is a regularity constraint and can produce a sparse structure. The intuition is that under $n \ll p$ situation, we do not have enough samples to fit complicated models. Although tending to produce biased estimates, simple model can have small variance owing to its simple structure. As a result the bias-variance tradeoff can often make simple model outperform the complicated model. Recently Bickel and Levina (2004) obtained some theoretical results supporting the benefits of assuming independence in classification when there are many more variables than observations.

From model (5) we can obtain the joint and conditional probabilities for $Y$

$$\Pr(Y = y, X_1, \cdots, X_p) = f(y, \theta) \prod_{j=1}^{p} f_j\{X_j, \theta_j(y)\},$$

$$\Pr(Y = y | X_1, \cdots, X_p) = \frac{f(y, \theta) \prod_{j=1}^{p} f_j\{X_j, \theta_j(y)\}}{\int_y f(y, \theta) \prod_{j=1}^{p} f_j\{X_j, \theta_j(y)\} dy}.$$

In the following, under a normal error regression model, we first derive the generalized
PLS (gPLS) model for continuous responses, which is closely related to the PLS regression procedure. The gPLS is then generalized naturally to model discrete outcomes.

3.1 Continuous response: linear regression with generalized PLS (gPLS)

Assume a normal distribution $N(\theta, \sigma^2)$ for $Y$ and the following normal error regression model

$$X_j = \alpha_j + \beta_j Y + \epsilon_j, \quad \text{where} \quad \epsilon_j \sim N(0, \sigma^2_j), \quad j = 1, \ldots, p. \quad (6)$$

We can show that conditional on the $(X_1, \ldots, X_p)$, $Y$ has a normal distribution $N(\tilde{\mu}, \tilde{\sigma}^2)$ with (see Appendix for details)

$$\tilde{\mu} = E(Y|X_1, \ldots, X_p) = \frac{\sigma^{-2}\theta + \sum_{j=1}^{p}\sigma_j^{-2}\beta_j(X_j - \alpha_j)}{\sigma^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\beta_j^2},$$

$$\tilde{\sigma}^2 = Var(Y|X_1, \ldots, X_p) = \frac{1}{\sigma^{-2} + \sum_{j=1}^{p}\sigma_j^{-2}\beta_j^2}. \quad (7)$$

The log likelihood for the observed data $(X, Y)$ is

$$-n(p + 1) \log \sqrt{2\pi} - \left\{ n \log \sigma + \frac{\sum_{i=1}^{n}(y_i - \theta)^2}{2\sigma^2} \right\} - \sum_{j=1}^{p} \left\{ n \log \sigma_j + \frac{\sum_{i=1}^{n}(x_{ij} - \alpha_j - \beta_j y_i)^2}{2\sigma_j^2} \right\}. \quad (8)$$

It is easily checked that the maximum likelihood estimates are

$$\hat{\beta}_j = \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2}, \quad \hat{\sigma}^2 = \frac{\|Y - \bar{y}\|^2}{n}, \quad \hat{\sigma}_j^2 = \frac{\|X_j - \hat{\alpha}_j - \hat{\beta}_j Y\|^2}{n},$$

$$\hat{\theta} = \bar{y}, \quad \hat{\alpha}_j = \bar{x}_j - \hat{\beta}_j \bar{y}, \quad \text{where} \quad \bar{x}_j = \frac{\sum_{i=1}^{n}x_{ij}}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n}y_i}{n}. \quad (9)$$

Plugging the maximum likelihood estimates into (7), we have the following predictive model for a future sample $X_h = (x_{h1}, \ldots, x_{hp})$ (see Appendix for details)

$$\hat{E}(Y_h|X_h) = \eta_0(X, Y) + \eta_1(X, Y)G_1(X_h; X, Y), \quad (8)$$

where

$$G_1(X_h; X, Y) = \sum_{j=1}^{p} \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|\hat{\sigma}_j^2}(x_{hj} - \bar{x}_j), \quad (9)$$
and
\[ \eta_0(X, Y) = \bar{y}, \quad \eta_1(X, Y) = \frac{n^{-1}\| Y - \bar{y} \|}{1 + \| Y - \bar{y} \|^2 \sum_{j=1}^n \hat{\beta}_j^2 \| X_j - \bar{x}_j \|} \cdot \frac{\| X_j - \bar{x}_j - \hat{\beta}_j (Y - \bar{y}) \|^{-2}}{\| X_j - \bar{x}_j \|}. \]

For the observed data, we have
\[
G_1(X; X, Y) = \sum_{j=1}^p \frac{Y^T (X_j - \bar{x}_j)}{\| Y - \bar{y} \|^2 \sigma_j^2} (X_j - \bar{x}_j). \tag{10}
\]

Compared to \( Z_1 \) (4), the main difference is the variance \( \sigma_j^2 \) estimation: for \( Z_1 \) the variance is estimated without using the \( Y \) information, while for \( G_1(X; X, Y) \) the variance can be estimated, e.g., by regressing \( X_j \) on \( Y \). The results should be similar if \( X_j \) and \( Y \) are unrelated, otherwise the regression approach might lead to better estimates. We define \( G_1(X; X, Y) \) as the **generalized PLS (gPLS) component (gPLSC)**.

Under the independence normal error regression model assumption (6), predictive model (8) is the optimal Bayesian rule minimizing the squared error loss. Instead of completely relying on the normal model assumptions for inference, we propose to estimate \((\eta_0, \eta_1)\) from the data using, e.g., the following least squares regression
\[
(\hat{\eta}_0, \hat{\eta}_1) = \arg \min_{\eta_0, \eta_1} \| Y - \eta_0 - \eta_1 G_1(X; X, Y) \|^2,
\]
which might make the procedure more robust. This approach has also been used in the PLS regression.

In summary, we propose the gPLS (for continuous outcome modeling) from a maximum likelihood estimation based framework assuming the independence normal error regression model. We also propose an additional estimation step to make the procedure more robust. For discrete outcome modeling, the approach can be readily generalized as follows.
3.2 Discrete response: logistic regression with gPLS

For a binary response $Y$, assume $\Pr(Y = 1) = \theta, \Pr(Y = 0) = 1 - \theta$, and the normal error regression model (6). It is easily checked that

$$\text{Logit}\{\Pr(Y = 1|X_1, \cdots, X_p)\} = \text{Logit}(\theta) - \sum_{j=1}^{p} \frac{\beta_j^2}{2\sigma_j^2} + \sum_{j=1}^{p} \frac{\beta_j}{\sigma_j^2} (X_j - \alpha_j),$$

(11)

where $\text{Logit}(\cdot)$ is the logit function: $\text{Logit}(\theta) = \log \frac{\theta}{1 - \theta}$. Given the observed data $(X, Y)$, it is easily checked that the maximum likelihood estimates are

$$\hat{\theta} = \bar{y}, \quad \hat{\alpha}_j = \bar{x}_j - \hat{\beta}_j \bar{y}, \quad \hat{\beta}_j = \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2}, \quad \hat{\sigma}_j^2 = \frac{\|X_j - \bar{x}_j - \hat{\beta}_j(Y - \bar{y})\|^2}{n}.$$

Similarly plugging in the maximum likelihood estimates, we obtain the following predictive model for a future sample $X_h = (x_{h1}, \cdots, x_{hp})$ (see Appendix for details)

$$\text{Logit}\{\Pr(Y_h = 1|X_h)\} = \gamma_0(X, Y) + \gamma_1(X, Y) G_1(X_h; X, Y),$$

(12)

where

$$G_1(X_h; X, Y) = \sum_{j=1}^{p} \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2} (x_{hj} - \bar{x}_j),$$

$$\gamma_0(X, Y) = \text{Logit}(\bar{y}) + \left(\bar{y} - \frac{1}{2}\right) \sum_{j=1}^{p} \frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2}, \quad \gamma_1(X, Y) = \frac{1}{\|Y - \bar{y}\|}.$$

Again for the observed data, we have the following gPLSC

$$G_1(X; X, Y) = \sum_{j=1}^{p} \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2} (X_j - \bar{x}_j).$$

Similarly instead of relying the normal model assumptions for inference, we propose to make the predictive model more robust by estimating $\gamma_0$ and $\gamma_1$ based on the data by, e.g., maximizing the conditional likelihood for $Y$

$$(\hat{\gamma}_0, \hat{\gamma}_1) = \arg \max_{\gamma_0, \gamma_1} \sum_{i=1}^{n} \left\{ y_i \log \frac{1}{1 + \exp(-\tau_i)} + (1 - y_i) \log \frac{1}{1 + \exp(\tau_i)} \right\},$$
where \( \tau_i = \gamma_0 + \gamma_1 G_1(X_i^T; X, Y) \).

Here we explicitly approach the high-dimensional data modeling from a supervised dimension reduction perspective. The constructed gPLSC (latent variable), \( G_1(X_h; X, Y) \), depends on both \( X \) and \( Y \). We emphasize that the strict model assumptions (6) are used mainly as a guidance to construct meaningful latent variables. We do not fully rely on them for inference. Instead, we treat the reduced dimension (latent variable) as a new variable and apply other analysis methods, e.g., the least squares or maximum likelihood estimation, to do inference.

Another advantage of putting dimension reduction into a model-based framework is that we can easily incorporate regularization into the estimation procedure, which is especially useful for modeling high-dimensional data with \( n \ll p \). In the following we discuss the use of \( L_1 \) penalized estimation to construct the gPLSC. The motivation is to do automatic variable selection and parameter estimation simultaneously.

### 3.3 Regularized generalized PLS (rgPLS)

Generally only a subset of all \( p \) predictors are useful for predicting the response. We can select some important variables using, e.g., the F-statistics, before computing the gPLSC. Notice that we only have \( n \) samples to estimate the regression parameters for all \( p \) predictors. Under the \( n \ll p \) situation, regression parameter estimates could be highly unstable. Shrinkage can often improve parameter estimation and model performance. For example, the nearest shrunken centroid classifier (NSC, also known as PAM, Tibshirani et al., 2002) applied soft-shrinkage to improve the performance of the nearest centroid classifier. It has been shown that \( L_1 \) penalty has the shrinkage and automatic variable selection properties (see Tibshirani, 1996; Efron et al., 2004, e.g.). Here we propose to adopt the following \( L_1 \) penalized estimation for predictor \( j \)

\[
\arg\max_{\beta_j, \sigma_j} \left( -n \log \sigma_j - \frac{1}{2} \sum_{i=1}^{n}(x_{ij} - \alpha_j - \beta_j y_i)^2 \right) - n\lambda_j |\beta_j|, 
\]  

(13)
where $\lambda_j > 0$ is a regularity parameter. Denote the parameter estimates as $\hat{\beta}_j(\lambda_j; X_j, Y)$ and $\hat{\sigma}_j(\lambda_j; X_j, Y)$. We define the regularized gPLS (rgPLS) for a new sample $X_h = (x_{h1}, \cdots, x_{hp})$ as

$$S_1(X_h; \lambda, X, Y) = \sum_{j=1}^{p} \frac{\hat{\beta}_j(\lambda_j; X_j, Y)}{\hat{\sigma}_j^2(\lambda_j; X_j, Y)} (x_{hj} - \bar{x}_j), \quad \text{where } \lambda = (\lambda_1, \cdots, \lambda_p).$$

And for the observed data we call

$$S_1(X; \lambda, X, Y) = \sum_{j=1}^{p} \frac{\hat{\beta}_j(\lambda_j; X_j, Y)}{\hat{\sigma}_j^2(\lambda_j; X_j, Y)} (X_j - \bar{x}_j)$$

the rgPLS component (rgPLSC).

### 3.4 Iterative algorithm for rgPLSC

For the penalized estimation (13), it is easy to see that $\hat{\alpha}_j = \bar{x}_j - \hat{\beta}_j \bar{y}$. We can use the following iterative algorithm to estimate $\beta_j$ and $\sigma_j$

(i) Initialize $\beta_j$ and $\sigma_j$ as

$$\beta_j^{(0)} = \frac{Y^T(X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2}, \quad \sigma_j^{(0)} = \|X_j - \bar{x}_j - \beta_j^{(0)}(Y - \bar{y})\|/\sqrt{n}.$$

(ii) Given the current estimate $\sigma_j^{(b)}$, model (13) reduces to a LASSO regression for $\beta_j$ (Tibshirani, 1996), hence

$$\hat{\beta}_j^{(b+1)} = \text{sign}(\beta_j^{(0)}) \left[ |\beta_j^{(0)}| - \frac{n \lambda_j}{\|Y - \bar{y}\|^2} \left\{ \sigma_j^{(b)} \right\}^2 \right]_+, \quad \hat{\sigma}_j^{(b+1)} = \frac{\|X_j - \bar{x}_j - \hat{\beta}_j^{(b+1)}(Y - \bar{y})\|}{\sqrt{n}},$$

where the subscript plus means the positive part: $z_+ = z$ if $z \geq 0$ and zero otherwise. \text{sign}(\cdot) is the sign function: \text{sign}(z) = 1 if $z > 0$, \text{sign}(z) = -1 for $z < 0$, and \text{sign}(z) = 0 when $z = 0$.

The following one-step estimation

$$\hat{\beta}_j^{(1)} = \text{sign}(\beta_j^{(0)}) \left\{ |\beta_j^{(0)}| - \lambda_j \|X_j - \bar{x}_j - \beta_j^{(0)}(Y - \bar{y})\|^2 \|Y - \bar{y}\|^2 \right\}_+$$

13
corresponds to first obtaining the maximum likelihood estimates of $\beta_j$ and $\sigma_j$, then applying the soft-thresholding to the regression parameters, which is the commonly used form of shrinkage.

For large $\lambda_j$, $\hat{\beta}_j$ will be shrunken down to zero, and hence does not contribute to the rgPLSC. So for $L_1$ penalized estimation, variable selection is an automatic product of estimation. More importantly, the regression coefficient estimates change continuously with the regularization parameter. Commonly used variable selection methods, e.g., the $t/F$-statistics, are hard-thresholding (see Fan and Li, 2001, e.g.): the parameter estimates are not continuous with respect to the cutoff values used for variable selection.

In the following we apply the proposed gPLS for cancer classification using some public microarray data. Specifically we will use the rgPLS, since it has built-in automatic gene selection, which is very important for analyzing the high-dimensional microarray data with $n \ll p$. Through the application, we will illustrate the relative performance of PCA and PLS based methods.

4 Application to cancer microarray data

We compare the following nine methods: PCA with one and two components (PCA.1, PCA.2); supervised PCA (SPCA); PLS with one and two components (PLS.1, PLS.2); PLS with one and two components using top 200 genes (PLS.s1, PLS.s2) selected by the multi-class comparison F-statistics (we tried different number of selected genes, the performance of the methods did not change too much); PAM; and rgPLS. The first seven methods applied linear discriminant analysis to the constructed PC and PLSC (see Nguyen and Rocke, 2002a, b, e.g.).

Classification errors are estimated by splitting the data into training and testing sets. The training set consists of two-thirds of the samples, used for fitting the classifier, where 5-fold cross validations are used to tune the classifier. The optimized classifier is then
used to predict the testing set. We repeat the random split 50 times and the error rates are averaged. To avoid selection bias, important genes are selected for each training set separately (Ambroise and McLachlan, 2002).

We have analyzed the following six public microarray data:

1) The leukemia gene expression data reported at Golub et al. (1999) contained the mRNA levels of 72 patients, among them 47 patients had Acute Lymphoblastic Leukemia (ALL) and 25 had Acute Myeloid Leukemia (AML). The data is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.


3) The prostate cancer microarray data (Singh et al., 2002) contained the expression profiles of 50 normal and 52 tumor prostate tissue samples, and were downloaded from http://www-genome.wi.mit.edu/MPR/prostate.

4) The breast cancer microarray data (West et al., 2001) measured the expression levels of 49 breast tumor samples. Each sample had a binary outcome describing the status of lymph node involvement in breast cancer. The data were downloaded from http://data.cgt.duke.edu/west.php.

5) The small round blue cell tumor (SRBCT) microarray data (Khan et al., 2001) contained the expression levels of 63 samples from four tumor types: burkitt lymphoma (BL, 8 samples), ewing sarcoma (EWS, 23 samples), neuroblastoma (NB, 12 samples), and rhabdomyosarcoma (RMS, 20 samples). The data were downloaded from http://research.nhgri.nih.gov/microarray/Supplement.
Table 1: Average classification errors (%) over 50 random splits for the five public microarray data. Listed within the parenthesis are the standard errors (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>Leukemia</th>
<th>Colon</th>
<th>Prostate</th>
<th>Breast</th>
<th>SRBCT</th>
<th>GCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA.1</td>
<td>3.48 (2.91)</td>
<td>33.80 (7.93)</td>
<td>42.06 (7.39)</td>
<td>61.13 (10.05)</td>
<td>61.68 (8.03)</td>
<td>79.07 (2.07)</td>
</tr>
<tr>
<td>PCA.2</td>
<td>2.52 (2.79)</td>
<td>25.00 (7.63)</td>
<td>40.97 (8.58)</td>
<td>50.00 (12.56)</td>
<td>57.16 (9.98)</td>
<td>68.93 (3.33)</td>
</tr>
<tr>
<td>SPCA</td>
<td>3.39 (3.08)</td>
<td>12.50 (5.27)</td>
<td>11.21 (5.94)</td>
<td>34.00 (13.13)</td>
<td>18.00 (7.96)</td>
<td>73.93 (4.53)</td>
</tr>
<tr>
<td>PLS.1</td>
<td>5.74 (6.17)</td>
<td>42.06 (7.39)</td>
<td>37.70 (10.37)</td>
<td>44.50 (10.07)</td>
<td>39.58 (10.96)</td>
<td>67.89 (4.56)</td>
</tr>
<tr>
<td>PLS.2</td>
<td>2.17 (2.37)</td>
<td>40.97 (8.58)</td>
<td>14.73 (4.82)</td>
<td>39.50 (10.45)</td>
<td>22.74 (8.47)</td>
<td>61.25 (3.72)</td>
</tr>
<tr>
<td>PLS.s1</td>
<td>2.17 (2.37)</td>
<td>14.00 (5.71)</td>
<td>15.15 (7.40)</td>
<td>41.25 (12.24)</td>
<td>20.53 (7.97)</td>
<td>62.04 (4.16)</td>
</tr>
<tr>
<td>PLS.s2</td>
<td>3.30 (3.23)</td>
<td>14.30 (5.80)</td>
<td>8.97 (4.15)</td>
<td>39.13 (12.10)</td>
<td>5.79 (5.02)</td>
<td>46.50 (4.56)</td>
</tr>
<tr>
<td>PAM</td>
<td>3.22 (2.89)</td>
<td>14.20 (6.17)</td>
<td>8.97 (4.11)</td>
<td>29.50 (16.12)</td>
<td>3.16 (3.01)</td>
<td>25.25 (3.70)</td>
</tr>
<tr>
<td>rgPLS</td>
<td>3.22 (2.89)</td>
<td>12.40 (5.55)</td>
<td>7.21 (3.96)</td>
<td>22.75 (11.69)</td>
<td>1.05 (3.01)</td>
<td>18.43 (4.23)</td>
</tr>
</tbody>
</table>

6) The GCM microarray data (Ramaswamy et al., 2001) contained the gene expression profiles of 198 samples from 14 major cancer types. There data were downloaded from http://www-genome.wi.mit.edu/MPR/GCM.html. Similar to (Ramaswamy et al., 2001, Figure 5), 8 metastatic samples were excluded and the analysis was based on 190 primary tumor samples.

Table 1 summarizes the results. In Figure 2 we have used the boxplots to summarize the 50 classification errors for the six public microarray data (see Appendix for details on tuning the rgPLS for classification).

We can see the general trend that methods based on PLS are better than those based on PCA. SPCA improves upon the plain PCA. For dimension reduction purpose, generally PCA and PLS based procedures could improve the accuracy by going from one to two dimensions. Using significant genes only can improve the classification accuracy. Overall we can see that the rgPLS has competitive performance compared to the other methods.

Next we conduct simulation studies to evaluate the proposed methods for modeling continuous outcomes.
Figure 2: Boxplot of the classification errors (%) over 50 random splits for the six public microarray data.
5 Simulation study

For continuous response we compare the proposed rgPLS to the supervised principal component analysis, gene shaving (Hastie et al., 2000), partial least squares regression and PCA regression (see, e.g., Bair et al., 2006 for very good introduction to these methods).

The two simulations conducted are based on the settings of Bair et al. (2004). Each simulated data consists of 5000 genes and 100 samples. Denote $x_{ij}$ as the expression values for gene $i$ and sample $j$. In the first study the data is generated as follows

$$x_{ij} = \begin{cases} 
3 + \epsilon_{ij} & \text{if } i \leq 50, j \leq 50 \\
4 + \epsilon_{ij} & \text{if } i \leq 50, j > 50 \\
3.5 + \epsilon_{ij} & \text{if } i > 50
\end{cases}$$

where the $\epsilon_{ij}$ are independent standard normal random variables. The response $Y$ is simulated as

$$y_j = \frac{\sum_{i=1}^{50} x_{ij}}{25} + \epsilon_j, \quad j = 1, \cdots, 100,$$

where $\epsilon_j$’s are independent normal random variables with mean 0 and variance 1.5.

The second simulation study was a “harder” one: the gene expressions are simulated as follows

$$x_{ij} = \begin{cases} 
3 + \epsilon_{ij} & \text{if } i \leq 50, j \leq 50 \\
4 + \epsilon_{ij} & \text{if } i \leq 50, j > 50 \\
3.5 + 1.5 \cdot I(u_{1j} < 0.4) + \epsilon_{ij} & \text{if } 51 \leq i \leq 100 \\
3.5 + 0.5 \cdot I(u_{2j} < 0.7) + \epsilon_{ij} & \text{if } 101 \leq i \leq 200 \\
3.5 - 1.5 \cdot I(u_{3j} < 0.3) + \epsilon_{ij} & \text{if } 201 \leq i \leq 300 \\
3.5 + \epsilon_{ij} & \text{if } i > 300
\end{cases}$$

Here the $u_{ij}$ are uniform random variables on $(0,1)$ and $I(\cdot)$ is the indicator function. The $\epsilon_{ij}$ are independent standard normal random variables. The response $Y$ is simulated as

$$y_j = \frac{\sum_{i=1}^{50} x_{ij}}{25} + \epsilon_j, \quad j = 1, \cdots, 100,$$

where $\epsilon_j$’s are independent normal random variables with mean 0 and variance 2.25. Compared to the first simulation, the second one has more noise in the gene expressions and the response.
Table 2: Results for the two simulation studies: the error rates are averaged over 10 simulations. The classifier is trained on the training set using 10-fold CV, which are then applied to the testing set. The standard errors are listed in parenthesis. The prediction methods are: (1) principal components regression (PCR), (2) principal components regression using only the first principal component (PCR.1), (3) partial least squares (PLS), (4) supervised principal component analysis (SPCA), (5) gene shaving, (6) regularized generalized partial least squares (rgPLS), (7) the oracle procedure. The error rates from Bair et al. (2004) are listed here for comparison. Overall the proposed rgPLS has competitive performance.

<table>
<thead>
<tr>
<th></th>
<th>First simulation (15)</th>
<th>Second simulation (16)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV Error</td>
<td>Test Error</td>
</tr>
<tr>
<td>(1) PCR</td>
<td>290.5 (10.18)</td>
<td>227.0 (7.36)</td>
</tr>
<tr>
<td>(2) PCR.1</td>
<td>315.3 (12.43)</td>
<td>252.1 (8.76)</td>
</tr>
<tr>
<td>(3) PLS</td>
<td>284.9 (10.04)</td>
<td>219.8 (7.43)</td>
</tr>
<tr>
<td>(4) SPCA</td>
<td>242.0 (10.32)</td>
<td>184.6 (7.36)</td>
</tr>
<tr>
<td>(5) Gene shaving</td>
<td>219.6 (8.34)</td>
<td>160.0 (4.34)</td>
</tr>
<tr>
<td>(6) rgPLS</td>
<td>156.8 (3.72)</td>
<td>157.2 (5.95)</td>
</tr>
<tr>
<td>(7) Oracle</td>
<td></td>
<td>153.1 (6.98)</td>
</tr>
</tbody>
</table>

Table 2 summarizes the results for the two simulation studies. The proposed rgPLS has competitive performance compared to the SPCA and the gene shaving. All of them have clear advantage over the regression methods based on PCA and PLS. The table also includes the prediction error for the “oracle” procedure, i.e., using the true model (see Appendix for details). When the data has less noise, the rgPLS and gene shaving can keep up with the oracle procedure. They suffer from the extra noise introduced in the second simulation.

6 Discussion

For microarray data, it is common to observe “large p small n”, and overfitting is very likely to occur. Regularity has proven useful in modeling such high-dimensional data. Through a maximum likelihood estimation based framework, we propose the generalized partial least squares (gPLS) model for analyzing both continuous and discrete outcomes in the high-dimensional data. For continuous outcome modeling, the proposed gPLS is closely related
to the commonly used partial least squares (PLS) regression model. While for modeling dis-
crete outcomes, the gPLS provides a more natural and appropriate approach than the PLS.
The proposed gPLS is developed from a regression model with the independence regularity
constraint. The maximum likelihood estimation based framework can also naturally incor-
porate shrinkage estimation by using the $L_1$ penalty to further regularize the model fitting.
Through simulation studies and applications to public microarray data, we have shown the
competitive performance of the proposed methods.

**Acknowledgements**

This research was partially supported by a University of Minnesota artistry and research
grant and a research grant from the Minnesota Medical Foundation. I would like to thank
the associate editor and two referees for their constructive comments, which have greatly
improved the presentation of the paper.

**References**

Broad patterns of gene expression revealed by clustering analysis of tumor and normal


comparative analysis of genome-scale expression data sets of two different organisms.
*PNAS*, 100, 3351–3356.


21


**APPENDIX**

24
A.1 Conditional mean of continuous response Y for normal error regression model

Assuming \( Y \sim N(\theta, \sigma^2) \) and the normal error regression model (6), we have

\[
\text{Pr}(Y, X_1, \cdots, X_p) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^p} \prod_{j=1}^{p} \sigma_j \exp \left\{ -\frac{(Y - \theta)^2}{2\sigma^2} - \sum_{j=1}^{p} \frac{(X_j - \alpha_j - \beta_j Y)^2}{2\sigma_j^2} \right\}
\]

\[
\propto \exp \left[ -\frac{1}{2} \left( \sigma^{-2} + \sum_{j=1}^{p} \frac{\beta_j^2}{\sigma_j^2} \right) Y^2 + \left\{ \frac{\theta}{\sigma^2} + \sum_{j=1}^{p} \frac{\beta_j(X_j - \alpha_j)}{\sigma_j^2} \right\} Y + Q(X_1, \cdots, X_p) \right],
\]

where \( Q(X_1, \cdots, X_p) \) is a quadratic function of \( (X_1, \cdots, X_p) \). It is easy to see that conditional on \( (X_1, \cdots, X_p) \), \( Y \) has a normal distribution \( N(\bar{\mu}, \bar{\sigma}^2) \). From the coefficients of the linear and quadratic terms of \( Y \), we can obtain that

\[
\bar{\mu} = \mathbb{E}(Y | X_1, \cdots, X_p) = \frac{\sigma^{-2} \theta + \sum_{j=1}^{p} \sigma_j^{-2} \beta_j(X_j - \alpha_j)}{\sigma^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \beta_j^2},
\]

\[
\bar{\sigma}^2 = \text{Var}(Y | X_1, \cdots, X_p) = \frac{1}{\sigma^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \beta_j^2}.
\]

A.2 gPLS model for continuous outcome

For a future sample \( X_h = (x_{h1}, \cdots, x_{hp}) \), plugging the maximum likelihood estimates into (7), we have

\[
\hat{\mathbb{E}}(Y_h | X_h) = \frac{\hat{\sigma}^{-2} \hat{\theta} + \sum_{j=1}^{p} \hat{\sigma}_j^{-2} \hat{\beta}_j(x_{hj} - \bar{x}_j + \hat{\beta}_j \bar{y})}{\hat{\sigma}^{-2} + \sum_{j=1}^{p} \hat{\sigma}_j^{-2} \hat{\beta}_j^2} = \bar{y} + \frac{1}{\hat{\sigma}^{-2} + \sum_{j=1}^{p} \hat{\sigma}_j^{-2} \hat{\beta}_j^2} \sum_{j=1}^{p} \frac{Y^T(X_j - \bar{x}_j)}{||Y - \bar{y}||^2} \frac{x_{hj} - \bar{x}_j}{\hat{\sigma}_j^2}.
\]

Hence

\[
G_1(X_h; X, Y) = \sum_{j=1}^{p} \frac{Y^T(X_j - \bar{x}_j)}{||Y - \bar{y}||^2} (x_{hj} - \bar{x}_j),
\]

\[
\eta_0(X, Y) = \bar{y},
\]

\[
\eta_1(X, Y) = \frac{||Y - \bar{y}||^{-1}}{\hat{\sigma}^{-2} + \sum_{j=1}^{p} \hat{\sigma}_j^{-2} \hat{\beta}_j^2} = \frac{n^{-1}||Y - \bar{y}||}{1 + ||Y - \bar{y}||^2 \sum_{j=1}^{n} \hat{\beta}_j^2 ||X_j - \bar{x}_j - \hat{\beta}_j(Y - \bar{y})||^{-2}}.
\]
where

\[ \hat{\beta}_j = \frac{Y^T (X_j - \bar{x}_j)}{\| Y - \bar{y} \|^2}, \quad \hat{\sigma}_j^2 = \frac{\| X_j - \bar{x}_j - \hat{\beta}_j (Y - \bar{y}) \|^2}{n}. \]

### A.3 gPLS model for binary response

For a future sample \( X_h = (x_{h1}, \ldots, x_{hp}) \), plugging the maximum likelihood estimates into (11), we have

\[
\text{Logit}\{ \Pr(Y_h = 1 | X_h) \} = \text{Logit}(\bar{y}) - \sum_{j=1}^p \frac{\hat{\beta}_j^2}{2\hat{\sigma}_j^2} + \sum_{j=1}^p \frac{\hat{\beta}_j}{\hat{\sigma}_j} (x_{hj} - \bar{x}_j + \hat{\beta}_j \bar{y}) \\
= \text{Logit}(\bar{y}) + (\bar{y} - 1/2) \sum_{j=1}^n \frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} + \sum_{j=1}^p \frac{\hat{\beta}_j}{\hat{\sigma}_j} (x_{hj} - \bar{x}_j).
\]

Hence

\[
G_1(X_h; X, Y) = \| Y - \bar{y} \| \sum_{j=1}^p \frac{\hat{\beta}_j}{\hat{\sigma}_j} (x_{hj} - \bar{x}_j) = \sum_{j=1}^p \frac{Y^T (X_j - \bar{x}_j)}{\| Y - \bar{y} \| \hat{\sigma}_j^2} (x_{hj} - \bar{x}_j),
\]

\[
\gamma_0(X, Y) = \text{Logit}(\bar{y}) + (\bar{y} - 1/2) \sum_{j=1}^n \frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2},
\]

\[
\gamma_1(X, Y) = \| Y - \bar{y} \|^{-1},
\]

where

\[ \hat{\beta}_j = \frac{Y^T (X_j - \bar{x}_j)}{\| Y - \bar{y} \|^2}, \quad \hat{\sigma}_j^2 = \frac{\| X_j - \bar{x}_j - \hat{\beta}_j (Y - \bar{y}) \|^2}{n}. \]

### A.4 Prediction error of the oracle procedure for the simulation studies

The “oracle” procedure is based on the true model, which can be simplified as

\[ Y = \alpha + \beta X + \epsilon, \quad X \sim N(0, \sigma_1^2), \quad \epsilon \sim N(0, \sigma_2^2), \]
\( n_1 = 100 \) training samples are used to estimate the regression coefficients \( \alpha \) and \( \beta \), which are then applied to predict \( n_2 = 100 \) testing samples.

\[
\text{Err} = \sum_{i=1}^{n_2} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2, \quad \mathbb{E}(\text{Err}) = n_2\mathbb{E}_1\{\mathbb{E}_2(Y - \hat{\alpha} - \hat{\beta}X)^2\},
\]

where the expectation \( \mathbb{E}_2 \) is over the testing data \((X, Y)\), and \( \mathbb{E}_1 \) is over the training data, i.e., over the distribution of \( \hat{\alpha}, \hat{\beta} \). We have

\[
\mathbb{E}(\text{Err}) = n_2\mathbb{E}_1\{\mathbb{E}_2\{\alpha - \hat{\alpha} + X(\beta - \hat{\beta}) + \epsilon\}^2\} = n_2\mathbb{E}_1\{(\alpha - \hat{\alpha})^2 + \sigma_1^2(\beta - \hat{\beta})^2 + \sigma_2^2\}
\]

\[
= n_2\sigma_2^2\mathbb{E}_X \left\{1 + \frac{1}{n_1} + \frac{\bar{x}^2 + \sigma_2^2}{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}\right\} = n_2\sigma_2^2 \left\{1 + \frac{1}{n_1} + \frac{\mathbb{E}(\chi^2) + n_1}{n_1} \mathbb{E}_2 \chi^2_{n_1-1}\right\}
\]

which is independent of the variance of \( X \), and the first and second order terms reflect the uncertainty in estimating the regression parameters from the training data.

Previous derivation has used the fact that for \( X \sim \chi^2_n \), we have

\[
E(X^m) = \prod_{k=1}^{m}(n - 2 + 2k), \quad E(X^{-m}) = \frac{1}{\prod_{k=1}^{m}(n - 2k)}.
\]

Similarly we can obtain the variance for the prediction error

\[
\text{Var} \{\text{Err}\} = \mathbb{E}_1\{\mathbb{V}ar_2(\text{Err})\} + \mathbb{V}ar_1\{\mathbb{E}_2(\text{Err})\}
\]

\[
= n_2\sigma_2^4 \left\{\sum_{k=1}^{n_1}(n_1 - 2k)\right\} + \frac{8}{n_1(n_1 - 3)} + \frac{26}{n_1(n_1 - 3)(n_1 - 5)} + \frac{36}{n_1(n_1 - 3)(n_1 - 5)}
\]

\[
= n_2\sigma_2^4 \left\{\sum_{k=1}^{n_1}(n_1 - 2k)\right\} + \frac{3}{n_1(n_1 - 3)} + \frac{5}{n_1(n_1 - 3)^2} + \frac{6}{n_1(n_1 - 3)^2} + \frac{21}{n_1(n_1 - 3)(n_1 - 5)} + \frac{68}{n_1(n_1 - 3)(n_1 - 5)^2} + \frac{131}{n_1(n_1 - 3)^2(n_1 - 5)^2}\}
\]

### A.5 Tuning rgPLS for microarray cancer classification

The \( L_1 \) penalized estimation for gene \( j \) is

\[
\arg\max_{\beta_j, \sigma_j} -n \log \sigma_j - \frac{\sum_{i=1}^{n}(x_{ij} - \alpha_j - \beta_j y_i)^2}{2\sigma_j^2} - n\lambda_j|\beta_j|.
\]
It is easy to see that $\hat{\alpha}_j = \bar{x}_j - \hat{\beta}_j \bar{y}$. So the previous model can be equivalently written as

$$\arg \max_{\beta_j, \sigma_j} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_{ij} - \bar{x}_j}{\sigma_j} - \beta_j \frac{y_i - \bar{y}}{\sigma_j} \right)^2 - n \lambda_j |\beta_j|.$$  

Let $s_y = \|Y - \bar{y}\| / \sqrt{n}$, which is a rough estimate of the outcome standard error. We have

$$\frac{1}{2} \sum_{i=1}^{n} \left( \frac{x_{ij} - \bar{x}_j}{\sigma_j} - \beta_j \frac{y_i - \bar{y}}{\sigma_j} \right)^2 + n \lambda_j |\beta_j| = \frac{1}{2} \sum_{i=1}^{n} \left( \frac{x_{ij} - \bar{x}_j}{\sigma_j} - \frac{y_i - \bar{y}}{s_y} \frac{\beta_j s_y}{\sigma_j} \right)^2 + n \frac{\lambda_j \sigma_j |\beta_j| s_y}{s_y \sigma_j}.$$  

Here $(X_j - \bar{x}_j)/\sigma_j$ and $(Y - \bar{y})/s_y$ can be treated as the standardized response and covariate. And we propose to use common values for $\lambda_j \sigma_j / s_y$ across genes, i.e., we assume $\lambda = \lambda_j \sigma_j / s_y$, and hence $\lambda_j = \lambda s_y / \sigma_j$, which is equivalent to standardizing all variables before doing the regularized estimation. Here $\sigma_j$ is not know, we propose to use the residuals from the ordinary least squares regression as a rough estimate

$$\hat{\sigma}_j^{(0)} = \frac{\|X_j - \bar{x}_j - \hat{\beta}_j^{(0)} (Y - \bar{y})\|}{\sqrt{n}}, \quad \text{where} \quad \hat{\beta}_j^{(0)} = \frac{Y^T (X_j - \bar{x}_j)}{\|Y - \bar{y}\|^2}.$$  

So for the following one-step approximate solution

$$\hat{\beta}_j^{(1)} = \text{sign}(\hat{\beta}_j^{(0)}) \left\{ |\hat{\beta}_j^{(0)}| - \lambda_j \|X_j - \bar{x}_j - \hat{\beta}_j^{(0)} (Y - \bar{y})\|^2 \|Y - \bar{y}\|^{-2} \right\}^+,$$

we have

$$\hat{\beta}_j^{(1)} = \text{sign}(\hat{\beta}_j^{(0)}) \left\{ |\hat{\beta}_j^{(0)}| - \lambda \|X_j - \bar{x}_j - \hat{\beta}_j^{(0)} (Y - \bar{y})\|/\|Y - \bar{y}\| \right\}^+.$$  

For multi-class classification, we employed the “one versus the others” approach. Specifically we treat the indicator of each class as the response and build a gPLS model. The built gPLS models can estimate the individual class probabilities, which are then normalized to obtain the final probability estimates.

The shrunken centroids used by PAM (Tibshirani et al., 2002) can also be derived from the $L_1$ penalized linear regression model, but it did not take into account the sample size variation (see Wu, 2006, e.g.). The penalized partial least squares proposed by Huang et al.
(2004) for modeling continuous responses first applied the PLS algorithm to obtain the final regression model, which was reorganized as a linear function of the original variables. Then the soft-shrinkage was directly applied to the individual coefficients. For the proposed gPLSC studied in this paper, the coefficients are functions of both the regression coefficient $\beta$ and the variance $\sigma^2$. The proposed $L_1$ penalized estimation directly shrink $\beta$. It would be of interest to empirically compare various shrinkage methods in future research.

A.6 Principal component analysis (PCA) of the prostate cancer microarray data

A.6.1 Mean scaled expression

We first analyzed the mean scaled expression data and followed the pre-processing procedure as Singh et al. (2002).

Figure A.1 shows the percentage of the explained variation by the first 10 principal components (PCs). The PCA has been applied to the raw data and after centering/standardization. For all situations, the first PC explained majority of the variation, especially so for the raw data.

Since the PCA is obtained without using any information from the response (sample status information), the first PC explaining the maximum variation does not necessarily have the best correlation with the response. Figure A.2 to A.4 plot the scatterplot matrix for the first 5 PCs from the three data, where the first class is labeled with black cross symbols and the second class with the red dots. We can see that the first PC does not have a clear separation for the two groups. Overall it looks like the third PC has the best correlation with the response (i.e., the best separation between the two groups), which only explained 1.7%, 1.7% and 3.9% of the total variation for the three data respectively.
Figure A.1: PCA of the prostate cancer microarray data: plotted are the percentage of explained variance for the first 10 PCs. The first plot is for the raw expression data, the second one is obtained after centering the expression values of each genes, while the third one further standardizes the expression values of each gene.

A.6.2 RMA expression summary

The expression CEL files are downloaded and processed using the robust multichip average (RMA; see Irizarry et al., 2003, e.g.). We repeat previous analysis.

Figure A.5 shows the percentage of the explained variation by the first 10 principal components (PCs). The PCA has been applied to the raw data and after centering/standardization. For all situations, the first PC explained majority of the variation, especially so for the raw data.

Figure A.6 to A.8 plot the scatterplot matrix for the first 5 PCs from the three data, where the first class is labeled with black cross symbols and the second class with the red dots. We can see that the first PC does not have a clear separation for the two groups. Overall it looks like the second or third PC has the best correlation with the response.
Figure A.2: Scatterplot of the first 5 PCs for the raw expression data. There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.
Figure A.3: Scatterplot of the first 5 PCs for the centered expression data. There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.
Figure A.4: Scatterplot of the first 5 PCs for the centered and scaled expression data. There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.
Figure A.5: PCA of the prostate cancer microarray data (RMA): plotted are the percentage of explained variance for the first 10 PCs. The first plot is for the raw expression data, the second one is obtained after centering the expression values of each genes, while the third one further standardizes the expression values of each gene.
Figure A.6: Scatterplot of the first 5 PCs for the raw expression data (RMA). There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.
Figure A.7: Scatterplot of the first 5 PCs for the centered expression data (RMA). There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.
Figure A.8: Scatterplot of the first 5 PCs for the centered and scaled expression data (RMA). There are two groups of samples, the first one being labeled with black cross symbols and the second one with the red dots.