Biostatistics (2010), **0**, 0, *pp.* 1–14
doi:10.1093/biostatistics/kxq069

Probabilistic classifiers with high-dimensional data

KYUNG IN KIM, RICHARD SIMON*

Biometric Research Branch, National Cancer Institute,
9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434, USA
rsimon@mail.nih.gov

SUMMARY

For medical classification problems, it is often desirable to have a probability associated with each class. Probabilistic classifiers have received relatively little attention for small *n* large *p* classification problems despite of their importance in medical decision making. In this paper, we introduce 2 criteria for assessment of probabilistic classifiers: well-calibratedness and refinement and develop corresponding evaluation measures. We evaluated several published high-dimensional probabilistic classifiers and developed 2 extensions of the Bayesian compound covariate classifier. Based on simulation studies and analysis of gene expression microarray data, we found that proper probabilistic classification is more difficult than deterministic classification. It is important to ensure that a probabilistic classifier is well calibrated or at least not "anticonservative" using the methods developed here. We provide this evaluation for several probabilistic classifiers and also evaluate their refinement as a function of sample size under weak and strong signal conditions. We also present a cross-validation method for evaluating the calibration and refinement of any probabilistic classifier on any data set.

Keywords: Gene expression analysis; High-dimensional data; Microarray; Probabilistic classification.

1. INTRODUCTION

Stimulated by the development of gene expression microarray technology, many types of classifiers of tumor samples have been developed (Dudoit *and others*, 2002; Wessels *and others*, 2005; Lai *and others*, 2006). Classification rules are typically learned using thousands of variables with at most hundreds of samples. However, for medical decision making, it is valuable to have a probabilistic classifier and to know which cases are clearly one class or the other and which are less well determined.

Medical decision making is complex. Misclassification costs are often asymmetric, difficult to quantify, and vary among physicians and patients. In this context, probabilistic classifiers that provide an estimate of the probability of membership in each class for new cases are more useful than classification rules that simply assign cases to a class. Probabilistic classifiers provide tools for a diverse audience of users who may use the probabilities in conjunction with other information such as treatment options and patient preferences for making complex integrated clinical decisions.

Probabilistic classifiers have received relatively little attention in the literature of small *n* large *p* classification problems where the number of candidate variables exceeds the number of cases available for

*To whom correspondence should be addressed.

model development. In this paper, we introduce 2 criteria for assessment of high-dimensional probabilistic classifiers: well-calibratedness and refinement (DeGroot and Fienberg, 1983).

Well-calibratedness is widely used in probabilistic forecasting fields (Dawid, 1986). We say a probabilistic classifier is "well calibrated" if for any predictive probability $0 < w < 1$, the relative frequency of the event which the probabilistic classifier predicts with probability w is w. That is, if $S(w) = \{x: \hat{\Pr}(C = 1|x) = w\}$, then $\Pr(C = 1|x \in S(w)) = w$. For example, in weather forecasting, predictions of $100w\%$ chance of rain are well calibrated if they result in rain $100w\%$ of the time.

Second, we introduce a measure based on refinement of probabilistic classifiers. The refinement measures the extent to which the classifier makes useful forecasts, rather than predictions for which the probability of being in class 1 is close to 0.5. In general, we define refinement as the expected value of $\Pr(C = 1|x \in S(w))(1 - \Pr(C = 1|x \in S(w)))$ with respect to the predictive probability w.

We evaluate several published high-dimensional probabilistic classifiers for their well-calibratedness and refinement. We consider the following widely used supervised probabilistic classification methods: prediction analysis of microarrays (PAM, Tibshirani *and others*, 2002), Bayesian compound covariate (BCC, Wright *and others*, 2003), 2 modifications of BCC (BCCm and BCCi), and penalized logistic regression with L_1 and L_2 penalties (PLR L1 and PLR L2, Friedman *and others*, 2010). For the probabilistic classifiers which are not well calibrated, we define conservativeness and anticonservativeness. Roughly, we say a predictor is conservative if it tends to predict with probability nearer to 0.5 than the true value.

In Section 2, we introduce the 6 probabilistic classifiers to be evaluated and define evaluation measures. In Sections 3 and 4, we evaluate the probabilistic classifiers using simulations. In Section 5, we evaluate the classifiers with 2 real data sets using a leave-one-out-cross-validation (LOOCV) framework, and last in Section 6, we discuss our results.

2. METHOD

2.1 *High-dimensional probabilistic classifiers*

We introduce the 6 high-dimensional probabilistic classifiers, PAM, BCC, BCCm, BCCi, PLR L1, and PLR L2 for 2-group classification problem.

Let x_i denote a p-dimensional feature vector and c_i a 0-1 valued class variable for the ith case. For example, in microarray data analysis, one could think that x_i represents the ith sample of gene expression data and c_i represents the ith cancer type or response to treatment. Given the training data set $\{(x_i, c_i)\}_{i=1,\ldots,n}$, we are interested in accurately predicting $\Pr(C = 1|x^*)$ for a new case with known feature vector x^*.

PAM probabilistic classifier. PAM employs a shrinkage technique called nearest shrunken centroids (Tibshirani *and others*, 2002). It should not be confused with the unsupervised "partition around medoids" method. The mean vector of each class is estimated by shrinking it toward the overall mean vector. The degree of shrinkage is determined by a tuning parameter that is usually optimized to minimize the cross-validated classification error estimate. In our use of PAM, we optimize the tuning parameter to maximize the cross-validated predictive likelihood.

Although PAM is usually used as a deterministic classifier, Tibshirani *and others* (2002) suggest computing a predicted probability of class 1 for a new case x^* as

$$\hat{p}(x^*) = \frac{\pi_1 \hat{\Pr}(x^*|C = 1)}{\pi_0 \hat{\Pr}(x^*|C = 0) + \pi_1 \hat{\Pr}(x^*|C = 1)}, \tag{2.1}$$

where π_0 and π_1 are the class prior probabilities, and the conditional probability of x^* in each class is approximated by a multivariate normal density. The shrunken class-specific sample means are used as the

mean vectors. The intra-class covariance matrix is taken as diagonal with the intra-class variance of the ith gene is estimated by $(s_i + s_0)^2$, where s_i is the pooled sample standard deviation of the ith gene and s_0 is a positive constant used for stabilization of variance estimates.

Bayesian composite probabilistic classifiers. The BCC method uses a dimension reduction technique for high-dimensional gene expression vectors in 2 group classification (Wright *and others*, 2003). BCC projects a p-dimensional gene expression vector x to a one-dimensional compound covariate $\mathbf{w}x$. The weights are based on the t-statistics $\mathbf{t}$ for comparing gene expression between the 2 classes. The weight equals the t-statistic if the gene is selected in the feature selection step and is zero otherwise. The feature selection step selects a fixed number of genes with t-statistics largest in absolute value, a number that can be optimized as a tuning parameter by cross-validation. Let us denote these weights by $\tilde{\mathbf{t}}$. Using Bayes theorem and normal approximation of the compound covariate $\tilde{\mathbf{t}}x^*$, the predictive probability of class 1 for a new case x^* is computed as

$$\hat{p}(x^*) = \frac{\pi_1 \phi(\tilde{\mathbf{t}}x^* | \hat{\mu}_1, \hat{\sigma}_1^2)}{\pi_0 \phi(\tilde{\mathbf{t}}x^* | \hat{\mu}_0, \hat{\sigma}_0^2) + \pi_1 \phi(\tilde{\mathbf{t}}x^* | \hat{\mu}_1, \hat{\sigma}_1^2)}, \tag{2.2}$$

where ϕ is the normal density function. The mean and variance of the compound covariates in class $k = 0, 1$ are estimated by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{c_i = k} \tilde{\mathbf{t}}x_i, \quad \hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{c_i = k} (\tilde{\mathbf{t}}x_i - \hat{\mu}_k)^2, \tag{2.3}$$

where n_k is the number of samples in the class.

Although this dimension reduction technique of BCC has been effectively applied to the classification of cancer, we have developed 2 modifications of the BCC. First, the class-specific means and variances of the compound covariate are biased because the same data are used for selecting the genes used in the dimension reduction and for estimating the means and variances. This results in predicted probabilities too close to 0 or 1. Wright *and others* (2003) noted this problem. They indicated that estimating prediction accuracy requires an independent test set, but provided no proposal for avoiding the overfitting. We propose replacing the $\tilde{\mathbf{t}}x_i$ values in the summations of (2.3) by $\tilde{\mathbf{t}}_{(-i)}x_i$ values, where $\tilde{\mathbf{t}}_{(-i)}$ is computed after omitting the ith case. The class-specific means and variances are computed based on projected values $\tilde{\mathbf{t}}_{(-i)}x_i$. We use the same formula as (2.2) for computing the predictive probability of class 1 for a new case x^*, but for the estimates of mean and variance of the compound covariates, we use

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{c_i = k} \tilde{\mathbf{t}}_{(-i)}x_i, \quad \hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{c_i = k} (\tilde{\mathbf{t}}_{(-i)}x_i - \hat{\mu}_k)^2. \tag{2.4}$$

We refer to this version of BCC as BCCm in the following sections.

Our second modification of BCC corrects for replacing unknown densities in application of Bayes theorem by densities with plug-in estimators of means and variances. The Bayes theorem–based probability that a new case is in class 1 should depend on the ratio of density of observing the $\tilde{\mathbf{t}}x^*$ value if the case is class 1 to the density for class 0, that is, the Bayes factor. It is not proper in a Bayesian sense to just estimate the unknown densities and ignore the uncertainty of the estimates. The proper Bayesian approach would be to compute the unknown densities as marginal densities with regard to the posterior distributions of the unknown means and variances. So the predictive posterior probability of $\tilde{\mathbf{t}}x^*$ for class k is computed as

$$\Pr(\tilde{\mathbf{t}}x^* | D_k) = \iint \phi(\tilde{\mathbf{t}}x^* | \mu_k, \sigma_k^2, D_k)\, p(\mu_k, \sigma_k^2 | D_k) \mathrm{d}\mu_k \, \mathrm{d}\sigma_k^2, \tag{2.5}$$

where $D_k = \{\tilde{\mathbf{t}}_{(-i)}x_i : c_i = k\}$ and $k = 0, 1$.

For small sample sizes, the estimators of class-specific means and variances may be very poorly determined yet are treated as known with certainty in using plug-in estimators to Bayes formula. In our modification, we have replaced the plug-in estimators by marginal distributions based on using noninformative priors for μ_k and σ_k^2. Following the derivation of Gelman *and others* (2003, Chapter 3, p. 77), the right-hand side of (2.5) has a t distribution of degree of freedom $n_k - 1$ with location parameter $\hat{\mu}_k$ and scale parameter $(1 + n_k^{-1})\hat{\sigma}_k^2$. So, the t probability density function for each class is used to compute the predictive probability instead of the normal density function in (2.2). We refer to this version of BCC as BCCi in the following sections. Note since BCCi uses $\tilde{\mathbf{t}}_{(-i)}x_i$ to estimate $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ instead of $\tilde{\mathbf{t}}x_i$, this modification includes the modification of BCCm.

Penalized logistic regression classifiers. Penalized logistic regression models use shrinkage techniques in estimating the coefficients of high-dimensional covariates. Penalties are imposed in the estimation step either on the sum of absolute values of coefficients (PLR L1) or on the sum of squared coefficients (PLR L2). Unlike PAM and BCC type classifiers, penalized logistic regression models predict class probability for a new case x^* as

$$\hat{p}(x^*) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x^*)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x^*)}, \tag{2.6}$$

where $\beta_1 = (\beta_{11}, \ldots, \beta_{1p})$. The regression coefficients are determined to maximize a penalized likelihood function. Either an L_1 or an L_2 penalty can be imposed. For example, with an L_1 penalty, the penalized log-likelihood function is

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} \{c_i \log p(x_i) + (1 - c_i) \log(1 - p(x_i))\} - \lambda \sum_{j=1}^{p} |\beta_{1j}|. \tag{2.7}$$

The penalty causes shrinkage of the regression coefficients. The degree of shrinkage can be determined by cross-validation to maximize the predictive likelihood function.

2.2 *Well-calibratedness, refinement, and evaluation measures*

We say that a probabilistic classifier is well calibrated if for any $0 < w < 1$,

$$\Pr(C = 1 \mid \hat{p}(x) = w) = w. \tag{2.8}$$

$\Pr(C = 1 \mid \hat{p}(x) = w)$ is an expectation over the set of x for which predictions $\hat{p}(x)$ of class 1 are w. We can interpret (2.8) as the proportion of the cases that the probabilistic classifier predicts with probability w that are actually class 1. Thus, if we collect cases for which the predicted probability of being class 1 equals w, then a proportion w of them can be expected to be in class 1. It does not mean that each of those cases has a true probability of w of being in class 1. One can easily show that the Bayesian probabilistic classifier obtained by the direct application of Bayes theorem when the class-specific densities are known is well calibrated. However, the Bayesian probabilistic classifier cannot be computed for real data where the class-specific densities are unknown. The probabilistic classifiers that have been introduced for high-dimensional problems have not been previously evaluated as to whether they are well calibrated.

The calibration curve, a smoothed curve of the predictive probability of class 1 versus its actual frequency is often drawn to see the degree of calibratedness of a probabilistic classifier. The true calibration curve defined in (2.8) of a well-calibrated classifier is a diagonal straight line by (2.8). When a probabilistic classifier is not well calibrated, it may be "conservative" or "anticonservative". When a class is

predicted with a probability greater than 0.5 ($w > 0.5$) and $\Pr(C = 1 \mid \hat{p}(x) = w) > w$, we say that the prediction is conservative in the sense that the predicted events (or class labels) happened more than predicted. When the frequency is smaller than the predicted probability, we say that the prediction is anticonservative.

We say that a probabilistic classifier is refined if the predictive probabilities tend to be close to 0 or 1 so that the variance of the frequency of the class labels is small. It is easily shown from the definition of the refinement in Section 1.

To assess the calibration and refinement quantitatively, we use the decomposition of average squared error between the class labels and predictive probabilities of class 1 (DeGroot and Fienberg, 1983).

For a given data set, (x_i, c_i) for $i = 1, \ldots, n$, the average squared error for a probabilistic classifier $\hat{p}$ is defined as $n^{-1} \sum_{i=1}^{n} (c_i - \hat{p}(x_i))^2$, where x_i is the ith feature vector and c_i is the corresponding binary class label and $\hat{p}(x_i)$ is the predictive probability of class 1 for x_i. Even for the true Bayesian probabilistic classifier, the average squared error is not zero unless the class-specific densities are widely separated.

Since for any finite set of predictions, there may be only a single prediction for any w value, in order to evaluate whether a classifier is well calibrated some binning or smoothing of predictions is necessary. Suppose we partition the unit interval into m equal subintervals or bins $B_k = ((k-1)/m, k/m]$ for $k = 1, \ldots, m$. For each bin B_k, we compute q_k, the proportion of predictions that fall into B_k, r_k the relative frequency of predictions in B_k for class 1, and u_k the center point of B_k. Then the average squared error is decomposed as sum of calibration score (CS) and refinement score (RS):

$$\mathrm{CS} = \sum_{k=1}^{m} (r_k - u_k)^2 q_k, \quad \mathrm{RS} = \sum_{k=1}^{m} r_k (1 - r_k) q_k. \tag{2.9}$$

The decomposition is similar to the bias–variance decomposition of the mean squared error.

The CS represents an average squared discrepancy between predictions and corresponding relative frequencies. The CS of a well-calibrated classifier is zero by definition. The Bayesian probabilistic classifier is also well calibrated. The RS represents the conditional variance of relative frequencies averaged over prediction bins. The well-calibrated classifiers do not necessarily have the smallest RS, but the Bayesian probabilistic classifier has the minimum RS among all classifiers (see a proof in the supplementary materials available at *Biostatistics* online).

Although both well-calibratedness and refinement are useful criteria in evaluating probabilistic classifiers, well-calibratedness is more important because it indicates that the predictions have aggregate validity. Hence, when selecting good probabilistic classifier, we would recommend using the most refined among well-calibrated classifiers.

3. Evaluation using multivariate normal models with single correlations

We evaluated the calibration and refinement of the high-dimensional classifiers described in the previous section and explored the factors that affect these measures using simulation.

We generated our training and test data sets using multivariate normal gene expression with class-specific mean vectors and common covariance matrices. Let x denote a p-dimensional gene expression vector and c a 0-1 valued class variable. For the ith case, c_i is generated from the Bernoulli distribution with probability 0.5 and vector x_i in class k is generated from multivariate normal distribution with mean μ_k for $k = 0, 1$ and common correlation matrix Σ. The simulations were performed with $p = 1000$ genes, the first $p_1 = 50$ of which were differentially expressed between classes with mean difference 1. The remaining 950 genes had the same mean for each class.

For the common correlation matrix Σ, we used the following 3 structures:

$$\text{Structure.1}: \begin{bmatrix} \Sigma_{11} & 0 \\ 0' & I_{p-p_1} \end{bmatrix}, \quad \text{Structure.2}: \begin{bmatrix} \Sigma_{11} & 0 \\ 0' & \Sigma_{22} \end{bmatrix}, \quad \text{Structure.3}: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22}, \end{bmatrix}$$

where Σ_{11} and Σ_{22} are $p_1 \times p_1$ and $(p-p_1)\times(p-p_1)$ intra-class correlation matrices for p_1 informative genes and for $p-p_1$ noninformative genes, respectively. Σ_{12} is the $p_1 \times (p-p_1)$ correlation matrix between informative and noninformative genes and I_{p-p_1} is the $(p-p_1)\times(p-p_1)$ identity matrix. For each structure, we compared the classifiers by imposing common pairwise correlations of either 0.25, 0.50, or 0.75. The case of full independence was also studied. The number of training samples was either 30, 60, or 120. We generated 1000 simulated training samples for each combination of conditions.

PAM and the penalized logistic regression methods incorporate a tuning parameter that determines the amount of shrinkage. For the former, the class means of each variable are shrunken; for the latter, the regression coefficients are shrunken. For the BCC methods, the tuning parameter is the number of variables selected for inclusion in the linear projection. In classification problems, such tuning parameters are usually optimized to minimize a cross-validated estimate of prediction error. Since we are interested in probabilistic prediction, we maximized the cross-validated predictive likelihood. That is, the cross-validated predictive likelihood is computed for each of a grid of values of the tuning parameter and the value with the largest cross-validated predictive likelihood is selected. We used LOOCV (Molinaro *and others*, 2005). The predictive likelihood of the training data can be written as

$$L(\lambda) = \prod_{i=1}^{n} \hat{p}_{(-i),\lambda}(x_i)^{c_i}(1-\hat{p}_{(-i),\lambda}(x_i))^{1-c_i}, \tag{3.1}$$

where $\hat{p}_{(-i),\lambda}(x)$ is computed in a training set omitting the ith case and using tuning parameter λ.

Before we summarize our results for the 1000 simulations for each condition, we first examine some results of a single simulation. Figure 1 shows results for a single test set with training sets of size $n = 30$, 60, and 120 when correlations among informative genes are 0.25 ("Structure.1"). A probabilistic classifier of each type has been developed for each of the 3 training sets. A single true model was used to generate these 3 training sets, and we sampled 5000 test sample points (x_i, c_i) for this model. For each test sample x_i vector, we compute the true probability that a sample with that x_i vector is from class 1 by employing Bayes theorem using the true class-specific densities used to generate the data. This true model is the Bayesian probabilistic classifier and these true posterior probability values are used for the vertical axes in Figure 1. For each test sample x_i, we also compute the predicted probability of the case being from class 1 for each of the 6 predictive classifiers developed in the training set. These values are used for the horizontal axes.

Figure 1 shows that evaluating probabilistic classification is more complex than for deterministic classification. For example, 2 classifiers having similar misclassification error rate can produce completely different graphical display. In the figure, the misclassification rates of BCC when $n = 30$ and PLR L2 when $n = 60$ are both close to 0.19 but the corresponding scatter plots are quite different.

Figure 1 also shows that the rate of convergence to the true Bayesian classifier varies dramatically among classification methods. It is desirable that as the sample size increases, the predictions of a classifier converge to the predictions of the true Bayesian probabilistic classifier and thus lie along the 45-degree line. PAM and BCC type classifiers show clear convergence, while PLR L1 appears to converge slowly and PLR L2 does not converge to the Bayesian probabilistic classifier. The dependency condition for Figure 1 is "Structure.1" with common correlation 0.25 among informative genes. The penalized logistic models do not perform well in this rather weak dependency condition.

The calibration curves in (2.8) corresponding to Figure 1 are shown in Figure 2. The calibration curves are smoothed with local regression with degree 1 and span 0.75 using R software loess package

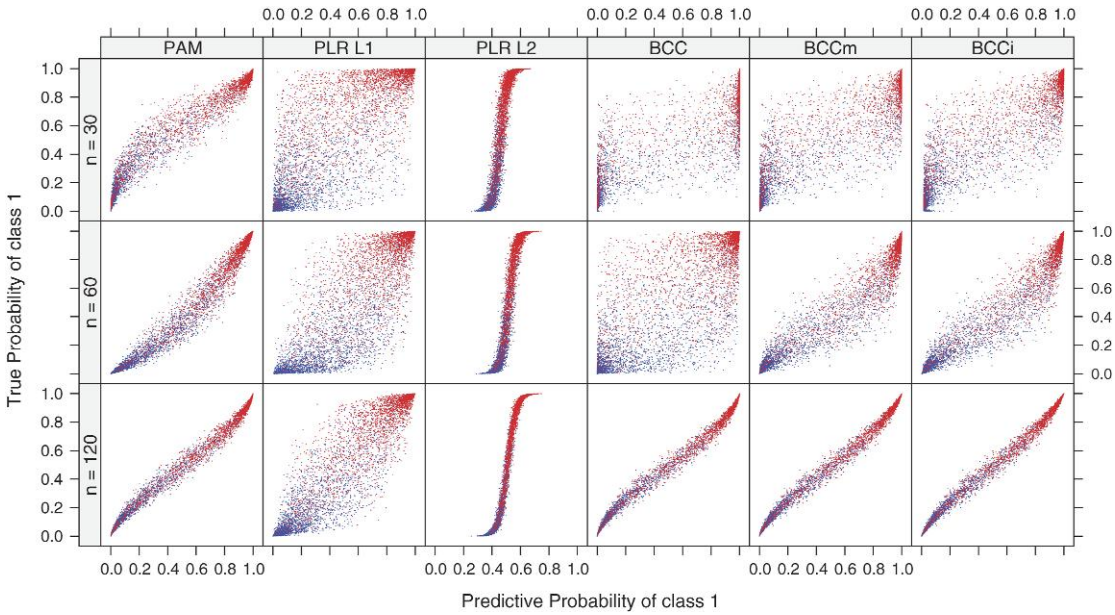

Fig. 1. A scatter plot for "Structure.1" condition. Rows represent 3 different training sample size ($n = 30, 60, 120$) and columns represent 6 probabilistic classifiers used to compute predictive probabilities. In each plot, x-axis represents predictive probability of class 1 and y-axis represents true probability of class 1. Each scatter consists of 5000 dots, where red ones are of class 1 and blue ones are of class 0. Equal prior class probability was assumed. Expression vectors are 1000-dimensional and normally distributed. Nonzero common pairwise correlation 0.25 was given only to informative 50 genes.

as in Venables and Ripley (2002) (Chapter 12, p. 350). CS and RS are based equal-sized bins of the unit interval as indicated in Section 2.2. At the right bottom of each figure, CS, RS and misclassification rates are reported.

In Figure 2, we see that evaluating probabilistic classifiers based only on misclassification rate can be misleading. When $n = 120$, PLR L2 has better misclassification rate 0.17 than the 0.20 for PLR L1. However, the probabilistic predictions of PLR L2 are hardly informative because most of them are centered around 0.5 regardless of true probabilities; PLR L2 is poorly calibrated and too conservative. In this case, PLR L1 is preferred.

Although the smoothed versions of many calibration curves in Figure 2 lie close to the diagonal line of the Bayesian probabilistic classifier, Figure 1 shows that their predictions are not necessarily in agreement with the true Bayesian probabilistic classifier in this high-dimensional setting. Approximation to the Bayesian probabilistic classifier is a more stringent requirement than well-calibratedness or refinement.

Figure 2 also shows conservativeness and anticonservativeness of probabilistic classifiers. The PLR L2 is conservative as mentioned above. On the other hand, BCC for $n = 30$ shows anticonservativeness. Its predictions tend to be closer to either 0 or 1 than the true class probabilities. This anticonservativeness is somewhat reduced in BCCi.

Based on the 1000 replications, we plotted average CS, RS and misclassification rates in Figure 3. For each of three structures and four correlations 0.0, 0.25, 0.5, 0.75 for each structure, we compared the six classifiers for training sample size $n = 30$. The table corresponding to Figure 3 is included in the Supplementary Materials.

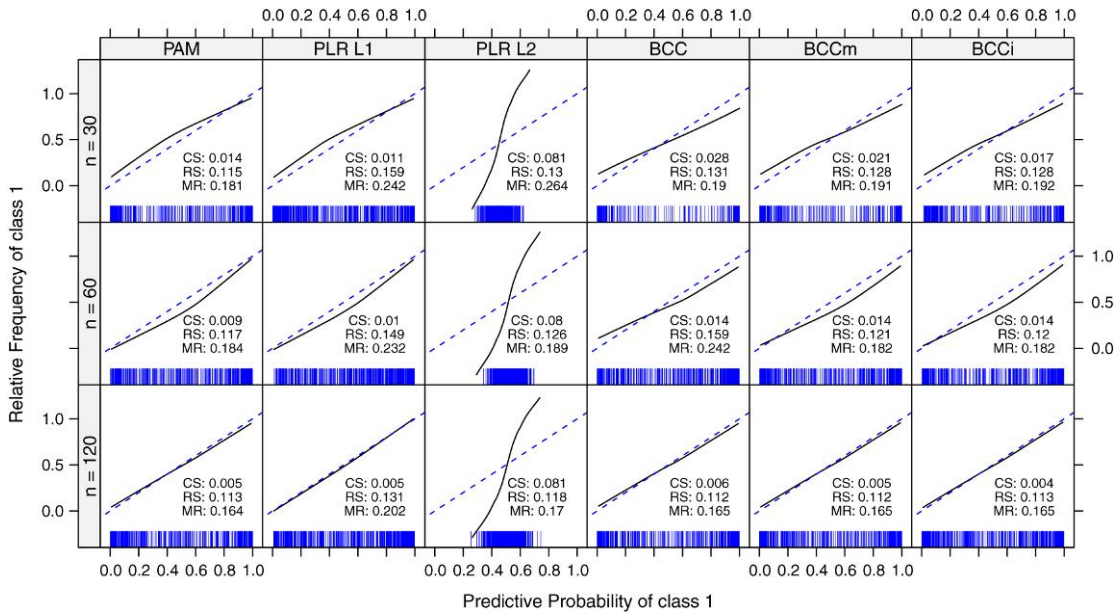

Fig. 2. A calibration curve plot for "Structure.1" condition. Rows represent 3 different training sample size ($n = 30, 60, 120$) and columns represent 6 probabilistic classifiers used to compute predictive probabilities. x-axis represents predictive probability of class 1 and y-axis represents the corresponding calibration curve, a smoothed version of actual relative frequency of class 1 for the 5000 testing samples of Figure 1. In each plot, the calibration curve is drawn by linear local regression using R loess package. Estimated CS, RS, misclassification rates (MR), and rug plot for prediction frequencies are added at the bottom of each plot.

Overall, PLR L2 appears poorly calibrated, whereas PAM, BCCm, and BCCi appear the best calibrated among the probabilistic classifiers evaluated. PAM appears most refined. PLR L1 and PLR L2 appear less refined than the others except on "Structure.3" with strong correlations where their RS are substantially better than the others. With regard to misclassification rates, BCC type classifiers and PAM are similar. The superiority of PLR L1 on "Structure.3" with strong correlation 0.75 condition is considerable.

Misclassification rates of PLR L1 are considerably lower than the other classifiers on "Structure.3". It seems that the stage-wise variable selection process (Efron *and others*, 2004) of PLR L1 enables the method to more efficiently exclude noise variables strongly correlated with informative genes than the other classifiers. However, it does not perform as well under other conditions with no correlations between informative and noninformative genes.

The original BCC generally has poorer CS and RS than BCCm and BCCi. PAM seems well calibrated and refined in most common correlation structures. However, the misclassification rate of PAM appears large for $n = 30$ when there are strong correlations among genes. The poor performance of PAM with strong correlations is not surprising because PAM assumes diagonal covariance structure. This misclassification rate decreases as the training sample size increases. Tables and Figures for $n = 60, 120$ are included in the supplementary material available at *Biostatistics* online.

4. EVALUATION USING FITTED MULTIVARIATE NORMAL MODELS FROM REAL DATA

We also evaluated the probabilistic classifiers in a simulation study based on the colon cancer data set of Dettling (2004). The simulated data were based on a multivariate normal model with class-specific

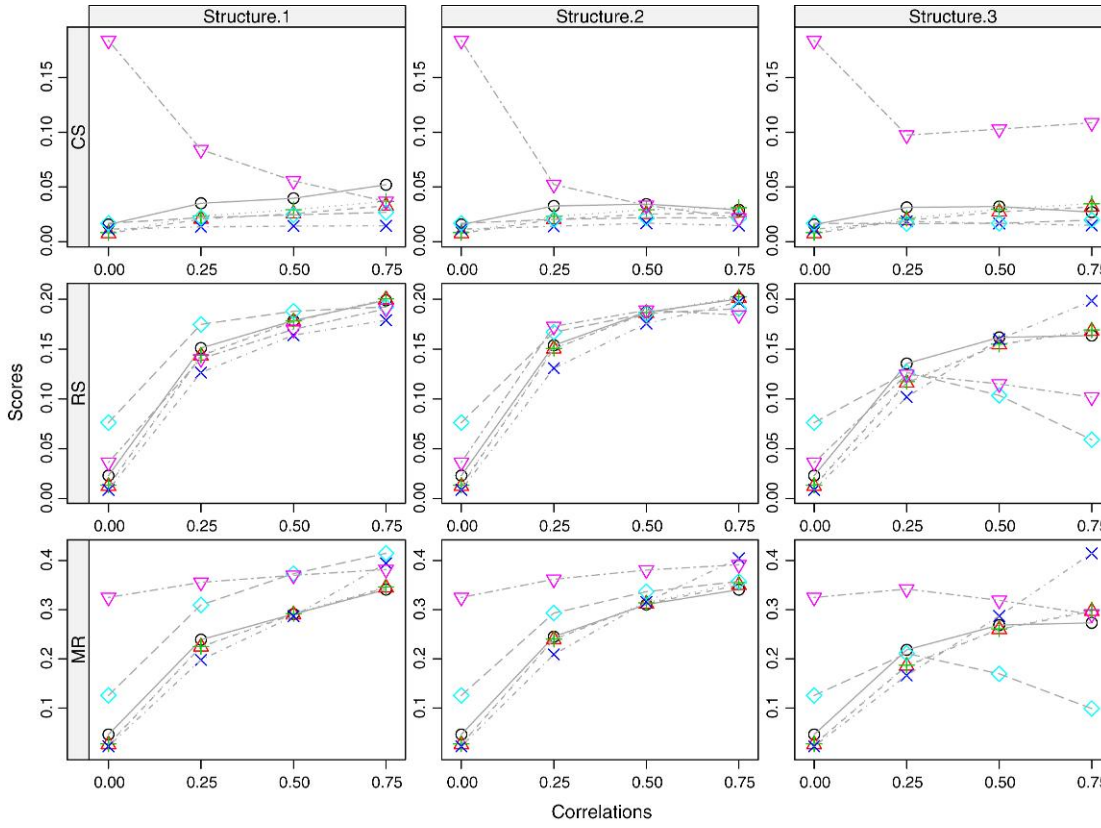

Fig. 3. Average CS, top row, RS, middle row, and MR, bottom row, over 1000 simulation replications for the 6 classifiers, PAM (×), PLR L1 (◇), PLR L2 (▽), BCC (◯), BCCm (+), and BCCi (△) with training sample size $n = 30$. Columns represent the 3 correlation structures: nonzero common pairwise correlation in the x-axis was applied only to informative genes ("Structure.1"), informative genes and noninformative genes ("Structure.2"), and all genes ("Structure.3") for each simulation.

mean vectors and common intra-class covariance matrix Σ. For the class-specific mean vectors, we first computed the sample mean vector for each class and shrunk the mean vectors by multiplying a shrinkage factor. A shrinkage factor of 1 means no shrinkage. For the common covariance matrix, we used robust estimations of Σ as described by Schäfer and Strimmer (2005) in which $\hat{\Sigma}$ is a convex combination of $(1 - \lambda)$ times a diagonal matrix and λ times the sample covariance matrix. $\lambda = 0$ indicates independence among genes. The colon cancer data set consist of 2000 genes and 62 samples. For the simulation, we generated 1000 replications using the fitted multivariate normal model with training sample size 100 and equal class prevalence.

Figure 4 shows results for a range of shrinkage factors and λ values. A small shrinkage factor yields small inter-class differences in gene expression. A large λ maximizes inter-gene correlations. The CS of PLR L2 in Figure 4 are much worse than with other classifiers except in the nearly null setting with a shrinkage factor of $1/4$.

The RS and misclassification rates of PLR L1 are smaller than those of other classifiers for strong dependency conditions ($\lambda > 0.5$). For weak dependency conditions ($\lambda < 0.5$), however, PLR L1 has larger scores than the other classifiers. As was seen for the simulation in Section 3, PLR L1 appears well

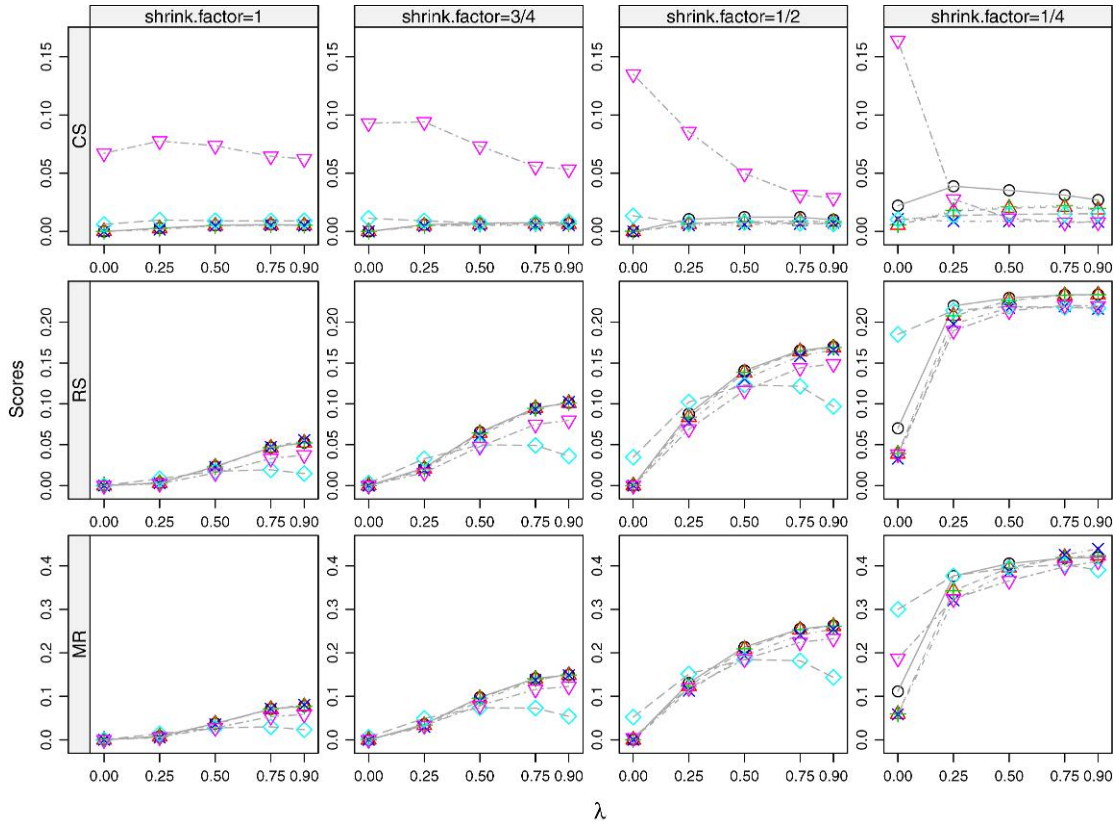

Fig. 4. Average CS, top row, RS, middle row, and MR, bottom row, with 1000 simulation replications for the 6 classifiers, PAM (×), PLR L1 (◇), PLR L2 (▽), BCC (◯), BCCm (+), and BCCi (△). The colon cancer data set in Dettling (2004) was used to fit class-specific sample mean vectors and intra-class sample covariance matrix of multivariate normal models. Columns represent the parameters for shrinking mean vectors: "shrink.factor=1" indicates no shrinkage and "shrink.factor=1/4" indicates shrinking the class-specific sample mean vectors by multiplying 1/4. λ values of x-axis represent the degree of shrinkage of sample covariance matrix. $\lambda = 0$ indicates the diagonal matrix and $\lambda = 1$ indicates sample covariance matrix.

adapted to strongly correlated conditions utilizing the stage-wise variable selection regression approach. However, it seems over-fit for independent or weakly correlated data.

PAM and the BCC type classifiers show similar CS, RS, and misclassification rates in most conditions. Their scores generally increase as λ increases. For the small effect size conditions, however, the original BCC appears to have larger scores than PAM and the 2 modifications of BCC. For more details, see Tables in the supplementary material available at *Biostatistics* online.

5. EVALUATION USING REAL DATA

We evaluated the probabilistic classifiers using 2 real data sets, the prostate cancer data set and colon cancer data set preprocessed in Dettling (2004). The prostate cancer data consists of 6033 genes and 102 samples (50 normal and 52 tumor samples) and the colon cancer data consists of 2000 genes and 62 samples (22 normal and 40 tumor samples). We did not make any model assumptions for the 2 data sets

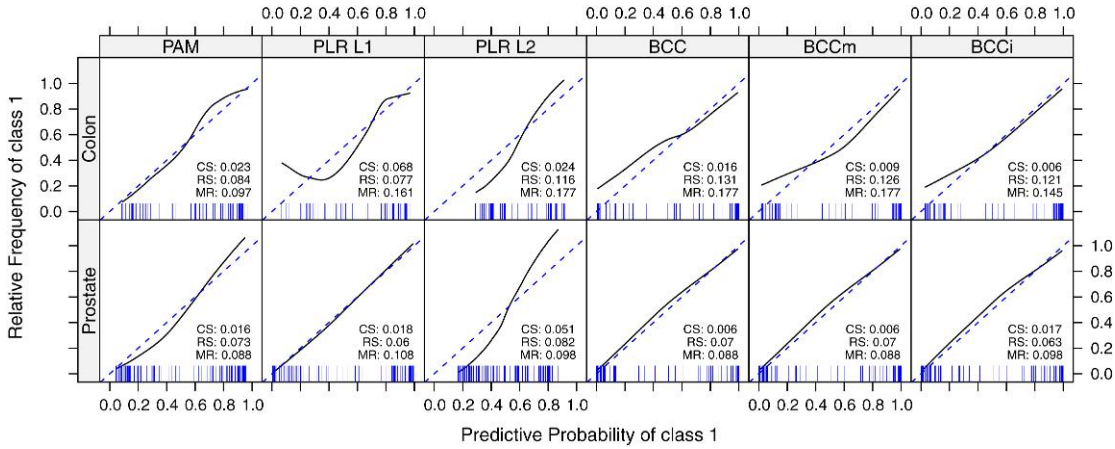

Fig. 5. Calibration curves and rugs of predictive probabilities for the colon and prostate cancer data by Dettling (2004). The colon cancer data set consists of 2000 genes with 62 (22 normal and 40 tumor) samples and the prostate cancer data set consist of 6033 genes with 102 (50 normal and 52 tumor) samples. Predictive probability of each sample was computed based on the rest samples as the training set for learning classifiers.

but used complete LOOCV to estimate CS and RS. For each sample left out, the training set consisted of the remaining specimens. For example, in the case of the prostate cancer data, we performed model development 102 times to compute all class membership probabilities for the cases. Within each leave-one-out training set, we optimized the tuning parameter using an inner loop of LOOCV to compute the predictive likelihood for a grid of tuning parameters.

Figure 5 shows the resulting cross-validated calibration curves and scatter plots of predictive probabilities (rugs). For the prostate cancer data, PLR L2 is poorly calibrated, consistent with the previously described simulations. BCC and BCCm are best calibrated and, with PAM, have the smallest misclassification rates. For the colon cancer data, BCCi and BCCm are best calibrated. PAM is reasonably well calibrated and has small RS and misclassification rate. Although PLR L1 has the smallest RS, it is poorly calibrated. Since the scatter of predictive probabilities are sparse in the unit interval, we used a small number of bins (the number of samples divided by 10) to compute CS and RS.

6. Discussion

Probabilistic classification is important for medical decision making. We introduced 2 evaluation measures for probabilistic classifiers, calibration and refinement. Like the well-known bias–variance decomposition of mean squared error, CS and RS indicate average accuracy and precision for probabilistic classifiers. Based on these 2 criteria, we evaluated several probabilistic classifiers in the high-dimensional setting and compared them with their misclassification rates.

Our simulation studies indicated that some published probabilistic classifiers may be very poorly calibrated under some conditions. They also showed that the degree of refinement and the rate of convergence to the true Bayesian classifier can vary substantially among classifiers. Our simulations based on defined covariance structures and based on real data were consistent and indicated that the PAM probabilistic classifier and the modified BCC probabilistic classifiers performed well under a wide range of conditions. The L1 penalized logistic regression was superior when correlations among genes, including between noise and informative genes, were strong but performed more poorly under other conditions. The L2 penalized

logistic regression performed very poorly in some conditions and cannot be recommended. The original BCC was poorly calibrated for small samples and the modified versions should be used instead.

When the common correlation matrix consists of a single correlation, the variance stabilizing constant s_0 seems to make PAM converge rapidly to the Bayesian probabilistic classifier in our simulations. However, with independent genes, the constant s_0 tends to make PAM overshrunken. In a small simulation, we modified PAM so that the s_0 value was selected by optimizing the predictive likelihood. We found that in the independent or weak correlation condition, s_0 tended to be smaller and the modified probabilistic classifier was better calibrated than PAM (see the supplementary material available at *Biostatistics* online).

While penalized logistic models performed very well when there were strong correlations, they did not do in other conditions. Since the penalized logistic models utilize correlations among covariates in model development, they appeared to overfit data sets in which the sample correlations were spurious (Troendle *and others*, 2004). The effect of correlations in developing the penalized logistic models needs further investigation.

The modifications we introduced improved the original BCC described by Wright *and others* (2003) especially when the effect size between 2 classes is small. The benefit of applying the explicit cross-validation methodology to avoid possible biases due to dependency between selection procedures and estimating steps is substantial. Both of these generalizations of the BCC can potentially be used in other settings. As previously noted, the assumption of normal class-specific densities of the projections $\tilde{\mathbf{t}}_{(-i)}x_i$ could be generalized. Our method of incorporating uncertainty in the class-specific densities when using Bayes' theorem is also more broadly applicable to probabilistic classification.

Well-calibratedness and refinement are useful criteria for evaluation of probabilistic classifiers in practice, and we have described a cross-validation–based method for estimating the calibration and refinement curves for a probabilistic classifier on a specific data set. It is impossible to compute the true Bayesian classifier without knowing the true model, and hence it is not possible to compare a probabilistic classifier to the true Bayesian classifier as in Figure 1 for a real data set. Although the Bayesian classifier has a CS of zero, probabilistic classifiers with small calibration and RS do not necessarily well approximate the true Bayesian classifier.

The probabilistic classifiers evaluated here can be easily extended to handle more than 2 classes. PAM and the penalized logistic models were originally developed to be able to deal with more than 2 classes. The predictive probabilities of BCC type classifiers are of the form of the Bayes formula in (2.2) so the polychotomous extension of BCC type classifiers is rather direct. The RS could be generalized for polychotomous classifiers by using entropy instead variance. Generalization of the CS is however more difficult because it is class-specific. A CS for each class can be computed from (2.9) using the probabilistic predictions for that class of the polychotomous classifier.

Our simulation studies assumed multivariate normality of gene expression data with a common covariance matrix. Since PAM and the BCC type classifiers are based on a normal class-specific densities, their performance may be affected by nonnormality. Although nonnormality–based simulation methods have been developed (e.g. Parrish *and others*, 2009), we found such methods to be too slow for our simulations with high-dimensional data and a 1000 replications. One could extend BCCm and BCCi to incorporate nonnormality by substituting heavy-tailed densities for the normal densities in (2.2) and (2.5), respectively. Simulated results based on common covariance structure could be also extended for class-specific covariance structure in future. The method based on the cross-validated scores in Section 5 can be used to evaluate classifiers without any model assumptions and should be applied to evaluate candidate probabilistic classifiers on the data sets of specific interest for use.

When the sample size is small, the average squared error (CS + RS) can also be a useful measure for selecting among candidate classifiers. CS and RS depend on the binning method, and when the sample size is small, computation of the scores can be unstable because of sampling variability and the small number of bins. Alternatively, one could use kernel density estimation instead of using bins for estimating

distribution of predictive probabilities. We included tables for the sum of CS and RS of the simulation studies in the supplementary material available at *Biostatistics* online.

When selecting good probabilistic classifier with moderate to large samples, well-calibratedness is the most important criteria because it measures the validity of the probabilistic forecast. Thus, in the application to real data, we recommend; first, compute CS and RS based on the cross-validated predictive probabilities as in Section 5. These estimates are free of any model assumptions. By comparing the scores and plots as in Figure 4, one may select the most refined among well-calibrated probabilistic classifiers.

7. Software

We used R packages pamr (Hastie *and others*, 2009) for PAM, glmnet (Friedman *and others*, 2010) for the penalized logistic models. R codes for BCC, BCCm, and BCCi are available upon request.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgment

Conflict of Interest: None declared.

References

Dawid, A. (1986). Probability forecasting. *Encyclopedia of Statistical Sciences* **7**, 210–218.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society, Series D (The Statistician)* **32**, 12–22.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20**, 3583–3593.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499. With discussion, and a rejoinder by the authors.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edition. London: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., Narasimhan, B. and Chu, G. (2009). *PAMR: PAM: Prediction Analysis for Microarrays*. R Package Version 1.33. http://cran.r-project.org/package=pamr.

Lai, C., Reinders, M. J., van't Veer, L. J. and Wessels, L. F. (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics* **7**, 235.

Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307.

Parrish, R. S., Spencer, III, H. J. and Xu, P. (2009). Distribution modeling and simulation of gene expression data. *Computational Statistical & Data Analysis* **53**, 1650–1660.

SCHÄFER, J. AND STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, Article32.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. AND CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567–6572.

TROENDLE, J. F., KORN, E. L. AND MCSHANE, L. M. (2004). An example of slow convergence of the bootstrap in high dimensions. *The American Statistician* **58**, 25–29.

VENABLES, W. N. AND RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

WESSELS, L. F. A., REINDERS, M. J. T., HART, A. A. M., VEENMAN, C. J., DAI, H., HE, Y. D. AND VEER, L. J. V. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* **21**, 3755–3762.

WRIGHT, G., TAN, B., ROSENWALD, A., HURT, E. H., WIESTNER, A. AND STAUDT, L. M. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9991–9996.

[*Received February 4, 2010; revised October 19, 2010; accepted for publication October 20, 2010*]