

SUPPLEMENTAL APPENDIX

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

SUPPLEMENTAL APPENDIX A

Estimate of $\boldsymbol{\beta}$ is obtained by maximum likelihood estimation (MLE). This appendix describes details of this MLE computation.

Likelihood Function for Polychotomous Regression

To obtain the likelihood function for N independent samples $(y_1, \mathbf{x}_1), \dots, (y_N, \mathbf{x}_N)$ under the polychotomous regression model we first define some notation. Let $c(\mathbf{x}_i; \boldsymbol{\beta}) = \log[1 + \sum_{t=1}^G \exp(g_t(\mathbf{x}_i))]$ and rewriting (6) we have $\pi(k|\mathbf{x}_i) = \exp[(g_k(\mathbf{x}_i) - c(\mathbf{x}_i; \boldsymbol{\beta}))]$. Thus,

$$\log\pi(k|\mathbf{x}_i) = g_k(\mathbf{x}_i) - c(\mathbf{x}_i; \boldsymbol{\beta}). \quad (1)$$

Also, for the i th observed response value y_i corresponding to explanatory values $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{ip})'$ (and $x_{i0} \equiv 1$) let $\mathbf{z}'_i = (z_{i0}, z_{i1}, \dots, z_{iG})$ be the row vector indicating whether y_i is in group $k \in \mathcal{O}$. That is $z_{ik} = I(y_i = k)$ where $I(A)$ is the indicator function for A . If \mathbf{Z} is the $N \times (G + 1)$ matrix consisting of rows \mathbf{z}'_i s then $\sum_{k=0}^G z_{ik} = 1$ (the row sums are one). Using the above notations, the likelihood for N independent samples (ignoring constants) is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^N [\pi(0|\mathbf{x}_i)^{z_{i0}} \pi(1|\mathbf{x}_i)^{z_{i1}} \dots \pi(G|\mathbf{x}_i)^{z_{iG}}]. \quad (2)$$

Hence, the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^N [z_{i0} \log\pi(0|\mathbf{x}_i) + z_{i1} \log\pi(1|\mathbf{x}_i) + \dots + z_{iG} \log\pi(G|\mathbf{x}_i)]. \quad (3)$$

Using (1) together with $\sum_{k=0}^G z_{ik} = 1$ for each i , the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^N [z_{i1}g_1(\mathbf{x}_i) + z_{i2}g_2(\mathbf{x}_i) + \dots + z_{iG}g_G(\mathbf{x}_i) - c(\mathbf{x}_i; \boldsymbol{\beta})]. \quad (4)$$

MLE for Polychotomous Regression Using Newton-Raphson

Estimation of $\boldsymbol{\beta}$ is obtained by maximum likelihood estimation (MLE). Iterative methods such as the Newton-Raphson method can be used to obtain the MLE $\hat{\boldsymbol{\beta}}$. This requires first and second order derivatives of $l(\boldsymbol{\beta})$. For convenience let $\pi_{ik} = \pi(k|\mathbf{x}_i; \boldsymbol{\beta})$. It is straight forward to obtain

$$\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_k} = \pi_{ik}(1 - \pi_{ik})\mathbf{x}_i \quad k = 1, \dots, G \quad (5)$$

$$\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_l} = -\pi_{ik}\pi_{il}\mathbf{x}_i \quad k = 0, 1, \dots, G. \quad (6)$$

Thus the derivative of $l(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_k$ is

$$\begin{aligned} S(\boldsymbol{\beta}_k) &= \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k} = \sum_{i=1}^N \left[z_{ik} \mathbf{x}_i + \frac{\partial}{\partial \boldsymbol{\beta}_k} c(\mathbf{x}_i; \boldsymbol{\beta}) \right] \\ &= \sum_{i=1}^N \mathbf{x}_i (z_{ik} - \pi_{ik}) \quad k = 1, \dots, G \end{aligned}$$

since $-c(\mathbf{x}_i; \boldsymbol{\beta}) = \log \pi_{i0}$ and $\partial \log \pi_{i0} / \partial \boldsymbol{\beta}_k = -\pi_{ik} \mathbf{x}_i$. The *score* vector is

$$S(\boldsymbol{\beta}) = \begin{bmatrix} S(\boldsymbol{\beta}_1) \\ \vdots \\ S(\boldsymbol{\beta}_G) \end{bmatrix}_{G(p+1) \times 1}. \quad (7)$$

The $G(p+1)$ squared *information* matrix $I(\boldsymbol{\beta}) = -E[\partial S(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}']$ requires second order derivatives of $l(\boldsymbol{\beta})$ and are given below

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_l'} = - \sum_{i=1}^N \mathbf{x}_i \left(\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_l} \right)' = \sum_{i=1}^N \pi_{ik} \pi_{il} \mathbf{x}_i \mathbf{x}_i' \quad (8)$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_k'} = - \sum_{i=1}^N \mathbf{x}_i \left(\frac{\partial \pi_{ik}}{\partial \boldsymbol{\beta}_k} \right)' = - \sum_{i=1}^N \pi_{ik} (1 - \pi_{ik}) \mathbf{x}_i \mathbf{x}_i'. \quad (9)$$

The asymptotic covariance matrix of the MLE of $\boldsymbol{\beta}$ is the inverse of $I(\boldsymbol{\beta})$. For iterative computation of the MLE using the Newton-Raphson method is it more concise to express $I(\boldsymbol{\beta})$ as follows. Define the following $N \times N$ diagonal matrices,

$$\begin{aligned} \mathbf{W}_{kk} &= \text{diag}\{\pi_{1k}(1 - \pi_{1k}), \dots, \pi_{Nk}(1 - \pi_{Nk})\}, \quad k = 1, \dots, G \\ \mathbf{W}_{kl} &= \text{diag}\{\pi_{1l}\pi_{1k}, \dots, \pi_{Nl}\pi_{Nk}\}, \quad l \neq k \end{aligned}$$

and letting $I_{kk}(\boldsymbol{\beta}) = \mathbf{X}' \mathbf{W}_{kk} \mathbf{X}$ and $I_{kl}(\boldsymbol{\beta}) = I_{lk}(\boldsymbol{\beta}) = -\mathbf{X}' \mathbf{W}_{kl} \mathbf{X}$, the information matrix can be express as

$$I(\boldsymbol{\beta}) = \begin{bmatrix} I_{11}(\boldsymbol{\beta}) & I_{12}(\boldsymbol{\beta}) & \cdots & I_{1G}(\boldsymbol{\beta}) \\ I_{21}(\boldsymbol{\beta}) & I_{22}(\boldsymbol{\beta}) & \cdots & I_{2G}(\boldsymbol{\beta}) \\ \vdots & \vdots & & \vdots \\ I_{G1}(\boldsymbol{\beta}) & I_{G2}(\boldsymbol{\beta}) & \cdots & I_{GG}(\boldsymbol{\beta}) \end{bmatrix}_{G(p+1) \times G(p+1)}. \quad (10)$$

For an initial value $\boldsymbol{\beta}^{(0)}$, the MLE of $\boldsymbol{\beta}$ is obtained iteratively through $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I^{-1}(\boldsymbol{\beta}^{(t)}) S(\boldsymbol{\beta}^{(t)})$. If the Newton-Raphson algorithm converges, then the vector of coefficients at convergence is denoted $\hat{\boldsymbol{\beta}}$ and it is the MLE of $\boldsymbol{\beta}$.

One disadvantage of using PD is when there is quasi-complete separation in the data (next Section). Detection of quasi-complete separation is numerically burdensome and classification is usually poor. Also, inversion problems can be encountered in the Newton-Raphson algorithm when searching for the MLE.

Existence of MLE for Polychotomous Regression Model

We briefly describe the conditions for existence of the MLE of β in the polychotomous regression model. The reader is referred to Albert and Anderson (1984) for details. Possible data configurations can be categorized into three mutually exclusive and exhaustive groups: (1) complete separation, (2) quasicomplete separation, and (3) overlap. The first two situations lead to parameter estimates often referred to as “infinite parameters.” Specifically for (1) there exists a vector β which correctly classify all observations to their class, i.e.

$$(\beta_k - \beta_j)' \mathbf{x}_i > 0 \quad k, j = 0, \dots, G (k \neq j)$$

for all $i \in C_k$, where C_k ($k = 0, \dots, G$) is an index set identifying all samples in class k . Here, the MLE does not exist and the $-2\log$ -likelihood decreases to zero. Empirical detection of complete separation is to stop iteration when the probability of correct classification is 1 for all samples. Many model fits with MPLS components reported here are of this type. Quasicomplete separation is when there is a vector β such that

$$(\beta_k - \beta_j)' \mathbf{x}_i \geq 0 \quad k, j = 0, \dots, G (k \neq j)$$

for all $i \in C_k$ and the equality holds for at least one (i, k, j) (one sample in each class). Again, the MLE does not exist for this data configuration. Empirical detection is based on monitoring the probability of correct classification approaching one and the dispersion matrix, which is unbounded. This was encountered often with PCs. For the third case, overlap, the MLE exist and is unique.

SUPPLEMENTAL APPENDIX B

PCA and univariate PLS

In PCA the goal is to extract gene components sequentially which maximize the total predictor (gene) variability, irrespective of how well the constructed gene components predict cancer classes. There is no a priori reason why gene components with high total gene (predictor) variability should predict cancer classes well. Formally, in PCA orthogonal linear combinations are constructed to maximize the variance of the linear combination of the gene expression values sequentially,

$$\mathbf{v}_k = \operatorname{argmax}_{\mathbf{v}'\mathbf{v}=1} \operatorname{var}(\mathbf{X}\mathbf{v}) \quad (11)$$

subject to the orthogonality constraint $\mathbf{v}_k' \mathbf{S} \mathbf{v}_j = 0$, for all $1 \leq j < k$ where $\mathbf{S} = \mathbf{X}'\mathbf{X}$, and \mathbf{X} is the $N \times p$ matrix of gene expression values. Note that the extracted PCs do not depend on the response vector \mathbf{y} , indicating cancer classes $0, 1, \dots, G$.

In contrast to PCA, PLS (orthogonal) components are constructed to maximize the sample covariance between the response values (\mathbf{y}) and the linear combination of the predictor or gene expression values (\mathbf{X}). That is, the components are constructed to maximize the objective criterion based on the sample covariance between \mathbf{y} and $\mathbf{X}\mathbf{w}$. Thus, we find the unit weight vector \mathbf{w} satisfying the following objective criterion,

$$\mathbf{w}_k = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} \operatorname{cov}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \quad (12)$$

subject to the orthogonality constraint $\mathbf{w}'_k \mathbf{S} \mathbf{w}_j = 0$ for all $1 \leq j < k$. For details on classical PLS see Höskuldsson (1988). An introduction to PCA can be found in Johnson and Wichern (1992). Detailed references on PLS and PCA can also be found in Nguyen et al. (2002).

SUPPLEMENTAL APPENDIX C

Other Analytical Issues: Finer Gene Selection and Normalization

Although not addressed in this work, issues of normalization have been recognized to be important in the analyses of microarray gene expression data. Since our focus is not on comparing normalization methods, we have used the original authors' normalization in all data sets considered. The reader is referred to the original references for details. A comprehensive search of factors affecting image results, quality and detection will likely result in more justifiable normalization methods. (This is a current research area which will be described elsewhere.)

Another important issue in the analysis of microarray gene expression is the selection of which gene expression, amongst the mass of mRNA expression data, to use. We used a rough preliminary gene screening procedure. For example, for the breast cancer data, 11 genes have all 3 pairwise absolute mean differences while 416 genes have at least two pairwise absolute mean differences. Various compromises between using the (small) 11 genes set and the (large) 416 genes set can be used. For example, a finer gene selection procedure can be as follows. Note that for the breast cancer data, 405 ($416 - 11$) genes have only 2 (out of 3 possible) pairwise differences. Among the 405 genes one can, for instance, include genes with significant differences between tumors with *BRAC1* and *BRCA2*, ignoring the sporadic tumors. Another strategy, perhaps more sensible, is to rank the genes by the magnitude of the pairwise differences. Then select $q\%$ (say 10%) of the genes with highest absolute difference. Depending on q , this set of genes combined with the 11 genes would give a compromise between the 11 and 416 genes set. Note that the choice of q is arbitrary. However, the selection of genes for prediction with respect to PLS can be naturally based on the weights given in equations (2) and (3), which were not used in this work. This is due to the complexity of the weights. The usefulness of this approach is currently under research.

SUPPLEMENTAL APPENDIX D

Comparisons to DQDA, DLDA, and Classification Tree

Comparing classification performances from various methods is often of interest. Although the emphasis in this study is more on the dimension reduction aspects we also compared classification using DQDA (diagonal quadratic discriminant analysis) and DLDA (diagonal linear discriminant analysis). In studying tumor classification using gene expression data Dudoit et al. (2000) found that DQDA and DLDA produced "impressively low misclassification rates" relative to other methods considered. Using the same gene components constructed via MPLS or PCA we compared the results to those from DQDA and DLDA. Note that

DQDA and DLDA are special cases of QDA with $\Sigma_g = \text{diag}\{\sigma_{g1}^2, \dots, \sigma_{gp}^2\}$ for $g = 0, 1, \dots, G$ and $\Sigma_g = \text{diag}\{\sigma_1^2, \dots, \sigma_p^2\}$ (not depending on cancer class g) respectively.

Classification results from DQDA and DLDA using algorithms A0, A1, and A2 are summarized in the right side of Tables 1-4 (A0, A1, and A2). Dudoit et al. (2000) noted that these simple methods performed very well and we find that this is true using gene expression components constructed via MPLS and PCA. Generally, the results are similar to those reported in earlier sections. The simpler classifier, DQDA-PCA, performed poorly, and is analogous to the poor performance of QDA-PCA noted earlier. The “simplest” classifier DLDA, particularly DLDA-MPLS, performed very well.

Recently, Zhang et al. (2001) demonstrated the use of the recursive partitioning (classification tree) method to classify 22 normal and 40 colon tumor tissues based on gene expression profiles. As pointed out by Zhang et al. (2001), recursive partitioning is applicable to multi-class response as well. The method appears promising and there are some appealing aspects to the method. For example, it is applicable to multi-class response, it selects the genes and performs prediction simultaneously, and it can handle a large number of genes. The reader is referred to Zhang et al. (2001) for details and further references. We used the publicly available software for recursive partitioning, RTREE (<http://peace.med.yale.edu/>), to perform multi-class classification on the four data sets. The overall leave-one-out cross-validated classification error for the leukemia, lymphoma, NCI, and BRCA data sets were 15.2%, 10.84%, 22.86% and 90.9%.

SUPPLEMENTAL APPENDIX E

Selecting the Number of Gene Components

We have chosen $K = 3$ gene components to fit classifiers. In this section we give some explanations for this choice. The choice of the number of gene components, K , to use for prediction can be chosen by cross-validation to minimize the predicted residual sum of squares (PRESS). However, this approach is not satisfactory because, often, the model which minimizes PRESS compared with a much simpler model (smaller K) will have only slight difference in PRESS scores. To compare across classification methods we fixed a $K = 3$ which is a reasonable balance between good prediction results and simplicity of models. Almost always $K = 1$, and often $K = 2$, is not satisfactory in terms of good prediction. However, often increasing K above 5 does not drastically improve classification results. $K = 3, 4$, or 5 often provide similar results, so the simpler choice of $K = 3$ is more desirable. For example, this can be seen by plotting the number of misclassifications versus K , $K = 1$ to 7 (see Supplemental Figure 9). The choice of $K = 3$ appear sufficient for PLS dimension reduction of the data sets at hand. Unless computational cost is very high, a study to select K , as given in Supplemental Figure 9, is useful. A more detailed study of the choice of K in PLS for prediction in the context of gene expression data will be given elsewhere (Nguyen et al. 2002b).

SUPPLEMENTAL APPENDIX F

Evaluation of Classification Results: Randomization Studies

The reliability of classification results is an important issue and we further address this issue in relation to the small sample size associated with microarray data, especially in cancer microarray data. As can be seen from the previous section, for the BRCA and NCI60 data, both with small sample sizes, the variation of the observed classification error rates can be large.

As a minimum, we checked to see whether the observed classification error is lower than classification on “random” data. That is, randomly assign the cancer labels (cancer group $0, 1, \dots$ or G) to each gene expression profile (sample) to generate a “new” permuted data set, say \mathbf{X}^* . We randomly generated $B = 100$ permuted data sets, $\mathbf{X}_{(1)}^*, \dots, \mathbf{X}_{(B)}^*$ and obtained corresponding classification error rates e_1, \dots, e_B using both algorithm A1 and A2. The observed classification error rate from the original (real) data set, say e_{obs} (given in Tables 1A1-4A2 and 1A2-4A2), can be compared to the distribution of error rates obtained from randomization (the e_i s). This was carried out on all four data sets for every subset of genes and methods combination used.

The results of the randomization study (Supplemental Figures 2-8) suggest that, in all cases, the observed classification error rate is significantly less than would be expected under randomization. We note that the simulation studies only suggest that the observed classification rates reported in Tables 1-4 are much lower than would be expected at random, which is obvious. As with any other analytical method, further validation based on new real data will shed more light on its usefulness.

REFERENCES

- Albert, A. and Anderson, J. A. (1984), “On the Existence of Maximum Likelihood Estimates in Logistic Models,” *Biometrika*, **71**, 1-10.
- Dudoit, S., Fridlyand, J. and Speed, T.P. (2000), “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data,” Technical Report # 576, Department of Statistics, University of California, Berkeley.
- Nguyen, D.V. and Rocke, D.M. (2002), “Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data,” *Bioinformatics*, **18**, 39-50.
- Nguyen, D.V., Wang, N., and Carroll, R.J. (2002b), “Missing Value Estimation for Cancer Microarray Gene Expression Data,” manuscript.
- Zhang, H.P., Yu, C., Singer, B. and Xiong, M. (2001), “Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data,” *Proceedings of the National Academy of Sciences, USA*, **98**, 6730-6735.