Bioconductor Project
Bioconductor Project Working Papers

Year 2004 *Paper* 5

Classification Using Generalized Partial Least Squares

Beiying Ding* Robert Gentleman†

*Medical Affairs, Amgen Inc., bding@amgen.com

†Department of Biostatistics, Harvard School of Public Health, rgentlem@hsph.harvard.edu

http://biostats.bepress.com/bioconductor/paper5

Classification Using Generalized Partial Least Squares

Beiying Ding and Robert Gentleman

Abstract

The advances in computational biology have made simultaneous monitoring of thousands of features possible. The high throughput technologies not only bring about a much richer information context in which to study various aspects of gene functions but they also present challenge of analyzing data with large number of covariates and few samples. As an integral part of machine learning, classification of samples into two or more categories is almost always of interest to scientists. In this paper, we address the question of classification in this setting by extending partial least squares (PLS), a popular dimension reduction tool in chemometrics, in the context of generalized linear regression based on a previous approach, Iteratively ReWeighted Partial Least Squares, i.e. *IRWPLS* (Marx, 1996). We compare our results with two-stage PLS (Nguyen and Rocke, 2002A; Nguyen and Rocke, 2002B) and other classifiers. We show that by phrasing the problem in a generalized linear model setting and by applying bias correction to the likelihood to avoid (quasi)separation, we often get lower classification error rates.

Classification Using Generalized Partial Least Squares

April 30, 2004

Abstract

Advances in computational biology have made simultaneous monitoring of thousands of features possible. The high throughput technologies not only bring about a much richer information context in which to study various aspects of gene functions but they also present challenge of analyzing data with large number of covariates and few samples. As an integral part of machine learning, classification of samples into two or more categories is almost always of interest to scientists. We address the question of classification in this setting by extending partial least squares (PLS), a popular dimension reduction tool in chemometrics, in the context of generalized linear regression, based on a previous approach, Iteratively ReWeighted Partial Least Squares, i.e. *IRWPLS* (Marx 1996). We compare our results with two-stage PLS (Nguyen and Rocke 2002a,b) and with other classifiers. We show that by phrasing the problem in a generalized linear model setting and by applying Firth's procedure to avoid (quasi)separation, we often get lower classification error rates.

Keywords: Cross-validation; Firth's procedure; Gene expression; Iteratively Reweighted Partial Least Squares; (Quasi)separation; Two-stage PLS.

1 Introduction

The wealth of gene expression data now available poses numerous statistical questions ranging from image analysis and variability analysis of gene expression levels (Chen et al. 1997, Newton et al. 2001), to the study of biochemical pathways. The huge number of genes relative to the moderate sample size renders many of the statistical modeling approaches inappropriate and hence efficient methods for dimension reduction and information extraction are of great interests. In this paper, we adapt a technology prevalent in chemometrics to the analysis of gene expression data. Our methodology easily extends to other settings, such as proteomic investigation through mass spectrometry or more classical problems such as Fisher's Iris data (Venables and Ripley 2002).

1.1 Partial Least Squares (PLS) in chemometrics

Similar data structures have been seen in the field of chemometrics, which has recently focused on analyzing observational data, originating mostly from organic and analytical chemistry, food research, and environmental studies. In these areas the number of observations tends to be many fewer than the number of measured variables and there is usually a high degree of collinearity among the variables, e.g. digitizations of analog signals, signals for different wavelengths in predicting chemical composition of a compound in spectroscopy. The similarity of these problems to those in computational biology suggests that the methodology developed for chemometrics may be appropriate for computational biology data.

Over the years, chemometricians have developed techniques for predictive modeling based on heuristic reasoning and the empirical evidence which have shown generally good performance. Both Partial Least Squares (PLS) and Principal Component Regression (PCR) have been popular regression methods in chemometrics (Wold 1975, Massy 1965). There are a wealth of articles on regression applications to chemical problems available in the *Journal of Chemometrics* (John Wiley) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier). An introduction to PLS regression is given by Geladi and Kowalski (1986) and the use of PLS in calibration can be found in Martens and Naes (1989). A statistical view of PLS along with other statistical methods is given by Frank and Friedman (1993).

1.2 Application of PLS for gene expression data

In this paper, we propose a procedure for two-group and multi-group (> 2 group) classification (prediction) of human tumor samples based on microarray gene expression data. The procedure involves incorporating PLS within the iteratively reweighted least squares (IRWLS) steps for multinomial or binary logistic regression. Our approach is based on Iteratively Reweighted Partial Least Squares (*IRWPLS*) first proposed by Marx (1996) and a procedure by Firth (1992a,b, 1993) which is applied to remedy and avoid the frequently encountered non-convergence and infinite parameter estimate

problems in logistic regression (Albert and Anderson 1984, Santner and Duffy 1986). This problem is usually present when the sample size is small relative to the number of parameters. Infinite parameter estimates can occur even when there is only one covariate which is highly predictive, hence the problem is due to the model rather than the ability to classify. For binary logistic regression, Heinze and Schemper (2002) showed that Firth's procedure gives finite parameter estimates.

More recently, more effort has been devoted to using penalized likelihood to tackle high dimensional problems, e.g. ridge penalized logistic regression (Eilers et al. 2001). Fort and Lambert-Lacroix (2003) also proposed combining PLS with logistic regression penalized with a ridge parameter. Comparisons of our results with their approaches are of interest and will be explored in future research.

2 Methods

We first introduce PLS in its original form, i.e. for a continuous response. We then consider the extension of PLS to generalized linear models (GPLS), specifically for categorical data in classification problems. A more detailed description is first devoted to the two-group classification problem where we also address separation problems in logistic regression. We then generalize the approach to multi-group classification.

2.1 Partial Least Squares (PLS)

Originating from general systems-analysis models and developed as a calibration method to predict chemical variables, PLS is usually presented as an algorithm.

Let $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2} \ldots, \mathbf{x_p})$ be the n by p matrix of predictors and $\mathbf{y}$ be the n by 1 response vector. $\mathbf{X}$ can often be written as a bilinear form (Kruskal 1978):

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}' + \mathbf{E_K} \\ &= \mathbf{t_1 p_1'} + \mathbf{t_2 p_2'} + \ldots + \mathbf{t_K p_K'} + \mathbf{E_K} \end{aligned}$$

where $\mathbf{T} = [\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_K}], \mathbf{P} = [\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_K}]$. The $\mathbf{t_k}$'s are called *latent variables* or *scores*, and the $\mathbf{p_k}$'s are called *loadings*. The $\mathbf{E_K}$ is the residual matrix and $\mathbf{K}$ is the number of PLS components. Moreover, we usually assume that the $\mathbf{X}$ matrix is standardized so that each column has mean 0 and standard deviation 1 (although the latter is not necessary). We further assume the following,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta \\ &= \mathbf{TQ} + \mathbf{f_K} \\ &= \mathbf{t_1}q_1 + \mathbf{t_2}q_2 + \ldots + \mathbf{t_K}q_K + \mathbf{f_K} \end{aligned}$$

where $\mathbf{Q} = [q_1, q_2, \ldots, q_K]$ and $\mathbf{f_K}$ is the residual. Thus, $\mathbf{X}$ and $\mathbf{y}$ are linked via the latent variables $\mathbf{T}$.

Usually the criterion for constructing components in PLS is to sequentially maximize the covariance between the response $\mathbf{y}$ and $\mathbf{Xg}$, subject to the constraint that $\mathbf{g'X'Xg} = 0$. The PLS components $\mathbf{t} = \mathbf{Xg}^*$ are orthogonal, where $\mathbf{g}^* = argmax_{\mathbf{g'g=1}}$ $\mathbf{Cov}(\mathbf{Xg}, \mathbf{y})$. If $\mathbf{K}$ is chosen to be the rank of $\mathbf{X}$ (i.e. minimum of the row rank and column rank of $\mathbf{X}$) and $\mathbf{X}$ is of full rank, then the PLS estimates of β are identical to ordinary least squares (OLS) estimates. However, since PLS is usually applied in cases where p is larger than n, a value of K smaller than the rank of $\mathbf{X}$ is often used. Hence, K can be viewed as a hyperparameter that also needs to be optimized. K is often selected by cross-validation as the number of PLS components for which the predicted sum of errors is minimized.

2.2 Two-stage PLS logistic regression

When the outcome variable is not continuous, the ordinary PLS method does not apply directly. Wang et al. (1999) proposed a probability-based multivariate algorithm combining partial least squares and logistic regression for identification of the development stages of oral cancer through analysis of autofluorescence spectra of oral tissues. Classification of the four stages of cancer development (normal, hyperplasia, dysplasia and

early cancer) is carried out in two steps (we will call this two-stage PLS regression in later sections). First the PLS components are obtained using the original covariates and coded response matrix (where 3 dummy variables of values 0 or 1 are used to represent the categorical response, which is treated as unordered). In the second step, they assume that

$$log(\frac{p_k}{p_1}) = \beta_{0k} + \boldsymbol{\beta}_k' \mathbf{s}_k$$

where $k = 1, \ldots, 4$, i.e. $p_k = P(class\ k|\mathbf{s}_k)$, and $\mathbf{s}_k$ is the vector of PLS components for an observation belonging to class k. MLE estimates for the regression coefficients can be obtained and the samples were classified into the category which has the highest predicted probability from the logistic regression based on the extracted components. The authors used leave-one-out cross-validation (LOOCV) to determine the number of PLS components and for evaluating the performance of the algorithm.

Nguyen and Rocke (2002b) applied a similar approach to problems of two-group tumor classification using two-stage PLS regression on microarray gene expression data. The original PLS procedure was first used for dimension reduction where the response variable is either 0 or 1, and logistic discrimination (LD) was applied to the chosen PLS components for classification. Quadratic discriminant analysis (*QDA*) was also tested as a comparison with LD. They applied their method to various data sets involving human tumor samples and stability of the classification results was assessed by re-randomization. They used a similar approach for multi-group classification (Nguyen and Rocke 2002a). Later we compare our results with theirs.

Although their results for two-group classification appear good, their approach may not be ideal since the original PLS algorithm is designed for a continuous outcome $\mathbf{y}$ with constant variance and a linear relationship with $\mathbf{X}$. Analogous to the development of generalized linear models to accommodate regressions of non-normal responses on a set of covariates, we consider the extension of PLS from the linear model to the generalized PLS setting.

2.3 Iteratively ReWeighted Partial Least Squares (*IRWPLS*)

McCullagh and Nelder (1989) showed that the maximum likelihood estimation of the parameters β of generalized linear models via Fisher scoring method can be rephrased as iteratively reweighted least squares (IRWLS) as follows,

$$\begin{aligned} \mathbf{A}\mathbf{b}^{t+1} &= \mathbf{A}\mathbf{b}^{t} + \mathbf{U} \\ &= \sum W\mathbf{x}\left(\eta + (y-\mu)\frac{\partial \eta}{\partial \mu}\right) \end{aligned}$$

where $\mathbf{b}^t$ are the regression coefficient estimates for β at t^{th} iteration, $\mathbf{U}$ is the score vector and $\mathbf{A}$ is the expected value of Fisher Information. The ij^{th} element of $\mathbf{A}$ is, $A_{ij} = -E(\frac{\partial U_i}{\partial \beta_j})$ where $i, j = 1, \ldots p$. The dependent variable here is a linearized form of the link function applied to the response variable,

$$z = \eta + (y-\mu)\frac{\partial \eta}{\partial \mu}$$

and the weights, W, are functions of the fitted values $\hat{\mu}$. Estimates are obtained by iteratively updating the adjusted dependent variable and weights until the convergence criterion is met.

Marx (1996) proposed an iteratively reweighted PLS algorithm which incorporates PLS into the framework of generalized linear models. That approach embeds the weighted PLS steps within the iterative steps, treating and updating the adjusted dependent variable $\mathbf{z}$ as the response rather than working with the original outcome. The two nested loops are iterated until the stopping criterion is satisfied. For more details refer to Marx (1996).

2.4 Firth's procedure

For classification problems using logistic regression, it is well-known that convergence poses a long-standing problem. Infinite parameter estimates can occur depending on the configuration of the sample points in the observation space (Albert and Anderson

1984, Santner and Duffy 1986). There are three categories of configurations of the sample points: complete separation, quasi-complete separation and overlap. For both complete and quasi-complete separation, there exists a vector $\mathbf{b}$ that correctly classifies all observations to their groups. Thus the MLE for $\boldsymbol{\beta}$ does not exist and the log-likelihood goes to zero and/or the dispersion matrix becomes unbounded as iterations proceed. Only under the *overlap* configuration do finite regression coefficient estimates exist. We note that although separation is an indication of perfect prediction and hence could be considered positively, it is problematic for logistic model fitting since it's in contradiction to the assumptions of the model. In high dimensional problems, such as analysis of gene expression data we find that separation is a commonly occurring problem. Since we want to make use of a logistic model as the basis for analyzing these data we must find some way to overcome the separation problem.

Firth (1992a,b, 1993) developed a procedure to remove the first-order term of the asymptotic bias of maximum likelihood estimates in GLMs based on a modification of the score function,

$$U(\beta_j)^* = U(\beta_j) + 0.5 \times trace\{I(\boldsymbol{\beta})^{-1}[\partial I(\boldsymbol{\beta})/\partial\beta_j]\} = 0,\ j = 1 \ldots p$$

where $U(\boldsymbol{\beta})$ is the original score function and $I(\boldsymbol{\beta})^{-1}$ is the inverse Fisher's information matrix evaluated at $\boldsymbol{\beta}$. When applying Firth's procedure to logistic regression, Heinze and Schemper (2002) showed that in the modified score function for logistic regression, each original observation, y_i, is split into two pieces, a *response* and a *nonresponse*. This guarantees finite estimates since for every covariate pattern there are some responses and some nonresponses which is the *overlap* configuration. So this procedure provides a solution to separation problems in logistic regression.

We can readily modify the original *IRWPLS*, by incorporating the Firth's procedure, to deal with two-group classification problems with large number of covariates (e.g.

genes). Specifically,

$$\begin{aligned} U(\beta_j)^* &= \sum_{i=1}^{n} \{(y_i + h_i/2) - p_i(1+h_i)\} x_{ij} \\ &= \sum_{i=1}^{n} (y_i^* - p_i^*) x_{ij} \\ &= \sum_{i=1}^{n} w_i^* x_{ij} \frac{\partial \eta_i}{\partial p_i^*} (y_i^* - p_i^*) \end{aligned}$$

where w_i^* is the i^{th} diagonal term of weight matrix

$$\begin{aligned} W^* &= W \times diag(h_i + 1) \\ y_i^* &= y_i + h_i/2 \\ p_i^* &= p_i \times (1 + h_i) \\ z_i^* &\doteq \eta_i + (y_i^* - p_i^*) \frac{\partial \eta_i}{\partial p_i^*}, \end{aligned}$$

now the pseudo response is z^*.

Although h_i's are functions of β, they are treated as fixed when derivatives of score functions with respect to β are taken to make the problem more tractable, hence the last equation above is actually an approximation. We call this approach *IRWPLSF* procedure in later sections.

Note that even though Marx (1996) didn't address the problem of separation, when it does occur, some ad-hoc criterions are usually used in order to get an estimate of the coefficient estimates. These rather arbitrary estimation criterions may actually invalidate other aspects of model fitting, e.g. selection of optimal K as an hyperparameter. The *IRWPLSF* procedure, however, avoids this problem. For example, it can properly evaluate the difference among convergent models with all values of K without any further data-dependent procedures.

2.5 *IRWPLS* for multi-group classification

In this section, we generalize the *IRWPLS* procedure to the multi-group classification scenario. It is a generalization of logit models for binary responses (see Fahrmeir et al. (2001) Chap. 3 for a discussion). In our application, we always treat the classes as nominal with no special ordering. But it would be straightforward to use any other model deemed appropriate, such as adjacent logit models, etc. Moreover, we assume that the counts at each configuration of the covariates are fixed, independent multinomials and we will refer to this model as the multinomial logit model. Let the categorical outcome Y have $C+1$ classes labeled $0, 1, \ldots, C$. We illustrate our model under the *common-baseline categorical* model that is commonly used. Suppose the baseline class is always labeled class 0, and for each $j = 1, \ldots C$, the logit model holds:

$$log(\frac{p_{ij}}{p_{i0}}) = \boldsymbol{\beta}_j' \mathbf{x}_i$$

where $i = 1, \ldots, n$, indexes samples. Note here that the most general case, i.e. different $\boldsymbol{\beta}_j$'s for each of the C logits, is considered. With the following constraints,

$$\sum_{j'=0}^{C} p_{ij'} = 1, \ \sum_{j'=0}^{C} y_{ij'} = 1$$

where $y_{ij} = I(sample\, i \in class\, j)$, with $I(\cdot)$ being an indicator function and using the standard form of the likelihood, the score functions are:

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &= \mathbf{X}'(\mathbf{Y} - \mathbf{P}) \\ &= \mathbf{X}'\mathbf{W}\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{P}}(\mathbf{Y} - \mathbf{P}). \end{aligned}$$

In the above formula, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_C)'$, where $\boldsymbol{\beta}_j$, $j = 1, \ldots, C$ is the $p \times 1$ regression coefficient vector corresponding to the j^{th} logit. $\mathbf{Y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n')'$ where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iC})'$, which is the $C \times 1$ response vector for the i^{th} sample. Similarly $\mathbf{P} = (\mathbf{p}_1', \mathbf{p}_2', \ldots, \mathbf{p}_n')'$ where $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{iC})'$, $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')'$

where

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i2} & \mathbf{0} & \dots & \mathbf{0} \\ & & \dots & & \\ \mathbf{0} & \mathbf{0} & \dots & 0 & \mathbf{x}_{iC} \end{pmatrix}$$

in which $\mathbf{x}_{ij}$ is the covariate vector corresponding to the j^{th} logit, $j = 1, \dots, C$. Usually $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \dots = \mathbf{x}_{iC} = \mathbf{x}_i$. Note that the baseline level, i.e. class 0, is now only implicitly modeled as a result of conforming to the linear constraints for uniqueness of the estimates. $\mathbf{W}$ is a block diagonal matrix with the i^{th} diagonal block being:

$$\mathbf{W}_i = \begin{pmatrix} p_{i1}(1-p_{i1}) & -p_{i1}p_{i2} & \dots & -p_{i1}p_{iC} \\ -p_{i1}p_{i2} & p_{i2}(1-p_{i2}) & \dots & -p_{i2}p_{iC} \\ & \dots & & \\ -p_{i1}p_{iC} & -p_{i2}p_{iC} & \dots & p_{iC}(1-p_{iC}) \end{pmatrix}$$

It can then be shown that the pseudo-response vector for the i^{th} sample is,

$$\mathbf{z}_i = \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{P}_i}(\mathbf{y}_i - \mathbf{p}_i).$$

Now the *IRWPLS* procedure can be carried out as before with the necessary changes, and we refer to this model fitting procedure as *MIRWPLS*.

For the multinomial logit model, the problem of (quasi-)complete separation still exists. Firth's procedure can be extended to multinomial case, which we will denote by *MIRWPLSF*. The pseudo response vector can be expressed in a similar form as before:

$$\mathbf{z}_i^* = \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{p}_i^*}(\mathbf{y}_i^* - \mathbf{p}_i^*)$$

where $\mathbf{p}_i^*$ and $\mathbf{y}_i^*$ are functions of the original $\mathbf{p}_i$ and $\mathbf{y}_i$ as in the description of *MIRWPLS*. The last equality illustrates that the problem can be rephrased as a multinomial logit model with an adjusted response vector $\mathbf{y}_i^*$ and mean vector $\mathbf{p}_i^*$ for each sample i. For a detailed derivation refer to Appendix A. Note, however, even though *MIRWPLSF*

tends to shrink β towards 0, and hence provides a more stable model than *MIRWPLS* (Firth 1993), finite estimates are no longer guaranteed due to the multiplicity of classes.

Prediction of outcome is simply based on polytomous discrimination (PD), which is essentially classifying a new observation into the class with highest predicted probability based on the fitted model. This prediction rule is also commonly referred as *softmax* (Ripley 1996).

2.6 Assessing Prediction

We use classification error rates to assess the performances of *IRWPLS*-based procedures and to compare with other classifiers. When a test set is available, the out-of-sample test set classification error rate is estimated using the model built on the training set. When a test set is not available, we use leave-one-out cross-validation (LOOCV). For each iteration, one of the n samples is reserved as a test sample and the remaining $n-1$ form the training samples. A model is constructed using the training samples, and the test sample, i.e. the one left out, is classified using the built model. After iterating through all n samples, the number of incorrect predictions divided by n is an estimate of the error rate.

The optimal number of PLS components is selected by choosing that value of K which minimizes LOOCV error rate for the training set. We also employ this for other procedures that involve hyperparameters, such as k-nearest neighbors (KNN).

3 Results

In this section, we compare the classification results from *IRWPLS*-based procedures with other classifiers including two-stage PLS, Fisher's linear discriminant analysis (*FLDA*), diagonal LDA (*DLDA*), quadratic discriminant analysis (*QDA*) (Dudoit et al. 2002), k-nearest neighbors (*KNN*), random forests (*RF*) (Breiman 2001, 2002) and support vector machines (*SVM*) (Furey et al. 2000, Guyon et al. 2002). Dudoit et al. (2002)

found that simple classifiers such as *DLDA* and *KNN* were better than other more sophisticated classifiers in a large scale comparison of discrimination methods based on their empirical distributions of misclassification error rates. All examples presented in this section were run on publicly available data. The data sources are listed in Appendix B. All software is available as R packages through either CRAN or the Bioconductor Project. See Appendix B for the appropriate URLs.

3.1 Two-group classification

3.1.1 Pima data

We first test our procedure on a simple data set where number of covariates is smaller than the number of samples available. We use the data as reported in Venables and Ripley (2002). There are 532 complete records, 200 of which are used as training set and the remaining 332 as the test set. The goal of the study was to establish a link between the 7 covariate measurements collected and whether or not the woman has diabetes (Smith et al. 1988). For both the training and test set, about 1/3 are from the diabetic group.

Results on applying *IRWPLS*-based procedures to the Pima diabetes data are shown in Table 1. K stands for the number of PLS components used. *Overall* refers to the overall misclassification rate whereas *N* and *D* refer to the class specific conditional error rates for nondiabetic and diabetic samples respectively. Due to the simple structure of the data and the abundance of samples, the best performance is achieved when using 7 PLS components, i.e. full rank, which is essentially logistic regression. LOOCV error for training set is 0.2350 and test set error is 0.1988 for both *IRWPLS* and *IRWPLSF*. The class specific training set CV error rates for *N* vs *D* are similar for *IRWPLS* and *IRWPLSF*, hence only detailed results for the latter are shown (Table 1). In general, the diabetic patients are more likely to be misclassified, possibly as a result of there being fewer cases in the data andor they may be more variable.

Table 1: % Misclassification for Pima Data using *IRWPLS(F)*

K	Training set (n=200) CV error				Test set (n=332) out-of-sample error			
	IRWPLS	*IRWPLSF*			IRWPLS	*IRWPLSF*		
	Overall	Overall	N	D	Overall	Overall	N	D
1	0.2500	0.2500	0.1364	0.4706	0.2199	0.2199	0.1345	0.4037
2	0.2500	0.2500	0.1515	0.4412	0.2078	0.2048	0.1166	0.3853
3	0.2800	0.2800	0.1591	0.5147	0.2229	0.2229	0.1031	0.3945
4	0.2600	0.2650	0.1515	0.4853	0.2048	0.2078	0.1031	0.3945
5	0.2600	0.2550	0.1439	0.4706	0.2078	0.2018	0.1031	0.3945
6	0.2650	0.2550	0.1364	0.4853	0.2018	0.2078	0.1031	0.3945
7	0.2350	0.2350	0.1288	0.4412	0.1988	0.1988	0.1031	0.3945

Table 2: Comparison of % Misclassification for Pima Data

Classifier	Training set (CV error)			Test set (out-of-sample error)		
	Overall	N	D	Overall	N	D
FLDA	0.2450	0.1364	0.4559	0.2018	0.1883	0.2294
DLDA	0.2400	0.2197	0.2794	0.2470	0.2242	0.5046
QDA	0.2750	0.1667	0.4853	0.2289	0.2108	0.2661
KNN	0.2550	0.1591	0.5735	0.2169	0.2063	0.2936
RF	0.2800	0.1667	0.4853	0.2469	0.1973	0.3119
SVM	0.2800	0.1515	0.5294	0.2319	0.2287	0.2385

The misclassification results from some other commonly used classification procedures are listed in Table 2. For *KNN*, $k = 6$ is the optimal number of k-nearest neighbor based on lowest LOOCV misclassification rate of the training set. The performance of *IRWPLS* and *IRWPLSF* is comparable with that of the other classifiers for this simple data set, both in terms of overall and class specific error rates. This is true even when *IRWPLS* or *IRWPLSF* are based on fewer than full rank PLS components. This example indicates that *IRWPLS*-based procedures are reasonable classifiers for standard low dimensional data.

3.1.2 Gene Filtering

Even though PLS can handle more covariates than there are samples, the number of genes in a gene expression dataset (often in the tens of thousands) is still too large

for practical use, especially given the fact that a considerable percentage of the genes do not show differential expressions across groups. Hence, filtering is often applied before classification to remove such genes. For the two-class problem reported here, we choose the m genes with the largest absolute t statistics. Gene selection is carried out as a part of the CV procedure, that is, every time we leave one sample out, both gene selection and model building are done using only the $n - 1$ samples and then prediction of the left out sample is done (Ambroise and McLachlan 2002).

3.1.3 Colon Data

The classification of colon cancer is discussed in Alon et al. (1999). Gene expression data from 40 tumor and 22 normal colon tissue samples were analyzed with an Affymetrix oligonucleotide array. Using two-way clustering, Alon et al. (1999) were able to cluster 19 normal and 5 tumor samples into one group and 35 tumor and 3 normal tissues into the other. Expression of the 2000 genes with highest minimal intensity across the 62 tissues were used in the analysis. Several EST's are replicated on the arrays and some replicates for the same EST have exactly the same expression measurements. Because of this, in all cases where there were replicate probe sets, we used the mean expression profile of the replicates, leaving 1911 nonredundant genes.

The number of misclassifications based on LOOCV for *IRWPLS* and *IRWPLSF* as well as *DLDA*, *KNN*, *RF* and *SVM*, are shown in Table 3. The numbers in brackets for *IRWPLS* and *IRWPLSF* are the optimal numbers of PLS components chosen by lowest LOOCV classification error rates of the two IRWPLS-based procedures respectively and those for *KNN* are the optimal numbers of nearest neighbors, again chosen by LOOCV.

The minimum number of 6 misclassifications is achieved by *IRWPLSF* with $m = 30$ genes and *KNN* with $m = 20$ genes. This result is comparable with Furey et al. (2000), who also misclassified 6 cases using a support vector machine (SVM). How-

Table 3: Comparison of Misclassification for Colon Data, n=62 (Tumor=40, Normal=22)

m	*IRWPLS*		*IRWPLSF*		*DLDA*	*KNN*		*RF*	*SVM*
5	11	(1)	12	(2)	11	12	(8)	16	13
10	9	(2)	8	(2)	8	8	(10)	12	10
20	7	(1)	7	(1)	7	6	(10)	8	9
30	8	(1)	6	(15)	8	8	(3)	9	8
40	8	(1)	8	(1)	8	9	(3)	9	9
50	8	(1)	8	(2)	7	8	(3)	9	8
100	8	(2)	7	(4)	11	7	(4)	10	10
200	8	(8)	6	(6)	14	8	(3)	10	9
500	10	(1)	7	(5)	18	8	(5)	10	10
1000	9	(4)	6	(5)	22	10	(2)	10	10

ever, Furey et al. (2000) used a slightly different feature selection procedure:

$$F(x_j) = \frac{|\mu_j^+ - \mu_j^-|}{\sigma_j^+ + \sigma_j^-},$$

where x_j is the gene expression for the j^{th} gene, μ_j^+, μ_j^- stand for the sample mean of the tumor and normal groups respectively, and σ_j^+, σ_j^- are the group standard deviations. This statistic is very similar to t-statistic assuming unequal group variance although summing over standard deviations rather than variances is rather unconventional. Applying the Furey filter, we find that the smallest misclassified number using *IRWPLSF* is 5 with $p = 40$, faring a little better than the SVM approach using the same gene selection procedure. This result also shows that gene selection influences the performance of classifiers and that it is quite important to make sure that comparison of classification methods is done by controlling nuisance factors such as the feature selection process.

Random splitting Due to the instability of LOOCV error rates for data with few samples and many covariates, comparison of various classifiers based solely on LOOCV classification errors may not be reliable. We now compare our *IRWPLS*-based procedures with the classifiers by randomly splitting the original dataset into a training set

and a test set. There is currently no consensus on how to choose the relative size of these randomly divided sets and we follow Dudoit et al. (2002) and choose the training set and test set size ratio to be 2:1. For each training and test set, we build the classifiers using the training set only and predict the test set data. The number of optimal PLS components for *IRWPLS*-based procedures and the optimal number of nearest neighbors for *KNN* are chosen by lowest CV error on the training set. Figure 1 shows the boxplots of the test set error rates for top $m = 10, 20, 30$ and 200 genes chosen by t statistic for each of the 6 classifiers based on $N = 100$ random splitting.

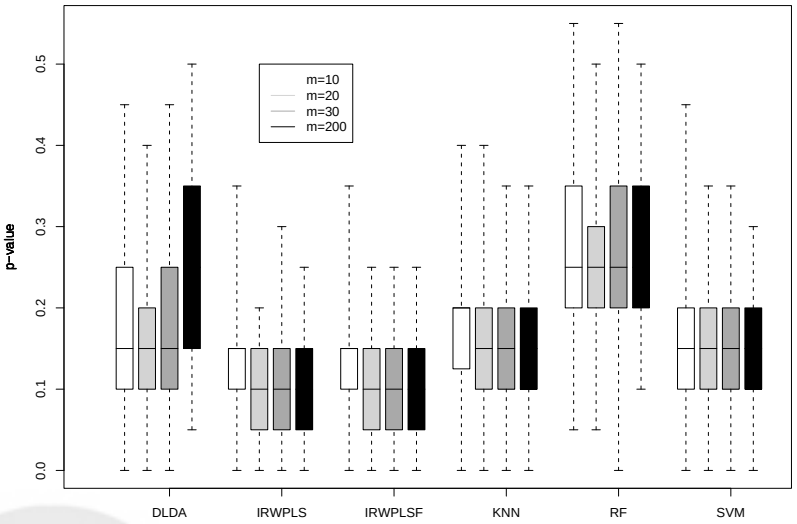

Figure 1: *Colon data 2:1 random splitting (N=100): boxplots of test set error rates for classifiers with top 10 (white), 20 (light grey), 30 (dark grey) and 200 (black) genes using a t-test filter.*

Figure 1 suggests that the error rates for *IRWPLS* and *IRWPLSF* are typically lower and less variable. There is no obvious difference between the distributions of error rates for *IRWPLS* and *IRWPLSF*.

3.2 Multi-group classification

3.2.1 Iris data

We now illustrate the *IRWPLS*-based multi-group classification approach, ie. *MIRWPLS* and *MIRWPLSF* and compare them with the other methods. We begin by using the well-studied iris data set (Fisher 1936).

Table 4: % Misclassification for Iris Data

K	*IRWPLS*	*IRWPLSF*
1	0.6667	0.6667
2	0.6800	0.6467
3	0.3800	0.2867
4	0.0333	0.0333
5	0.0267	0.0400
6	0.0400	0.0267
7	0.0200	0.0267
8	0.0200	0.0267
9	0.0200	0.0200
10	0.0200	0.0267

From Table 4, the minimum CV misclassification rate is 0.02 (3 out of 150) for both *MIRWPLS* (for K from 7 to 10) and *MIRWPLSF* with K=9. Compared with error rates from other standard classification procedures we mentioned before, e.g. *FLDA* (0.0200), *DLDA* (0.0400), *QDA* (0.0200), *KNN* (0.0400), *RF* (0.0400) and *SVM* (0.0333), we can see that the *MIRWPLS* procedures achieve the same minimum error rate (0.0200) with less than full rank than the other classifiers achieve using full data information.

In the following section, we compare results from *MIRWPLS* and *MIRWPLSF* with the multi-class PLS (*MPLS*) classification approach in Nguyen and Rocke (2002a) as well as other classifiers. The procedure in Nguyen and Rocke (2002a) is a natural extension of the two-stage PLS logistic regression, where the first stage of PLS component extraction is in principle the same as before only that now instead of univariate

response, C response variables are needed to uniquely represent $C + 1$ groups. The univariate PLS procedure can be extended to accommodate this situation (Höskuldsson 1988, Helland 1988, Garthwaite 1994). The second stage of *MPLS* uses the *common-baseline* multinomial logit regression model and polytomous prediction (PD), *QDA*, *DQDA* or *DLDA*.

3.2.2 Gene filtering

Analogous to using *t*-tests for two-group gene filtering, here we apply the *all-pairwise t-filter* used by Nguyen and Rocke (2002a), for choosing genes which have a large number of significant pairwise differences across groups.

3.2.3 NCI60 data

This study involves using cDNA microarrays to study the gene expression profiling among 60 cell lines from the NCI60 Cancer Microarray Project (Ross et al. 2000, Scherf et al. 2000). Data on 10 tumor cell lines (the numbers in brackets are the numbers of samples in that cell line): breast (7), central nervous system (CNS, 6), colon (7), leukemia (6), melanoma (8), non-small-cell-lung-carcinoma (NSCLC, 9), ovarian (6), prostate (2), renal (8), unknown (1), are available. To compare with results from *MPLS* reported in Nguyen and Rocke (2002a), we use only 5 of the cancer types: CNS, colon, leukemia, melanoma, and renal. Furthermore, a subset of 1415 genes (1375 genes and 40 drug targets) which were specifically studied by Scherf et al. (2000) were used. Missing values exist for some of the samples and expression values for genes with 2 or fewer missing values were imputed using the median expression across samples for that gene. Genes with more than 2 missing values were excluded from our analysis. This reduces the number of genes to 1299. Classification results are reported in Table 5.

Using the all pairwise t-filter, the misclassification rates for *IRWPLS*-based procedures are considerably lower compared with those of *MPLS-PD* (Table 5) (numbers in parentheses are the optimal PLS component numbers). The number of misclassifications drops from 15 (out of 35) for *MPLS-PD* to 2 for both *MIRWPLS* and *MIRWPLSF*,

Table 5: Comparison of Misclassification for subgroups of NCI60 data using all pairwise t-filter (CNS=6, colon=7, leukemia=6, melanoma=8, renal=8)

m	MPLS		*MIRWPLS*	*MIRWPLSF*	*DLDA*	*KNN*	*RF*	*SVM*
	PD	DLDA						
41-54-69 ($\geq$8)	15	5	2 (3)	2 (3)	2	2 (8)	3	3
148-159-189 ($\geq$7)	9	3	0 (3)	1 (4)	2	1 (4)	2	3

when genes having at least 8 significant pairwise absolute mean difference are used. Whereas, with even more genes, i.e. genes with at least 7 significant pairwise scores, almost perfect classification can be achieve using *IRWPLS*-based procedures (*MIRWPLS* = 0, *MIRWPLSF* = 1) compared with 9 misclassifications for *MPLS-PD*. Although error rates for *MPLS-DLDA* improved quite a bit over those of *MPLS-PD*, they are still consistently larger than those of the *MIRWPLS* and *MIRWPLSF*'s. In this case, all the other classifiers have comparable performances compared with *MIRWPLS(F)*, especially *DLDA* and *KNN* (numbers in parentheses are the optimal number of nearest neighbors).

There are no obvious explanations of the high error rate observed for *MPLS* in Table 5. However, intuitively it is unappealing to treat the binary elements, coded as dummy variables as continuous and to use their covariance (or correlation) with the **X**'s to construct PLS components. This point was also made in the two-group PLS classification results of section 2. Secondly, for *MPLS*, the objective criterion is to maximize the covariance between **Xw** and **Yc**, i.e. linear combinations of **X** and **Y** matrices respectively, until convergence. The interpretation of a linear combinations of elements in the response matrix is problematic. The issues with using the two-stage approach may be even more serious in the multi-group case than the two-group case, where **Y** is a vector and the second problem is not encountered. Moreover, we feel that since *MPLS* might suffer from convergence problems in the first stage of the PLS component construction, and (quasi-)complete separation in the second stage, its accuracy and predictive power is questionable. These interpretations, however, do not necessarily relate directly to why *MPLS* fares poorly.

To summarize, overall for the two-group case, the *IRWPLSF* procedure tends to be more stable, with its finite regression coefficients relative to *IRWPLS* in terms of classification as well as model fitting. For the multi-group classification problems, we have shown that both *MIRWPLS* and *MIRWPLSF* represent quite a substantial improvement over *MPLS*. There is no consistent evidence in our experiments to favor either one of the *IRWPLS*-based procedures for multi-group classification in terms of prediction error rate. Moreover, we have found that *IRWPLS* based procedures are, in general, comparable with other popular classifiers such as *FLDA*, *DLDA*,*QDA*, *KNN*, *RF* and *SVM*, etc. in terms of LOOCV error rates for the training set and test set error rates when test set is available, e.g. Pima data. When no test set is available, we also show that with random splitting into training and set sets, the distributions of test set error rates for *IRWPLS*-based procedures tend to be lower and less variable compared with other popular classifiers.

4 Discussion

With the introduction of high throughput microarray technology, data on the expression level of thousands of genes can be obtained simultaneously. This has provided a wealth of information as well as a challenge to develop efficient analytical methods, especially from a statistical point of view. We have in our efforts found a solution to one important aspect of machine learning, class prediction, via partial least squares regression. In comparison with the two-stage PLS approach (Wang et al. 1999, Nguyen and Rocke 2002b,b), we seek alternatives in the context of generalized linear models. We reintroduced the iteratively reweighted partial least squares (*IRWPLS*) first proposed by Marx (1996). We also resolve (quasi-)complete separation problems by applying Firth's procedure, which guarantees finite regression coefficients for binary logistic regression (Firth 1992a,b, 1993, Heinze and Schemper 2002). We further extended the *IRWPLS* procedure to multinomial logit case where more than 2 groups exist, *MIRWPLS*. A second multi-class model, *MIRWPLSF*, which incorporates bias reduction into multi-group classification is also derived.

We have shown that *IRWPLS*-based procedures have comparable classification efficiency with some of the classic approaches such as *FLDA, DLDA,QDA, KNN, RF* and *SVM* when standard data with relatively simple structure (i.e. $n > p$) is encountered. We have also observed that for high dimensional microarray expression datasets that *IRWPLS*-based procedures achieve lower classification error rates than the two-stage PLS approach especially for multi-group classification. They also tend to give fewer misclassifications compared with *DLDA, KNN, RF* and *SVM*. *(M)IRWPLSF* has similar performance as that of *(M)IRWPLS* but provides a more stable model.

Model-based classifiers, such as the *IRWPLS*-based procedures, may not be as flexible as algorithm-based ones. However, algorithmic classifiers, such as *SVM*, are often blackbox tools, with tuning parameters that are not necessarily intuitive for users, e.g. choice of kernel functions, scale factor, etc, whereas *IRWPLS*-based procedures provide us with a well-established framework not only for class prediction but also for good interpretation, stability and statistical inference. For example, one can interpret the latent variables, i.e. $\mathbf{t}$'s, similarly as one would with the principle components. Also each t_i is a linear combination of the original $\mathbf{X}$, this could suggest which covariates (genes) are important based on their weights. These issues relate more to the variable selection and gene importance, which is another important topic that will be addressed separately.

Moreover, even though we formulated multi-group classification IRWPLS for nominal classes under *common-baseline* logit model, it can easily be extended to handle ordinal classes, e.g. cumulative logit model, adjacent logit model, etc when cancer stages are naturally ordered. Such flexibility of model formulation in regression based methods, to best reflect the nature of the application, is usually not offered by the other classifiers. So even though sometimes we may not see substantial improvement of *MIRWPLS(F)* over the other algorithmic classifiers such as DLDA, KNN, etc. under a particular formulation of the logit model, the former really offers a much more powerful and flexible tool for problems of higher complexity.

Also in the PLS context, more research is needed to develop methods for determining the optimal number of components, estimating standard errors, and so on. We plan to explore these issues in our future work.

Acknowledgement

This research is supported in part by Jerry Ritz and the Connell O'Reilly Cell Manipulation and Gene Transfer Laboratory at DFCI. The authors are obliged to the Anestis Antoniadis, Gersende Fort, Sophie Lambert-Lacroix and the referees for helpful comments.

Appendix

A *MIRWPLSF pseudo-response derivation*

As we have mentioned in section 2.4, the Firth-modified score function is:

$$\begin{aligned} U(\boldsymbol{\beta})^* &= U(\boldsymbol{\beta}) + 0.5\frac{\partial}{\partial \boldsymbol{\beta}} log(|I(\boldsymbol{\beta})|) \\ &= U(\boldsymbol{\beta}) + 0.5\frac{\partial}{\partial \boldsymbol{\beta}} log(|\mathbf{X}'\mathbf{W}\mathbf{X}|) \\ &= U(\boldsymbol{\beta}) + 0.5\, trace\{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\frac{\partial \mathbf{W}}{\partial \boldsymbol{\beta}}\mathbf{X}\} \\ &= U(\boldsymbol{\beta}) + 0.5\, trace\{\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\frac{\partial \mathbf{W}}{\partial \boldsymbol{\beta}}\mathbf{W}^{-1}\} \\ &= U(\boldsymbol{\beta}) + 0.5\, trace\{\mathbf{H}\frac{\partial \mathbf{W}}{\partial \boldsymbol{\beta}}\mathbf{W}^{-1}\} \\ &= U(\boldsymbol{\beta}) + 0.5\, \mathbf{X}'\mathbf{H}_{\mathbf{w}} \end{aligned}$$

where $\mathbf{H} = \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ is the *hat matrix* and H_w is a $nC \times 1$ vector, each element of which is a function of the corresponding term of $\mathbf{P}$ and diagonal elements of $\mathbf{H}$.

$$
\begin{aligned}
\mathbf{H}_w &= \begin{pmatrix} h_{11} \\ h_{12} \\ \vdots \\ h_{1C} \\ \vdots \\ h_{n1} \\ h_{n2} \\ \vdots \\ h_{nC} \end{pmatrix} - \begin{pmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{1C} \\ \vdots \\ p_{n1} \\ p_{n2} \\ \vdots \\ p_{nC} \end{pmatrix} \cdot \begin{pmatrix} h_{1.} + h_{11} \\ h_{1.} + h_{12} \\ \vdots \\ h_{1.} + h_{1C} \\ \vdots \\ h_{n.} + h_{n1} \\ h_{n.} + h_{n2} \\ \vdots \\ h_{n.} + h_{nC} \end{pmatrix} \\
&= diag(\mathbf{H}) - \mathbf{P} \cdot \mathbf{H}^*
\end{aligned}
$$

where h_{ij} corresponds to the $((i-1) * C + j)^{th}$ diagonal term of $\mathbf{H}$ ($i = 1, \ldots n$, $j = 1, \ldots C$) and $h_{i.} = \sum_{j'=1}^{C} h_{ij'}$. $\mathbf{H}^*$, a column vector of length nC, is introduced here for notational convenience. $\mathbf{H}^* = (\mathbf{h}_1^{*T}, \mathbf{h}_2^{*T}, \ldots, \mathbf{h}_n^{*T})'$ and $\mathbf{h}_i^* = (h_{i.} + h_{i1}, h_{i.} + h_{i2}, \ldots, h_{i.} + h_{iC})'$. Similarly, $\mathbf{H}_w = (\mathbf{h}'_{w\,1}, \mathbf{h}'_{w\,2}, \ldots, \mathbf{h}'_{w\,n})'$ where $\mathbf{h}_{w\,i} = (h_{w\,i1}, h_{w\,i2}, \ldots, h_{w\,iC})'$.

Continuing the above derivation,

$$
\begin{aligned}
U(\boldsymbol{\beta})^* &= \mathbf{X}'\mathbf{W}\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{P}}(\mathbf{Y} - \mathbf{P}) + 0.5\,\mathbf{X}'\mathbf{W}\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{P}}\mathbf{H}_\mathbf{w} \\
&= \mathbf{X}'\mathbf{W}\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{P}}(\mathbf{Y} - \mathbf{P} + 0.5\,\mathbf{H}_\mathbf{w}) \\
&= \mathbf{X}'\mathbf{W}^*\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{P}^*}(\mathbf{Y} - \mathbf{P} + 0.5\,\mathbf{H}_\mathbf{w}),
\end{aligned}
$$

where similar to binary outcome case,

$$
\begin{aligned}
\mathbf{W}^* &= \mathbf{W}\,diag(\mathbf{H}_w/2 + 1) \\
\mathbf{P}^* &= \mathbf{P} \cdot (\mathbf{H}_w/2 + 1).
\end{aligned}
$$

Now the pseudo response vector can be expressed as:

$$\begin{aligned}
\mathbf{z}_i &= \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{P}_i^*}(\mathbf{y}_i - \mathbf{p}_i + 0.5\,\mathbf{h}_{w_i}) \\
&= \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{P}_i^*}\{(\mathbf{y}_i + 0.5\,\mathbf{h}_i) - (1 + 0.5\,\mathbf{h}_i^*)\mathbf{p}_i\} \\
&= \boldsymbol{\eta}_i + \frac{\partial \boldsymbol{\eta}_i}{\partial \mathbf{P}_i^*}(\mathbf{y}_i^* - \mathbf{p}_i^*)
\end{aligned}$$

where $\mathbf{h}_i = (h_{i1}, h_{i2}, \ldots, h_{iC})'$ and $\mathbf{h}_i^* = (h_{i1}^*, h_{i2}^*, \ldots, h_{iC}^*)'$.

B URLs

- Package URLs
  - CRAN: *http://cran.us.r-project.org/*
    - *FLDA, QDA*: *MASS*
    - *DLDA sma*
    - *KNN*: *class*
    - *RF*: *randomForest*
    - *SVM*: *e1071*
  - IRWPLS (S-PLUS): *http://www.stat.lsu.edu/bmarx*
  - Bioconductor: *http://www.bioconductor.org*
    - *(M)IRWPLS(F)*: *gpls*
- Data URLs
  - Pima data: R *MASS* package: *Pima.tr* (training set) and *Pima.te* (test set)
  - Alon colon data: *http://microarray.princeton.edu/oncology/affydata/index.html*
  - Iris data: R *MASS* package: *iris*
  - NCI60:
    - Information: *http://genome-www.stanford.edu/nci60*
    - Data: *http://discover.nci.nih.gov/datasetsNature2000.jsp*

References

Albert, A. and Anderson, J. A. (1984). "On the existence of maximum likelihood estimates in logistic regression models". *Biometrika*, 71:1–10.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". *Proc Natl Acad Sci U. S. A.*, 96:6745–6750.

Ambroise, Christophe and McLachlan, Geoffrey (2002). "Selection bias in gene extraction on the basis of microarray gene-expression data". *PNAS*, 99 (10):6562-6566.

Breiman, L. (2001). "Random Forests". *Machine Learning* 45(1):5–32.

Breiman, L. (2002). *Manual on setting up, using, and understanding Random Forests V3.1*. Dept. of Statistics, UC Berkeley.

Chen, Y., dougherty, E., and Bitterner, M. (1997). "Ratio-based decisions and the quantitative analysis of cDNA microarray images". *Journal of Biomedical Optics*, 2:364–374.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data". *JASA*, 97 (457):77-87.

Eilers, P.H., Boer, J.M., van Ommen, Gert-Jan and van Houwelingen, Hans C. (2001). "Classification of microarray data with penzlied logistic regression". In *Proceedings of SPIE, progress in biomedical optics and image*, Volumne 4266, pages 187-198.

Fahrmeir, L., Tutz, G., (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* Springer Series in Statistics.

Firth, D. (1992a). "Bias reduction, the jeffreys prior and glim". In Fahrmeir, L., Francis, B., Gilchrist, R., and Tutz, G., editors, *Advances in GLIM and Statistical Modelling*, pages 91–100. Springer-Verlag.

Firth, D. (1992b). "Generalized linear models and jeffreys priors: an iterative weighted least-squares approach". In Dodge, Y. and Whittaker, J., editors, *Computational statistics*, volume 1, pages 553–557. Physica-Verlag.

Firth, D. (1993). "Bias reduction of maximum likelihood estimates" (Corr: 95V82 p667). *Biometrika*, 80:27–38.

Fisher, R. (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics*, 7:179–88.

Frank, I. E. and Friedman, J. H. (1993). "A statistical view of some chemometrics regression tools" (Disc: p136-148). *Technometrics*, 35:109–135.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D. and. Schummer, M., and Haussler, D. (2000). "Support vector machine classification and validation of cancer tissue samples using microarray expression data". *Bioinformatics*, 16:906–914.

Garthwaite, P. H. (1994). "An interpretation of partial least squares". *Journal of the American Statistical Association*, 89:122–127.

Geladi, P. and Kowalski, b. (1986). "Partial least squares regression: A tutorial". *Analytica chimica Acta*, 185:1–17.

Fort, Gersende and Lambert-Lacroix, Sophie. (2003). "Classification using partial least squares with penalized logistic regression" *IAP-statistics* TR0331.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). "Gene selection for cancer classification using support vector machines". *Machine Learning*, 46:389–422.

Heinze, G. and Schemper, M. (2002). "A solution to the problem of separation in logistic regression". *Statistics in Medicine*, 21:2409–2419.

Helland, I. S. (1988). "On the structure of partial least squares regression". *Communications in Statistics, Part B – Simulation and Computation [Split from: @J(CommStat)]*, 17:581–607.

Höskuldsson, A. (1988). "PLS regression methods". *Journal of Chemometrics*, 2:211–228.

Kruskal, J. (1978). "Factor analysis and principal components i. bilinear methods". In *International Encyclopedia of Statistics*. Collier Macmillan Publishers.

Martens, H. and Naes, T. (1987). "Multivariate calibration by data compression". In Willaims, P. and Norris, K., editors, *Near-Infraread Technology for the Agricultural and Food Industries*. American Association of Cereal Chemists.

Martens, H. and Naes, T. (1989). *Multivariate calibration*. John Wiley & Sons.

Marx, B. D. (1996). "Iteratively reweighted partial least squares estimation for generalized linear regression". *Technometrics*, 38:374–381.

Massy, W. F. (1965). "Principal components regression in exploratory statistical research". *Journal of the American Statistical Associaton*, 60:234–246.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, 2 edition.

Newton, M., Kendziorski, C., Richmond, C., Blatterner, F., and Tsui, K. (2001). "On differentialvariability of expression ratios: Improving statistical inference about gene expression changes from microarray data". *Journal of Computational Biology*, 8:37–52.

Nguyen, D. V. and Rocke, D. M. (2002a). "Multi-class cancer classification via partial least squares with gene expression profiles". *Bioinformatics*, 18:1216–1226.

Nguyen, D. V. and Rocke, D. M. (2002b). "Tumor classification by partial least squares using microarray gene expression data". *Bioinformatics*, 18:39–50.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Ross, D., Scherf, U., Eisen, M., Perou, C., Rees, C.and Spellman, P., Iyer, V., Jeffrey, S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D.,

Shalon, D.and Myers, T., Weinstein, J., Botstein, D., and Brown, P. (2000). "Systematic variation in gene expression patterns in human cancer cell lines". *Nat Genet*, 24:227–35.

Santner, T. J. and Duffy, D. E. (1986). "A note on a. albert and j. a. anderson's conditions for the existence of maximum likelihood estimates in logistic regression models". *Biometrika*, 73:755–758.

Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000). "A gene expression database for the molecular pharmacology of cancer nature genetics". *Nature Genetics*, volume 24:236–244.

Smith, J. W., Everhart, J. E., Dickson, W. C., and Knowler, W. C.and Johannes, R. S. (1988). "Using the adap learning algorithm to forecast the onset of diabetes mellitus". In Greenes, R. A., editor, *Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988)*, pages 261–265. IEEE Computer Society Press.

Venables, W. and Ripley, B. (2002). *Modern applied statistics with S*. Springer, 4 edition.

Wang, C.-Y., Chen, C.-T., Chiang, C.-P., Young, S.-T., Chow, S.-N., and Chiang, H. K. (1999). "A probability-based multivariate statistical algorithm for autofluoresence spectroscopic identification of oral carcinogenesis". *Photochemistry and Photobiology*, 69(4):471–477.

Wold, H. (1975). "Soft modeling by latent variables: the nonlinear iterative partial least squares approach". In Gani, J., editor, *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, London. Academic Press.