Partial Least Squares Discriminant Analysis of Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia Phenotypes Using DNA Expression Microarray

Jonathan Bryan

April 30, 2018

Abstract

Gene-phenotype association studies using high-dimensional DNA microarray present a classical problem of insufficient degrees of freedom to estimate a regression model using classical approaches such as ordinary least squares. We use a partial least squares discriminant analysis (PLS-DA) method to first find a lower dimensional representation of the gene expression data and then regress responses on the lower dimensional component predictors to perform discriminant analysis for classification. We can conveniently map our lower dimensional predictors back to the original gene expression data to get estimated coefficients for individual genes. Data from Golub et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" was used to perform PLS-DA to classify 73 cases of either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) on 7129 gene expression covariates. The PLS-DA model achieved 100 percent accuracy with only two components for the training set and three components for the test set which suggested a more parsimonious solution compared to the 50 genes used in the reference study. Using the reverse mapping method, we identified the top five genes associated with ALL and AML. "M19507_at" encodes for myeloperoxidase had the highest coefficient associated with ALL while gene "D49824_s_at" encodes for a serotype of class 1 major histocompatibility complex (MHC) molecules and had the largest association with AML.

Introduction

Classification models with linear and additive predictors are ill-posed when the number of predictors in the model is greater than the number of samples. This problem has infinitely many solutions unless constraints are made to the objective function. Alternative regression approaches such as principal component and partial least squares regression have been developed that find a lower dimensional space to which the response is regressed. This paper reviews both PCR and penalized classification approaches before explaining the partial least squares discriminant analysis approach used.

Principal components classification

The approach of principal components regression (PCR) is to factor the design matrix into a pre-specified or cross-validated number of lower dimensional components used to model the response.[1] These lower dimensional components can then be used as predictors in a classifier. The principal component scores and loadings can also be transformed back into the original predictors to estimate the regression coefficients of the full model. The model is shown below, where $X_{n\times p}$ is a design matrix, $W_{a\times p}$ is a loading matrix, $T_{n\times a}$ is a score matrix, $\beta_{1\times p}$ is a vector of coefficients, $y_{n\times 1}$ is a vector of responses, ϵ is the error term, $g()$ is a link function and L is an exponential likelihood function.

[1] Wehrens, Ron, and B-H. Mevik. "The pls package: principal component and partial least squares regression in R." (2007).

$$T = XW \text{ (PCR step)}$$
$$g(E[y]) = T\beta + \epsilon \text{ (Linear classifier)}$$
$$arg\ \max_{\beta}\ L(\beta; y, T) \text{ (MLE)}$$

A key limitation of PCR is the loss of information between the response variable and the newly projected subspace of the predictors as the dimensional reduction takes place independently of the response variation. This limits the predictive ability of the PCR model.

Penalized classification

Another popular approach is LASSO regularization, which uses an L1 penalty to coerce parameter estimates to zero, inducing sparsity on the model.[2] However, when $p > n$, LASSO classifiers can select a maximum of n predictors, as well as selecting only one covariate from a set of highly correlated predictors. In addition, if the design matrix contains discrete or discontinuous variables then LASSO classifier may lack a unique solution, requiring methods to assess predictor-inclusion uncertainty.[3] Another approach, the elastic net penalty overcomes the situations when the LASSO optimization is not strictly convex by combining the L1 and L2 (Ridge regression) regularizers to enforce strict convexity.[4] The elastic net has the benefit of producing a unique solution, inducing shrinkage, and enforcing congruence for coefficient estimates of highly correlated variables rather than just selected one and dropping the others. In the model specification shown below, λ is the tuning parameter that is usually optimized using cross-validation.

$$g(E[y]) = X\beta + \epsilon \text{ (Linear classifier)}$$
$$arg\ \max_{\beta}\ L(\beta; y, X) + \lambda_1||\beta||_1 \text{ (Lasso)}$$
$$arg\ \max_{\beta}\ L(\beta; y, X) + \lambda_1||\beta||_1 + \lambda_2||\beta||^2 \text{ (Elastic net)}$$

Phenotype-gene expression data

Phenotype-gene expression studies provide a natural setting for the application of dimension reduction techniques and shrinkage models.[5] Gene expression data can record continuous numbers of reads for thousands of unique DNA and RNA sequences from an individual. Often researchers are interested in relationships between the normalized genetic expression data and observed phenotypes, such as protein levels and disease states. Golub et al. utilized neighborhood analysis on 6817 genes to isolate 1100 genes with higher correlations to either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) than expected by pure chance.[6] The authors then further isolated 50 informative genes based again upon their correlation with the leukemia classes. Each gene was then given a weighted vote based upon expression level and correlation with the class distinction. A prediction strength was established a priori and the data was run through the model and summed to determine the predicted class. Leave-one-out cross-validation was used to assess the accuracy of the predictors. While the authors achieved 100 percent accuracy, the choice of 50 genes was arbitrary, and in fact predictors based on between 10 and 200 genes all had 100 percent accuracy. Such ad-hoc methods are beneficial for proof-of-concept, but more robust methods for model selection, such as dimension reduction and regularization methods are preferred.

[2]Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological) (1996): 267-288.

[3]Tibshirani, R. J. (2013). The lasso problem and uniqueness. Electronic Journal of Statistics, 7, 1456-1490.

[4]Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[5]Pérez-Enciso, M., & Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Human genetics, 112(5-6), 581-592.

[6]Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.

Methods

Partial least squares regression in combination with discriminant analysis (DA) was used to model the leukemia class of the individuals in the sample. The modeling objective is to collapse the gene expression data into orthogonal components in a way that preserves information between the latent factors as the response. The latent factors are then used as predictors in the DA classifier.

Partial least squares discriminant analysis

In comparison to PCR, partial least squares discriminant analysis (PLS-DA) seeks to maximize the covariance of latent factors and the response variable.[7] PLS-DA produces X-scores and loadings, much like PCR, but also Y-scores and loadings that explain the given response space (assuming there is more than one response variable, if not then Y is just a column vector). The PLS-DA algorithm will pass information from the Y response space to the X predictor space by swapping the score blocks at each iteration of calculating the modified principal components.

The resulting modified principal components are not orthogonal (although weights can be used to coerce them to be orthogonal) but are more strongly associated with variation in the response space and provide more accurate predictions .[8] PLS-DA has several advantages over other methods. It is advantageous when the number of predictors is larger than the sample size and collinearity is high among the predictors. PLS-DA is also a robust prediction method because it reduces out of sample variance of residual errors and noise in the data in comparison to common linear classification algorithms. It important to note that the design matrix should be normalized because the algorithm is sensitive to scale. In the equation below for a binary classifier, the $T_{n\times a}$ and $U_{n\times 2}$ matrices are the lower dimensional scores matrices for X and Y, respectively. The $P_{a\times p}$ and $Q_{a\times 2}$ matrices are each orthogonal loading matrices while the $E_{n\times p}$ and $F_{n\times p}$ are error terms.

$$
\begin{aligned}
&X = TP^T + E \text{ (X decomposition)}\\
&Y = UQ^T + F \text{ (Y decomposition)}\\
&Y = TBQ^T + F\\
&arg\ \min_{\beta} ||Y - TBQ^T||
\end{aligned}
$$

PLS-DA predicts unknown observations by multiplying the new matrix $X^{(new)}$ by the weighted loading matrices P^* obtained from the trained model. $X^{(new)}$ is decomposed into the specified number of components and predictive scores values are calculated. The classification boundary is usually the closest group mean score. PLS-DA can be conceptualized as a penalized canonical correlation analysis, where the separation boundary is a function of the among-groups variability, where in comparison PCA discrimination can fail if within-group variation overwhelms among-group variation.[9]

[7] Liland, Kristian Hovde, and Ulf Geir Indahl. "Powered partial least squares discriminant analysis." Journal of Chemometrics23.1 (2009): 7-18.

[8] Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. Chemometrics and intelligent laboratory systems, 33(1), 47-61.

[9] Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics: A Journal of the Chemometrics Society, 17(3), 166-173.

Results

Model accuracy

The 10-fold cross-validated PLS-DA model contained 33 lower dimensional components of the gene expression covariates. Twenty-one components appear adequate enough to explain much of the variation in the gene expression data (Figure 1). The model achieved 100 percent accuracy on the training set with two components and with three components on the test set (Figure 2). Additional, component predictors on the test set increases the error rate to 3 percent. Figures 2 and 3 highlights the separation of the feature space in relation to the leukemia class. Note that all numbers including and the left of 29 are AML classes while those to the right are all ALL classes.

Figure 1. Component Percentage of Gene Expression Variance

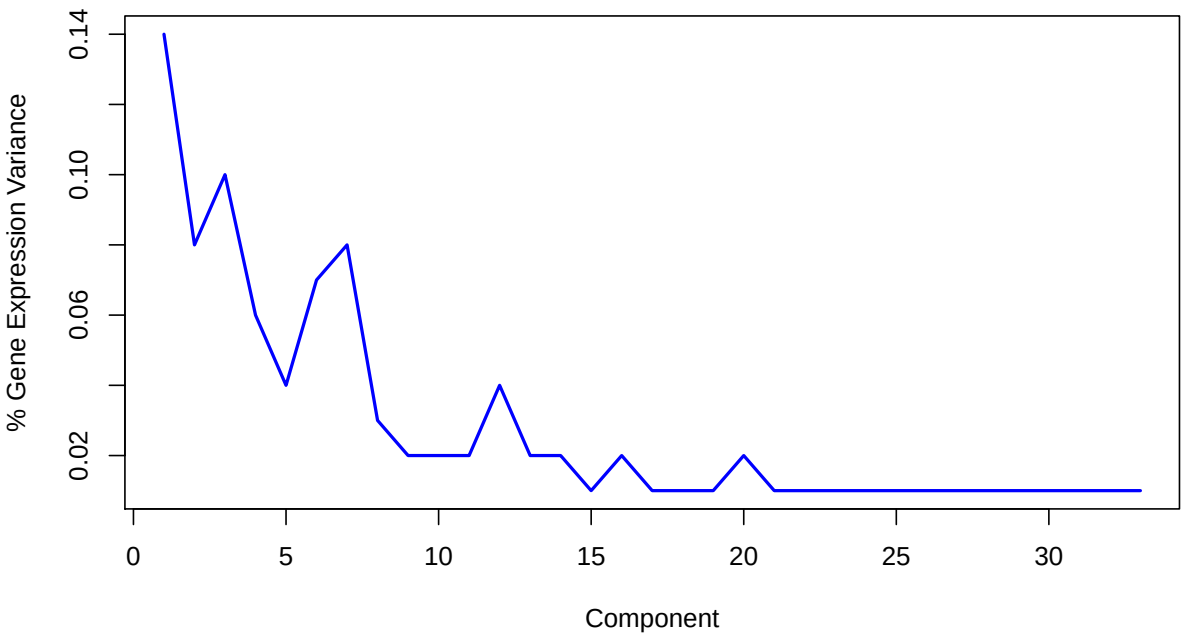


Figure 2. PLS−DA Classification Error

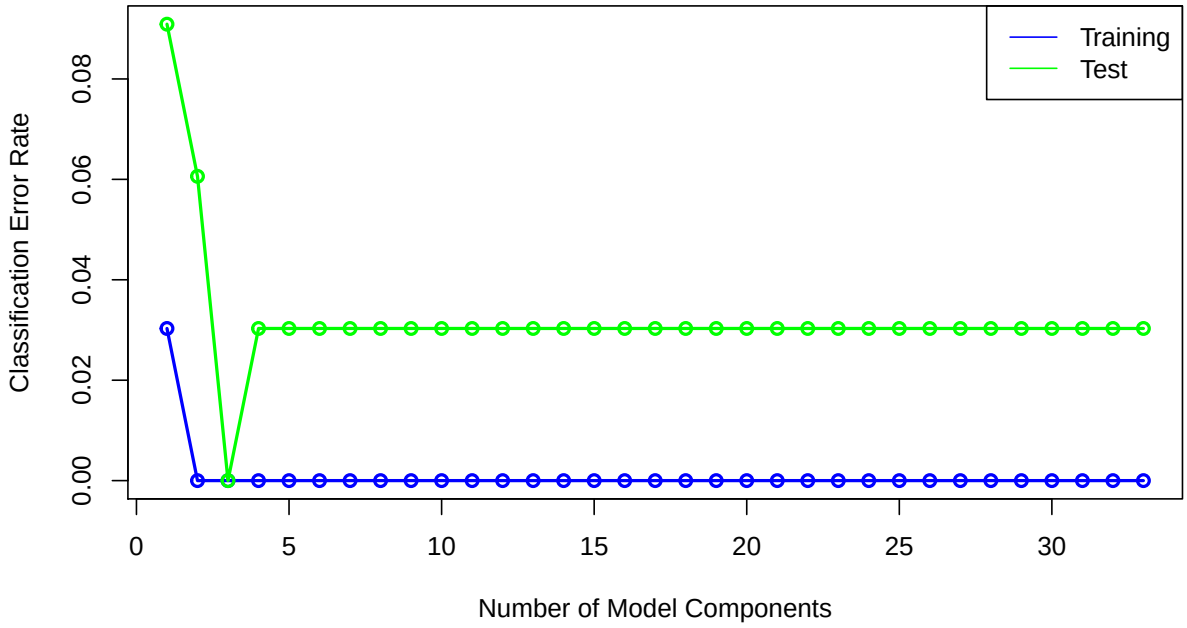


Figure 3. Score Plot of Component 1 and 2 for Training Set

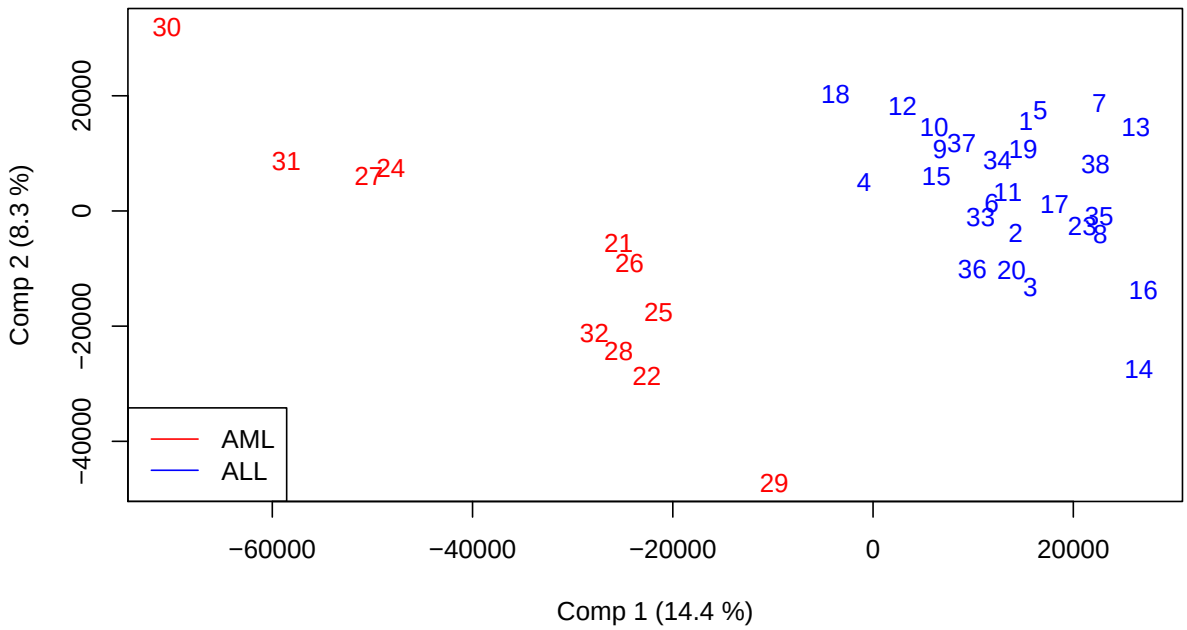


Figure 4. Score Plot of Component 1 and 2 for Test Set

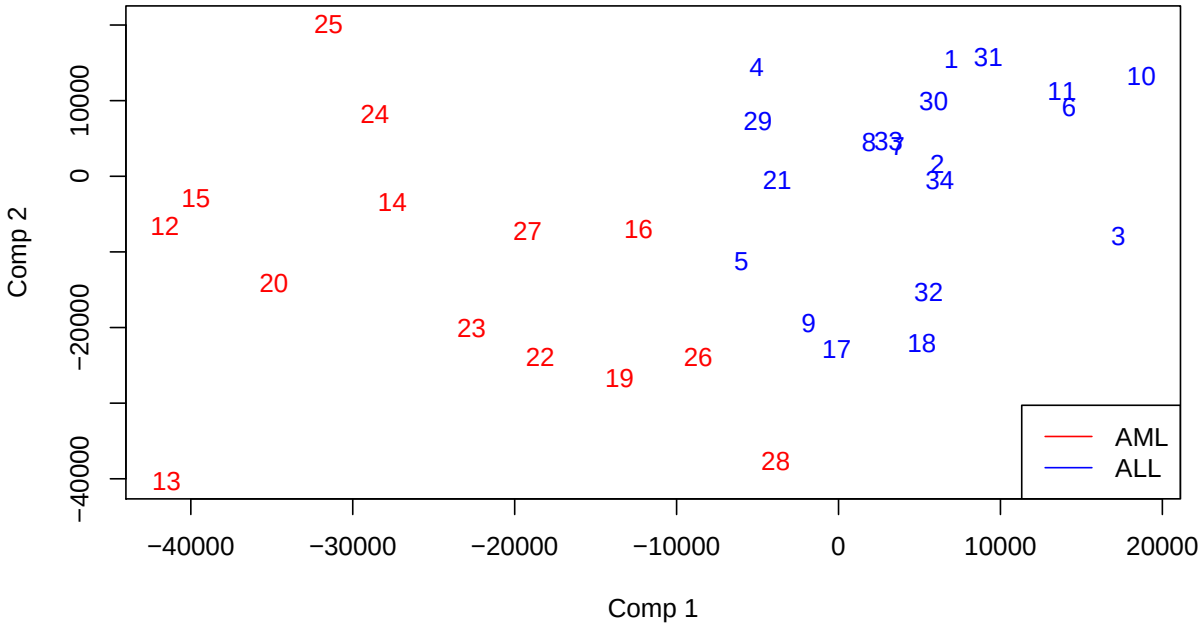


Table 1: Top 5 Genes Associated with ALL

	Description	Coefficient
Z19554_s_at	type III intermediate filament protein	3.5e-06
M25079_s_at	hemoglobin subunit beta	3.6e-06
M91036_rna1_at	hemoglobin subunit gamma-1	3.6e-06
M96326_rna1_at	azurocidin	3.9e-06
M19507_at	myeloperoxidase	4.2e-06

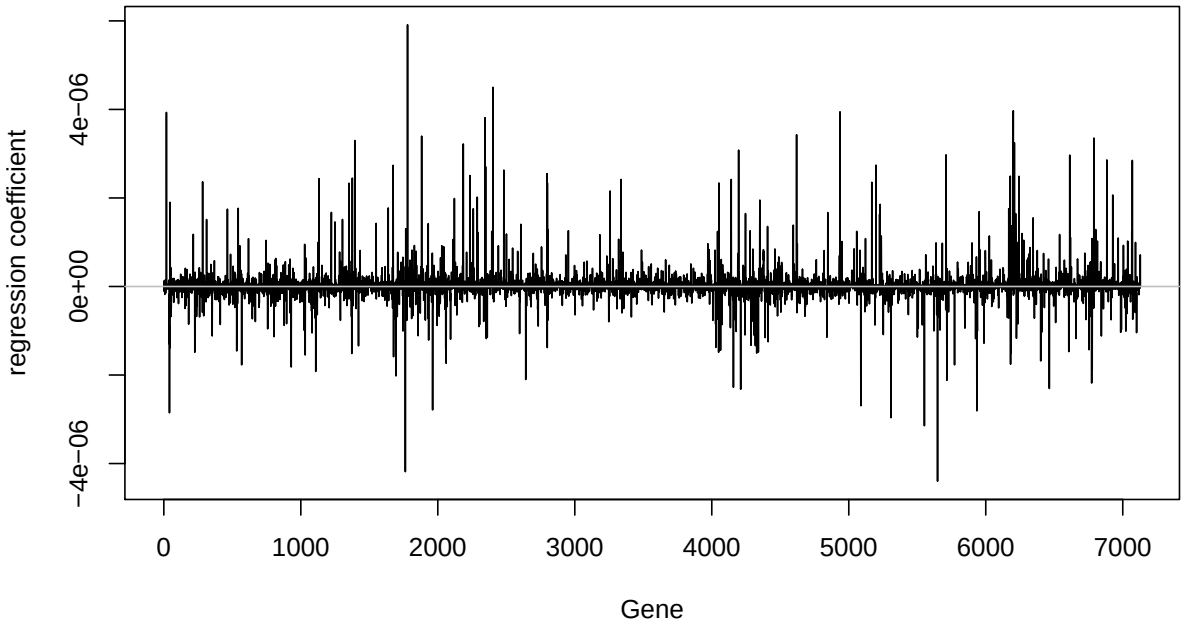
Table 2: Top 5 Genes Associated with AML

	Description	Coefficient
L20688_at	GDP-dissociation inhibitor protein	-2.5e-06
M17733_at	thymosin beta-4	-2.8e-06
M14483_rna1_s_at	prothymosin alpha	-2.9e-06
L06797_s_at	orphan G protein-coupled receptor	-3.0e-06
D49824_s_at	class 1 major histocompatibility complex	-4.0e-06

Influential Predictors

Figure 4 is the transformed components to the original gene expression coefficients. It is clear that the overall effect sizes are small and most genes do not contribute to the predictive ability of the classifier. Tables 1 and 2 display the top five genes with the largest coefficients associated with ALL and AML. We observe that gene "M19507_at", which encodes for myeloperoxidase, with the largest coefficient associated with ALL. In contrast, gene "D49824_s_at", which encodes for a serotype of class 1 major histocompatibility complex (MHC) molecules, had the largest association with AML. Only the gene that codes for Azurocidin was cross-listed with the 50 genes that Golub et al. isolated.

Figure 5. PLS−DA Component Coefficients



Discussion

In comparison to the Golub et al. approach, PLS-DA was able to achieve similar levels of accuracy with a substantially smaller amount of predictors selected in a rigorous fashion. This approach also allows us to work backwards to determine the predictive power of all genes on the response rather than removing them from the model. However, PLS-DA requires cross-validation or out-of-sample testing to determine the optimal number of components to use for out-of-sample testing.

The nonexistence of a unique solution when the number of predictors is larger than the sample size of the data is a challenging problem for linear classifiers, especially when multicollinearity and missing data are also issues. Other methods that seek to reduce the dimensions of the predictors, such as principal component analysis (PCR), allow for noise reduction and solve the collinearity problem. However, useful information for accurate prediction can be lost because PCR explains the useful directional information in the predictor space, which may not be sufficiently linked to the space of the new observed responses. PLS-DA in this paper has been shown to produce equally accurate parsimonious gene expression models for predicting cancer phenotype while allowing for a full model of the gene expression parameter space.