Sample size planning for developing classifiers using high-dimensional DNA microarray data

KEVIN K. DOBBIN*, RICHARD M. SIMON

Biometric Research Branch, National Cancer Institute, 6130 Executive Boulevard, Rockville, MD 20852, USA
dobbinke@mail.nih.gov

SUMMARY

Many gene expression studies attempt to develop a predictor of pre-defined diagnostic or prognostic classes. If the classes are similar biologically, then the number of genes that are differentially expressed between the classes is likely to be small compared to the total number of genes measured. This motivates a two-step process for predictor development, a subset of differentially expressed genes is selected for use in the predictor and then the predictor constructed from these. Both these steps will introduce variability into the resulting classifier, so both must be incorporated in sample size estimation. We introduce a methodology for sample size determination for prediction in the context of high-dimensional data that captures variability in both steps of predictor development. The methodology is based on a parametric probability model, but permits sample size computations to be carried out in a practical manner without extensive requirements for preliminary data. We find that many prediction problems do not require a large training set of arrays for classifier development.

Keywords: Gene expression; Microarrays; Prediction; Predictive inference; Sample size.

1. INTRODUCTION/MOTIVATION

The goal of many gene expression studies is the development of a predictor that can be applied to future biological samples to predict phenotype or prognosis from expression levels (Golub *and others*, 1999). This paper addresses the question of how many samples are required to build a good predictor of class membership based on expression profiles. Determining appropriate sample size is important as available clinical samples are often either very limited or costly to acquire and assay. As microarray studies move from the laboratory toward the clinic, the reason for developing the predictor is increasingly to assist with medical decisions (Paik *and others*, 2004), and the consequences of having a predictor that performs poorly because the sample size was too small can be serious.

Few methods have been published for developing genomic classifiers. Most publications on sample size determination for microarray studies are limited to the objective of identifying genes which are differentially expressed among the pre-defined classes. These have been reviewed by Dobbin and Simon (2005).

---

*To whom correspondence should be addressed.

Hwang *and others* (2002) addressed the objective of sample size planning for testing the global null hypothesis that no genes are differentially expressed, which is equivalent to testing the null hypothesis that no classifier performs better than chance. However, a sample size sufficient for rejecting the global null hypothesis may not be sufficient for identifying a good classifier. Mukherjee *and others* (2003) developed a learning curve estimation method that is applicable to the development of predictors but requires that extensive data already be available so that the learning curve parameters can be estimated. Fu *and others* (2005) developed a martingale stopping rule for determining when to stop adding cases, but it assumes that the predictor is developed sequentially one case at a time and does not provide an estimate of the number of cases needed.

The high dimensionality of microarray data, combined with the complexity of gene regulation, make any statistical model for the data potentially controversial. This has led some authors to avoid modeling the expression data directly, and instead model the general abstract learning process (Mukherjee *and others*, 2003; Fu *and others*, 2005). But a model for gene expression does not need to be exactly correct to be useful, and to provide insights into the classification problem that other more abstract approaches do not. We do not assert that the model presented here is exactly correct, but a useful oversimplification. Such oversimplifications are not uncommon in sample size determination methodologies: for example, models used to estimate sample sizes in clinical trials are often simpler than the planned data analysis. The simpler model is likely to have lower sensitivity and specificity, resulting in conservative sample size estimates. The simpler model also has the advantage that the resulting calculations will be more transparent, whereas sample size calculations based on the more complex model may be opaque and unconvincing.

A novel contribution of this paper is the integration of dimension-reduction into the framework of the normal model to calculate sample size for high-dimensional genomic data. We develop a novel methodology for calculating the significance level α to be used in gene selection that will produce a predictor with the best resulting expected correct classification rate. We present methods for sample size calculation when one class is under-represented in the population. We also present novel results on how the size of the fold-change for differentially expressed genes, the noise level, and the number of differentially expressed genes, affect predictor development and performance.

Section 2 presents the predictive objective that will be used to drive sample size calculation. Section 3 presents the probability model for microarray data and the optimal classification rates. In Section 4, sample size algorithms are developed. In Section 5, the accuracy of the approximation formulas used in Section 4 are assessed, as well as their robustness to violations of model assumptions; also, the effect of model parameter combinations on sample size requirements and correct classification rates are examined. In Section 6, the robustness of the methodologies are assessed by application to a number of synthetic and real-world datasets that violate the model assumptions. In Section 7, results and recommendations are summarized.

2. THE SAMPLE SIZE OBJECTIVE

In a traditional class comparison study, the sample size objective is to achieve a certain power, say 95%, to detect a specified difference between the class means when testing at an α significance level. Under the usual class comparison model assumptions, an adequate sample size will exist because the power goes to 1 as the sample size goes to infinity.

By analogy to the class comparison case, one might wish in the class prediction setting to establish a sample size large enough to ensure that the probability of correct classification (PCC) will be above, say, 95%. There are at least two problems with this objective. The first problem is that the probability of correct classification will depend on what samples are chosen for the training set; in other words, the probability of correct classification will not be a fixed quantity, but will be a random variable with some variance. Hence, for any sample size there will likely be some positive probability that the correct classification

rate is below 95%. So we will instead consider the "expected value" of this random variable, where the expectation is taken over all possible sample sets of the same size in the population.

The second problem is that, unlike the power in a class comparison study (which always goes to 1 as the sample size goes to infinity), the PCC in a class prediction study will not necessarily go to 1 as the sample size goes to infinity. This is because for any two populations there may be areas in the predictor space where samples from each class overlap, so that class membership cannot be determined with confidence in these areas. An extreme example would be two identically distributed populations, where no predictor can be expected to do better than a coin toss (50%). Lachenbruch (1968) solved this problem by framing the question as: how large a sample size is required for the resulting predictive function "to have an error rate within γ of the optimum value?" For example, a γ of 0.10 ensures that the expected probability of correct classification for the predictor will be within 10% of the best possible predictor. We will use an objective equivalent to Lachenbruch's, namely: determine the sample size required to ensure that the expected[1] correct classification probability for a predictor developed from training data is within γ of the optimal expected correct classification probability for a linear classification problem[2].

We would also note that although we will focus attention on this objective, the formulas developed here could also be used to ensure sensitivity and/or specificity above a specified target.

3. THE PROBABILITY MODEL

The general framework for the class prediction problem is that in some population of interest, individuals can be split into k disjoint classes, $C_1, C_2, \ldots, C_k$. The classes may correspond to different outcome groups (e.g. relapse in 5 years versus no relapse in 5 years), different phenotypes (e.g. adenocarcinoma versus squamous cell carcinoma), etc. A predictive model will be developed based on a training set T. For each individual in the training set, one observes that individual's class membership, and a data vector $\mathbf{x}$, the gene expression vector; the goal of the training set experiment is to develop a predictor of class membership based on gene expression, and possibly an estimate of the predictor's performance. Our goal is to determine a sample size for the training set.

Consider a two-class problem, with the gene expression data vector denoted by $\mathbf{x}$, which consists of normalized, background-corrected, log-transformed gene expression measurements. To simplify notation and presentation, let the first m elements of the data vectors represent the differentially expressed genes[3], the remaining $(p - m)$ elements of the data vectors represent undifferentially expressed genes, and each differentially expressed gene be centered around zero. The probability model for this two-class prediction problem is

$$\mathbf{x} \sim \begin{cases} N(\mu, \Sigma): \mathbf{x} \in C_1, \\ N(-\mu, \Sigma): \mathbf{x} \in C_2. \end{cases} \tag{3.1}$$

The mean vector has the form $\mu = (\mu_1, \mu_2, \ldots, \mu_m, 0, \ldots, 0)^T$, where $\mu_1, \ldots, \mu_m$ represent the differentially expressed gene means.

The model stipulates that within each class, expression vectors follow a multivariate normal distribution with the same covariance matrix. It will be also assumed that differentially expressed genes are independent of undifferentially expressed genes. If Σ is singular, then some genes are linear combinations of other genes (see, e.g. Rao, 1973, pp 527–8). Put another way, there are "redundant" genes with expression that is completely determined by the expression of other genes. Having these type of redundant

---

[1] The expectation is taken over all training samples of size n in the population.
[2] Under the assumptions of the homogeneous variance multivariate normal model.
[3] A differentially expressed gene is defined as a gene with different average expression in the different classes.

genes is analogous to having an overparameterized linear model, and a model reduction transformation (Hocking, 1996, p 81) can eliminate the redundant genes, resulting in a non-singular covariance matrix. We will imagine these redundant genes have been eliminated, so that the covariance matrix Σ is non-singular[4]. Marginal normality of the gene expression vectors may be considered reasonable for properly transformed and normalized data, although multivariate normality may be more questionable. However, in taking a model-based approach one must make some assumption and one weaker than multivariate normality is unlikely to lead to a tractable solution. The assumption of independence between differentially expressed and non-differentially expressed genes is not critical and is mainly made for mathematical convenience. Violations of this assumption will be evaluated.

A key issue is the size of the μ_i relative to the biological and technical variation present. A relatively small μ_i would correspond to a differentially expressed gene that is nearly uninformative—i.e. that will be of little practical use for prediction. We discuss these types of genes with numerous examples below and show that nearly uninformative genes, even if there are many of them (50–200 among thousands of genes), are generally of little use for predictor construction and sample size estimates should be based on more informative genes. If the biological reality is that all the differentially expressed genes are nearly uninformative, then we will see that no good predictor will result from the microarray analysis.

It will simplify presentation to assume that each differentially expressed gene has a common variance, and each undifferentially expressed gene has a common variance. In practice, genes are likely to have different variances. But the relationship between fold-change and gene variance determines the statistical power in the gene selection phase of predictor development. In order to keep this relationship intuitive, it is important to have a single variance estimate rather than a range of variance estimates. The single variance parameter can be considered a mean or median variance, in which case the targeted power will be achieved on average, with higher power for some genes and lower for others. More conservatively, the 90th percentile of the gene variances can be used, in which case the targeted power will be achieved even for genes exhibiting the highest variation (which may be the ones of most potential interest) across the population.

3.1   Notation

Throughout, PCC(n) will denote the marginal expected probability of correct classification taken over all samples of size n in the population. PCC(∞) $= \lim_{n \nearrow \infty}$PCC($n$) will denote the expected PCC for an optimal linear classifier.

In the population of interest, the proportion in class C_1 is p_1, and the proportion in class C_2 is $p_2 = 1 - p_1$. The covariance matrix can be written in a partitioned form, with notation $\Sigma = \begin{pmatrix} \sigma_I^2 \Sigma_I & 0 \\ 0 & \sigma_U^2 \Sigma_U \end{pmatrix}$ where Σ_I is an $m \times m$ correlation matrix for the differentially expressed genes and Σ_U is the $(p - m) \times (p - m)$ correlation matrix for the undifferentially expressed genes.

3.2   Optimal classification rates for the model

The optimal classification rule and rate will depend on the proportion in class C_1 in the population. For a two-class problem with equal covariance matrices, if the model parameters are known, then the optimal normal-based linear discriminant rule, that is, the Bayes rule, is known and the classification rate of this classifier can be determined. In Appendix A, it is shown that

$$\text{PCC}(\infty) = p_1 \Phi \left(\frac{\mu' \Sigma^{-1} \mu - \frac{1}{2} \ln \left(\frac{1 - p_1}{p_1} \right)}{\sqrt{\mu' \Sigma^{-1} \mu}} \right) + (1 - p_1) \Phi \left(\frac{\mu' \Sigma^{-1} \mu + \frac{1}{2} \ln \left(\frac{1 - p_1}{p_1} \right)}{\sqrt{\mu' \Sigma^{-1} \mu}} \right),$$

---

[4]Note that assuming Σ is non-singular is not the same as assuming the estimated covariance matrix S is non-singular. S will usually be singular because of the "large p small n" issue, i.e. because there are many more genes than samples.

where ln is the natural logarithm, and Φ is the cumulative distribution function for a standard normal random variable. When $p_1 = \frac{1}{2}$, so that each class is equally represented in the population, this simplifies to

$$\text{PCC}(\infty) = \Phi(\sqrt{\mu'\Sigma^{-1}\mu}). \tag{3.2}$$

Note that $2\sqrt{\mu'\Sigma^{-1}\mu}$ is the Mahalanobis' distance between the class means, making this result closely related to that of Lachenbruch (1968).

In the special case when $\mu_i = \delta, i = 1, 2, \ldots, m$, and $\sigma_I^2 = \sigma_U^2 = \sigma^2$, it is shown in Appendix A that an upper bound on the the best probability of correct classification is

$$\text{PCC}(\infty) \leqslant \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_I^*}}\right), \tag{3.3}$$

where λ_I^* is the smallest eigenvalue of Σ_I, which is 1 if genes are independent.

4. METHODS

Consider linear classifiers of the form: classify in class C_1 if $l(\mathbf{x}) = \mathbf{w}'\mathbf{x} > k$ and in class C_2 otherwise. The vector $\mathbf{w}$ is estimated from the training set, and will depend on how genes are selected for inclusion in the predictor, and what prediction method is used. We will take a simple approach to predictor construction which does not weight the importance of the individual genes in the predictor. Each element of $\mathbf{w}$ is 0 or 1; 1 indicates a gene determined to be differentially expressed by the hypothesis test of H_0: $\mu_i = 0$ versus H_1: $\mu_i \neq 0$; 0 indicates a gene determined not to be differentially expressed. This simple predictor is likely to have lower sensitivity and specificity than more sophisticated ones that assign weights to individual genes and we would by no means recommend people to use it. But, the sample sizes that we calculate this way should tend to be conservative (large).

Consider the hypothesis tests for gene selection described in the previous paragraph. These are tests of differential expression for each of the p genes. With each hypothesis test is associated a specificity, which will be denoted $1 - \alpha$, and is the probability of correctly identifying a gene that is not differentially expressed; and also a sensitivity or power, which will be denoted $1 - \beta$, and is the probability of correctly identifying a gene as differentially expressed when in fact it is differentially expressed by a specified amount (δ). These hypothesis tests could be based on many different statistics. The calculations here will use two-sample t-tests.

4.1 *Formulas for* PCC(n)

Each differentially expressed gene will be assumed to be differentially expressed by δ, and a common variance $\sigma^2 = \sigma_I^2 = \sigma_U^2$ for genes will be assumed. An approximate lower bound for the expected probability of correct classification is derived in Appendix B, and is

$$\text{PCC}(n) \geqslant p_1\Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{\delta m(1-\beta) - \frac{1}{2}\ln\left(\frac{1-p_1}{p_1}\right)}{\sqrt{m(1-\beta) + (p-m)\alpha}}\right)$$

$$+ (1 - p_1)\Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{\delta m(1-\beta) + \frac{1}{2}\ln\left(\frac{1-p_1}{p_1}\right)}{\sqrt{m(1-\beta) + (p-m)\alpha}}\right),$$

where λ_1 is the largest eigenvalue of the population correlation matrix. When the other parameters are fixed, PCC(n) reaches a minimum at $p_1 = \frac{1}{2}$. When $p_1 = \frac{1}{2}$, so that the two classes are equally represented, this simplifies to (Appendix B)

$$\text{PCC}(n) \geqslant \Phi\left(\frac{\delta}{\sigma\sqrt{\lambda_1}}\sqrt{m}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{m(1-\beta)+(p-m)\alpha}}\right).$$

In the special case when $\Sigma = \sigma^2 I$, $\lambda_1 = 1$.

Note that $1 - \beta$ is the power associated with the gene-specific hypothesis tests that each gene is not differentially expressed among the classes, and the term under the final root sign is the true discovery rate (TDR)[5].

In fact, power calculations can be used to eliminate β from the equation (see Appendix C).

4.2 Sample size determination

Recall that the objective is to find a sample size that will ensure that $\text{PCC}(\infty) - \text{PCC}(n) < \gamma$, where γ is a pre-specified constant. The calculation can be based on the general formula

$$\text{PCC}(\infty) - \text{PCC}(n)$$
$$\leqslant p_1 \Phi\left(\frac{1}{\sigma\sqrt{\lambda^*}}\frac{\delta m - \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)}{\sqrt{m}}\right) + (1-p_1)\Phi\left(\frac{1}{\sigma\sqrt{\lambda^*}}\frac{\frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right) + \delta m}{\sqrt{m}}\right)$$
$$- p_1 \Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{\delta m(1-\beta) - \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)}{\sqrt{m(1-\beta)+(p-m)\alpha}}\right) - (1-p_1)\Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{\frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right) + \delta m(1-\beta)}{\sqrt{m(1-\beta)+(p-m)\alpha}}\right).$$
$$(4.1)$$

Note that this formula will only guarantee that the overall PCC is within the specified bound, but the PCC for the individual classes may differ. In particular, the rarer subgroup may have much poorer PCC. A more stringent approach is discussed below in Section 4.4. If we assume that $p_1 = \frac{1}{2}$ and that genes are independent, then the simpler formula

$$\text{PCC}(\infty) - \text{PCC}(n) \approx \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\right) - \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{m(1-\beta)+(p-m)\alpha}}\right) \qquad (4.2)$$

can be used to determine the sample size. Ideally, one would want to eliminate m from the equation, so that the number of differentially expressed genes need not be stipulated. One can do this by maximizing the difference over m. As m gets large, the distance between the class means increases, and $\text{PCC}(\infty)-\text{PCC}(n)$ goes to zero (but the difference is not always strictly decreasing). So the maximum m value should be a low integer, and most likely $m = 1$. To ensure that the difference is less than γ, one can use

$$\text{PCC}(\infty) - \text{PCC}(n)$$
$$\leqslant \max_{1\leqslant m\leqslant p}\left\{\Phi\left(\frac{\delta}{\sigma}\sqrt{m}\right) - \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{m(1-\beta)+(p-m)\alpha}}\right)\right\}. \qquad (4.3)$$

[5]Technically, this is the approximate TDR, the expected value of the true discovery proportion (TDP), which is one minus the false discovery proportion (FDP). Let FD_n be the number of false discoveries when the sample size is n, and TD_n the number of true discoveries. Then, $\text{TDR} = E[\text{TDP}] = 1 - E[\text{FDP}] \approx 1 - \frac{E[\text{FD}_n]}{E[\text{TD}_n]+E[\text{FD}_n]} = 1 - \frac{(p-m)\alpha}{m(1-\beta)+(p-m)\alpha}$.

This gives rise to the following algorithm for sample size determination:

1. Initialize $n = 0$.
2. $n = n + 2$.
3. For $m \in 1, 2, \ldots, p$, use (C.1) to find the optimal α for each m, by a linear search over $\alpha \in [0, 1]$. Then plug these αs into (4.3) to get an upper bound on $\text{PCC}(\infty) - \text{PCC}(n)$, call this U_n.
4. Is $U_n \leqslant \gamma$? If no, return to step 2. If yes, continue to step 5.
5. Use a sample of size n, with $n/2$ from each class.

If we do not assume gene independence, then it will be necessary to estimate extreme eigenvalues of the correlation matrix. Based on simulation with block diagonal compound symmetric covariance matrices (not shown), we suggest the method of Ledoit and Wolf (2004), which seems to perform reasonably well compared to others we examined. One could then plug these estimates into the equation

$$
\begin{aligned}
&\text{PCC}(\infty) - \text{PCC}(n) \\
&\leqslant \underset{1 \leqslant m \leqslant p}{\text{Max}} \, \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_I^*}}\right) - \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_1}}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{m(1-\beta)+(p-m)\alpha}}\right).
\end{aligned} \tag{4.4}
$$

4.3 *Sample size for a test set*

Once a predictor has been developed, an estimate of its performance is required. Such estimates are calculated either using a separate testing set of samples that were not used at all in the predictor development or by cross-validation applied to the training set. Advantages to having an independent test set are discussed in Simon *and others* (2003).

How best to split a collection of samples into a test set and training set is addressed in Molinaro *and others* (2005). Alternatively, one can use the methods developed here to determine the sample size required for the training set, then base the sample size for the test set on the width of the resulting confidence interval for the PCC. This is a valid approach because the independence of the test set implies that a binomial random variable correctly models the number of correct classifications. If L is the interval length, then the sample size formula[6] for the test set is just $n_{\text{test}} = \frac{4z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{L^2}$, where $\hat{p}$ is the estimated correct classification rate. For example, if $\hat{p} = 0.90$, then $L = 0.19, 0.17, 0.15, 0.12$ results in sample sizes of 40, 50, 60, 100, respectively.

4.4 *Controlling the PCC in each class*

We have presented methods for controlling the overall probability of correct classification. In some cases, one may want to control the PCC for each class individually. If one class is under-represented, then the PCC using the optimal cut-point will be lower in the under-represented class.

$$
\text{PCC}_{\text{under-represented class}}(n) \geqslant \Phi\left(\frac{1}{\sigma\sqrt{\lambda}} \cdot \frac{\delta \cdot m \cdot \text{pow}(\alpha, n, p_{\min}) - \frac{1}{2}\ln\left[\frac{1-p_{\min}}{p_{\min}}\right]}{\sqrt{m \cdot \text{pow}(\alpha, n, p_{\min}) - (p-m)\alpha}}\right),
$$

where $p_{\min}$ is the proportion in the under-represented class and $\text{pow}(\alpha, n, p_{\min})$ is the power to detect a differentially expressed gene given α, n, and $p_{\min}$. This formula can be used to develop methods similar to those we have presented to determine sample size.

---

[6]The formula is valid when $n\hat{p} \geqslant 5$ and $n(1 - \hat{p}) \geqslant 5$, which must be verified.

A simpler rough approach is to let $n_{0.5}$ be the sample size calculated by the methods we have presented for the case when half the population is from each class, for example, based on (4.3). Then, use a sample size of

$$n_{p_1} = \frac{n_{0.5}}{2 \min(p_1, 1 - p_1)}$$

to ensure that the expected number from the under-represented class is the same as it would be if both classes were equally represented.

If p_1 is close to 0 or 1, then this approach may lead to very large sample sizes. One option in this case is to oversample from the rarer class, so that for example half the samples are from each class. But resulting estimates of classification error will depend on external estimates of the proportion in each class, so oversampling is most appropriate for preliminary studies.

5. RESULTS

We first ran an extensive simulation to verify that the approximations used in the course of the derivation of the equation for the PCC produced good estimates. Table 1 shows the estimated probabilities of correct classification based on (4.1), and compares these to estimates of the population values based on Monte Carlo for a variety of combinations of values of effect size, $2\delta/\sigma$, number of genes affected, m, and sample size, n. For the Monte Carlo-based estimates, data were generated according to the model specifications, then predictors developed as outlined in Section 4, and finally prediction accuracy assessed on an independent test set. As can be seen, the equation-based estimates are close to the Monte Carlo-based population estimates and, when different, tend to be conservative.

We next consider choice of α, the significance level for gene selection. α can be chosen to maximize the resulting PCC. Plots of the PCC(n) as a function of α for varying values of n are shown in Figure 1. The plots show that there is an optimal α value on the interval. As the sample size n increases, the α that maximizes the probability of correct classification decreases. This trend is intuitively reasonable because α determines the cutoff value used in the gene selection step, and as the sample size gets large the power associated with small α's will increase. Also note from the plots that the larger the effect size ($2\delta/\sigma$), the smaller the α value that will maximize PCC(n) for fixed n. This makes sense since larger effect sizes will be easier to detect, so that one can afford to use more stringent α's to reduce the false-positive rate.

Figure 2 shows the sample sizes required as a function of $2\delta/\sigma$, for $m = 1$ and $m = 2$ differentially expressed genes. Here, $p_1 = 1/2$. The sample sizes were optimized[7] over α, and will ensure the expected PCC(n) is within 0.05 of the best possible prediction rule. For a fixed effect size $2\delta/\sigma$ per gene, $m = 2$ is an easier classification problem than $m = 1$, because the distance between the class means is $2\frac{\delta}{\sigma}\sqrt{m}$. So the sample size requirements are smaller for $m = 2$. For an effect size of 1.5 (e.g. $2\delta = 1$ for a two-fold change in expression, and $\sigma = 0.71$), a sample size around 60 appears adequate.

We next turned to the question of the relationship between effect size, $2\delta/\sigma$, the number of differentially expressed genes m, and the Mahalonobis distance between the class means (which determines the best possible classification rate). Table 2 shows examples where the effect size is small, so that each gene is nearly uninformative. For example, when the effect size is 0.2, and the number of informative genes is 200 or less, then one of two bad things will happen: either even an optimal predictor will have a poor PCC(∞), so that predictions will be unreliable or development of a predictor with PCC(∞) $-$ PCC(n) within some reasonable bound will require prohibitively large sample sizes (500 or more). In the latter case, while a good predictor exists, the problem is too hard practically to work out with this technology. The table also shows that as the effect size gets larger, these issues go away. Hence, it is critical that at

---

[7]For these plots, the optimal α was calculated for all even sample sizes from 2 to 100, and then for each possible sample size the corresponding optimal α was used to estimate the mean PCC.

Table 1. *The expected PCC estimate is generally either accurate or conservative (underestimates the true PCC). Here, the number of genes is $p = 10\,000$, $2\delta/\sigma$ is the effect size for differentially expressed genes, n is the number of samples in the training set, m is the number of differentially expressed genes, and α is the significance level used for gene selection. Monte Carlo PCC estimates were calculated by generating a predictor based on a simulated sample, then generating 100 datasets from the same populations and calculating the prediction error rate of the predictor; this entire process was repeated 100 times and the average correct classification proportions appear in the table*

$2\frac{\delta}{\sigma}$	n	m	α	$\widehat{\text{PCC}}$ equation	$\widehat{\text{PCC}}$ Monte Carlo	$\widehat{\text{Bias}}$
4	32	1	0.0001	0.92	0.96	−0.04
4	32	1	0.001	0.72	0.78	−0.06
4	32	1	0.01	0.58	0.59	−0.01
4	32	1	0.05	0.54	0.59	−0.05
4	32	1	0.1	0.53	0.54	−0.01
4	32	10	0.0001	>0.99	1.00	≈0
4	32	10	0.001	>0.99	1.00	≈0
4	32	10	0.01	0.97	0.99	−0.02
4	32	10	0.05	0.81	0.99	−0.18
4	32	10	0.1	0.74	0.78	−0.04
3	70	1	0.0001	0.86	0.92	−0.06
3	70	1	0.001	0.67	0.81	−0.14
3	70	1	0.01	0.56	0.59	−0.03
3	70	1	0.05	0.53	0.52	+0.01
3	70	1	0.10	0.52	0.52	−0.02
3	70	10	0.0001	>0.99	1.00	≈0
3	70	10	0.001	>0.99	1.00	≈0
3	70	10	0.01	0.92	0.98	−0.06
3	70	10	0.05	0.75	0.73	+0.02
3	70	10	0.10	0.68	0.74	−0.06

least some genes have a reasonably large effect size in order to develop a predictor. Therefore, for sample size calculations, one should assume that some genes do have reasonable effect sizes.

These considerations also lead to the following recommendation: when using conservative sample size approaches with $m = 1$, the estimated effect size $2\delta/\sigma$ should correspond to the estimated largest effect size for informative genes. This is the approach we used in Section 6 and seemed to perform adequately. Note also that we assume that p is in the 1000 to 22 000 range, that there are $m = 1$–200 or so genes differentially expressed. When either p or m values fall outside these assumptions, then our conclusions may not be valid. For example, microarrays with 10's or 100's of genes represented would require amendment of this approach.

Figure 3 shows plots of the probability of correct classification as a function of n. The lines represent different values of $2\delta/\sigma$. One can see the PCC(n) values approaching their asymptotic PCC(∞) values as n approaches 100.

6. EXAMPLES USING SYNTHETIC AND REAL MICROARRAY DATA

We tested the robustness of these methods using synthetic microarray data that violates the model assumptions, and real microarray data that likely do as well. Results are presented in Table 3. See table caption for detailed description of analysis.
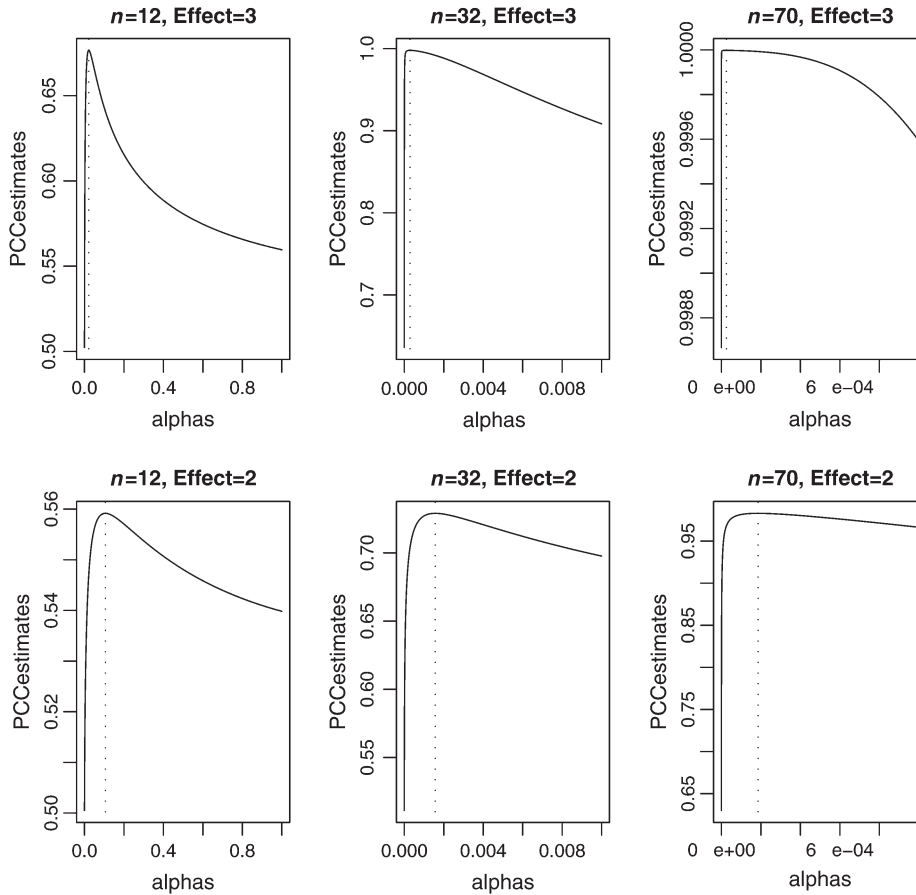
Fig. 1. Plots of the estimated PCC as a function of α, plotted for various values of n, based on (C.1). In each plot, "Effect" is defined as $\frac{2\delta}{\sigma}$, $m = 10$ is the number of differentially expressed genes, and $p = 10\,000$ is the number of genes in vector **x**. The sample sizes are, from left to right, $n = 12$, $n = 32$, and $n = 70$. The maximum points on the plots (indicated by vertical lines) are, for the top row, 0.02, 0.0003, and 3×10^{-5} and for the bottom row, 0.11, 0.002, and 0.0002.

The estimates of $\text{PCC}(\infty) - \text{PCC}(n)$ and $\text{PCC}(n)$ based on our method are uniformly conservative across all datasets. The $\text{PCC}(\infty) - \text{PCC}(n)$ estimates are very conservative when applied to the Golub dataset. This might be expected since there are likely many genes with large effect sizes in these different tumor types, so that our assumption that there is only $m = 1$ informative gene makes the problem much harder than it really is. The $\text{PCC}(n)$ estimates are extremely conservative not only for the Golub dataset but also for several others as well. This is offset somewhat in the $\text{PCC}(\infty) - \text{PCC}(n)$ estimates since both quantities are biased in the same direction, causing some of the bias to cancel out of the difference.

One must be somewhat careful in the interpretation of Table 3, because although for smaller sample sizes like $n = 24$ the $\text{PCC}(\infty) - \text{PCC}(n)$ and $\text{PCC}(n)$ estimates, which are based on means over multiple simulations, appear uniformly conservative, there was also significant variation observed in the performance of classifiers in different simulations. For example, while the mean for the synthetic dataset on the first row of the table was 0.92 using SVMs, the worst classifier had an estimated correct classification rate of 0.79.
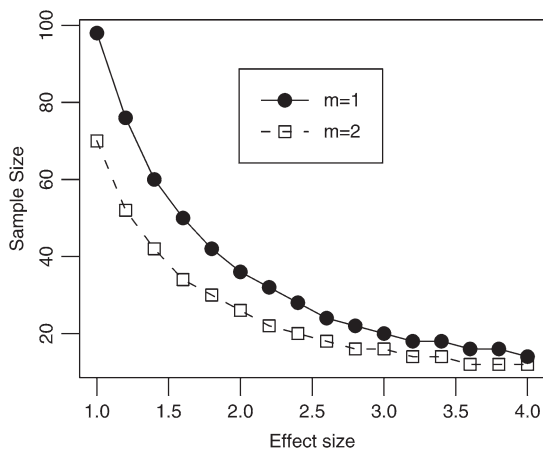
Fig. 2. Plot of effect size $(2\delta/\sigma)$ versus sample size. Optimal α used for determination of sample size. The sample sizes ensure that the average PCC is within $\gamma = 0.05$ of the best possible correct classification probability. Gene independence is assumed. Half the population is from each group, so that $p_1 = 1/2$.

Table 2. *Impact of small effect sizes on the correct classification rate. PCC*(∞) *is the best possible correct classification rate, PCC*(500) *and PCC*(200) *are the correct classification rates with samples of sizes* 500 *and* 200, *respectively. For small effect size* 0.2, *a strong classifier exists only if many genes are differentially expressed; but even with many differentially expressed genes and a sample size of* 500, *the estimator performs poorly. For somewhat larger effect size* 0.4, *fewer differentially expressed genes are required for a good classifier to exist, but even* $n = 200$ *samples do not result in estimates within* 10% *of the optimal classification rate. For effect size* 0.6, *the situation is more amenable if running a large sample is feasible. For reference, if* $\sigma = 0.5$, *then an effect size of* 0.6 *corresponds to a* $2^{2\delta} = 2^{0.3} = 1.23$ *fold-change. Gene independence assumed here*

Differentially expressed genes	Effect size $2\delta/\sigma$	PCC(∞)	PCC(500)	PCC(200)
50	0.2	0.76	0.60	0.54
100	0.2	0.84	0.64	0.56
150	0.2	0.89	0.67	0.57
200	0.2	0.92	0.70	0.59
10	0.4	0.74	0.71	0.62
20	0.4	0.81	0.79	0.68
30	0.4	0.86	0.84	0.72
40	0.4	0.89	0.88	0.75
10	0.6	0.83	0.83	0.79
20	0.6	0.91	0.91	0.88
30	0.6	0.95	0.95	0.93
40	0.6	0.97	0.97	0.95

We applied both the gene independence-based method and the Ledoit and Wolfe eigenvalue estimation method to two other real microarray datasets (not shown) and found that resulting sample size estimates were similar in the two approaches, and in the 40–60 samples range.

In conclusion, our method tends to be conservative in estimating PCC(n), PCC(∞), and PCC(∞) − PCC(n) for the datasets we examined, sometimes very conservative. So the method should be lead to
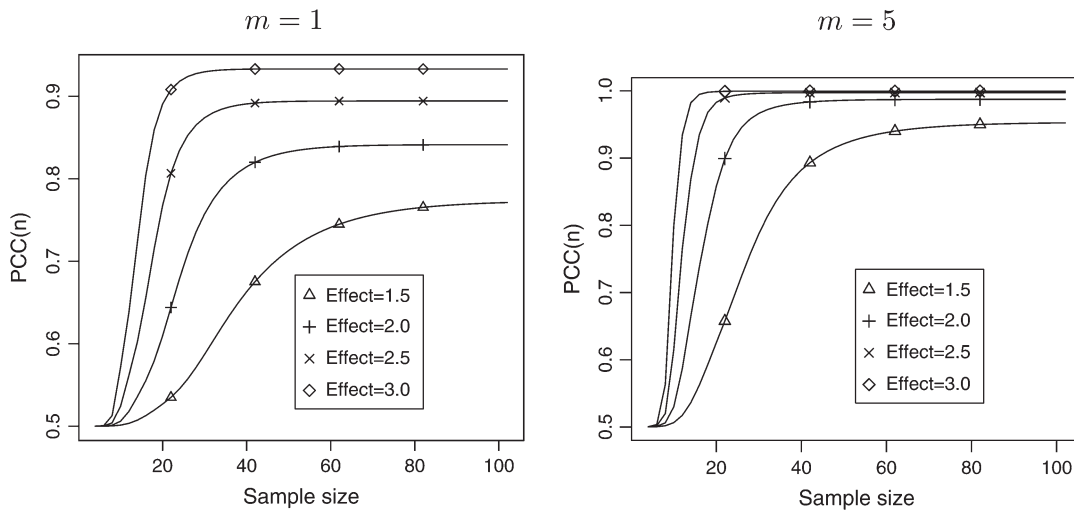
Fig. 3. Left plot is $m = 1$ and right plot is $m = 5$. $p = 10\,000$. Plot of sample size versus PCC for various values of the effect size $2\delta/\sigma$. Gene independence is assumed. PCC(n) uses optimal α. Population assumed evenly split between the classes, so $p_1 = 1/2$.

adequate sample sizes while sometimes producing larger sample size estimates than are truly required. For example, our formulas are likely to significantly overestimate the required sample size for classification problems involving morphologically diverse tissues that are expected to have many differentially expressed genes with large effect sizes. In these cases, it may be advisable to follow the guidelines in Mukherjee *and others* (2003) instead.

7. Conclusion

We have presented novel methods for determining sample size for building predictors of class membership from high-dimensional microarray data. These methods take into account variability both in predictor construction and gene selection. These methods require only that two quantities be specified: the size of the fold-change relative to the variation in gene expression and the number of genes on the microarrays. We presented an alternative approach based on eigenvalue estimation. We investigated the robustness of our method on synthetic datasets that violated our model assumptions and on publicly available microarray datasets.

We found that sample sizes in the range of 20–30 per class may be adequate for building a good predictor in many cases. These results are similar to Mukherjee *and others* (2003). In general, we found that the sample size requirements for prediction are relatively small. We showed that the reason for this is that if a good gene-expression-based predictor exists in the population, then it is likely that some genes exhibit significant differential expression between the classes relative to the biological noise present. Hence, adequate power to identify these genes can be achieved without a large number of microarrays. One drawback of our approach is that it controls the expected probability of correct classification to be within some tolerance of the optimal, but does not control the actual probability of correct classification; for small sample sizes, the probability of correct classification may be highly variable depending on what samples are chosen for the training set, so that our method may not give adequate control and should be used with caution in these situations.

We identified scenarios in which either no good predictor exists or it is practically not feasible to construct one. No good predictor may exist when differential expression is small relative to biological

Table 3. *Robustness evaluation of PCC estimates. Gene selection based on* $\alpha = 0.001$, *with support vector machine (SVM) and nearest neighbor (1-NN). The sample size is n, with n/2 taken from each class. All equation-based* $\widehat{PCC}$ *estimates use* $m = 1$. *Synthetic datasets: generated from multivariate normal distribution with three differentially expressed genes with effect sizes given.* $p = 1000$ *genes with block diagonal correlation structure compound symmetric with* 20 *blocks of* 50 *genes each and within-block correlation* ρ, *and each gene with variance* 1. *True PCC(n) for each classification algorithm estimated using n for classifier development and remainder for PCC(n) estimation, with* 10 *replicates of a sample size of* 200, *then averaging correct classifications. PCC(∞) estimated using cross-validation on a sample size of* 200. *Real datasets: taken from Golub and others* (1999) *and Rosenwald and others* (2002). *Predictor developed using n for classifier development and remainder for PCC(n) estimation. Process repeated* 30(20) *times for Golub (Rosenwald) datasets, and average probabilities of correct classification presented. Effect size* $2\delta/\sigma$ *was estimated from the complete data using empirical Bayes methods. See Appendix E for details. PCC(∞) estimated based on cross-validation on complete datasets. Analyses were performed using BRB ArrayTools developed by Dr. Richard Simon and Amy Peng Lam*

Synthetic data								
Reality			$\widehat{PCC}(\infty) - \widehat{PCC}(n)$			$\widehat{PCC}(n)$		
$2\delta/\sigma$	ρ	n	Our	SVM[†]	1-NN[†]	Our	SVM[†]	1-NN[†]
3, 2, 1	0.8	24	0.07	0.00	0.00	0.86	0.92	0.92
1.5, 1.5, 1.5	0.8	100	0.07	0.00	0.00	0.70	0.90	0.85
1.5, 1.5, 1.5	0.8	24	0.16	0.13	0.12	0.61	0.77	0.73

Real data								
Reality			$\widehat{PCC}(\infty) - \widehat{PCC}(n)$			$\widehat{PCC}(n)$		
Dataset	Classes	n	Our	SVM[†]	1-NN[†]	Our	SVM[†]	1-NN[†]
Golub	AML[‡]/ALL[§]	40	0.21	0.01	0.00	0.75	0.96	0.96
Golub	AML/ALL	20	0.27	0.02	0.03	0.69	0.94	0.94
Golub	AML/ALL	10	0.45	0.09	0.10	0.51	0.87	0.87
Rosenwald	2-year vital	100	0.06	0.00	0.02	0.50	0.55	0.54
Rosenwald	GCB[¶]/non-GCB	200	0.07	0.00	0.03	0.67	0.96	0.91
Rosenwald	GCB/non-GCB	100	0.07	0.02	0.04	0.67	0.94	0.90
Rosenwald	GCB/non-GCB	50	0.09	0.06	0.07	0.64	0.90	0.87

[†]Test set-based estimates.
[‡]AML: acute myeloid leukemia.
[§]ALL: acute lymphoblastic leukemia.
[¶]GCB: germinal-center B-cell-like lymphoma.

variation—and this may be the case even when as many as 100 genes are differentially expressed. We further found that even if enough genes are differentially expressed to construct a reasonable predictor (in theory), if the fold-changes for differentially expressed genes are uniformly small relative to the biological variation, then identification of differentially expressed genes and construction of a good predictor will probably not be feasible.

We investigated both the cases when each class is equally represented in the population and when they are not equally represented. We presented methods for controlling the overall PCC and for controlling the PCC in the worst group in these situations (sensitivity and specificity).

The eigenvalue-based estimation method presented is quite preliminary and we would in general recommend that the independence assumption approach be used instead. One problem with the eigenvalue

approach is that there will be uncertainty about the eigenvalue estimates. Another problem is that these formulas are theoretical worst-case-scenarios bounds that may in fact be much worse than reality, and therefore lead to too conservative estimate for the sample size. Additionally, Table 3 suggests that the gene independence method may be conservative enough already, even when it is violated.

A method for finding the optimal significance level α to use for gene selection when developing a predictor was presented, and approaches to determining sample size for a testing set discussed. The optimal significance levels α tend to be on par with those generally used in microarray experiments (e.g. $\alpha = 0.001$). We further showed that the probability of correct classification depends critically on the power and the true discovery rate, so that gene selection methods that control the false discovery rate should produce good predictors.

APPENDICES

A.1   *Appendix A*

Assume (3.1) where Σ is positive definite, and has the form $\Sigma = \begin{pmatrix} \Sigma_I & 0 \\ 0 & \Sigma_U \end{pmatrix}$, where Σ_I indicates an $m \times m$ covariance matrix. For an x randomly selected from the population of interest $\mathcal{P}$, define $p_1 = P(x \in C_1)$. Then, $1 - p_1 = P(x \in C_2)$ follows.

The linear prediction rule which results in the best probability of correct classification classifies x in C_1 if (see, e.g. Enis and Geisser, 1974)

$$2x'\Sigma^{-1}\mu > \log\left(\frac{1-p_1}{p_1}\right)$$

and classifies x in C_2 otherwise. If $p_1 = \frac{1}{2}$, then the rule reduces to: classify x in C_1 if $2x'\Sigma^{-1}\mu > 0$ and classify x in C_2 otherwise.

The vector $\mu'\Sigma^{-1}x$ is a linear combination of normal random variables, and so is normally distributed:

$$\mu'\Sigma^{-1}x \sim \begin{cases} N(\mu'\Sigma^{-1}\mu, \mu'\Sigma^{-1}\mu): C_1, \\ N(-\mu'\Sigma^{-1}\mu, \mu'\Sigma^{-1}\mu): C_2. \end{cases}$$

Therefore, the PCC for this optimal classifier is

$$\text{PCC}(\infty) = p_1 P(CC|x \in C_1) + (1 - p_1)P(CC|x \in C_2)$$

$$= p_1 P\left(\mu'\Sigma^{-1}x > \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)\Big|C_1\right) + p_1 P\left(\mu'\Sigma^{-1}x < \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)\Big|C_2\right)$$

$$= p_1 P\left(z > \frac{\frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right) - \mu'\Sigma^{-1}\mu}{\sqrt{\mu'\Sigma^{-1}\mu}}\right) + (1 - p_1)P\left(z < \frac{\frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right) + \mu'\Sigma^{-1}\mu}{\sqrt{\mu'\Sigma^{-1}\mu}}\right)$$

$$= p_1 \Phi\left(\frac{\mu'\Sigma^{-1}\mu - \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)}{\sqrt{\mu'\Sigma^{-1}\mu}}\right) + (1 - p_1)\Phi\left(\frac{\mu'\Sigma^{-1}\mu + \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)}{\sqrt{\mu'\Sigma^{-1}\mu}}\right).$$

If $p_1 = \frac{1}{2}$, then this simplifies to

$$\text{PCC}(\infty) = \Phi\left(\sqrt{\mu'\Sigma^{-1}\mu}\right).$$

An extremal property of eigenvalues is that $\mu'\Sigma_I^{-1}\mu \leqslant \mu'\mu\frac{1}{\lambda_I^*}$ (Schott, 1997, Theorem 3.15). If $\mu_i = \delta, i = 1, \ldots, m$ and $\sigma^2 = \sigma_I^2 = \sigma_U^2$, then it follows that

$$\text{PCC}(\infty) \leqslant \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_I^*}}\right).$$

If we further assume that the covariance matrix for the differentially expressed genes has the form $\sigma^2 I$, so that differentially expressed genes are independent, then

$$\text{PCC}(\infty) = \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\right).$$

A.2 *Appendix B*

Assume that each gene has the same variance σ^2, and all differentially expressed genes have the same fold-change, $\mu_i = \delta, i = 1, \ldots, m$. The linear predictor w developed on some training set consists of zeros and ones. To simplify notation, let $k = \frac{1}{2}\log\left(\frac{1-p_1}{p_1}\right)$.

$$\text{PCC}(n) = E_w[p_1 P(w'x > k|w, x \in C_1) + (1 - p_1)P(w'x < k|w, x \in C_2)]$$

$$= p_1 E_w\left[\Phi\left(\frac{w'\mu - k}{\sqrt{w'\Sigma w}}\right)\right] + (1 - p_1)E\left[\Phi\left(\frac{k + w'\mu}{\sqrt{w'\Sigma w}}\right)\right]$$

$$\approx p_1\Phi\left(\frac{E[w'\mu] - k}{\sqrt{E[w'\Sigma w]}}\right) + (1 - p_1)\Phi\left(\frac{k + E[w'\mu]}{\sqrt{E[w'\Sigma w]}}\right).$$

Now, $E[w'\mu] = \sum_{i=1}^{m}\delta E[w_i] = m\delta(1 - \beta)$. Also, if λ_1 is the largest eigenvalue of the correlation matrix of the genes, then $w'\Sigma w \leqslant \lambda_1\sigma^2 w'w$ (see, for example, Schott, 1997, Theorem 3.15). So $E[w'\Sigma w] \leqslant \lambda_1\sigma^2\sum_{i=1}^{p}E[w_i^2] = \lambda_1[\sigma^2 m(1 - \beta) + \sigma^2(p - m)\alpha]$. Thus,

$$\text{PCC}(n) \geqslant p_1\Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{\delta m(1 - \beta) - k}{\sqrt{m(1 - \beta) + (p - m)\alpha}}\right) + (1 - p_1)\Phi\left(\frac{1}{\sigma\sqrt{\lambda_1}}\frac{k + \delta m(1 - \beta)}{\sqrt{m(1 - \beta) + (p - m)\alpha}}\right).$$

If $p_1 = \frac{1}{2}$, then $k = 0$ and

$$\text{PCC}(n) \geqslant \Phi\left(\frac{\delta}{\sigma\sqrt{\lambda_1}}\frac{m(1 - \beta)}{\sqrt{m(1 - \beta) + (p - m)\alpha}}\right).$$

If genes are also independent, so that $\Sigma = \sigma^2 I$, then $E[w'\Sigma w] = \sigma^2(m(1 - \beta) + (p - m)\alpha)$ and

$$\text{PCC}(n) \approx \Phi\left(\frac{\delta}{\sigma}\frac{m(1 - \beta)}{\sqrt{m(1 - \beta) + (p - m)\alpha}}\right).$$

A.3   *Appendix C*

In this appendix, we present a method for finding the optimal α for fixed n.

Fix a sample size n, and assume this is an even number. First, note that the distance between the class means under the current model is 2δ. Now, use the normal approximation sample size formula to solve for β. Recall that notationally $t_{\beta,n-2} = T_{n-2}^{-1}(\beta)$ where T_{n-2}^{-1} is the inverse cumulative distribution function for a central T distribution with $n-2$ degrees of freedom

$$n \approx 4\frac{\sigma^2}{(2\delta)^2}[t_{n-2,\alpha/2} + t_{n-2,\beta}]^2$$

$$\beta \approx T_{n-2}\left(-\frac{\delta}{\sigma}\sqrt{n} - t_{n-2,\alpha/2}\right)$$

$$= 1 - T_{n-2}\left(\frac{\delta}{\sigma}\sqrt{n} + t_{n-2,\alpha/2}\right)$$

$$1 - \beta \approx T_{n-2}\left(\frac{\delta}{\sigma}\sqrt{n} + t_{n-2,\alpha/2}\right).$$

This results in

$$\text{PCC}(n, \alpha) \approx \Phi\left(\frac{\frac{\delta}{\sigma}m T_{n-2}\left(\frac{\delta}{\sigma}\sqrt{n} + t_{n-2,\alpha/2}\right)}{\sqrt{m T_{n-2}\left(\frac{\delta}{\sigma}\sqrt{n} + t_{n-2,\alpha/2}\right) + \alpha(p - m)}}\right), \tag{C.1}$$

where T_{n-2} is the cumulative distribution function for a Student's t distribution with $n-2$ degrees of freedom. Equation (C.1) can be used to select an α significance level that will maximize $\text{PCC}(n)$.

A.4   *Appendix D*

In Appendix A, it was shown that

$$\text{PCC}(\infty) \leqslant \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_I^*}}\right).$$

In Appendix B, it was shown that

$$\text{PCC}(n) \geqslant \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_1}}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{\sqrt{m(1-\beta) + (p-m)\alpha}}}\right),$$

so that

$$\text{PCC}(\infty) - \text{PCC}(n) \leqslant \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_I^*}}\right) - \Phi\left(\frac{\delta}{\sigma}\sqrt{\frac{m}{\lambda_1}}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{\sqrt{m(1-\beta) + (p-m)\alpha}}}\right).$$

Under the assumption $\Sigma = \sigma^2 I$ and $\mu_i = \delta, 1 \leqslant i \leqslant m$, we have

$$\text{PCC}(\infty) - \text{PCC}(n) = \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\right) - \Phi\left(\frac{\delta}{\sigma}\sqrt{m}\sqrt{1-\beta}\sqrt{\frac{m(1-\beta)}{m(1-\beta) + (p-m)\alpha}}\right).$$

A.5 *Appendix E*

The empirical Bayes method used in Table 3 is presented. Let $\tau^2 = \frac{2}{n}\hat{\sigma}^2_{\text{median}}$ where $\hat{\sigma}^2_{\text{median}}$ is the median (across genes) of the estimated pooled within-class variance. Let s^2 be the sample variance of the estimated effect sizes, $\text{var}(2\widehat{\delta})$. Then, $\hat{B} = \frac{\tau^2}{\tau^2 + (s^2 - \tau^2)^+}$ and $\widehat{\delta/\sigma}_g = \bar{\delta}/\sigma + (1 - \hat{B})(\widehat{\delta/\sigma}_g - \bar{\delta}/\sigma)$. See, e.g. Carlin and Louis (1996). The largest estimated effect size was used.

REFERENCES

CARLIN, B. P. AND LOUIS, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

DOBBIN, K. AND SIMON, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* **6**, 27–38.

ENIS, P. AND GEISSER, S. (1974). Optimal predictive linear discriminants. *Annals of Statistics* **2**, 403–10.

FU, W. J., DOUGHERTY, E. R., MALLICK, B. AND CARROLL, R. J. (2005). How many samples are needed to build a classifier: a general sequential approach. *Bioinformatics* **21**, 63–70.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A. *and others*. (1999). Molecular classification of cancer: class discovery and class prediction by expression monitoring. *Science* **286**, 531–7.

HOCKING, R. R. (1996). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. New York: John Wiley and Sons.

HWANG, D., SCHMITT, W. A., STEPHANOPOULOS, G. AND STEPHANOPOULOS, G. (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics* **18**, 1184–93.

LACHENBRUCH, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics* **24**, 823–34.

LEDOIT, O. AND WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365–411.

MOLINARO, A. M., SIMON, R. AND PFEIFFER, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–7.

MUKHERJEE, S., TAMAYO, P., ROGERS, S., RIFKIN, R., ENGLE, A., CAMPBELL, C., GOLUB, T. R. AND MESIROV, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10**, 119–42.

PAIK, S., SHAK, S., TANG, G., KIM, C., BAKER, J., CRONIN, M., BAEHNER, F. L., WALKER, M. G., WATSON, D., PARK, T. *and others*. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* **351**, 2817–26.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edition. New York: John Wiley and Sons.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MULLER-HERMELINK, H. K., SMELAND, E. B., GILTNANE, J. M. *and others*. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 1937–47.

SCHOTT, J. R. (1997). *Matrix Analysis for Statistics*. New York: John Wiley and Sons.

SIMON, R., RADMACHER, M. D., DOBBIN, K. AND MCSHANE, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute USA* **95**, 14–8.